Overview of the NTCIR-17 FairWeb-1 Task

Sijie Tao, Nuo Chen, Tetsuya Sakai
Waseda University, Japan
tsjmailbox@ruri.waseda.jp
pleviumtan@toki.waseda.jp
tetsuyasakai@acm.org

Zhumin Chu
Tsinghua University, P.R.C.
chuzm19@mails.tsinghua.edu.cn

Hiromi Arai
RIKEN AIP, Japan
hiromi.arai@riken.jp

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Maria Maistro
University of Copenhagen, Denmark
mm@di.ku.dk

ABSTRACT

This paper provides an overview of the NTCIR-17 FairWeb-1 Task. FairWeb-1 is an English web search task which seeks more than an ad-hoc web search task. Our task considers not only document relevance but also group fairness. We designed three types of search topics for this task: researchers (R), movies (M), and Youtube contents (Y). For each topic type, attribute sets are defined for considering group fairness. We utilise a deduped version of the Chuweb21 corpus as the target corpus. We received 28 runs from six teams, including six runs from the organisers team. In this paper, we describe the task, the test collection construction and the official evalution results of the submitted runs.

1 INTRODUCTION

This paper presents an overview of FairWeb-1, a pilot task at NTCIR-17. Unlike Ad-hoc document retrieval tasks such as We Want Web (WWW) tasks [21] at previous NTCIR conferences where the participant teams only need to consider document relevance, FairWeb-1 considers not only document relevance from a viewpoint of search engine users but also group fairness from a viewpoint of entities that are being sought. Three entity types are considered at FairWeb-1: researchers (R), movies (M), and Youtube contents (Y). Search topics are designed to describe information needs about the three types of entities. For each type of entities, we have one or two *attribute sets*, which contain either *nominal* or *ordinal groups* defined for considering group fairness. Moreover, a target distribution is provided for each attribute set. Given the search topics and the attribute sets, the participant teams are expecetd to submit SERPs[1] that not only contain relevant documents at the top ranks but also are group fair with respect to each attribute set. The submitted runs are evaluated with a suite of evaluation measures called GFR (Group Fairness Relevance), which considers both relevance and group fairness [19].

Table 1 shows the timeline of the NTCIR-17 FairWeb-1 task. Table 2 shows the information of the participant teams and the numbers of submitted runs. Nine teams resgistered for the task, and by the submission due, we received 28 runs from six participant teams, including six runs from the organisers team. More information about the task are available at http://sakailab.com/fairweb1/. Details of the runs can be found in the participants' papers [3, 4, 7, 23].

[1]Search Engine Result Pages

The rest of this paper is organised as follows: Section 2 describes the task specification, evaluation measures and the pilot data. Section 3 introduces the construction of the test collection, including information regarding the target corpus, the submitted runs, and the entity annotation. Section 4 shows the details of the official evaluation results of the submitted runs. Finally, Section 5 conludes the paper.

Table 1: Timeline of FairWeb-1 at NTCIR-17

Time	Content
October 3 2022	Sample topics released
February 15 2023	Pilot runs released
March 15 2023	Topics released
May 22 2023	Run submissions due
May 23 - July 31 2023	Entity annotations
August 1 2023	Evaluation results released
September 1 2023	Draft participant paper submissions due
November 1 2023	Camera-ready paper submissions due
December 2023	NTCIR-17 Conference@NII, Tokyo, Japan

Table 2: FairWeb-1 participants and the number of submitted runs.

Team name	Institution	#runs
[undisclosed]∗	[undisclosed]	2
RSLFW [7]	Waseda University, Japan	5
THUIR [23]	Tsinghua University, P.R.C.	5
UDinfo [3]	University of Delaware, U.S.A.	5
rmit_ir[4]	RMIT University, Australia	5
ORGANISERS		6
Total #runs		28

∗The runs from this team contained bugs and are not useful for our analysis.

2 TASK

FairWeb-1 is an English web search task which considers group fairness. Imagine if we are serving as chairs of an IR conference, and we want to hire diverse IR researchers as organisers. There

are several deminsions we may need consider: different career stages (include junior researchers, not just famous researchers), different genders, and different nationalities. However, if we search "information retrieval researchers" on present search engines, they may return only famous people with a poor gender balance. The motivation of this task is to encourage researchers develop search algorithms that retrieve not only relevant but also group-fair results.

2.1 Task Specification

FairWeb-1 considers not only relevance but also group fairness from a viewpoint of entities that are being sought. Four entity types were considered when the task was lauched: researhers (R), movies (M), Twitter accounts (T), and Youtube contents (Y). Based on the results of pilot data experiments, we decided to drop Twitter (T) topics as it was difficult to locate a sufficient number of Twitter account from retrieved documents. Therefore, we construct the test collections based on three entity types. In our task, relevant entities are entities that satisfy the condition specified in the topic description. Topics are developed based on the four entity types, so entity types can also be considered as topic types at FairWeb-1. For example, for an M type topic "Daniel Craig 007 movies", No Time to Die is a relevant entity, while GoldenEye is an M type entity but it is not relevant to this topic.

For each entity/topic type, one or two attribute sets are defined to consider group fairness. Table 3 shows the attribute sets for each topic type. For researchers topics (R topics), we have two attribute sets: HINDEX (ordinal) and GENDER (nominal). HINDEX contains groups of researchers' based on Google scholar h-indexes. Take x as a researcher's h-index, we consider four groups in HINDEX: $x < 10$, $10 \leq x < 30$, $30 \leq x < 50$, and $50 \leq x$. For GENDER, we consider three groups for convenience: *he*, *she*, and *other* [6, 10]. This attribute is collected based soly on what pronoun is used in one of the researcher's official biographies that we have located. Note that this label does not reflect our view on the gender of each researcher.

For movies topics (M topics), we have RATINGS (ordinal) and ORIGIN (nominal). RATINGS contains groups based on the *number of ratings* on the movie's IMDb page.[2] Take x as the number of ratings, four groups are considered in this set: $x < 100$, $100 \leq x < 10,000$, $10,000 \leq x < 1,000,000$, and $1,000,000 \leq x$. ORIGIN contains eight geographic regions: Africa, America, Antractica, Asia, Caribbean, Europe, Middle East, and Oceania. For each movie, the "country of origin" field on its IMDb page (which may contain multiple country/region names) is mapped to these eight regions. The mapping from countries/regions to the ORIGIN groups can be found in this excel file: https://waseda.box.com/ORIGIN-golddistribution.

For Youtube topics (Y topics), we have one attribute set named SUBSCS (ordinal). Retrieved Youtube contents are divided into four groups based on the *numbers of subscribers* of the content creator. The grouping strategy is the same as RATINGS of M topics.

For each attribute set, we used a uniform distribution as the target distribution, except for ORIGIN. For ORIGIN, we defined a target distribution based on how many countries/regions each ORIGIN group contains; the distribution can also be found in the ORIGIN-golddistribution excel file.

[2]https://www.imdb.com/

Given a certain type of search topic and the attribute sets with target distributions, the participants are expected to return a SERP that contains not only relevant documents near top ranks, but also is group fair with respect to each attribute set. The submitted runs are evaluated with a framework named GFR (Group Fairness Relevance), which is a combination of relavance-based evaluation measures and group fairness measures. At FairWeb-1, we use ERR and iRBU for relevance evaluation. The group fairness of a run is evaluated by measuring the similarity between the achieved distribution and the target distribution. The details of evaluation measures will be given in the next section.

Slides introducing the task can be found at https://waseda.box.com/fairweb1intro2023feb.

Table 3: Attribute sets for each topic type

Topic type	Attribute sets
R	HINDEX (ordinal, 4 groups) GENDER (nominal, 3 groups)
M	RATINGS (ordinal, 4 groups) ORIGIN (nominal, 8 groups)
Y	SUBSCS (ordinal, 4 groups)

2.2 Evaluation Measures

For evaluating the runs based on relevance and group fairness, we utilise the GFR (Group Fairness and Relevance) framework [19]; the details can be found in the paper. Below, we briefly explain how the measures are computed for the FairWeb-1 task.

Let A be an attribute set containing attribute values (or *groups*) a_i ($i = 1, \ldots, |A|$), with a target distribution p_* with group membership probabilities $p_*(a_i)$. For example, for A = GENDER = {he, she, other}, we have $p_*(\text{he}) = p_*(\text{she}) = p_*(\text{other}) = 1/3$.

Given a ranked list L (i.e., a SERP) of documents to be evaluated, we require a group membership vector with respect to A for each document. More specifically, for the document ranked at k, its group membership vector contains as its elements $G(L, k, a_i)$, the probability that the document belongs to group a_i, s.t. $\sum_i G(L, k, a_i) = 1$. (Section 3.5 explains how the vectors are derived from relevant entity annotations in our task.) For example, a document may have a group membership vector of $(1, 0, 0)$ for GENDER, which means that the document is 100% about researchers who uses the "he" pronoun.

From the group membership vectors, we can compute the *achieved distribution* $p_{L,k}$ at rank k of L with respect to A, where the probability for each group is given by $p_{L,k}(a_i) = \sum_{j=1}^{k} G(L, j, a_i)/k$, i.e., the average probability across ranks 1 through k. Then, for every rank k where there is a relevant document, we can compute a *distribution similarity* between the achived distribution at k and the target distribution:

$$\text{DistrSim}(L, k) = \text{DistrSim}(p_{L,k} \parallel p_*) = 1 - \text{Divergence}(p_{L,k} \parallel p_*), \quad (1)$$

where the Divergence function is JSD (Jensen-Shannon Divergence) for attribute sets containing nominal groups (i.e., GENDER and ORIGIN in our case), and either NMD (Normalised Match Distance)

or RNOD (Root Normalised Order-aware Divergence) for those containing ordinal groups (i.e., HINDEX, RATINGS, and SUBSCS) in our case. The details and properties of these divergence measures are discussed elsewhere [15, 16]. It should be noted that divergences such as JSD, Kullback-Leibler Divergence, and Mean Absolute Error are not suitable for comparing distributions over ordinal groups [13].

For each topic type, the GFR score is computed as follows.

$$\mathrm{GFR}(L)@l = \sum_{k=1}^{l} \mathrm{Decay}(L,k) \left(w_0 \mathrm{Utility}(L,k) + \sum_{m=1}^{M} w_m \mathrm{DistrSim}(L,k) \right), \tag{2}$$

where l is the document cutoff (we let $l = 20$ for the official evaluation), M is the number of attribute sets considered ($M = 2$ for M topics and R topics; $M = 1$ for Y topics), $w_0, \ldots, w_M$ are the weights for weighted averaging (we employ unweighted averaging and therefore $w_0 = w_1 = w_2 = 1/3$ for M and R topics, and $w_0 = w_1 = 1/2$ for Y topics).

The Decay function in Eq. 2 is based on ERR (Expected Reciprocal Rank) [2, 12]. That is, the satisfaction probability at rank k ($p^{\mathrm{sat}}_{L,k}$) is defined to be 3/4 for L2-relevant documents and 1/4 for L1-relevant documents, and the Decay is computed as:

$$\mathrm{Decay}(L,k) = p^{\mathrm{sat}}_{L,k} \prod_{j=1}^{k-1} (1 - p^{\mathrm{sat}}_{L,j}) \quad (k > 1) \tag{3}$$

and $\mathrm{Decay}(L,1) = p^{\mathrm{sat}}_{L,1}$.

As for the Utility function in Eq. 2, we consider an ERR-based function ($\mathrm{Utility}(L,k) = 1/k$) as well as an iRBU (intentwise Rank-Biased Utlity [22])-based function ($\mathrm{Utility}(L,k) = \phi^k$, with $\phi = 0.99$ throughout our evaluation). The ERR-based Decay value decreases rapidly as we go down the SERP (1,0.5, 0.33, 0.025, etc.), while the iRBU-based one decreases very gradually (1, 0.99, 0.98, 0.97, etc.), and therefore the latter is probably more suitable for our task as our topics are informational rather than navigational.

Eq. 2 generalises the Sakai/Robertson NCU (Normalised Cumulative Utility) framework [18], which only had the Decay and Utility components. From this viewpoint, GFR represents the expected user experience under the ERR user model, where the experience of each user group (i.e., those who abandon the SERP at a particular rank) is given as the weighted average of Utility and DistrSim scores [19].

All scores were computed using `NTCIREVAL` version 230130.[3] Information on how to use this tool for the FairWeb-1 evaluation can be found in the raw official results (See Section 4).

2.3 Pilot Data

In February 2023, we released a pilot data set, which provides details on how our baseline runs are evaluated on a pilot topic set, which contains one topic per each entity/.topic type (including a Twitter type that was later droppped).[4]

3 TEST COLLECTION CONSTRUCTION

3.1 Target Corpus

We adopted Chuweb21D [5] as the target corpus. Chuweb21D was constructed based on the Chuweb21 collection [5] through de-duplication. Similar to Colossal Clean Crawled Corpus (c4) [11] and CC-News-En [8] datasets, the Chuweb21 collection was derived from the 2021-17 (April, 2021) block of the Common Crawl dataset [5]. The original 2021-17 block of Common Crawl data contains about 3.1 billion web pages crawled from 10th to 23rd in April 2021. After a series of filtering procedures including language detection, domain filtering and content length constraint, the cleaned Chuweb21 collection contains 82.4 million web pages or 1.69 TiB of compressed content.

The Chuweb21 collection was first used in the NTCIR-16 WWW-4 task [21]. During the pooling set annotation stage, the task organizers found a fairly severe document duplication issue in the Chuweb21 collection, which reduces the efficiency of annotation resource usage and ranking system comparison. To solve this problem, Chu et al. [5] used the Simhash algorithm [9] to carry out massive de-duplication work on the Chuweb21 collection, and released the de-duplicated Chuweb21D collection. The Chuweb21D collection contains two different releases obtained under different de-duplication thresholds, Chuweb21D-60 and Chuweb21D-70, which represent the scale of documents retained after de-duplication is 60% and 70% respectively. Specifically, we used Chuweb21D-60 as the target corpus, which contains about 49.8 million web pages and is now freely available online [6].

3.2 Topics

The first five authors of this paper constructed 45 test topics (15 R topics, 15 M topics, and 15 Y topics) based on our actual interests, so that they can serve as *gold* annotators (i.e., the annotators/assessors are the ones that have the actual information needs and therefore they get to decide what is relevant or not) [1, 20]. The topic set can be found at https://waseda.box.com/fairweb1topics.

3.3 Submitted Runs

We received 28 runs from five participant teams, including six runs from the organisers. Table 4 shows the system descriptions of each run.

3.4 Entity Annotation

The entity annotation is done by utilising a assessment tool named FAIRE (FAIRE stands for Annotation Interface for Relevant Entities). This annotation interface is developed based on a tool named PLY, which is a relevance assessment tool used at the WWW tasks. At FairWeb-1, the organisers who developed the test topics served as annotators, and each topic is assessed by the organiser who created it. Each annotator was given a user account to access the interface. Once the annotator logs in, the interface shows a list of assigned topics and the annotation progress for each topic. Figure 1 is a screenshot of FAIRE where a user just logs in. The annotator can choose any topic she/he likes from the list to start annotation.

[3] https://research.nii.ac.jp/ntcir/tools/ntcireval-en.html
[4] https://waseda.box.com/fairweb1pilotdata
[5] https://commoncrawl.org/blog/april-2021-crawl-archive-now-available
[6] https://github.com/chuzhumin98/Chuweb21D

Table 4: The SYSDESC field of each run. For brevity, the baseline run file names run.*.{query,description} are hereafter shown as run.*.{Q,D}, respectively. See the participants' papers for more details [3, 4, 7, 23].

Run	SYSDESC field
undisclosed-1	BM25 with query-field run
undisclosed-2	BM25 with description-field run
RSLFW-Q-MN-1	RSLFW Baseline using COIL with manual query operation
RSLFW-Q-MN-2	RSLFW Baseline + pm1 rerank algorithm
RSLFW-Q-MN-3	RSLFW Baseline + pm2 rerank algorithm
RSLFW-Q-RR-4	Official Baseline + pm2 rerank algorithm
RSLFW-Q-RR-5	Re-rank official baseline using Python library reranking
THUIR-D-RR-5	xQuAD + sparse relevance score
THUIR-QD-RG-1	RRF, k = 60
THUIR-QD-RG-2	Ultilize lightgbm to fuse 4 dense & 12 sparse relevance features
THUIR-QD-RR-3	Ultilize prompts to calculate feature scores
THUIR-QD-RR-4	PM2 + dense relevance score
UDinfo-D-RR-1	re-rank run.bm25-D60-D_ver0313.txt leveraging gender and geo location embeddings
UDinfo-D-RR-3	re-rank run.qld-D60-D_ver0313.txt leveraging gender and geo location embeddings
UDinfo-D-RR-5	re-rank run.qljm-D60-D_ver0313.txt leveraging gender and geo location embeddings
UDinfo-Q-RR-2	re-rank run.bm25-D60-Q_ver0313.txt leveraging gender and geo location embeddings
UDinfo-Q-RR-4	re-rank run.qld-D60-Q_ver0313.txt leveraging gender and geo location embeddings
rmit_ir-D-RR-1	Linear combination of top 50 relevance and fairness with lambda= 0.9
rmit_ir-D-RR-2	PM2 with lambda= 0.9
rmit_ir-D-RR-3	PM2 on top 50 with lambda= 0.9
rmit_ir-D-RR-4	Linear combination of relevance and fairness with lambda= 0.9
rmit_ir-Q-RR-5	Linear combination of top 50 relevance and fairness with lambda= 0.5
run.bm25-depThre3-D	BM25 description-field run, with the default parameters of Anserini
run.bm25-depThre3-Q	BM25 query-field run, with the default parameters of Anserini
run.qld-depThre3-D	qld (query likelihood with Dirichlet smoothing) description-field run, with the default parameters of Anserini
run.qld-depThre3-Q	qld (query likelihood with Dirichlet smoothing) query-field run, with the default parameters of Anserini
run.qljm-depThre3-D	qljm (query likelihood with Jelinek-Mercer smoothing) description-field run, with the default parameters of Anserini
run.qljm-depThre3-Q	qljm (query likelihood with Jelinek-Mercer smoothing) query-field run, with the default parameters of Anserini

Figure 2 is a screenshot of FAIRE where the pool file for Topic R005 has been loaded. The top area shows the current topic and description. The annotator can change to another topic by clicking the prev or the next button. By clicking the topic menu, the annotator can also change to any other assigned topics from a dropdown list. The left panel of the page shows the document pool for the current topic. The user can annotate the pool in any order by clicking the document in the left panel, and once the annotation of a document is done, a coloured tag will be placed beside the document ID. In the middle of the page, there is a document viewer and several text fields for the annotator to read the document and fill the schema of the top three relevant entities. If there is no relevant entity found in the document, or the document is not shown properly (ex. FAIRE returns an error page, or the texts are unreadable), the annotator can click the corresponding checkboxes below the text fields. Moreover, a comment field is shown at the bottom for the annotator to put any remarks or commends regarding the document or the entities. Once the annotation of the document is done, the annotator can click the save button to store the result to the database. The results are allowed to be updated if the annotator wants to make any change.

3.5 Deriving Document Relevance and Group Membership

After the entity annotation is finished, the annotation results are utilised to derive document relevance and group membership. The document relevance level $g(d) \in \{0, 1, 2\}$ of document d is defined as follows:

$$g(d) = \begin{cases} 0 & (E(d) = \emptyset); \\ \max_{e \in E(d)} r(e) & (otherwise). \end{cases} \quad (4)$$

where E(d) is the set of relevant entities extracted from document d, and $r(e) \in \{1, 2\}$ is the relevance level of a relevant entity $e \in E(d)$. In other words, the relevance level of a document is simply defined as the max entity relevance level within the document.

As for the group membership probabilities of d, we conducted two strategies to derive the probability vectors based on different attribute sets. The first one is called hard group membership for entities. This is for the scenario where an entity can be mapped to exactly one group of an attribute set. Let $C = \{C_1, ..., C_{|C|}\}$ be an attribute set, and $F(e, C_i)$ be a flag that maps an entity e to exactly one group in C. For example, for a researcher entity e whose bio says "he" ($C = GENDER$), $F(e, C_1) = 1, F(e, C_2) = F(e, C_3) = 0$. Then for a document d with relevant entities $E(d)$, the group membership probabilities of d is defined as:

$$P(d, C_i) = \begin{cases} 1/|C| & (E(d) = \emptyset); \\ \frac{|\sum_{e \in E(d)} F(e, C_i)|}{|\sum_i \sum_{e \in E(d)} F(e, C_i)|} & (otherwise). \end{cases} \quad (5)$$

The second method to derive document group membership is named soft group membership, which is designed for the ORIGIN attribute set for movie entities. For a movie entity, it can be mapped to more than one group of the ORIGIN attribute set. Let

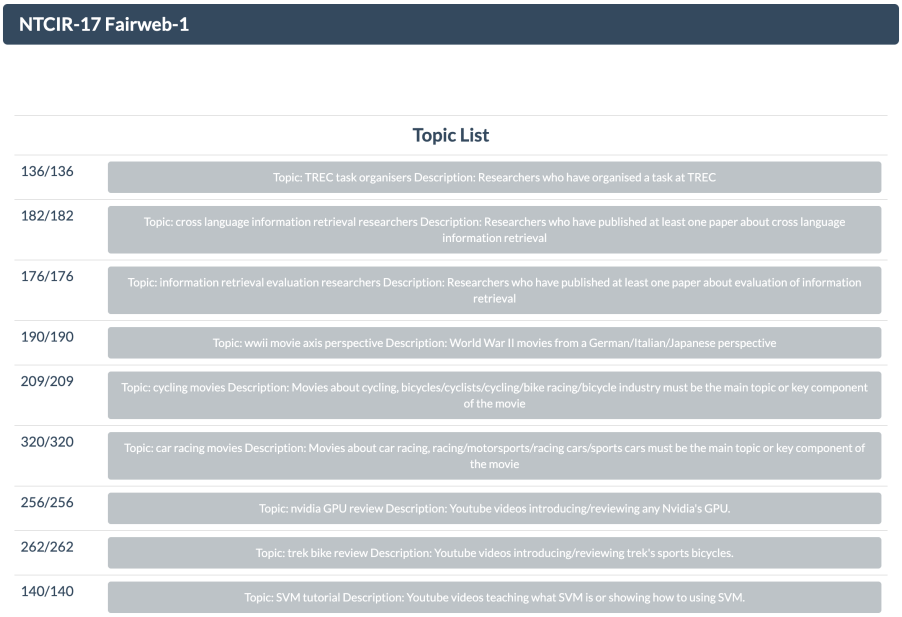

Figure 1: Screenshot of FAIRE, showing the assigned topics of an annotator

$ORIGIN(e)(\subseteq C)$ be a set of geographical regions for a movie entity $e \in E(d)(m = |ORIGIN(e)|, m \geqq 1)$. For example, if e's countries of origin is UK and China, then $ORIGIN(e) = \{Asia, Europe\}$. For movie entities, the group membership with respect to ORIGIN is defined as follows:

$$P(d, C_i) = \begin{cases} 1/|C| & (E(d) = \emptyset); \\ \frac{|\sum_{e \in E(d)} G(e,C_i)|}{|\sum_i \sum_{e \in E(d)} G(e,C_i)|} & (otherwise). \end{cases} \quad (6)$$

where

$$G(e, C_i) = \begin{cases} 1/m & (C_i \in ORIGIN(e)); \\ 0 & (otherwise). \end{cases} \quad (7)$$

4 OFFICIAL RESULTS

The raw official result files and details can be found at https://waseda.box.com/ntcir17fairweb1officialpublic . Please go through the README file first.

4.1 Relevance: ERR and iRBU for the Full Topic Set

Table 5 shows the run rankings according to Mean ERR and Mean iRBU over the entire topic set, respectively. The table indicates clusters of runs in terms of statistical significance: for example, THUIR-QD-RR-4 is the top ranker with respect to ERR, it outperforms statistically significantly the runs ranked from 18 to 28. And THUIR-QD-RR-3 and THUIR-QD-RG-2 form the top cluster in terms of iRBU, by outperforming the runs ranked at 22 through 28.

4.2 Relevance and Group Fairness for Movie (M) Topics

Tables 6, 7 and 8 show the mean scores averaged across the 15 M topics, with statistical test results for those involving group fairness. As can be seen from the results, THUIR-QD-RR-3 is still the top performer in terms of both relevance and group fairness. It also outperforms runs ranked at 26, an organisers' run statistically significantly with respect to RATINGS.

Figures 3 and 4 visualise the mean scores of each system. From Figure 3, we can see that the top three runs from team THUIR have a relatively huge advantage to the other runs in terms of mean iRBU. Figure 4 shows the intersectional group fairness over the 15 M topics. It can be seen that all the runs are generally "equally fair" to both attributes, and there are no obvious bias towards either attribute. The advantage of the three THUIR runs can also be observed from Figure 4, they are the only runs have mean GF scores larger than 0.5 and forms the top cluster in the figure.

Figure 5 visalises the difficulty of each M topic in terms of scores averaged over all runs. Topic M004 (Star Wars parody movies), M014 (biographical movies), and M015 (fictional creature movies) have much lower scores than the other M topics. The difficulty of M004 probably comes from the limited correct answer to the query, while the reason why M014 and M015 have fewer relevant entities remains further investigation.

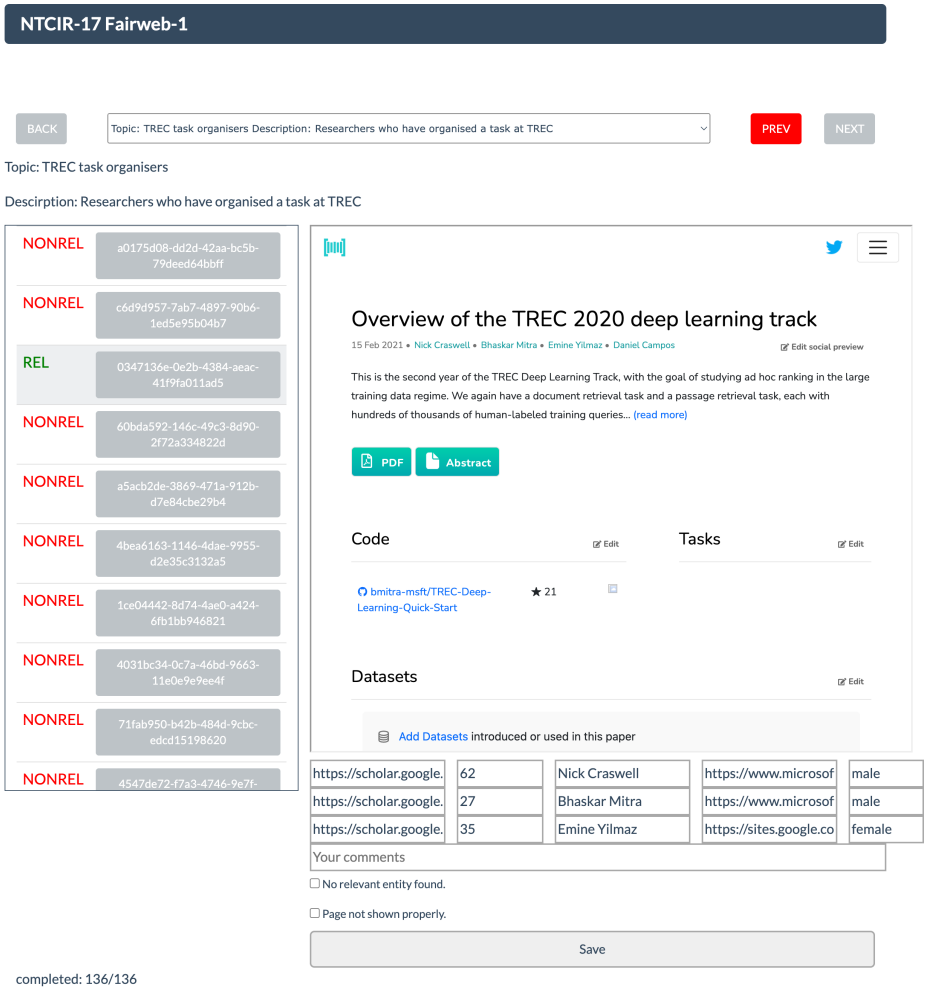

Figure 2: Screenshot of FAIRE, showing the annotation page

4.3 Relevance and Group Fairness for Researcher (R) Topics

Tables 9, 10 and 11 show the mean scores averaged across the 15 R topics, with statistical test results for those involving group fairness. The top 24 runs outperform the three zero-score runs, and there are no other statistically significant differences amoung the runs.

Figures 6 and 7 visualise the mean scores of each system. It can be observed that THUIR-QD-RG-2 and THUIR-QD-RG-1 are the best two runs over the R topics, and they form the top cluster in Figure 7. Figure 7 shows the intersectional group fairness over the R topics. As can be seen, no runs have obvious bias towards either attribute, while compared with what was observed in Figure 4, all the runs tend to be have slightly higher GENDER GF scores than HINDEX.

Figure 8 visalises the difficulty of each R topic in terms of scores averaged over all runs. It can be found that topic R004 (AIRS authors) is the most difficult topic amoung all R topics. The high ambguity of term "AIRS" probably results in retrieving many irrelevant documents and leads to limited relavant hits.

4.4 Relevance and Group Fairness for YouTube (Y) Topics

Tables 12, 13 and 14 show the mean scores averaged across the 15 Y topics, with statistical test results for those involving group fairness. In terms of GF and GFR scores, the top 17 runs outperfrom the two zero-score runs, and there are no other statistical significant differences observed.

Figures 9 and 10 visualises the mean scores of each system.

Figure 10 visalises the difficulty of each Y topic in terms of scores averaged over all runs. From the figure, we can see that there are six Y topics that have low scores, which is more than M and R topics. The limited number of relevant documents in the corpus possibly results in higher difficulty of these Y topics.

Table 5: Mean ERR and iRBU scores for each run over the 45 topics. ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$ [14])." For example, in terms of ERR, THUIR-QD-RR-4 statistically significantly outperforms the runs ranked at 18 through 28.

Rank	Run	Mean ERR	Rank	Run	Mean iRBU
1	THUIR-QD-RR-4	0.2139 (>18-28)	1	THUIR-QD-RG-2	0.5744 (>22-28)
2	THUIR-QD-RR-3	0.2074 (>20-28)	2	THUIR-QD-RR-4	0.5740 (>22-28)
3	THUIR-QD-RG-2	0.2021 (>22-28)	3	THUIR-QD-RR-3	0.5651 (>23-28)
4	RSLFW-Q-RR-5	0.1847 (>26-28)	4	UDinfo-Q-RR-2	0.4977 (>26-28)
5	UDinfo-Q-RR-2	0.1788 (>26-28)	5	RSLFW-Q-RR-5	0.4973 (>26-28)
6	rmit_ir-Q-RR-5	0.1685 (>26-28)	6	UDinfo-Q-RR-4	0.4838 (>26-28)
7	THUIR-QD-RG-1	0.1542 (>27-28)	7	THUIR-QD-RG-1	0.4813 (>26-28)
8	run.qljm-depThre3-Q	0.1495 (>27-28)	8	rmit_ir-Q-RR-5	0.4787 (>26-28)
9	RSLFW-Q-RR-4	0.1485 (>27-28)	9	THUIR-D-RR-5	0.4772 (>26-28)
10	UDinfo-Q-RR-4	0.1465 (>27-28)	10	RSLFW-Q-RR-4	0.4737 (>26-28)
11	run.bm25-depThre3-Q	0.1390 (>27-28)	11	UDinfo-D-RR-3	0.4377 (>26-28)
12	rmit_ir-D-RR-4	0.1379 (>27-28)	12	run.qljm-depThre3-Q	0.4336 (>26-28)
13	UDinfo-D-RR-5	0.1343 (>27-28)	13	run.qld-depThre3-Q	0.4330 (>26-28)
14	UDinfo-D-RR-3	0.1310 (>27-28)	14	rmit_ir-D-RR-1	0.4304 (>26-28)
15	rmit_ir-D-RR-1	0.1306 (>27-28)	15	UDinfo-D-RR-1	0.4293 (>26-28)
16	UDinfo-D-RR-1	0.1306 (>27-28)	16	run.bm25-depThre3-Q	0.4242 (>26-28)
17	RSLFW-Q-MN-1	0.1245 (>27-28)	17	UDinfo-D-RR-5	0.4237 (>26-28)
18	run.qld-depThre3-Q	0.1226 (>27-28)	18	rmit_ir-D-RR-4	0.4181 (>26-28)
19	THUIR-D-RR-5	0.1218 (>27-28)	19	rmit_ir-D-RR-3	0.4017 (>27-28)
20	run.qljm-depThre3-D	0.1152 (>27-28)	20	run.qljm-depThre3-D	0.3889 (>27-28)
21	run.qld-depThre3-D	0.1126 (>27-28)	21	run.qld-depThre3-D	0.3872 (>27-28)
22	run.bm25-depThre3-D	0.1113 (>27-28)	22	rmit_ir-D-RR-2	0.3769 (>27-28)
23	rmit_ir-D-RR-3	0.1084 (>27-28)	23	run.bm25-depThre3-D	0.3624 (>27-28)
24	RSLFW-Q-MN-2	0.1067 (>27-28)	24	RSLFW-Q-MN-1	0.3510 (>27-28)
25	rmit_ir-D-RR-2	0.1029 (>27-28)	25	RSLFW-Q-MN-2	0.3298 (>27-28)
26	RSLFW-Q-MN-3	0.0770	26	RSLFW-Q-MN-3	0.2271 (>27-28)
27	undisclosed-1	0.0000	27	undisclosed-1	0.0000
28	undisclosed-2	0.0000	28	undisclosed-2	0.0000

4.5 A Case Study: THUIR-QD-RG-2 vs. run.qld-depThre3-D (for M topics)

In Table 8, THUIR-QD-RG-2 is the top performer on average (for M topics) in terms of both Mean GF^{JSD} (ORIGIN) and Mean GF^{RNOD} (RATINGS), whereas the worst performer in the same cluster (in terms of all three measures) is the baseline run.qld-depThre3-D. Also, in Table 5 where the entire topic set is used for computing relevance-based measures, THUIR-QD-RG-2 is the top performer on average in terms of Mean iRBU, and it statistically significantly outperforms run.qld-depThre3-D which is ranked at 23. In this section, we compare THUIR-QD-RG-2 and run.qld-depThre3-D over the M topics to illustrate how the GFR framework works.

Figures 11-13 visualise the per-topic measure scores for THUIR-QD-RG-2 and run.qld-depThre3-D. It can be observed that the iRBU, GF^{RNOD} (for RATINGS), and GF^{JSD} (for ORIGIN) scores are generally similar. Hereafter, we shal focus on Topic M012 ("cartoon movies"), as THUIR-QD-RG-2 is about twice as effective as run.qld-depThre3-D in terms of all three measures.

Tables 15 and 16 show how group fairness scores are computed for the SERPs from the two runs for Topic M012. First, it can be observed that while the THUIR run manages to retrieve as many as 12 L1-relevant documents, the baseline run retrieves only 2 L1-relevant documents; the ERR-based abandoning probabilities $P_{\mathrm{ERR}}(r)$ are also shown. Next, the RATINGS columns show that the group membership for each L0 document (i.e., documents that do not contain any relevant entities) is considered to be uniform over the four groups of the RATINGS attribute set, while that for each L1-relevant document is computed based on the number of ratings of each movie entity found within the document. Moreover, for example, in Table 15, the RNOD-based DistrSim for the 7th document is computed as follows.

(1) By averaging over the group membership vectors from Ranks 1-7, the achieved distribution at Rank 7 is considered to be $(0.2619, 0.3095, 0.2143, 0.2143)$.
(2) By comparing the above with the uniform gold distribution $(0.25, 0.25, 0.25, 0.25)$, the RNOD-based DistrSim is computed to be 0.9519.

Finally, GF^{RNOD} (RATINGS) is obtained as the sum of $P_{\mathrm{ERR}}(r) * \mathrm{DistrSim}(r)$, that is, 0.8867.

Similarly, the ORIGIN columns shows that the group membership for each L0 document is considered to be uniform over the eight groups of the ORIGIN attribute set, namely, Africa, America,

Table 6: Mean ERR and iRBU scores for each run over the 15 M topics.

Run	Mean ERR	Run	Mean iRBU
THUIR-QD-RR-3	0.2653	THUIR-QD-RR-3	0.7230
THUIR-QD-RR-4	0.2518	THUIR-QD-RG-2	0.6923
rmit_ir-Q-RR-5	0.2434	THUIR-QD-RR-4	0.6859
THUIR-QD-RG-2	0.2280	run.qljm-depThre3-Q	0.6026
run.qljm-depThre3-Q	0.2114	rmit_ir-Q-RR-5	0.5819
RSLFW-Q-RR-5	0.2044	RSLFW-Q-RR-5	0.5674
UDinfo-Q-RR-2	0.2017	UDinfo-Q-RR-2	0.5668
UDinfo-Q-RR-4	0.1957	UDinfo-Q-RR-4	0.5582
RSLFW-Q-MN-1	0.1893	RSLFW-Q-RR-4	0.5463
rmit_ir-D-RR-4	0.1887	THUIR-D-RR-5	0.5316
rmit_ir-D-RR-1	0.1722	rmit_ir-D-RR-4	0.5239
run.bm25-depThre3-Q	0.1712	UDinfo-D-RR-5	0.5136
run.qld-depThre3-Q	0.1653	run.bm25-depThre3-Q	0.5035
RSLFW-Q-RR-4	0.1620	run.qld-depThre3-Q	0.4958
THUIR-QD-RG-1	0.1608	run.qljm-depThre3-D	0.4883
UDinfo-D-RR-1	0.1582	rmit_ir-D-RR-1	0.4862
run.bm25-depThre3-D	0.1564	rmit_ir-D-RR-3	0.4741
UDinfo-D-RR-5	0.1499	UDinfo-D-RR-1	0.4585
RSLFW-Q-MN-3	0.1489	rmit_ir-D-RR-2	0.4579
RSLFW-Q-MN-2	0.1489	THUIR-QD-RG-1	0.4400
run.qljm-depThre3-D	0.1478	run.bm25-depThre3-D	0.4337
rmit_ir-D-RR-2	0.1472	RSLFW-Q-MN-1	0.4268
rmit_ir-D-RR-3	0.1466	RSLFW-Q-MN-3	0.4250
THUIR-D-RR-5	0.1223	RSLFW-Q-MN-2	0.4250
UDinfo-D-RR-3	0.1222	UDinfo-D-RR-3	0.4069
run.qld-depThre3-D	0.1187	run.qld-depThre3-D	0.3728
undisclosed-1	0.0000	undisclosed-1	0.0000
undisclosed-2	0.0000	undisclosed-2	0.0000

Antarctica, Asia, Caribbean, Europe, Middle East, Oceania, and that, not surprisingly, the majority of the relevant documents are biased towards movies from America. In Table 15, the JSD-based DistrSim for the 7th document is computed as follows.

(1) By averaging over the group membership vectors from Ranks 1-7, the achieved distribution at Rank 7 is considered to be $(0.1071, 0.1786, 0.1071, 0.1786, 0.1071, 0.1071, 0.1071, 0.1071)$. Note that the probabilities for America and Asia are slightly higher than the rest due to the group membership vector of the 7th document.
(2) By comparing the above with the gold distribution, namely, $(0.2447, 0.1118, 0.0084, 0.1540, 0.0992, 0.2004, 0.08020.1013)$, the JSD-based DistrSim is computed to be 0.9259.

Finally, GF^{JSD} (ORIGIN) is obtained as the sum of $P_{ERR}(r) * \text{DistrSim}(r)$, that is, 0.8630.

Table 7: Mean group fairness scores for each run over the 15 M topics. ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$."

Rank	Run	Mean GF^{JSD} (ORIGIN)	Rank	Run	Mean GF^{NMD} (RATINGS)	Rank	Run	Mean GF^{RNOD} (RATINGS)
1	THUIR-QD-RG-2	0.5684 (>27-28)	1	THUIR-QD-RR-3	0.6433 (>26-28)	1	THUIR-QD-RG-2	0.5788 (>27-28)
2	THUIR-QD-RR-3	0.5391 (>27-28)	2	THUIR-QD-RG-2	0.6330 (>27-28)	2	THUIR-QD-RR-3	0.5683 (>27-28)
3	THUIR-QD-RR-4	0.5332 (>27-28)	3	THUIR-QD-RR-4	0.6118 (>27-28)	3	THUIR-QD-RR-4	0.5435 (>27-28)
4	THUIR-D-RR-5	0.4900 (>27-28)	4	run.qljm-depThre3-Q	0.5462 (>27-28)	4	THUIR-D-RR-5	0.4983 (>27-28)
5	RSLFW-Q-RR-4	0.4768 (>27-28)	5	THUIR-D-RR-5	0.5307 (>27-28)	5	run.qljm-depThre3-Q	0.4871 (>27-28)
6	run.qljm-depThre3-Q	0.4716 (>27-28)	6	RSLFW-Q-RR-4	0.5169 (>27-28)	6	RSLFW-Q-RR-4	0.4758 (>27-28)
7	UDinfo-Q-RR-4	0.4601 (>27-28)	7	UDinfo-Q-RR-4	0.5161 (>27-28)	7	UDinfo-Q-RR-4	0.4750 (>27-28)
8	UDinfo-D-RR-5	0.4543 (>27-28)	8	UDinfo-Q-RR-2	0.5132 (>27-28)	8	UDinfo-Q-RR-2	0.4706 (>27-28)
9	UDinfo-Q-RR-2	0.4493 (>27-28)	9	RSLFW-Q-RR-5	0.5124 (>27-28)	9	RSLFW-Q-RR-5	0.4693 (>27-28)
10	RSLFW-Q-RR-5	0.4479 (>27-28)	10	rmit_ir-Q-RR-5	0.5043 (>27-28)	10	UDinfo-D-RR-5	0.4488 (>27-28)
11	run.qld-depThre3-Q	0.4275 (>27-28)	11	UDinfo-D-RR-5	0.4888 (>27-28)	11	rmit_ir-Q-RR-5	0.4480 (>27-28)
12	run.qljm-depThre3-D	0.4273 (>27-28)	12	rmit_ir-D-RR-4	0.4784 (>27-28)	12	run.qld-depThre3-Q	0.4351 (>27-28)
13	rmit_ir-D-RR-4	0.4211 (>27-28)	13	run.qld-depThre3-Q	0.4668 (>27-28)	13	run.bm25-depThre3-Q	0.4283 (>27-28)
14	rmit_ir-Q-RR-5	0.4177 (>27-28)	14	run.bm25-depThre3-Q	0.4623 (>27-28)	14	rmit_ir-D-RR-4	0.4281 (>27-28)
15	run.bm25-depThre3-Q	0.4135 (>27-28)	15	run.qljm-depThre3-D	0.4606 (>27-28)	15	rmit_ir-D-RR-3	0.4234 (>27-28)
16	rmit_ir-D-RR-3	0.3989 (>27-28)	16	rmit_ir-D-RR-3	0.4529 (>27-28)	16	run.qljm-depThre3-D	0.4211 (>27-28)
17	rmit_ir-D-RR-1	0.3842 (>27-28)	17	rmit_ir-D-RR-1	0.4502 (>27-28)	17	rmit_ir-D-RR-1	0.4062 (>27-28)
18	rmit_ir-D-RR-2	0.3772 (>27-28)	18	rmit_ir-D-RR-2	0.4309 (>27-28)	18	rmit_ir-D-RR-2	0.4035 (>27-28)
19	UDinfo-D-RR-1	0.3672 (>27-28)	19	UDinfo-D-RR-1	0.4279 (>27-28)	19	UDinfo-D-RR-1	0.3913 (>27-28)
20	RSLFW-Q-MN-3	0.3514 (>27-28)	20	THUIR-QD-RG-1	0.4025 (>27-28)	20	THUIR-QD-RG-1	0.3684 (>27-28)
21	RSLFW-Q-MN-2	0.3514 (>27-28)	21	run.bm25-depThre3-D	0.3993 (>27-28)	21	run.bm25-depThre3-D	0.3630 (>27-28)
22	UDinfo-D-RR-3	0.3476 (>27-28)	22	UDinfo-D-RR-3	0.3876 (>27-28)	22	UDinfo-D-RR-3	0.3569 (>27-28)
23	run.bm25-depThre3-D	0.3401 (>27-28)	23	RSLFW-Q-MN-3	0.3847 (>27-28)	23	RSLFW-Q-MN-3	0.3503 (>27-28)
24	THUIR-QD-RG-1	0.3395 (>27-28)	24	RSLFW-Q-MN-2	0.3847 (>27-28)	24	RSLFW-Q-MN-2	0.3503 (>27-28)
25	RSLFW-Q-MN-1	0.3179 (>27-28)	25	RSLFW-Q-MN-1	0.3718 (>27-28)	25	RSLFW-Q-MN-1	0.3296 (>27-28)
26	run.qld-depThre3-D	0.3122 (>27-28)	26	run.qld-depThre3-D	0.3507 (>27-28)	26	run.qld-depThre3-D	0.3208 (>27-28)
27	undisclosed-1	0.0000	27	undisclosed-1	0.0000	27	undisclosed-1	0.0000
28	undisclosed-2	0.0000	28	undisclosed-2	0.0000	28	undisclosed-2	0.0000

Table 8: Mean GFR score for each run over the 15 M topics: relevance based on iRBU; group fairness for RATINGS based on RNOD. (Group fairness for ORIGIN is based on JSD.) ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$."

Rank	Run	Mean GFR
1	THUIR-QD-RG-2	0.6132 (> 27-28)
2	THUIR-QD-RR-3	0.6101 (> 27-28)
3	THUIR-QD-RR-4	0.5875 (> 27-28)
4	run.qljm-depThre3-Q	0.5205 (> 27-28)
5	THUIR-D-RR-5	0.5066 (> 27-28)
6	RSLFW-Q-RR-4	0.4996 (> 27-28)
7	UDinfo-Q-RR-4	0.4977 (> 27-28)
8	UDinfo-Q-RR-2	0.4956 (> 27-28)
9	RSLFW-Q-RR-5	0.4949 (> 27-28)
10	rmit_ir-Q-RR-5	0.4825 (> 27-28)
11	UDinfo-D-RR-5	0.4722 (> 27-28)
12	rmit_ir-D-RR-4	0.4577 (> 27-28)
13	run.qld-depThre3-Q	0.4528 (> 27-28)
14	run.bm25-depThre3-Q	0.4484 (> 27-28)
15	run.qljm-depThre3-D	0.4456 (> 27-28)
16	rmit_ir-D-RR-3	0.4321 (> 27-28)
17	rmit_ir-D-RR-1	0.4255 (> 27-28)
18	rmit_ir-D-RR-2	0.4129 (> 27-28)
19	UDinfo-D-RR-1	0.4057 (> 27-28)
20	THUIR-QD-RG-1	0.3827 (> 27-28)
21	run.bm25-depThre3-D	0.3789 (> 27-28)
22	RSLFW-Q-MN-3	0.3756 (> 27-28)
23	RSLFW-Q-MN-2	0.3756 (> 27-28)
24	UDinfo-D-RR-3	0.3705 (> 27-28)
25	RSLFW-Q-MN-1	0.3581 (> 27-28)
26	run.qld-depThre3-D	0.3353 (> 27-28)
27	undisclosed-1	0.0000
28	undisclosed-2	0.0000

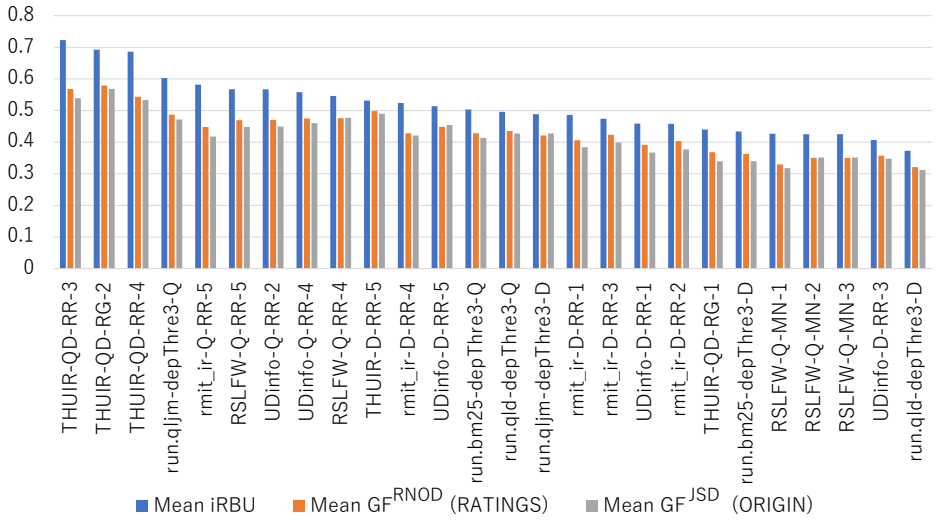

Figure 3: Visualisation of mean scores over the 15 M topics.

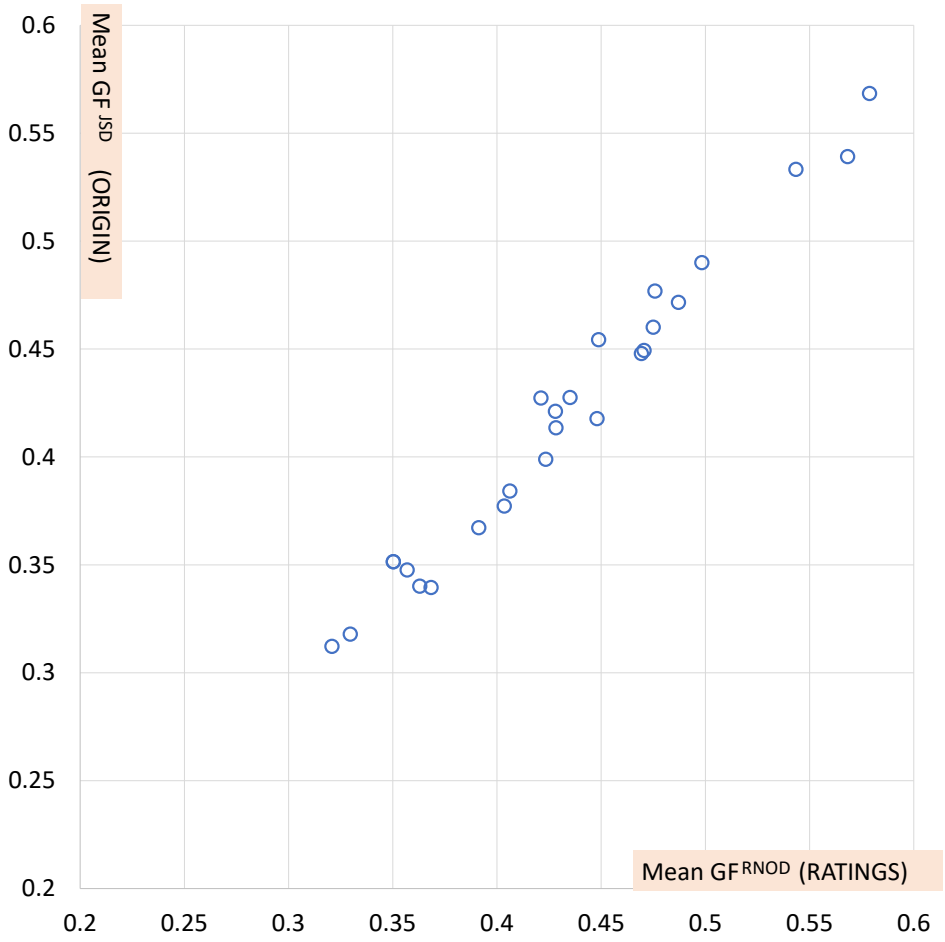

Figure 4: Intersectional group fairness: mean scores over the 15 M topics.

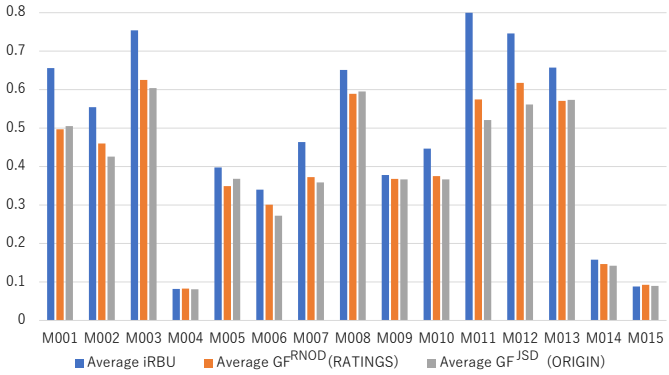

Figure 5: Visualisation of topic difficulty in terms of average score across all runs (M topics).

Table 9: Mean ERR and iRBU scores for each run over the 15 R topics

Run	Mean ERR	Run	Mean iRBU
THUIR-QD-RG-2	0.2638	THUIR-QD-RG-2	0.6560
THUIR-QD-RR-4	0.2460	THUIR-QD-RG-1	0.6013
THUIR-QD-RR-3	0.2276	THUIR-QD-RR-4	0.5957
RSLFW-Q-RR-5	0.2131	THUIR-QD-RR-3	0.5804
run.qljm-depThre3-Q	0.2104	run.qld-depThre3-D	0.5695
run.bm25-depThre3-Q	0.1989	UDinfo-D-RR-3	0.5582
UDinfo-Q-RR-2	0.1989	run.qld-depThre3-Q	0.5518
THUIR-QD-RG-1	0.1918	run.bm25-depThre3-Q	0.5489
run.qld-depThre3-D	0.1749	UDinfo-Q-RR-2	0.5489
UDinfo-D-RR-3	0.1731	RSLFW-Q-RR-5	0.5488
run.qld-depThre3-Q	0.1567	UDinfo-Q-RR-4	0.5367
UDinfo-Q-RR-4	0.1550	THUIR-D-RR-5	0.5351
UDinfo-D-RR-1	0.1532	UDinfo-D-RR-1	0.5055
run.bm25-depThre3-D	0.1509	RSLFW-Q-RR-4	0.4974
RSLFW-Q-RR-4	0.1478	run.qljm-depThre3-Q	0.4971
run.qljm-depThre3-D	0.1459	rmit_ir-Q-RR-5	0.4880
UDinfo-D-RR-5	0.1432	rmit_ir-D-RR-1	0.4816
THUIR-D-RR-5	0.1421	run.bm25-depThre3-D	0.4801
rmit_ir-D-RR-4	0.1374	run.qljm-depThre3-D	0.4361
rmit_ir-Q-RR-5	0.1328	rmit_ir-D-RR-3	0.4133
rmit_ir-D-RR-1	0.1309	UDinfo-D-RR-5	0.4100
rmit_ir-D-RR-3	0.1036	rmit_ir-D-RR-4	0.4063
RSLFW-Q-MN-1	0.1021	RSLFW-Q-MN-1	0.3700
RSLFW-Q-MN-2	0.0890	rmit_ir-D-RR-2	0.3551
rmit_ir-D-RR-2	0.0867	RSLFW-Q-MN-2	0.3084
RSLFW-Q-MN-3	0.0000	RSLFW-Q-MN-3	0.0000
undisclosed-1	0.0000	undisclosed-1	0.0000
undisclosed-2	0.0000	undisclosed-2	0.0000

Table 10: Mean group fairness scores for each run over the 15 R topics. ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$."

Rank	Run	Mean GF^{JSD} (GENDER)	Rank	Run	Mean GF^{NMD} (HINDEX)	Rank	Run	Mean GF^{RNOD} (HINDEX)
1	THUIR-QD-RG-2	0.5831 (>26-28)	1	THUIR-QD-RG-2	0.5841 (>26-28)	1	THUIR-QD-RG-2	0.5352 (>26-28)
2	THUIR-QD-RG-1	0.5823 (>26-28)	2	THUIR-QD-RG-1	0.5569 (>26-28)	2	THUIR-QD-RG-1	0.5257 (>26-28)
3	run.qld-depThre3-D	0.5497 (>26-28)	3	run.qld-depThre3-D	0.5306 (>26-28)	3	run.qld-depThre3-D	0.4975 (>26-28)
4	UDinfo-D-RR-3	0.5374 (>26-28)	4	THUIR-QD-RR-3	0.5247 (>26-28)	4	THUIR-QD-RR-3	0.4875 (>26-28)
5	run.qld-depThre3-Q	0.5356 (>26-28)	5	UDinfo-D-RR-3	0.5195 (>26-28)	5	UDinfo-D-RR-3	0.4866 (>26-28)
6	THUIR-D-RR-5	0.5351 (>26-28)	6	THUIR-QD-RR-4	0.5164 (>26-28)	6	THUIR-D-RR-5	0.4841 (>26-28)
7	UDinfo-Q-RR-4	0.5190 (>26-28)	7	run.qld-depThre3-Q	0.5152 (>26-28)	7	run.qld-depThre3-Q	0.4807 (>26-28)
8	run.bm25-depThre3-Q	0.5096 (>26-28)	8	THUIR-D-RR-5	0.5080 (>26-28)	8	THUIR-QD-RR-4	0.4720 (>26-28)
9	UDinfo-Q-RR-2	0.5096 (>26-28)	9	UDinfo-Q-RR-4	0.4994 (>26-28)	9	UDinfo-Q-RR-4	0.4650 (>26-28)
10	THUIR-QD-RR-4	0.5086 (>26-28)	10	run.bm25-depThre3-Q	0.4977 (>26-28)	10	run.bm25-depThre3-Q	0.4605 (>26-28)
11	THUIR-QD-RR-3	0.4987 (>26-28)	11	UDinfo-Q-RR-2	0.4977 (>26-28)	11	UDinfo-Q-RR-2	0.4605 (>26-28)
12	RSLFW-Q-RR-5	0.4986 (>26-28)	12	RSLFW-Q-RR-5	0.4942 (>26-28)	12	RSLFW-Q-RR-4	0.4562 (>26-28)
13	UDinfo-D-RR-1	0.4985 (>26-28)	13	RSLFW-Q-RR-4	0.4815 (>26-28)	13	RSLFW-Q-RR-5	0.4556 (>26-28)
14	rmit_ir-Q-RR-5	0.4927 (>26-28)	14	rmit_ir-Q-RR-5	0.4778 (>26-28)	14	rmit_ir-Q-RR-5	0.4530 (>26-28)
15	RSLFW-Q-RR-4	0.4886 (>26-28)	15	rmit_ir-D-RR-1	0.4751 (>26-28)	15	rmit_ir-D-RR-1	0.4509 (>26-28)
16	rmit_ir-D-RR-1	0.4819 (>26-28)	16	UDinfo-D-RR-1	0.4682 (>26-28)	16	UDinfo-D-RR-1	0.4434 (>26-28)
17	run.bm25-depThre3-D	0.4694 (>26-28)	17	run.bm25-depThre3-D	0.4400 (>26-28)	17	run.bm25-depThre3-D	0.4155 (>26-28)
18	run.qljm-depThre3-Q	0.4315 (>26-28)	18	run.qljm-depThre3-Q	0.4362 (>26-28)	18	run.qljm-depThre3-Q	0.3999 (>26-28)
19	rmit_ir-D-RR-3	0.4125 (>26-28)	19	run.qljm-depThre3-D	0.4038 (>26-28)	19	run.qljm-depThre3-D	0.3824 (>26-28)
20	run.qljm-depThre3-D	0.4120 (>26-28)	20	rmit_ir-D-RR-3	0.4006 (>26-28)	20	rmit_ir-D-RR-3	0.3815 (>26-28)
21	rmit_ir-D-RR-4	0.3861 (>26-28)	21	rmit_ir-D-RR-4	0.3858 (>26-28)	21	rmit_ir-D-RR-4	0.3613 (>26-28)
22	UDinfo-D-RR-5	0.3829 (>26-28)	22	UDinfo-D-RR-5	0.3765 (>26-28)	22	UDinfo-D-RR-5	0.3554 (>26-28)
23	RSLFW-Q-MN-1	0.3771 (>26-28)	23	RSLFW-Q-MN-1	0.3563 (>26-28)	23	RSLFW-Q-MN-1	0.3395 (>26-28)
24	rmit_ir-D-RR-2	0.3572 (>26-28)	24	rmit_ir-D-RR-2	0.3420 (>26-28)	24	rmit_ir-D-RR-2	0.3255 (>26-28)
25	RSLFW-Q-MN-2	0.3080	25	RSLFW-Q-MN-2	0.3040	25	RSLFW-Q-MN-2	0.2916
26	RSLFW-Q-MN-3	0.0000	26	RSLFW-Q-MN-3	0.0000	26	RSLFW-Q-MN-3	0.0000
27	undisclosed-1	0.0000	27	undisclosed-1	0.0000	27	undisclosed-1	0.0000
28	undisclosed-2	0.0000	28	undisclosed-2	0.0000	28	undisclosed-2	0.0000

Table 11: Mean GFR score for each run over the 15 R topics: relevance based on iRBU; group fairness for HINDEX based on RNOD. (Group fairness for GENDER is based on JSD.) ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$."

Rank	Run	Mean GFR
1	THUIR-QD-RG-2	0.5914 (>26-28)
2	THUIR-QD-RG-1	0.5698 (>26-28)
3	run.qld-depThre3-D	0.5389 (>26-28)
4	UDinfo-D-RR-3	0.5274 (>26-28)
5	THUIR-QD-RR-4	0.5254 (>26-28)
6	run.qld-depThre3-Q	0.5227 (>26-28)
7	THUIR-QD-RR-3	0.5222 (>26-28)
8	THUIR-D-RR-5	0.5181 (>26-28)
9	UDinfo-Q-RR-4	0.5069 (>26-28)
10	run.bm25-depThre3-Q	0.5064 (>26-28)
11	UDinfo-Q-RR-2	0.5064 (>26-28)
12	RSLFW-Q-RR-5	0.5010 (>26-28)
13	UDinfo-D-RR-1	0.4824 (>26-28)
14	RSLFW-Q-RR-4	0.4807 (>26-28)
15	rmit_ir-Q-RR-5	0.4779 (>26-28)
16	rmit_ir-D-RR-1	0.4714 (>26-28)
17	run.bm25-depThre3-D	0.4550 (>26-28)
18	run.qljm-depThre3-Q	0.4428 (>26-28)
19	run.qljm-depThre3-D	0.4101 (>26-28)
20	rmit_ir-D-RR-3	0.4025 (>26-28)
21	rmit_ir-D-RR-4	0.3846 (>26-28)
22	UDinfo-D-RR-5	0.3828 (>26-28)
23	RSLFW-Q-MN-1	0.3622 (>26-28)
24	rmit_ir-D-RR-2	0.3459 (>26-28)
25	RSLFW-Q-MN-2	0.3027
26	RSLFW-Q-MN-3	0.0000
27	undisclosed-1	0.0000
28	undisclosed-2	0.0000

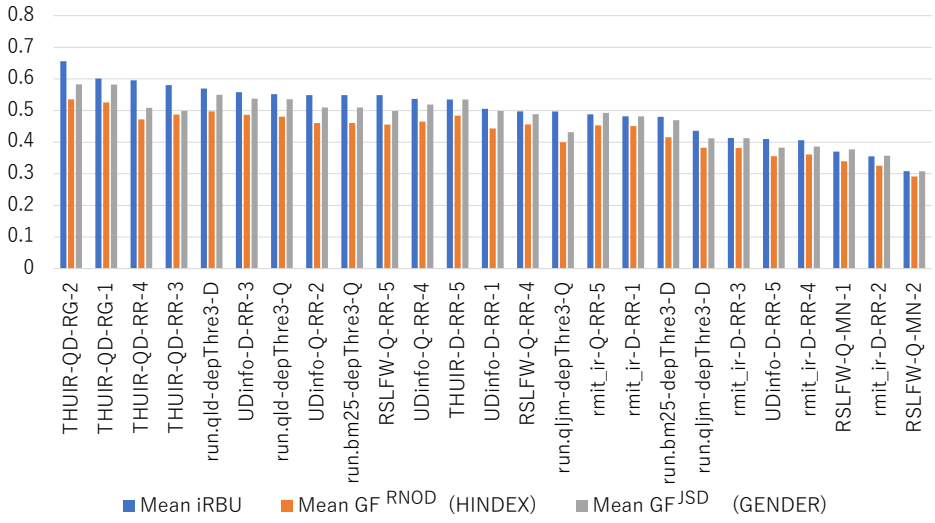

Figure 6: Visualisation of mean scores over the 15 R topics.

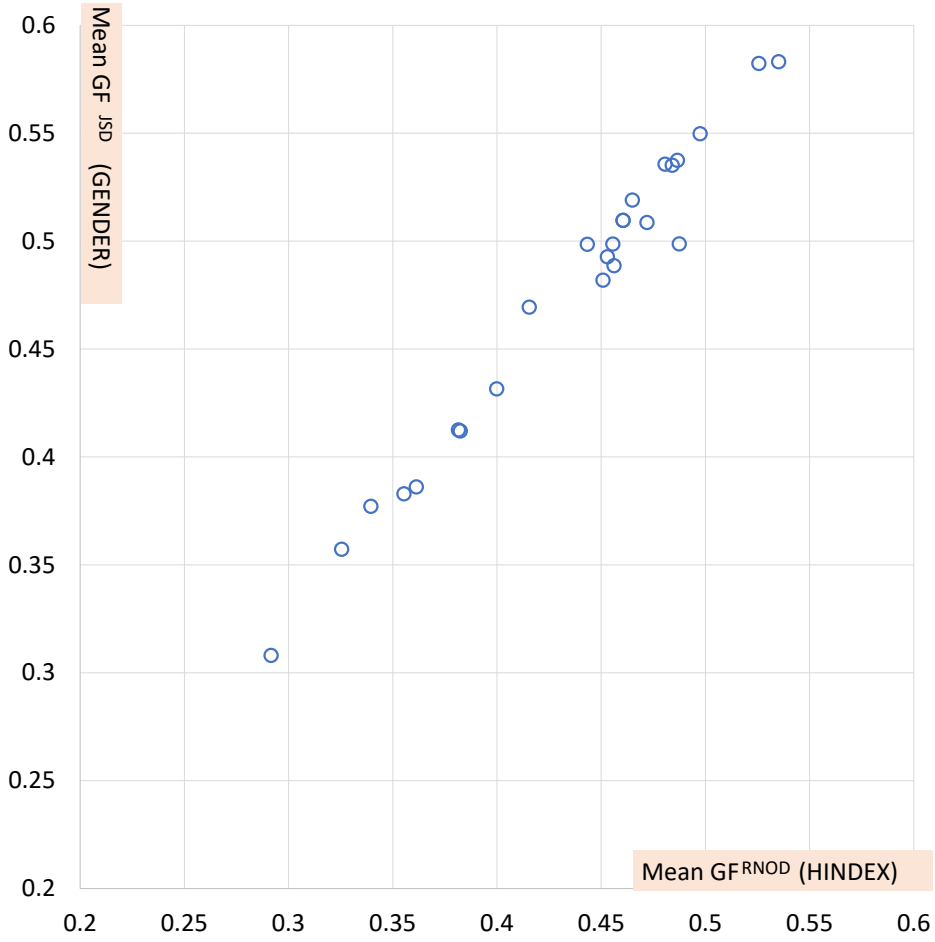

Figure 7: Intersectional group fairness: mean scores over the 15 R topics.

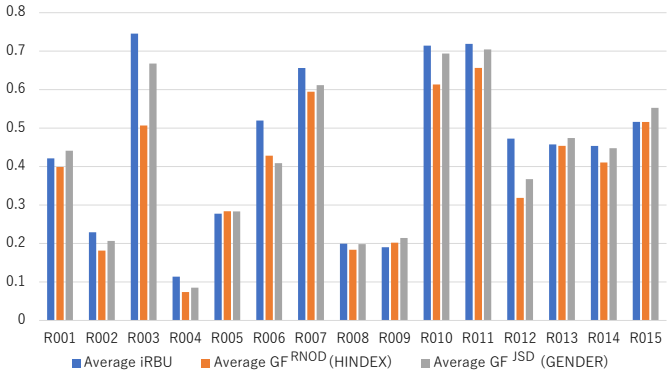

Figure 8: Visualisation of topic difficulty in terms of average score across all runs (R topics).

Table 12: Mean ERR and iRBU scores for each run over the 15 Y topics

Run	Mean ERR	Run	Mean iRBU
THUIR-QD-RR-4	0.1438	THUIR-QD-RR-4	0.4404
RSLFW-Q-RR-5	0.1365	THUIR-QD-RG-1	0.4026
UDinfo-Q-RR-2	0.1357	THUIR-QD-RR-3	0.3919
RSLFW-Q-RR-4	0.1357	UDinfo-Q-RR-2	0.3775
rmit_ir-Q-RR-5	0.1294	RSLFW-Q-RR-4	0.3775
THUIR-QD-RR-3	0.1293	RSLFW-Q-RR-5	0.3757
THUIR-QD-RG-2	0.1144	THUIR-QD-RG-2	0.3749
THUIR-QD-RG-1	0.1099	rmit_ir-Q-RR-5	0.3662
UDinfo-D-RR-5	0.1098	THUIR-D-RR-5	0.3649
THUIR-D-RR-5	0.1009	UDinfo-Q-RR-4	0.3564
UDinfo-D-RR-3	0.0978	UDinfo-D-RR-3	0.3479
rmit_ir-D-RR-1	0.0888	UDinfo-D-RR-5	0.3476
UDinfo-Q-RR-4	0.0887	rmit_ir-D-RR-4	0.3240
rmit_ir-D-RR-4	0.0877	UDinfo-D-RR-1	0.3239
RSLFW-Q-MN-3	0.0822	rmit_ir-D-RR-1	0.3233
RSLFW-Q-MN-2	0.0822	rmit_ir-D-RR-3	0.3177
RSLFW-Q-MN-1	0.0822	rmit_ir-D-RR-2	0.3177
UDinfo-D-RR-1	0.0804	RSLFW-Q-MN-3	0.2562
rmit_ir-D-RR-3	0.0749	RSLFW-Q-MN-2	0.2562
rmit_ir-D-RR-2	0.0749	RSLFW-Q-MN-1	0.2562
run.qljm-depThre3-D	0.0520	run.qld-depThre3-Q	0.2514
run.bm25-depThre3-Q	0.0471	run.qljm-depThre3-D	0.2424
run.qld-depThre3-Q	0.0459	run.bm25-depThre3-Q	0.2202
run.qld-depThre3-D	0.0442	run.qld-depThre3-D	0.2194
run.qljm-depThre3-Q	0.0266	run.qljm-depThre3-Q	0.2010
run.bm25-depThre3-D	0.0266	run.bm25-depThre3-D	0.1735
undisclosed-1	0.0000	undisclosed-1	0.0000
undisclosed-2	0.0000	undisclosed-2	0.0000

Table 13: Mean group fairness scores for each run over the 15 Y topics. ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$."

Rank	Run	Mean GF^{NMD} (SUBSCS)	Rank	Run	Mean GF^{RNOD} (SUBSCS)
1	THUIR-QD-RR-4	0.4112 (>27-28)	1	THUIR-QD-RR-4	0.3809 (>27-28)
2	THUIR-QD-RG-1	0.3830 (>27-28)	2	THUIR-QD-RG-1	0.3638 (>27-28)
3	THUIR-QD-RR-3	0.3601 (>27-28)	3	THUIR-D-RR-5	0.3396 (>27-28)
4	THUIR-D-RR-5	0.3550 (>27-28)	4	THUIR-QD-RR-3	0.3297 (>27-28)
5	THUIR-QD-RG-2	0.3423 (>27-28)	5	UDinfo-Q-RR-4	0.3157 (>27-28)
6	UDinfo-Q-RR-2	0.3315 (>27-28)	6	THUIR-QD-RG-2	0.3141 (>27-28)
7	RSLFW-Q-RR-4	0.3315 (>27-28)	7	UDinfo-D-RR-3	0.3091 (>27-28)
8	UDinfo-Q-RR-4	0.3309 (>27-28)	8	UDinfo-D-RR-5	0.3083 (>27-28)
9	RSLFW-Q-RR-5	0.3292 (>27-28)	9	UDinfo-Q-RR-2	0.3081 (>27-28)
10	UDinfo-D-RR-5	0.3279 (>27-28)	10	RSLFW-Q-RR-4	0.3081 (>27-28)
11	UDinfo-D-RR-3	0.3228 (>27-28)	11	RSLFW-Q-RR-5	0.3058 (>27-28)
12	rmit_ir-Q-RR-5	0.3169 (>27-28)	12	rmit_ir-D-RR-3	0.3025 (>27-28)
13	UDinfo-D-RR-1	0.3157 (>27-28)	13	rmit_ir-D-RR-2	0.3025 (>27-28)
14	rmit_ir-D-RR-3	0.3146 (>27-28)	14	UDinfo-D-RR-1	0.3017 (>27-28)
15	rmit_ir-D-RR-2	0.3146 (>27-28)	15	rmit_ir-D-RR-4	0.2945 (>27-28)
16	rmit_ir-D-RR-4	0.3100 (>27-28)	16	rmit_ir-D-RR-1	0.2928 (>27-28)
17	rmit_ir-D-RR-1	0.3084 (>27-28)	17	rmit_ir-Q-RR-5	0.2915 (>27-28)
18	run.qld-depThre3-Q	0.2451	18	run.qld-depThre3-Q	0.2391
19	run.qljm-depThre3-D	0.2425	19	run.qljm-depThre3-D	0.2329
20	RSLFW-Q-MN-3	0.2339	20	RSLFW-Q-MN-3	0.2201
21	RSLFW-Q-MN-2	0.2339	21	RSLFW-Q-MN-2	0.2201
22	RSLFW-Q-MN-1	0.2339	22	RSLFW-Q-MN-1	0.2201
23	run.qld-depThre3-D	0.2155	23	run.qld-depThre3-D	0.2100
24	run.bm25-depThre3-Q	0.2112	24	run.bm25-depThre3-Q	0.2039
25	run.qljm-depThre3-Q	0.2071	25	run.qljm-depThre3-Q	0.2038
26	run.bm25-depThre3-D	0.1777	26	run.bm25-depThre3-D	0.1731
27	undisclosed-1	0.0000	27	undisclosed-1	0.0000
28	undisclosed-2	0.0000	28	undisclosed-2	0.0000

Table 14: Mean GFR score for each run over the 15 Y topics: relevance based on iRBU; group fairness for SUBSCS based on RNOD. ">" means "stastistically significantly outperforms (according to a randomised Tukey HSD test with $B = 5,000$ trials, $\alpha = 0.05$."

Rank	Run	Mean GFR
1	THUIR-QD-RR-4	0.4107 (>27-28)
2	THUIR-QD-RG-1	0.3832 (>27-28)
3	THUIR-QD-RR-3	0.3608 (>27-28)
4	THUIR-D-RR-5	0.3523 (>27-28)
5	THUIR-QD-RG-2	0.3445 (>27-28)
6	UDinfo-Q-RR-2	0.3428 (>27-28)
7	RSLFW-Q-RR-4	0.3428 (>27-28)
8	RSLFW-Q-RR-5	0.3408 (>27-28)
9	UDinfo-Q-RR-4	0.3361 (>27-28)
10	rmit_ir-Q-RR-5	0.3288 (>27-28)
11	UDinfo-D-RR-3	0.3285 (>27-28)
12	UDinfo-D-RR-5	0.3280 (>27-28)
13	UDinfo-D-RR-1	0.3128 (>27-28)
14	rmit_ir-D-RR-3	0.3101 (>27-28)
15	rmit_ir-D-RR-2	0.3101 (>27-28)
16	rmit_ir-D-RR-4	0.3092 (>27-28)
17	rmit_ir-D-RR-1	0.3080 (>27-28)
18	run.qld-depThre3-Q	0.2453
19	RSLFW-Q-MN-3	0.2381
20	RSLFW-Q-MN-2	0.2381
21	RSLFW-Q-MN-1	0.2381
22	run.qljm-depThre3-D	0.2377
23	run.qld-depThre3-D	0.2147
24	run.bm25-depThre3-Q	0.2121
25	run.qljm-depThre3-Q	0.2024
26	run.bm25-depThre3-D	0.1733
27	undisclosed-1	0.0000
28	undisclosed-2	0.0000

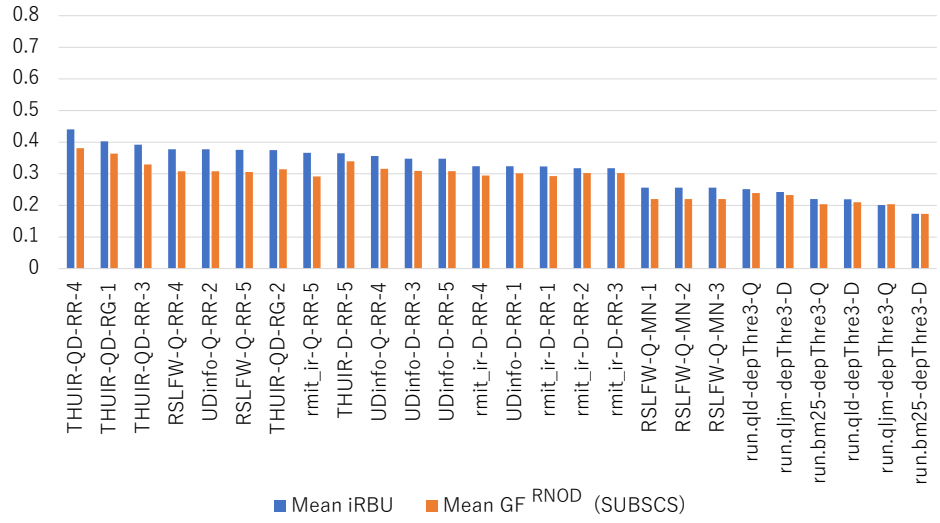

Figure 9: Visualisation of mean scores over the 15 Y topics.

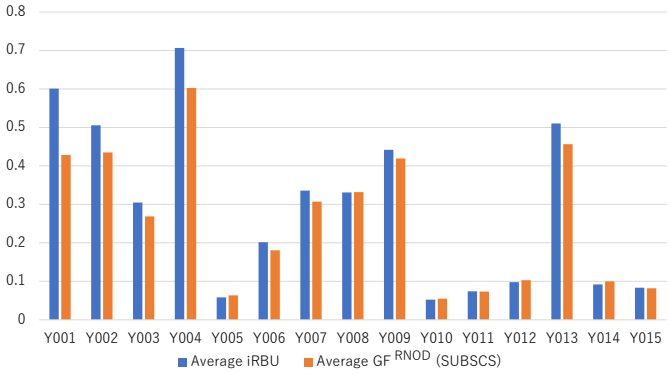

Figure 10: Visualisation of topic difficulty in terms of average score across all runs (Y topics).

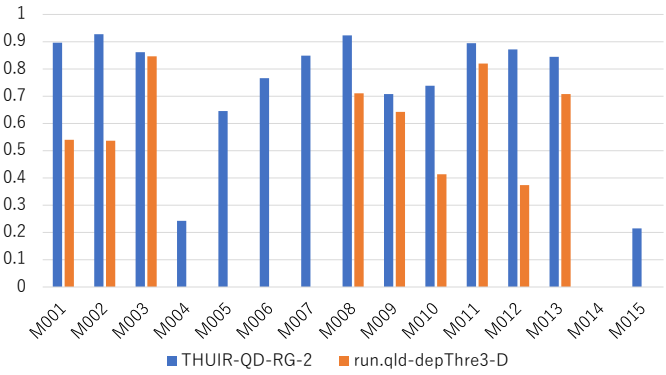

Figure 11: THUIR-QD-RG-2 vs. run.qld-depThre3-D over the M topics (iRBU).

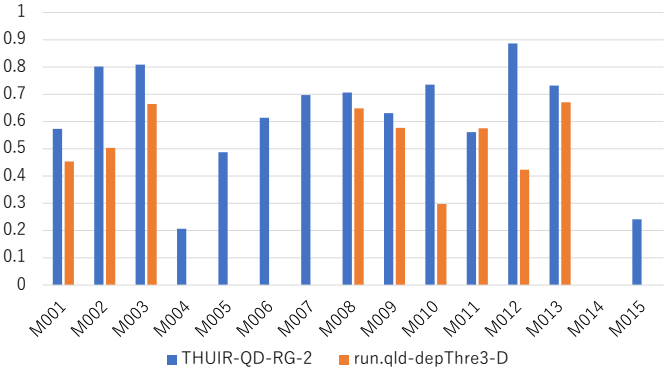

Figure 12: THUIR-QD-RG-2 vs. run.qld-depThre3-D over the M topics (GF^{RNOD} for RATINGS).

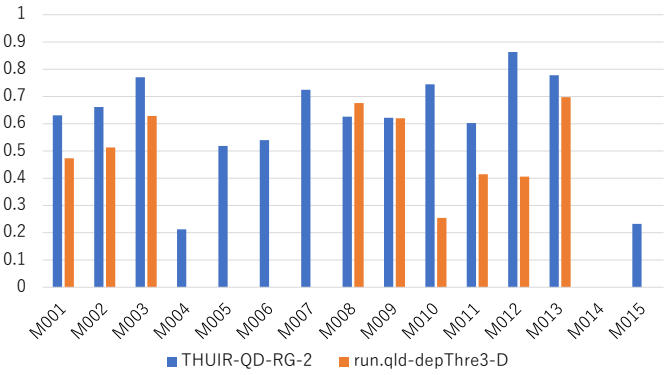

Figure 13: THUIR-QD-RG-2 vs. run.qld-depThre3-D over the M topics (GF^{JSD} for ORIGIN).

Table 15: Evaluating the SERP of THUIR-QD-RG-2 for Topic M012.

Rank r	Relevance level	$P_{\text{ERR}}(r)$	RATINGS Membership	RATINGS DistrSim(r) (RNOD-based)	ORIGIN Membership	ORIGIN DistrSim(r) (JSD-based)
1	L0	0	0.25 per group		0.125 per group	
2	L0	0	0.25 per group		0.125 per group	
3	L0	0	0.25 per group		0.125 per group	
4	L0	0	0.25 per group		0.125 per group	
5	L0	0	0.25 per group		0.125 per group	
6	L0	0	0.25 per group		0.125 per group	
7	L1	0.2500	0.33/0.67/0/0	0.9519	0/0.5/0/0.5/0/0/0/0	0.9259
8	L0	0	0.25 per group		0.125 per group	
9	L1	0.1875	0/0.67/0.33/0	0.9315	0/0.67/0/0/0.33/0/0	0.9249
10	L1	0.1406	0/0.33/0.67/0	0.9182	0/1/0/0/0/0/0/0	0.9031
11	L1	0.1055	0/1/0/0	0.8833	0/1/0/0/0/0/0/0	0.8799
12	L1	0.0791	0/0/1/0	0.8805	0/0.78/0/0.11/0/0.11/0/0	0.8668
13	L1	0.0593	0/0.33/0.67/0	0.8666	0/0.83/0/0.17/0/0/0/0	0.8511
14	L0	0	0.25 per group		0.125 per group	
15	L1	0.0445	1/0/0/0	0.8963	0/1/0/0/0/0/0/0	0.8427
16	L1	0.0334	1/0/0/0	0.9005	0/1/0/0/0/0/0/0	0.8253
17	L1	0.0250	0/0.50/0.50/0	0.8926	0/1/0/0/0/0/0/0	0.8089
18	L1	0.0188	1/0/0/0	0.8895	0/1/0/0/0/0/0/0	0.7935
19	L1	0.0141	0/0/1/0	0.8846	0/1/0/0/0/0/0/0	0.7789
20	L1	0.0106	1/0/0/0	0.8783	0/1/0/0/0/0/0/0	0.7653
			GF^{RNOD} (RATINGS)	0.8867	GF^{JSD} (ORIGIN)	0.8630

Table 16: Evaluating the SERP of run.qld-depThre3-D for Topic M012.

Rank r	Relevance level	$P_{\text{ERR}}(r)$	RATINGS Membership	RATINGS DistrSim(r) (RNOD-based)	ORIGIN Membership	ORIGIN DistrSim(r) (JSD-based)
1	L0	0	0.25 per group		0.125 per group	
2	L0	0	0.25 per group		0.125 per group	
3	L0	0	0.25 per group		0.125 per group	
4	L0	0	0.25 per group		0.125 per group	
5	L0	0	0.25 per group		0.125 per group	
6	L0	0	0.25 per group		0.125 per group	
7	L0	0	0.25 per group		0.125 per group	
8	L0	0	0.25 per group		0.125 per group	
9	L0	0	0.25 per group		0.125 per group	
10	L0	0	0.25 per group		0.125 per group	
11	L0	0	0.25 per group		0.125 per group	
12	L0	0	0.25 per group		0.125 per group	
13	L0	0	0.25 per group		0.125 per group	
14	L1	0.2500	0/0/1/0	0.9628	0/0.78/0/0.11/0/0.11/0/0	0.9276
15	L0	0	0.25 per group		0.125 per group	
16	L0	0	0.25 per group		0.125 per group	
17	L0	0	0.25 per group		0.125 per group	
18	L1	0.1875	0.33/0.67/0/0	0.9733	0/0.5/0/0.5/0/0/0/0	0.9273
19	L0	0	0.25 per group		0.125 per group	
20	L0	0	0.25 per group		0.125 per group	
			GF^{RNOD} (RATINGS)	0.4232	GF^{JSD} (ORIGIN)	0.4058

5 CONCLUSIONS

This paper reported on the official results of the NTCIR-17 FairWeb-1 task. The most remarkable results are as follows.

- In terms of relevance measures, the THUIR runs (e.g. THUIR-QD-RR-4 - a "PM2 and dense relevance" run [23]) perform well on the entire topic set, and on each of the three topic subsets.
- In terms of group fairness (GF) measures, the THUIR runs perform well as well, on all three topic subsets (M, R, Y), with all attribute sets.
- Hence, in terms of GFR as well, the THUIR runs are the winners (in terms of average performance) on all three topic subsets.

We also demonstrated how the GFR framework [19] works using a THUIR run and a baseline run.

We plan to propose the FairWeb-2 task for NTCIR-18 and to continue studying the group-fair web search problem, in particular to see whether we can reproduce the best THUIR runs on new data and even improve upon them. Furthermore, we plan to introduce a group-fair *conversational search* subtask, which requires systems to return textual turns instead of ranked lists of web pages. More details can in be found in Sakai [17].

DISCLAIMER

Certain companies and products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products or companies identified are necessarily the best available for the purpose.

REFERENCES

[1] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter?. In *Proceedings of ACM SIGIR 2008*. 667–674.

[2] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of ACM CIKM 2009*. 621–630.

[3] Fumian Chen and Hui Fang. 2023. UDInfoLab at the NTCIR-17 FairWeb-1 Task. In *Proceedings of NTCIR-17*. https://doi.org/10.20736/0002001300

[4] Sachin Pathiyan Cherumanal, Kaixin Ji, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 2023. RMIT_IR at the NTCIR-17 FairWeb-1 Task. In *Proceedings of NTCIR-17*. https://doi.org/10.20736/0002001315

[5] Zhumin Chu, Tetsuya Sakai, Qingyao Ai, and Yiqun Liu. 2023. Chuweb21D: A Deduped English Document Collection for Web Search Tasks. In *In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. ACM.

[6] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the TREC 2022 Fair Ranking Track. In *Proceedings of TREC 2022 (NIST Special Publication)*.

[7] Fan Li, Kaize Shi, Nuo Chen Kenta Inaba, Sijie Tao, and Tetsuya Sakai. 2023. RSLFW at the NTCIR-17 FairWeb-1 Task. In *Proceedings of NTCIR-17*. https://doi.org/10.20736/0002001303

[8] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A large English news corpus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3077–3084.

[9] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*. 141–150.

[10] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D. Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 269–279. https://doi.org/10.1145/3576840.3578316

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints* (2019). arXiv:1910.10683

[12] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*. 116–163.

[13] Tetsuya Sakai. 2018. Comparing Two Binned Probability Distributions for Information Access Evaluation. In *Proceedings of ACM SIGIR 2018*. 1073–1076.

[14] Tetsuya Sakai. 2018. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer. https://link.springer.com/book/10.1007/978-981-13-1199-4

[15] Tetsuya Sakai. 2020. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online). Association for Computational Linguistics, 2759–2769.

[16] Tetsuya Sakai. 2021. A Closer Look at Evaluation Measures for Ordinal Quantification. In *Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management (CIKM 2021)*. https://ceur-ws.org/Vol-3052/paper21.pdf

[17] Tetsuya Sakai. 2023. Fairness-based Evaluation of Conversational Search: A Pilot Study. In *Proceedings of EVIA 2023*.

[18] Tetsuya Sakai and Noriko Kando. 2008. Are Popular Documents More Likely to be Relevant? A Dive into the ACLIA IR4QA Pools. In *Proceedings of EVIA 2008*. 8–9.

[19] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A Versatile Framework for Evaluating Ranked Lists in terms of Group Fairness and Relevance. *ACM TOIS* (2023). https://doi.org/10.1145/3589763

[20] Tetsuya Sakai, Sijie Tao, Nuo Chen, Yujing Li, Maria Maistro, Zhumin Chu, and Nicola Ferro. 2023. On the Ordering of Pooled Web Pages, Gold Assessments, and Bronze Assessments. *ACM TOIS* (2023). https://dl.acm.org/doi/pdf/10.1145/3600227

[21] Tetsuya Sakai, Sijie Tao, Zhumin Chu, Maria Maistro, Yujing Li, Nuo Chen, Nicola Ferro, Junjie Wang, Ian Soboroff, and Yiqun Liu. [n.d.]. Overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) Task. In *Proceedings of NTCIR-16*. 234–235. https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/01-NTCIR16-OV-WWW-SakaiT.pdf

[22] Tetsuya Sakai and Zhaohao Zeng. 2020. Retrieval Evaluation Measures that Agree with Users' SERP Preferences: Traditional, Preference-based, and Diversity Measures. *ACM TOIS* 39, 2 (2020), Article 14.

[23] Yiteng Tu, Haitao Li, Zhumin Chu, Qingyao Ai, and Yiqun Liu. 2023. THUIR at the NTCIR-17 FairWeb-1 Task: An Initial Exploration of the Relationship Between Relevance and Fairness. In *Proceedings of NTCIR-17*. https://doi.org/10.20736/0002001317