RMIT_IR at the NTCIR-17 FairWeb-1 Task

Sachin Pathiyan Cherumanal
RMIT University
Melbourne, Australia
s3874326@student.rmit.edu.au

Kaixin Ji
RMIT University
Melbourne, Australia
kaixin.ji@student.rmit.edu.au

Danula Hettiachchi
RMIT University
Melbourne, Australia
danula.hettiachchi@rmit.edu.au

Johanne R. Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

ABSTRACT

This report describes the participation of the RMIT IR group at the NTCIR-17 FairWeb-1 task. We submitted five runs with the aim of exploring the role of explicit search result diversification (SRD) and ranking fusion to generate fair rankings considering multiple fairness attributes. We also explored the use of a linear combination-based technique (LC) to take into consideration the relevance while re-ranking. In this report, we compared results from all our submitted runs against each other and the retrieval baselines along each topic type separately (i.e., Researcher, Movie, YouTube). Overall, our results show that neither the SRD-based runs nor the linear combination-based runs show any statistically significant improvement over the retrieval baselines. The source code of the framework for generating group memberships is made available at https://github.com/rmit-ir/fairweb-1.

KEYWORDS

information retrieval, diversification, fairness

TEAM NAME

RMIT_IR

SUBTASKS

FairWeb-1 Task

1 INTRODUCTION

Information access systems such as search engines have been effective in assisting users with information needs in decision-making processes. Given the impact of such systems in daily life, search engines not only must provide relevant information to the user, but also fair exposure to diverse information [8]. For instance, the general chairs of an information retrieval conference aiming to curate a diverse organizing committee should obtain search results for the query "information retrieval researchers" that are diverse in a number of dimensions associated with protected attributes (e.g., gender and geographical location).

Recent efforts in organizing evaluation campaigns such as the TREC 2022 Fair Ranking Track and the NTCIR-17 FairWeb-1 Task[1] address the problem of fairness-aware information retrieval. While the TREC 2022 Fair Ranking Track focuses on multi-attribute fairness (i.e., achieving a balance of relevance and fairness across multiple attributes), the NTCIR-17 FairWeb-1 Task goes a step further and considers the characteristics of each attribute. The differences between the evaluation metrics used at the TREC 2022 Fair Ranking Track and this track have been discussed by Sakai et al. [16]. For instance, some attributes may not just be categorical but also ordinal, as in the case of a researcher's *h*-index when a user searches for a specific researcher's profile [18].

This paper describes the participation of the RMIT IR Group at the NCTIR-17 FairWeb-1 Task. Our runs aim to explore the role of Search Result Diversification (*SRD*) and ranking fusion in the context of multi-attribute group fairness. In information retrieval, fairness and diversity have been studied side-by-side over the recent years [9, 13, 16]. *SRD* can be seen as a mechanism to strike a balance between diversity and relevance in a ranked list [12], so following some of our previous work, we wanted to explore how *SRD* techniques achieved the balance between group fairness and relevance along nominal and ordinal fairness attributes. For this task, we use a proportionality-based explicit SRD technique namely *PM-2*. We also explored a linear combination-based technique (LC) inspired by an implicit *SRD* method called Maximal Marginal Relevance (MMR) [3]. Furthermore, this task includes multiple fairness attributes for two of the topic types – i.e., Researcher-related Topic (gender and h-index) and Movie-related Topic (origin and rating). For each of these topic types, we first re-rank using *SRD* and then perform ranking fusion using Reciprocal Ranking Fusion (RRF) [5] to combine the diversified rankings into one final fairness-aware ranking.

The rest of the paper is organized as follows. Section 2.1 discusses the method we used to create a membership file for each fairness attribute. Next, in Section 2.2, we provide detailed information on the SRD techniques, the rationale behind using them for this task, the retrieval runs used for re-ranking, and the ranking fusion used. In total, we submitted five systems of varying parameters and in Section 3 we discuss the results of these submitted runs and draw conclusions.

2 PROPOSED APPROACH

Before creating our runs for NTCIR FairWeb-1, we had to generate the membership files for each fairness group that belonged to one of the three topic types: Researcher-related (R), Movie-related (R), and YouTube-related (R). This file contained the association of a document in the ranking to the different fairness aspects of a group. For further details about the fairness attributes, please refer to the task overview paper [18].

[1] Hereafter, we use 'track' and 'task' interchangeably.

2.1 Membership Generation

We used `BeautifulSoup4`[2] to clean the document HTML files. We have made available a custom-written framework to perform entity extraction and generate the membership files for this task.

2.1.1 Researcher-related Topic (R). The Researcher-related topic type included two fairness attributes, *gender* (nominal) and the *h-index* (ordinal). Given the nature of the topic and the attributes, we first extracted entities from the documents that referred to a 'PERSON'. To do this, we used `SpaCy`[3] to extract entity text that belonged to the entity label 'PERSON'. The downstream task used these entity texts to extract gender and *h*-index (see Figure 1).

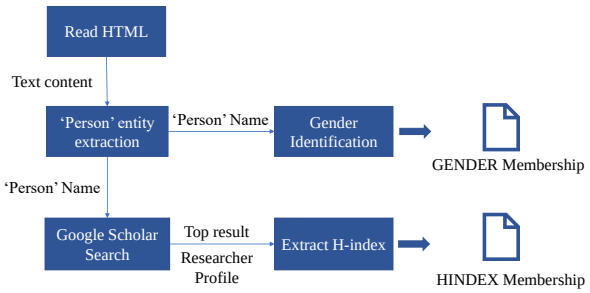

Figure 1: Membership Generation Process for R-Topic.

Gender. The *gender* attribute involves the following three groups: *he*, *she*, and *other*. To identify the gender representations in each document, we used the Python package *gender_guesser*[4], which predicts the gender using the first name. It may be worth noting that using gender_guesser may have its own ethical implications.

h-index. The *h-index* attribute involves the following five groups: $x < 10$, $10 \leq x < 30$, $30 \leq x < 50$, $50 \leq x$. If no results were found, we assigned $x = 0$. To extract the *h*-index, we used the Python package scholarly[5] that facilitates retrieval of author and publication information from Google Scholar. We used the first search result and the associated *h*-index for simplicity.

2.1.2 Movie-related Topic (M).

Origin. The *origin* attribute involves the following eight groups: Africa, America, Antarctica, Asia, the Caribbean, Europe, the Middle East, and Oceania. We use regular expressions to search for the following pattern:

r "\ n\n (?: Country|Countries of origin):\ s *(.*?)\ n\n".

We use `pycountry`[6] to normalize country names and extract ISO 3166-1 alpha-2 codes, and `pycountry-convert`[7] to convert the codes to continent codes and names. In this process we needed to manually correct some country names written differently than in the package, e.g., USA has converted to US and UK, England, United Kingdom has converted to GB (the alpha-2 codes for the United Kingdom). Besides, we have combined the continents North America and South America into America.

Rating. The *rating* attribute involves the following four groups: $x < 100$, $100 \leq x < 10k$, $10k \leq x < 1m$, $1m \leq x$. We first parse the HTML file using Python package `BeautifulSoup4`,[8] then find the object has either an itemprop attribute with the value 'ratingCount' or a class attribute with the value 'wpd-rating-value'.

For the former case, because the corresponding HTML elements are different for sites, we extract the number from content if it is a <meta> element or from text if it is a element. For the latter, we find another inner object with a class attribute with the value 'wpdrc'. It is important to note that the resulting text must be formatted only to include integers such as 1000 instead of '1,000'.

2.1.3 YouTube-related Topic (Y).

Subscriber. The *subscriber* attribute involves the following four groups: $x < 100$, $100 \leq x < 10k$, $10k \leq x < 1m$, $1m \leq x$. We first parse the HTML file using the `BeautifulSoup4`, similarly as in the previous topics. We identified three conditions to find the number of subscribers: *followers*, *comments*, and *ratings*.

(1) **Followers:** Search if it contains a class_name attribute with any of the following values,
   - 'essb–fc–network–total_followers'
   - 'follower_count'
   - 'followers –num'
   - 'jetpack – subscribe – count'
   - 'js –question–follower–count'
   - 'subscribe –counter'
   - 'mh–comment–count–link'

   Then we look for *'follower|subscrib|other'* in the found object text and extract the associated value (more details are available in the code).
(2) **Comments:** Find the total number of comments as a reference for the subscribers. To find the value, we search within the contents of a script element and assess the values with predefined keys ['@graph'][3]['commentCount'].
(3) **Ratings:** The number of ratings. Instead of parsing and searching the HTML elements, we simply use a regular expression pattern r '" ratingCount ":"(\ d +)",') within the text string.

We believe that the sparsity of the memberships may affect our downstream ranking stage. So, we also report the measure of sparsity [2] in our generated membership file using the following metric $S = \frac{N_0}{N}$, where S is the sparsity measure, N_0 is the number of zero in the file and N is the total number of elements. Table 1 shows that the most populated membership was for the R-GENDER and the least populated was M-ORIGIN. We believe that evaluating the sparsity or membership associations between the submitted runs and those used to evaluate the runs may help us level the playing field and compare the different systems submitted to this task at a much granular level.

[2] https://pypi.org/project/beautifulsoup4/
[3] https://pypi.org/project/spacy/
[4] https://pypi.org/project/gender-guesser/
[5] https://pypi.org/project/scholarly/
[6] https://pypi.org/project/pycountry/
[7] https://pypi.org/project/pycountry-convert/
[8] https://pypi.org/project/beautifulsoup4/

Table 1: Sparsity of membership files for each attribute. Higher values of S indicate a more sparse membership file with a larger proportion of zeroes or missing values.

Attribute	Sparsity (S)
R-GENDER	41.8009
R-HINDEX	55.6857
M-ORIGIN	87.2221
M-RATINGS	75.0000
Y-SUBSCS	75.0000

2.2 Re-Ranking

For the retrieval stage, we used the BM25 (description) and BM25 (query) based runs provided by the organizers of NTCIR Fairweb-1. Since some of the topic types involved multiple attributes, we followed Pathiyan Cherumnal et al. [15] and fused the diversified ranking using Reciprocal Ranked Fusion (RRF) [5] from *polyfuse*.[9]

2.2.1 Search Results Diversification. We wanted to understand to what extent explicit search results diversification methods can diversify a ranking given a fairness attribute. Given a fairness attribute, values for that particular attribute are treated as *aspects* or subtopics, aligned with popular literature on *SRD* [6, 7, 10, 19]. For instance, in this task, we treat *gender* as an attribute with *he*, *she*, and *other* as the aspects of the attribute. We were interested in achieving statistical parity through our re-ranking, so we decided to investigate the proportionality-based SRD technique, i.e., PM-2 proposed by Dang and Croft [7]. PM-2 iteratively picks the best aspect that maintains overall proportionality and then selects the best document for each position in the diversified list [7]. Using PM-2, we then performed diversification of the candidate list along each of the attributes. Then we fused them using RRF for each of the topic types *Researcher*, *Movie*, and *YouTube* (as shown by the Eqs. 1, 2 and 3 respectively).

$$P_R = \text{RRF}\left(\text{PM-2}(r, \text{GENDER}), \text{PM-2}(r, \text{HINDEX})\right) \tag{1}$$

$$P_M = \text{RRF}\left(\text{PM-2}(r, \text{ORIGIN}), \text{PM-2}(r, \text{RATINGS})\right) \tag{2}$$

$$P_Y = \text{PM-2}(r, \text{SUBSCS}) \tag{3}$$

2.2.2 Linear Combination. For the second technique we used, we took inspiration from an implicit SRD technique called Maximal Marginal Relevance (MMR) [3] similar to yet different from the adaptations proposed by McDonald et al. [11], Pathiyan Cherumanal et al. [14]. MMR aimed to maximise a ranked list's novelty, diversity, and relevance. However, we adapted MMR to maximise the fairness and relevance of a ranked list.

In this re-ranking method, we use a linear combination (LC) approach, which tries to maximise the fairness and relevance of a ranked list. Similar to MMR, we use a parameter λ that takes a value in the range of [0,1].

$$\text{LC} = ((1 - \lambda) * R) + (\lambda * F) \tag{4}$$

[9] https://github.com/rmit-ir/polyfuse

Where R is the normalised relevance score for a document from the retrieved ranked list. λ is the parameter that helps us set preferred weights. This means the LC technique would give minimal weight to fairness when $\lambda = 0$ and maximised fairness when $\lambda = 1$. The F in Eq. 4 refers to the distance we measure between the membership distribution (m_d) and the target distribution (T_d). In Eq. 5, RNOD refers to the Root Normalised Order-aware Divergence, and NMD refers to the Normalised Match Distance.

$$F = \begin{cases} 1 - \text{mean}(\text{RNOD}, \text{NMD}) & \text{if A is ordinal} \\ \text{JSD} & \text{otherwise} \end{cases} \tag{5}$$

When the attribute,A is ordinal, F would be computed as the mean of RNOD(m_d, T_d) and NMD(m_d, T_d). When A is nominal, F would be computed as the Jensen-Shannon Distance (JSD) between m_d and T_d. The metrics have been discussed in further detail in Sakai et al. [16]. The Eqs. 6, 7 and 8 denote the RRF-based fusion we performed across the three topic types *Researcher*, *Movie*, and *YouTube*.

$$L_R = \text{RRF}\left(\text{LC}(r, \text{Gender}), \text{LC}(r, \text{HINDEX})\right) \tag{6}$$

$$L_M = \text{RRF}\left(\text{LC}(r, \text{ORIGIN}), \text{LC}(r, \text{RATINGS})\right) \tag{7}$$

$$L_Y = \text{LC}(r, \text{SUBSCS}) \tag{8}$$

3 RESULTS AND DISCUSSION

In this section, we report the results of the five submitted runs and compare them against the retrieval baselines.

- **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 $\lambda = 0.5$

We report the official effectiveness measures used in the NTCIR FairWeb-1. Effectiveness in terms of Relevance is measured via Expected Reciprocal Rank (ERR) [4] and intentwise RBU (iRBU) [17] – an adaptation for the ad-hoc retrieval scenario of the Rank-Biased Utility (RBU) proposed by Amigó et al. [1]. Fairness is evaluated for each topic type using GFR Score[10] that comprises of JSD (for nominal attributes), NMD and RNOD (for ordinal attributes). To calculate statistical significance, the organizers used a randomized Tukey HSD test with $B = 5,000 trials$, $\alpha = 0.05$. In the results shown in Tables 2, 3, 4 and 5, we could not observe statistically significant improvement for the submitted systems across multiple fairness and relevance measures.

For the fairness scores, we discuss the results from our submitted runs for each topic type, i.e., (R, M, and Y) separately, as done by the organizers in the overview paper. We used two baselines for our runs, BM25(Q) and BM25(D). We discuss how our submitted runs compare against them.

[10] More details about the evaluation measures available in the overview paper [18].

Table 2: Submitted run and relevance scores. The highlighted values represent the highest scores achieved. The statistical significance was calculated using a randomized Tukey HSD test with $B = 5,000 trials$, $\alpha = 0.05$. No statistically significant differences were observed.

System	Mean ERR	Mean iRBU
BM25(D)	0.1113	0.3624
BM25(Q)	0.1390	0.4242
rmit_ir-D-RR-1	0.1306	0.4304
rmit_ir-D-RR-2	0.1029	0.3769
rmit_ir-D-RR-3	0.1084	0.4017
rmit_ir-D-RR-4	0.1379	0.4181
rmit_ir-Q-RR-5	**0.1685**	**0.4787**

Table 3: R Topic. The highlighted values represent the best scores achieved. The statistical significance was calculated using a randomized Tukey HSD test with $B = 5,000 trials$, $\alpha = 0.05$. No statistically significant differences were observed.

System	Mean GF^{JSD}	Mean GF^{NMD}	Mean GF^{RNOD}
BM25(D)	0.4694	0.4400	0.4155
BM25(Q)	0.5096	**0.4977**	**0.4605**
rmit_ir-D-RR-1	0.4819	0.4751	0.4509
rmit_ir-D-RR-2	0.3572	0.3420	0.3255
rmit_ir-D-RR-3	0.4125	0.4006	0.3815
rmit_ir-D-RR-4	0.3861	0.3858	0.3613
rmit_ir-Q-RR-5	**0.4927**	0.4778	0.4530

Table 4: M Topic. The highlighted values represent the best scores achieved. The statistical significance was calculated using a randomized Tukey HSD test with $B = 5,000 trials$, $\alpha = 0.05$. No statistically significant differences were observed.

System	Mean GF^{JSD}	Mean GF^{NMD}	Mean GF^{RNOD}
BM25(D)	0.3401	0.3993	0.3630
BM25(Q)	0.4135	0.4623	0.4283
rmit_ir-D-RR-1	0.3842	0.4502	0.4062
rmit_ir-D-RR-2	0.3772	0.4309	0.4035
rmit_ir-D-RR-3	0.3989	0.4529	0.4234
rmit_ir-D-RR-4	**0.4211**	0.4784	0.4281
rmit_ir-Q-RR-5	0.4177	**0.5043**	**0.4480**

Relevance Scores. Table 2 shows the mean relevance scores of the submitted systems, and we see that only *rmit_ir-Q-RR-5* outperforms the baseline retrieval runs that we used.

Regarding R-Topic (from Table 3), we see that *rmit_ir-Q-RR-5* outperforms the baseline only for the Mean GF^{JSD}. It is worth noting that *rmit_ir-Q-RR-5* does not outperform the baseline along Mean GF^{NMD} and Mean GF^{RNOD}. However, this may have been due to errors propagating from our membership generation step for this particular topic-type.

Table 5: Y Topic. The highlighted values represent the highest scores achieved. The statistical significance was calculated using a randomized Tukey HSD test with $B = 5,000 trials$, $\alpha = 0.05$. No statistically significant differences were observed.

System	Mean GF^{NMD}	Mean GF^{RNOD}
BM25(D)	0.1777	0.1731
BM25(Q)	0.2112	0.2039
rmit_ir-D-RR-1	0.3084	0.2928
rmit_ir-D-RR-2	0.3146	**0.3025**
rmit_ir-D-RR-3	0.3146	**0.3025**
rmit_ir-D-RR-4	0.3100	0.2945
rmit_ir-Q-RR-5	**0.3169**	0.2915

Regarding M-Topic (from Table 4), we see that all of our submitted runs outperform their respective baseline runs across all the fairness scores. *rmit_ir-Q-RR-5* seems to be the best performing across Mean GF^{NMD} and and Mean GF^{RNOD} and *rmit_ir-D-RR-4* was the best performing along Mean GF^{JSD}.

In the case of the Y-Topic (from Table 5), it constitutes only an ordinal attribute, so only Mean GF^{NMD} and Mean GF^{RNOD} was used for the evaluation. We see that *rmit_ir-Q-RR-5* does outperform our baseline runs along both measures. We see that *rmit_ir-D-RR-2* and *rmit_ir-D-RR-3* (i.e., PM-2 diversification-based runs) were the best among our submitted systems. Our PM-2 based diversification runs (i.e., rmit_ir-D-RR-2 and rmit_ir-D-RR-3) do not outperform the retrieval baselines in the R and M topics. However, this is not the case with Y-topic. This warrants further investigation in both the membership generation phase as well as the ranking stage.

4 CONCLUSION

This report described the different runs we submitted to the NTCIR-17 FairWeb-1 task. We started by discussing our motivation behind exploring diversification-based techniques and the specific pre-processing techniques we employed for generating the membership files. Subsequently, we compared our submitted runs against the baselines along multiple relevance and fairness-aware measures used in this task. Although our Linear Combination runs indicate improvement over PM-2 based diversification, the gains are not statistically significant. A more detailed examination of both the membership generation stage and the ranking stage may provide insights into the reasons behind certain anomalies. Additionally, it is worth mentioning that our Linear Combination technique can be customized to include any other distance measures as needed.

ACKNOWLEDGMENTS

The authors would like to acknowledge Country. This research has been carried out on the unceded lands of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin nation. We pay our respects to their Ancestors and Elders, past, present, and emerging. We respectfully acknowledge their connection to land, waters, and sky. This work is partially supported by the Australian Research Council (DE200100064, CE200100005).

REFERENCES

[1] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/3209978.3210024

[2] KR Bindu, Rhama Lalgudi Visweswaran, PC Sachin, Kundavai Devi Solai, and Soundarya Gunasekaran. 2017. Reducing the Cold-User and Cold-Item Problem in Recommender System by Reducing the Sparsity of the Sparse Matrix and Addressing the Diversity-Accuracy Problem. In *Proceedings of International Conference on Communication and Networks: ComNet 2016*. Springer, 561–570. https://doi.org/10.1007/978-981-10-2750-5_58

[3] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 335–336. https://doi.org/10.1145/290941.291025

[4] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[5] Gordon V. Cormack, Charles L.A. Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[6] Van Dang and Bruce W. Croft. 2013. Term Level Search Result Diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 603–612. https://doi.org/10.1145/2484028.2484095

[7] Van Dang and W. Bruce Croft. 2012. Diversity by Proportionality: An Election-Based Approach to Search Result Diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/2348283.2348296

[8] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Foundations and Trends in Information Retrieval* 16, 1-2 (2022), 1–177. https://doi.org/10.1561/1500000079

[9] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* 57, 1 (2020), 102138. https://doi.org/10.1016/j.ipm.2019.102138

[10] Shangsong Liang, Emine Yilmaz, Hong Shen, Maarten De Rijke, and W. Bruce Croft. 2017. Search Result Diversification in Short Text Streams. *ACM Trans. Inf. Syst.* 36, 1, Article 8 (jul 2017), 35 pages. https://doi.org/10.1145/3057282

[11] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2022. Search Results Diversification for Effective Fair Ranking in Academic Search. *Information Retrieval Journal* 25, 1 (2022), 1–26. https://doi.org/10.1007/s10791-021-09399-z

[12] Sachin Pathiyan Cherumanal. 2022. Fairness-Aware Question Answering for Intelligent Assistants. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3492. https://doi.org/10.1145/3477495.3531682

[13] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2021. Evaluating Fairness in Argument Retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 3363–3367. https://doi.org/10.1145/3459637.3482099

[14] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2022. RMIT at TREC 2021 Fair Ranking Track. In *Proc. TREC*. https://trec.nist.gov/pubs/trec30/papers/RMIT-IR-F.pdf

[15] Sachin Pathiyan Cherumnal, Marwah Alaofi, Reham Abdullah Altalhi, Elham Naghizade, Falk Scholer, and Damiano Spina. 2023. RMIT CIDDA IR at the TREC 2022 Fair Ranking Track. In *Proceedings of TREC 2022*.

[16] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A Versatile Framework for Evaluating Ranked Lists in Terms of Group Fairness and Relevance. *ACM Trans. Inf. Syst.* 42, 1, Article 11 (aug 2023), 36 pages. https://doi.org/10.1145/3589763

[17] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 595–604. https://doi.org/10.1145/3331184.3331215

[18] Sijie Tao, Nuo Chen, Tetsuya Sakai, Zhumin Chu, Hiromi Arai, Ian Soboroff, Nicola Ferro, and Maria Maistro. 2023. Overview of the NTCIR-17 FairWeb-1 Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*. https://doi.org/10.20736/0002001318

[19] Lakshmi Vikraman, W. Bruce Croft, and Brendan O'Connor. 2018. Exploring Diversification In Non-Factoid Question Answering. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval* (Tianjin, China) *(ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 223–226. https://doi.org/10.1145/3234944.3234973