RMIT CIDDA IR at the TREC 2022 Fair Ranking Track

SACHIN PATHIYAN CHERUMANAL, RMIT University, Australia
MARWAH ALAOFI, RMIT University, Australia
REHAM ABDULLAH ALTALHI, RMIT University, Australia
ELHAM NAGHIZADE, RMIT University, Australia
FALK SCHOLER, RMIT University, Australia
DAMIANO SPINA, RMIT University, Australia

This report describes the participation of the RMIT CIDDA IR[1] group at the TREC 2022 Fair Ranking Track (Task 1). We submitted 8 runs with the aim to explore the role of explicit search result diversification, ranking fusion, and the use of a multi-criteria decision-making method to generate fair rankings considering multiple protected attributes.

1 INTRODUCTION

The quality of articles in Wikipedia greatly depends on the level of editorial effort that is deployed to maintain them. In an ideal world, every document that requires editing would get the attention needed by Wikipedia editors to improve its quality. Unfortunately, the demand for editorial effort is substantially higher than the available editorial resources. To meet this limitation, Wikipedia coordinators/editors use search engines to find the Wikipedia documents that will be edited. Articles and topics in Wikipedia have different characteristics in terms of protected attributes. If the search engines used in the editorial process do not consider these attributes, different societal groups may receive disparate outcomes, eventually leading to negative social and economic impact.

The Fair Ranking Track at TREC 2022 aims to address the problem of providing fair rankings of Wikipedia articles for coordinators/editors. This report describes the participation of the RMIT CIDDA IR group in Task 1 of the track, which consists of providing a fair ranking of Wikipedia articles when multiple protected attributes need to be considered. We addressed this challenge by proposing a three-stage approach that includes ad hoc retrieval, explicit Search Results Diversification (SRD), and ranking fusion. We also explored the use of a multi-criteria decision-making method (Analytic Hierarchy Process, AHP) as a mechanism to allow stakeholders to have control of the way that different attributes could be combined in the fusion stage.

2 TASK 1: FAIR RANKINGS TO SUPPORT WIKIPROJECT COORDINATORS

2.1 Goal

The goal of Task 1 at the Fair Ranking Track at TREC 2022[2] is to support WikiProject coordinators in producing a ranked list of articles that editors can then consult when looking for work to do. Given a query (i.e., keywords in a WikiProject title) the system needs to produce a single ranking consisting of 500 articles retrieved from a collection of Wikimedia articles (a dump of the English Wikipedia). The quality of systems is evaluated in two dimensions: *relevance*,

[1] http://rmit.edu.au/cidda
[2] https://fair-trec.github.io/docs/Fair_Ranking_2022_Participant_Instructions.pdf [Accessed: 3 Nov 2022]

Authors' addresses: Sachin Pathiyan Cherumanal, RMIT University, Melbourne, Australia, sachin.pathiyan.cherumanal@student.rmit.edu.au; Marwah Alaofi, RMIT University, Melbourne, Australia, marwah.alaofi@student.rmit.edu.au; Reham Abdullah Altalhi, RMIT University, Melbourne, Australia, s3805747@student.rmit.edu.au; Elham Naghizade, RMIT University, Melbourne, Australia, e.naghizade@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.

based on relevance assessments for the articles derived from Wikipedia data; and *fairness*, based on the exposure of a set of fairness/protected attributes associated with the articles, listed in Section 2.2.

2.2 Fairness/Protected Attributes

Wikipedia articles in the collection are characterized with 11 fairness/protected attributes:

topic_countries: The geographical location (countries) associated with the article topic.
topic_regions: The geographical location (sub-continental regions) associated with the article topic.
sources_countries: The geographical location (countries) associated with the article based on the article's sources.
sources_regions: The geographical location (sub-continental regions) associated with the article based on the article's sources.
gender: The gender of the individual about whom the biography pertains, mapped into four distinct categories: Man, Woman, Non-binary, and Unknown (missing data or not a biography).
topic_age: Age of the topic: how old the subject of the article is (in years), mapped to four distinct categories: Unknown, Pre-1900s, 20th century, and 21st century.
occupations: The occupation of the subject of an article. An article can have 0 (unknown) to many occupations associated with it.
alphabetical: The first letter of the article, mapped to four discrete categories: a–d, e–k, l–r, and s–z.
creation_date: Year in which the article was created. The date is mapped to one of four discrete categories: 2001–2006, 2007–2011, 2012–2016, and 2017–2022.
pageviews: Popularity of the article quantified as the number of times the page was viewed in February 2022, normalized and mapped to four discrete categories: Low, Medium–Low, Medium–High, and High.
languages: The number of other Wikipedia language articles that the article is replicated in, mapped to three discrete categories: English-only, 2–4 languages, and 5+ languages.

2.3 Evaluation Measures

Three official evaluation measures were used in the task: Normalized Discounted Cumulative Gain nDCG [6] for relevance, Attention-Weighted Rank Fairness (AWRF) [14] for fairness, and Score, which is the product of nDCG and AWRF. Both nDCG and AWRF are computed using a logarithmic decay function. For all measures, higher scores indicate better performance. Rankings are truncated at depth 500.

3 APPROACH

Our approach consists of three stages: ad hoc retrieval to produce an initial ranking (Section 3.1); a re-ranking stage based on *SRD* (Section 3.2); and a final ranking fusion stage to generate the final ranking (Section 3.3).

3.1 Initial Ranking: Ad Hoc Retrieval

The first stage consists of retrieving Wikipedia articles given a WikiProject topic. We used the `keywords` field of the topic (i.e., the parsed title provided by the organizers). To index the collection, we used the `plain` collection, i.e., the full article text, without Wiki markup (plain file only). Retrieval was carried out using the `Pyserini` [9] implementation of Okapi BM25 with default parameters ($k1 = 0.9$, $b = 0.4$).

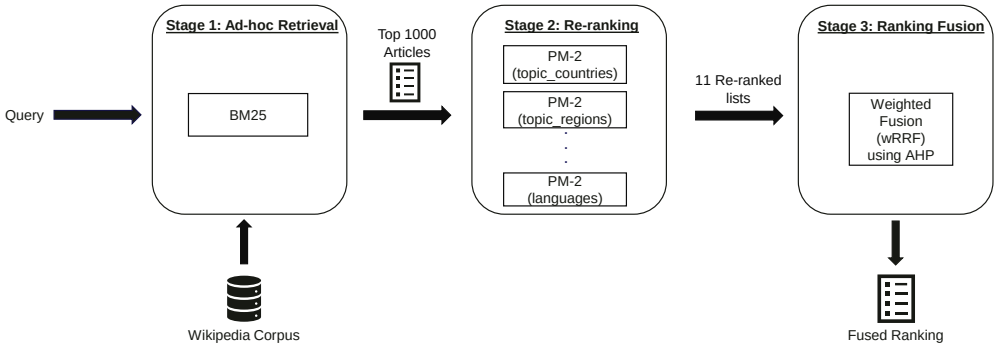

Fig. 1. Architecture diagram depicting the different components used in our approach.

3.2 Re-ranking: Search Result Diversification

We are interested in understanding the relation between *SRD* and fairness-aware rankings[11, 12]. In the fairness scenario proposed for Task 1 of the Fair Ranking Track at TREC 2022, fairness/protected attributes are known – in contrast to the TREC 2021 Fair Ranking Track [4], which had one of the attributes hidden from the participants. Hence, the problem can be addressed as an *SRD* problem. In this edition, we wanted to understand to what extent *SRD* methods can be used to diversify a ranking given a protected/fairness attribute. Given a fairness attribute, values for that particular attribute are treated as *aspects* or sub-topics, aligned with literature on *SRD* [2, 3, 8, 16].

As we are interested in considering statistical parity to diversify our ranking, we decided to explore a proportionality-based *SRD* method, i.e., the PM-2 model proposed by Dang and Croft [3], with $\alpha = 0.5$.

PM-2 is applied for each of the 11 fairness categories listed in Section 2.2. We use PM-2 (*attribute*) to denote the diversification obtained using each attribute when reporting our results.

3.3 Final Ranking: Ranking Fusion

As the task includes multiple fairness/protected attributes, we obtain a set of 11 diversified rankings as the output of the (previous) re-ranking stage. In the final stage, ranking fusion is used to combine the diversified rankings and obtain the final ranking (as shown in Figure 1).

We used the `polyfuse`[3] [7] implementation of Reciprocal Ranking Fusion (RRF)[1] to fuse the diversified rankings. Given the initial ranking R, the final ranking L is obtained by applying RRF to the 11 diversified rankings obtained with PM-2:

$$L = \text{RRF}\left(\text{PM-2}(R, \text{topic\_countries}), \text{PM-2}(R, \text{topic\_regions}), \ldots, \text{PM-2}(R, \text{languages})\right) \quad (1)$$

Analytic Hierarchic Process. Equation 1 assumes that all protected attributes have the same importance (or weight) when fusing the rankings. However, some fairness attributes may be more important than others and different stakeholders (e.g., WikiProject coordinators) may have different ways of modeling those priorities w.r.t. different fairness/protected attributes. In this section, we aim to explore a mechanism to allow stakeholders to weight different attributes to our ranking pipeline.

In operational research, a number of multi-criteria decision-making methods have been proposed [10]. We opted to explore the feasibility (and sensitivity) of using Analytic Hierarchy Process (AHP) by Saaty [13] to enable our weighted

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

reciprocal ranking fusion. AHP provides a mechanism that allows stakeholders to define a weighting schema for complex criteria by defining pairwise importance relationships, where higher weights imply higher relative importance. The application of AHP spans across multiple areas, including education, manufacturing, politics, and engineering, among others [15]. For instance, AHP had been used for evaluating journal quality by consolidating multiple criteria related to the quality of a journal into an integrated measure [5]. To our knowledge, this is the first attempt to apply AHP in fairness-aware rankings.

Table 1 shows an example of the input for AHP applied to our scenario. By checking the consistency across the relationships, an iterative method produces a weighting schema that adheres to the constrains/relationships defined by the stakeholder.

Table 1. An example of the input for Analytic Hierarchy Process (AHP) .

	topic_countries	topic_regions	sources_countries	sources_regions	gender	topic_age	occupations	alphabetical	creation_date	pageviews	languages
topic_countries	1	1	1	1	1	0.33	1	0.25	0.2	0.14	0.2
topic_regions	1	1	1	1	1	0.33	1	0.25	0.2	0.14	0.2
sources_countries	1	1	1	1	1	0.33	1	0.25	0.2	0.14	0.2
sources_regions	1	1	1	1	1	0.33	1	0.25	0.2	0.14	0.2
gender	1	1	1	1	1	0.33	1	0.25	0.2	0.14	0.2
topic_age	3	3	3	3	3	1	3	4	0.33	1	0.25
occupations	1	1	1	1	1	0.33	1	0.2	0.2	0.14	0.2
alphabetical	4	4	4	4	4	0.25	5	1	0.33	1	4
creation_date	5	5	5	5	5	3	5	3	1	0.33	4
pageviews	7	7	7	7	7	1	7	1	3	1	5
languages	5	5	5	5	5	4	5	0.25	0.25	0.2	1

Table 2. Weights obtained after performing AHP. AHP_1 consists of the weights obtained for the input given in Table 1.

Attribute	AHP_1	AHP_2	AHP_3	AHP_4	AHP_5	AHP_6	AHP_7	AHP_8	AHP_9	AHP_{10}	AHP_{11}	AHP_{12}	AHP_{13}	AHP_{14}
topic_countries	0.028	0.028	0.029	0.031	0.023	0.054	0.044	0.037	0.029	0.026	0.020	0.034	0.032	0.031
topic_regions	0.028	0.028	0.029	0.031	0.040	0.054	0.044	0.037	0.029	0.026	0.024	0.018	0.017	0.021
sources_countries	0.028	0.028	0.029	0.031	0.045	0.054	0.044	0.037	0.029	0.026	0.030	0.024	0.023	0.025
sources_regions	0.028	0.028	0.029	0.031	0.040	0.054	0.044	0.037	0.029	0.026	0.030	0.032	0.030	0.031
gender	0.028	0.028	0.029	0.029	0.034	0.054	0.044	0.037	0.029	0.026	0.034	0.045	0.043	0.035
topic_age	0.124	0.118	0.125	0.031	0.078	0.117	0.101	0.085	0.144	0.098	0.097	0.104	0.104	0.057
occupations	0.028	0.027	0.028	0.036	0.074	0.054	0.065	0.055	0.131	0.087	0.086	0.066	0.102	0.026
alphabetical	0.139	0.133	0.141	0.198	0.176	0.104	0.125	0.105	0.144	0.098	0.097	0.104	0.104	0.109
creation_date	0.199	0.267	0.216	0.213	0.178	0.326	0.320	0.264	0.144	0.279	0.272	0.230	0.236	0.348
pageviews	0.239	0.186	0.212	0.192	0.177	0.065	0.064	0.219	0.144	0.208	0.214	0.238	0.206	0.188
languages	0.131	0.128	0.134	0.133	0.136	0.065	0.108	0.090	0.144	0.098	0.097	0.104	0.104	0.127

In our runs, 14 different weighting schemas were obtained by arbitrarily defining consistent relationships across the 11 fairness/protected attributes. In our results, we denote these variations with AHP_1,. . ., AHP_{14}.[4] Table 2 lists the weighting schemas obtained for our experiments. The second row (AHP_1) corresponds to the weighting schema obtained for the relationships defined in Table 1.

Weighted Reciprocal Ranking Fusion. We adapted RRF [1] to accept weights for the rankings to be fused. The score of the document in the final ranking is a linear combination of the reciprocal ranks. Hereafter, we use wRRF to refer to this variant. Given the set of documents D in the n rankings to be fused $\mathcal{R} = \{R_1, \ldots, R_n\}$, the weights corresponding for those rankings $W = \{w_1, \ldots, w_n\}$, and $k = 60$:

$$\text{wRRF}(d \in D) = w_1 \cdot \frac{1}{k + R_1(d)} + \ldots + w_n \cdot \frac{1}{k + R_n(d)} \tag{2}$$

We used wRRF (Equation 2) to fuse the diversified rankings taking into account the weights produced by AHP.

4 SUBMITTED RUNS

A total of 8 runs were submitted:

BM25 (rmit_cidda_ir_1): Initial ranking obtained by BM25 as described in Section 3.1.

BM25 + PM-2 (languages) (rmit_cidda_ir_2): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for the *languages* attribute.

BM25 + PM-2 (all) + wRRF (AHP_5) (rmit_cidda_ir_3): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for all 11 fairness attributes; ranking fusion using wRRF. The weighting schema used in wRRF is obtained using AHP_5.

BM25 + PM-2 (topic_age) (rmit_cidda_ir_4): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for the *topic_age* attribute.

BM25 + PM-2 (gender) (rmit_cidda_ir_5): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for the *gender* attribute.

BM25 + PM-2 (all) + wRRF (AHP_2) (rmit_cidda_ir_6): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for all 11 fairness attributes; ranking fusion using wRRF. The weighting schema used in wRRF is obtained using AHP_2.

BM25 + PM-2 (all) + wRRF (AHP_9) (rmit_cidda_ir_7): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for all 11 fairness attributes; ranking fusion using wRRF. The weighting schema used in wRRF is obtained using AHP_9.

BM25 + PM-2 (all) + wRRF (AHP_{14}) (rmit_cidda_ir_8): Ad hoc retrieval using BM25; re-ranking using explicit search result diversification using PM-2 for all 11 fairness attributes; ranking fusion using wRRF. The weighting schema used in wRRF is obtained using AHP_{14}.

5 RESULTS

In this section we report the results obtained in the training topics (Section 5.1) and the results of the 8 submitted runs submitted in the evaluation topics (Section 5.2). For statistical significance, we performed Tukey's Honestly Significant Difference (HSD) test to correct for multiple comparisons.

5.1 Results for Training Topics

Table 3 shows the results for the training topics. The table includes the scores for the three stages of our proposed pipeline. Compared to ad hoc retrieval, the rest of the runs seem to have equivalent performance in terms of relevance. In terms of fairness, diversifying

[4]Note that the validity of these relationship w.r.t. to the needs of the stakeholders is out of the scope of this work, as it cannot be assessed without consultation with domain experts (e.g., WikiProject coordinators and editors).

Table 3. Results for the runs tested using the training topics and the official evaluation measures for relevance (nDCG), fairness (AWRF), and both (Score, i.e., the product of nDCG and AWRF. The best score for each measure is in boldface.

Run	Id	nDCG	AWRF	Score
BM25	rmit_cidda_ir_1	0.5691	0.4994	0.2885
BM25				
+ PM-2 (creation_date)		0.5407	0.5062	0.2789
+ PM-2 (alphabetical)		0.5555	0.4956	0.2820
+ PM-2 (gender)	rmit_cidda_ir_5	**0.5726**	0.5004	**0.2908**
+ PM-2 (languages)	rmit_cidda_ir_2	0.5438	**0.5105**	0.2840
+ PM-2 (occupations)		0.5687	0.4999	0.2888
+ PM-2 (topic_countries)		0.5653	0.4930	0.2827
+ PM-2 (topic_regions)		0.5603	0.4953	0.2817
+ PM-2 (pageviews)		0.5483	0.4998	0.2781
+ PM-2 (sources_countries)		0.5655	0.4861	0.2777
+ PM-2 (sources_regions)		0.5613	0.4942	0.2820
+ PM-2 (topic_age)	rmit_cidda_ir_4	0.5617	0.4987	0.2861
BM25 + PM-2 (All attributes)				
+ wRRF (AHP_1)		0.5517	0.5051	0.2838
+ wRRF (AHP_2)	rmit_cidda_ir_6	0.5487	0.5063	0.2832
+ wRRF (AHP_3)		0.5509	0.5054	0.2837
+ wRRF (AHP_4)		0.5495	0.5052	0.2829
+ wRRF (AHP_5)	rmit_cidda_ir_3	0.5543	0.5049	0.2848
+ wRRF (AHP_6)		0.5535	0.5040	0.2839
+ wRRF (AHP_7)		0.5512	0.5055	0.2837
+ wRRF (AHP_8)		0.5509	0.5048	0.2829
+ wRRF (AHP_9)	rmit_cidda_ir_7	0.5580	0.5046	0.2866
+ wRRF (AHP_{10})		0.5512	0.5056	0.2835
+ wRRF (AHP_{11})		0.5516	0.5054	0.2836
+ wRRF (AHP_{12})		0.5527	0.5047	0.2838
+ wRRF (AHP_{13})		0.5537	0.5050	0.2844
+ wRRF (AHP_{14})	rmit_cidda_ir_8	0.5451	0.5066	0.2814

for one fairness atrribute tends to perform better than fusing multiple diversified rankings. Varying the input to AHP (i.e., simulating different stakeholders indicating different weights for the fairness attributes) does not seem to have a substantial impact on the fusion.

According to Tukey's HSD test, only the diversified ranking using the *gender* attribute (Row 5, rmit_cidda_ir_5) has statistically significant differences (with p-value < 0.05) w.r.t. the diversified rankings obtained with the pageviews and topic_countries attributes (Rows 8 and 10), for the three metrics. Other differences reported in the table are not statistically significant.

5.2 Results for Submitted Runs

Table 4 shows the results for the runs submitted using the evaluation topics. In terms of relevance, any further manipulation of the ranking obtained by BM25 results in lower effectiveness scores (refer Figure 2). In terms of fairness, we observe the opposite, as both re-ranking and fusion increase AWRF scores. The run diversifying using the *gender* attribute obtains the highest fairness scores and the best trade-off between relevance and fairness. Also, Figure 2 compares our submitted runs against the best submission at Task 1 of TREC 2022 Fair Ranking Track, where the best submission is the run with the highest Score (i.e., the product of nDCG and AWRF).

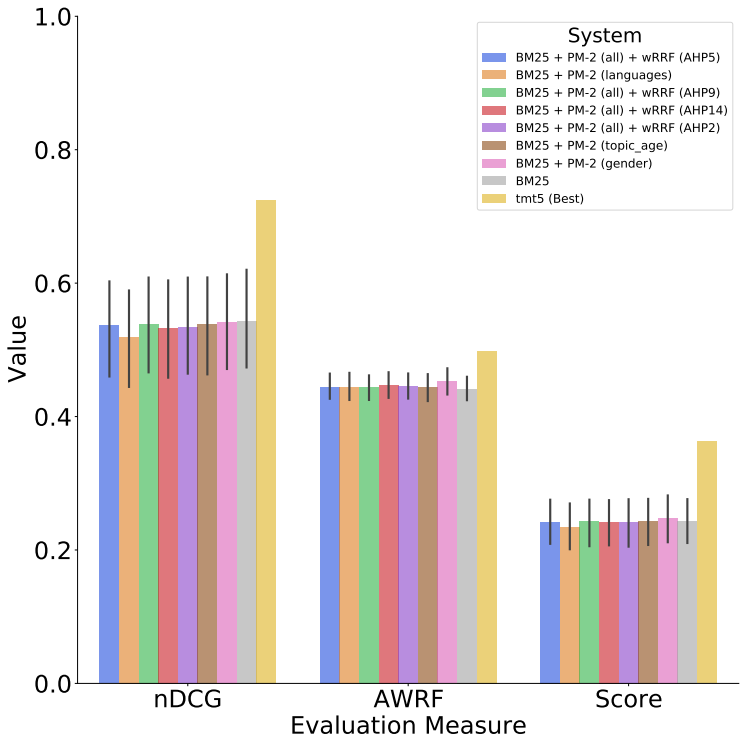

Fig. 2. Results for the 8 runs submitted by RMIT CIDDA IR compared against the run submitted to TREC Fair Ranking Track 2022 Task 1 that obtained the highest Score - tmt5 (Best). Results are obtained using evaluation topics and the official evaluation measures for relevance (nDCG), fairness (AWRF), and both (Score, i.e., the product of nDCG and AWRF). Error bars indicate 95% confidence intervals.

Table 4. Results for the 8 runs submitted by RMIT CIDDA IR to TREC Fair Ranking Track 2022 Task 1 using the evaluation topics and the official evaluation measures for relevance (nDCG), fairness (AWRF), and both (Score, i.e., the product of nDCG and AWRF). The best score for each measure is in boldface.

Run	Id	nDCG	AWRF	Score
BM25	rmit_cidda_ir_1	**0.5438**	0.4416	0.2433
BM25 + PM-2 (languages)	rmit_cidda_ir_2.	0.5197	0.4443	0.2345
BM25 + PM-2 (topic_age)	rmit_cidda_ir_4	0.5388	0.4435	0.2431
BM25 + PM-2 (gender)	rmit_cidda_ir_5	0.5417	**0.4525**	**0.2485**
BM25 + PM-2 (all) + wRRF (AHP_5)	rmit_cidda_ir_3	0.5365	0.4447	0.2420
BM25 + PM-2 (all) + wRRF (AHP_2)	rmit_cidda_ir_6	0.5343	0.4457	0.2418
BM25 + PM-2 (all) + wRRF (AHP_9)	rmit_cidda_ir_7	0.5382	0.4443	0.2426
BM25 + PM-2 (all) + wRRF (AHP_{14})	rmit_cidda_ir_8	0.5322	0.4469	0.2415

According to Tukey's HSD test, the following differences are statistically significant ($p < 0.05$). Diversifying with *languages* is significantly worse than the initial ranking BM25 and than the diversified run using *gender*. The difference between these two diversified rankings is also statistically significant for the Score results. In terms of AWRF, the scores obtained by BM25 + PM-2 (gender) are statistically significantly better than all the other runs besides BM25 + PM-2 (all) + wRRF (AHP_9).

6 CONCLUSIONS

In this work we explored the relation between fairness-aware rankings and *SRD* in the context of multiple fairness/protected attributes. Reciprocal Rank Fusion (RRF), a competitive ranking fusion method, was used to combine the multiple diversified rankings. We also explored the use of Analytic Hierarchy Process (AHP), a multi-criteria decision-making method, to simulate a human-in-the-loop fusion stage, where stakeholders could have control on how to weight different fairness attributes. The results demonstrated little sensitivity, suggesting that the way we incorporated diversification and fusion have little impact on the way the initial ranking in the ad hoc retrieval stage is modified.

These initial experiments have a number of limitations that we plan to address in future work. First, each stage of the pipeline was only instantiated with one method. Although they are all competitive in terms for the three stages (i.e., ad hoc retrieval, diversification, and ranking fusion), a more comprehensive study with other methods is needed. Moreover, we believe it is worth analyzing supervised versions of our approach (e.g., learning the weights for ranking fusion). Finally, we plan to explore ways of embedding multi-criteria decision-making methods at earlier stages of the pipeline (e.g., to inform the hierarchical diversification), in the hope of finding more effective ways of enabling stakeholders to control the way multiple fairness attributes should be combined to generate fair rankings.

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. The authors would like to thank Dr. Luke Gallagher for kindly extending `polyfuse` with wRRF, a weighted version of Reciprocal Ranking Fusion. This work is partially supported by the Australian Research Council (DE200100064, CE200100005).

REFERENCES

[1] 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

[2] 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

[3] 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

[4] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the TREC 2021 Fair Ranking Track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*.

[5] Guisseppi A. Forgionne and Rajiv Kohli. 2001. A Multiple Criteria Assessment of Decision Technology System Journal Quality. *Information & Management* 38, 7 (2001), 421–435. https://doi.org/10.1016/S0378-7206(00)00079-3

[6] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[7] Oren Kurland and J. Shane Culpepper. 2018. Fusion in Information Retrieval: SIGIR 2018 Half-Day Tutorial. 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, 1383–1386. https://doi.org/10.1145/3209978.3210186

[8] 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

[9] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2356–2362. https://doi.org/10.1145/3404835.3463238

[10] Nolberto Munier and Eloy Hontoria. 2021. *Uses and Limitations of the AHP Method*. Springer. https://doi.org/10.1007/978-3-030-60392-2

[11] 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* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 3363–3367. https://doi.org/10.1145/3459637.3482099

[12] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2022. RMIT at TREC 2021 Fair Ranking Track. In *Proceedings of TREC 2021*. NIST, 4 pages. https://trec.nist.gov/pubs/trec30/papers/RMIT-IR-F.pdf

[13] Thomas L. Saaty. 1988. What is the Analytic Hierarchy Process? In *Mathematical models for decision support*. Springer, 109–121. https://doi.org/10.1007/978-3-642-83555-1_5

[14] Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/3308560.3317595

[15] Omkarprasad S. Vaidya and Sushil Kumar. 2006. Analytic Hierarchy Process: An Overview of Applications. *European Journal of Operational Research* 169, 1 (2006), 1–29. https://doi.org/10.1016/j.ejor.2004.04.028

[16] 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