A Meta-learning Approach to Fair Ranking

Yuan Wang
Santa Clara University
Santa Clara, CA, USA
ywang4@scu.edu

Zhiqiang Tao
Santa Clara University
Santa Clara, CA, USA
ztao@scu.edu

Yi Fang
Santa Clara University
Santa Clara, CA, USA
yfang@scu.edu

ABSTRACT

In recent years, the fairness in information retrieval (IR) system has received increasing research attention. While the data-driven ranking models achieve significant improvements over traditional methods, the dataset used to train such models is usually biased, which causes unfairness in the ranking models. For example, the collected imbalance dataset on the subject of the expert search usually leads to systematic discrimination on the specific demographic groups such as race, gender, etc, which further reduces the exposure for the minority group. To solve this problem, we propose a Meta-learning based Fair Ranking (MFR) model that could alleviate the data bias for protected groups through an automatically-weighted loss. Specifically, we adopt a meta-learning framework to explicitly train a meta-learner from an unbiased sampled dataset (meta-dataset), and simultaneously, train a listwise learning-to-rank (LTR) model on the whole (biased) dataset governed by "fair" loss weights. The meta-learner serves as a weighting function to make the ranking loss attend more on the minority group. To update the parameters of the weighting function and the ranking model, we formulate the proposed MFR as a bilevel optimization problem and solve it using the gradients through gradients. Experimental results on several real-world datasets demonstrate that the proposed method achieves a comparable ranking performance and significantly improves the fairness metric compared with state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; • **Social and professional topics** → **Codes of ethics**.

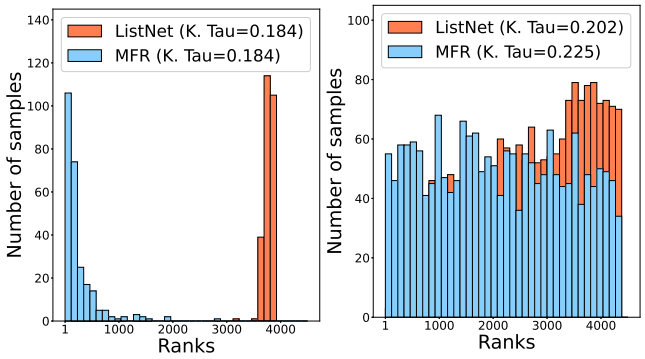
KEYWORDS

Fairness-aware IR, Meta-learning, Learning-to-Rank

ACM Reference Format:
Yuan Wang, Zhiqiang Tao, and Yi Fang. 2022. A Meta-learning Approach to Fair Ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531892

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531892



(a) Law Students (race) **(b) Law Students (gender)**

Figure 1: Illustration of the predicted rankings distribution of the protected groups (*female students, African American students*) on the two different datasets. We report Kendall's Tau as the ranking metric. The proposed MFR model ranks the items from the protected groups higher compared to ListNet [5], which indicates that the MFR improves the protected attribute's exposure with unbiased ranking performance.

1 INTRODUCTION

Recently, the fairness in information retrieval (IR) system has attracted more and more attention [17, 18, 23]. The ranking models aim to give the relevant scores for the items under the query, and the top items with the highest scores will be delivered to the users. These ranking models are generally data-driven, which means the models will observe particular patterns in the training dataset and make predictions based on them. However, when the subject of the ranking problem is about the expert search or the job recommendation, the *systematic biases* from the dataset – usually stemming from a biased data distribution – will introduce *unfairness* in the trained model. For example, the traditional LTR model such as ListNet [5] will "discriminately" assign lower weights to the minority group due to the data bias (see Fig. 1). As addressed by Friedman [10], the historic discrimination to the socially underrepresented group in the dataset will make its way into the model as the pattern will be observed during the training process. The unfairness problem could be summarized as the disparate exposure [21] as the disadvantaged protected group is not treated as equally as the advantaged group in the dataset. This disparate exposure could lead to a negative impact on many real-world ranking problems, such as the unequal opportunity in the job market for the underrepresented group.

To solve the unfairness problem, tremendous research efforts have been made in designing fairness-aware algorithms, among which, the fairness ranking models can be categorized as the score-based and supervised ones. For score-based models, there are the

Rank-aware proportional representation [18], the Constrained ranking maximization [6], etc. Some score-based models aim to correct the bias in the training data, and the others aim to adjust the prediction scores for better fairness. There are also supervised models, such as DELTR [21], FA*IR [20], etc, which could learn a fair model from the biased dataset. In general, the ranking models focus on different mitigation points such as the post-, in-, and pre-processing of the model training. Although the in-processing models have achieved good performance on the fairness metric, there is still the limitation as the model is learned from the biased dataset. Thus, the meta-learning could benefit the aforementioned problem by training a meta-learner on a meta-dataset. The meta-dataset is collected uniformly without any bias, which would train a fair meta-learner so that the ranking model could learn from it. For general fairness problems such as training the classification model on a biased dataset, researchers have applied the Model-Agnostic Meta-Learning (MAML) [9]. For example, the Meta-Weight-Net [13] proposed to explicitly learn a weighting function from the meta-dataset which is updated simultaneously with the classifier. However, meta-learning is still under-explored for the fairness-aware ranking problems.

In this study, we propose a meta-learning framework to formulate the fairness-aware ranking task as a bilevel optimization problem, where the upper-level is the meta-trainer and the lower-level is the ranking model. That is, we can train a meta-learner on the meta-dataset which could help the ranking model to learn fairly on the biased dataset. The meta-dataset is a small unbiased dataset, which is collected by uniformly sampling from the training dataset under all queries for both the protected group and the unprotected group. In detail, at each training iteration, we use the ranking model and the ranking loss function to compute the loss values for each data sample from the training dataset. Then we train a multi-layer neural network as the weighting function to re-weight the loss values, and the weighting function is optimized by the weighted loss values on the meta-dataset. Since the weighting function which is the meta-learner is subject to the ranking models, our goal is to optimize the loss' weights (given by the meta-learner) to achieve fairness on the training dataset. Intuitively, we can see the loss' weight as the hyperparameter which could be learned, and we train a meta-learner to tune the hyperparameter on the meta-dataset. Such the training process could also be referred to as the bilevel optimization as the learned parameters of the ranking model depend on the parameters of the meta-learner. To the best of our knowledge, we propose the first meta-learning approach to fair ranking. In summary, this paper makes the following contributions:

- We propose a general meta-learning framework for the fairness ranking called Meta-learning based Fair Ranking (MFR) that addresses the data bias by automatically re-weighting the ranking losses.
- We formulate the MFR as a bilevel optimization problem and solve it using gradients through gradients.
- Experiments on the real-world datasets demonstrate that the proposed method achieves a comparable ranking performance and significantly improves the fairness metric compared with state-of-the-art methods.

2 RELATED WORK

Fairness on Ranking. Among the ranking models, Zehlike et al. [23] summarised them into the score-based models and supervised learning models. For the score-based models, Yang et al. [17, 18], Celis et al. [6], and Stoyanovich et al. [15] proposed different ways to intervene on the score outcomes to reduce unfairness. Kleinberg et al. [11] proposed models to intervene in the score distribution which correct the scores in the training data for the bias. Also, Asudeh et al. [1] designed a fair ranking function that takes an ordering of the items as input and outputs the fairness metric results.

For the supervised models, Lahoti et al. [12] proposed a pre-processing approach to learn the fair training data, which would help to train an unbiased model. For the in-processing models, DELTR [21] is a listwise LTR loss function with the unfairness measure so that the model is optimized for both the ranking metrics and the fairness metric. Specifically, DELTR [21] aims to address potential issues of discrimination and unequal opportunity in rankings at training time. Beutel et al. [2] proposed a pairwise fairness metric and uses it as a regularizer to encourage improving this metric during the ranking model training. For the post-processing models which usually re-rank the model's prediction, Zehlike et al. [20] proposed the FA*IR to assure the number of candidates from the protected group is above the minimum requirement in the ranking. Zehlike et al. [22] also proposed the CFAθ which enables a continuous interpolation between different fairness definitions. Biega et al. [3] proposed an algorithm that optimizes equity of user attention with the relevance loss. All of the fairness ranking models mentioned above are designed in the traditional machine learning fashion, whereas our proposed method takes advantage of the meta-learning. With the meta-learning, we formulate the MFR as a bilevel optimization problem and solve it using gradients through gradients. In this way, we utilize a uniformly sampled meta-dataset to train a meta-learner which could help the ranking model to learn fairly on the biased training dataset.

Meta-Learning on Fairness. Zhao et al. [25] proposed the Follow the Fair Meta Leader (FFML) which could learn an online fair classification model's primal with good accuracy and the dual parameters that are associated with fairness. Then Zhao et al. [26] proposed the Primal-Dual Fair Meta-learning to learn a good initialization of the model through meta-learning so that the model adapts to the new fair learning task quickly. For the unbiased multi-class classification problem, Zhao et al. [24] developed a few-shot discrimination prevention learning model based on the MAML. To learn fairly from minimal data on a new task, Slack et al. [14] proposed the Fair-MAML to expand the famous meta-learning framework MAML such that each task includes a fairness regularization term and fairness hyperparameter in the task losses. In addition, Chen et al. [7] proposed the AutoDebias that unifies various biases from the risk discrepancy perspective and applies meta-learning on the recommender system to optimize the debiasing parameters using a set of uniform data. However, the model only focused on the recommender system, and the meta-learning on fairness ranking is not a well-researched field. Unlike AutoDebias, our proposed method focuses on the ranking problem, which provides a general framework to address the data bias.

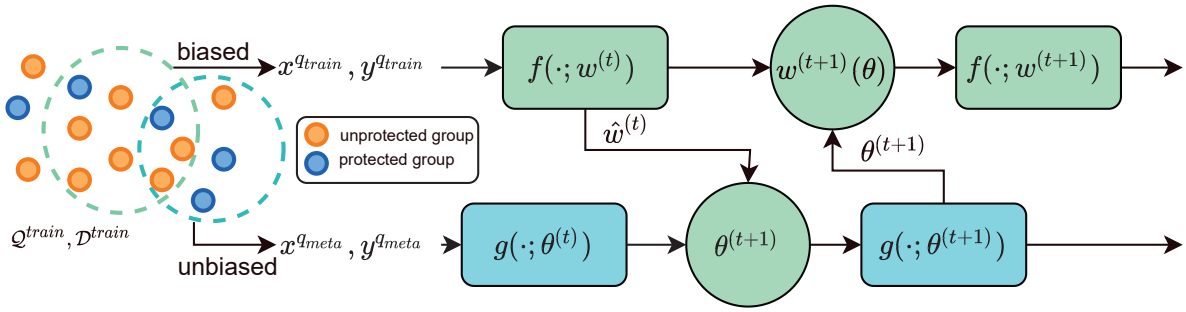


Figure 2: MFR learning algorithm flowchart (steps 4 and 6 in Algorithm 1). Note that $f(\cdot; w)$ is the ranking model, $g(\cdot; \theta)$ is the meta-learner, b is the batch size for the training dataset, d is the batch size for the meta-dataset, and α and β are the learning rates. At each iteration, we firstly update θ in the meta-learner using Eq. (8) with the meta-dataset, and then we update w in the ranking model using Eq. (9) with the training dataset.

3 META-LEARNING BASED FAIR RANKING

We aim to train a fairness-aware ranking model that could achieve good performance on both utility and fairness metrics. To do that, we tune the ranking model's loss weights values to make the model emphasize more on the protected group than the unprotected one during the ranking inference. Instead of using the fixed weights, we utilize a meta-dataset which is sampled from the original training dataset with an unbiased distribution and smaller size to train a meta-learner as a weighting function. The meta-learner could guide the ranking model to learn fairly.

Given the training dataset with a set of queries $\mathcal{Q}^{train}$ with $|\mathcal{Q}^{train}| = m$ and a set of items $\mathcal{D}^{train}$ with $|\mathcal{D}^{train}| = n$. Each query q from $\mathcal{Q}^{train}$ is associated with a list of item candidates $d^{(q)}$ from $\mathcal{D}^{train}$, and each item is represented as a feature vector $x_i^{(q)}$. For each query q, the feature vector $x^{(q)}$ is associated with the relevance score $y^{(q)}$. Let $f(x^{(q)}; w)$ be the ranking model and w represent all the learnable parameters in f. Then the output of the ranking model could be denoted as $\hat{y}^{(q)} = f(x^{(q)}; w)$. Generally, we learn the optimized parameters w^* by $\min_w \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(y_i^{(q)}, \hat{y}_i^{(q)})$ and $\mathcal{L}$ could be used as any ranking loss functions. However, equally treating $\mathcal{L}$ to each sample could lead the ranking model f unfair to minority groups since the heavy data bias issue in the training dataset. To address this challenge, we introduce a meta-learner $g(\cdot; \theta)$, parameterized by θ, to adaptively tune loss weights for each sample to achieve a fair exposure over diversity. Thus, we rewrite the training loss as the following:

$$\mathcal{L}^{train}(w; \theta) = \frac{1}{m}\sum_{i=1}^{m}\phi_i\mathcal{L}_i(w) = \frac{1}{m}\sum_{i=1}^{m}\phi_i\mathcal{L}(y_i^{(q)}, \hat{y}_i^{(q)}), \quad (1)$$

where $\hat{y}_i^{(q)} = f(x_i^{(q)}; w)$ represents the model output, and $\phi_i \in [0, 1]$ represents the i-th sample's loss weight given by the proposed meta-learner $g(\cdot; \theta)$. Notably, $\mathcal{L}^{train}(w; \theta)$ governed by the meta-learner's output weights is conditioning on a fixed θ and used for updating the ranking model's parameter w. For convenience, we denote $\mathcal{L}_i(w)$ as the original loss value of the i-th training data sample output from the ranking loss $\mathcal{L}$. Following [13], we develop the meta-learner g as a multi-layer neural network, which takes as input a loss value, and instantiate g as

$$\phi_i = g(\mathcal{L}_i(w); \theta) = g\Big(\mathcal{L}_i\big(y^{(q)}, f(x^{(q)}; w)\big); \theta\Big), \quad (2)$$

where i could be a sample from either the training dataset or the meta-dataset. We set the last-layer's activation function in g as a `sigmoid` so that the range of the output lies between 0 and 1. Eventually, we define a meta training loss function as

$$\mathcal{L}^{meta}\big(w(\theta)\big) = \frac{1}{s}\sum_{i=1}^{s}\mathcal{L}_i(w(\theta)). \quad (3)$$

Here we update the parameters of the ranking network by doing the gradient decent on a batch of a training data with the loss function in Eq. (1), and we can define $w(\theta)$ as:

$$\hat{w}^{(t)}(\theta) = w^{(t)} - \alpha\frac{1}{b}\sum_{i=1}^{b} g\big(\mathcal{L}_i^{train}(w^{(t)}); \theta\big)\nabla_w\mathcal{L}_i^{train}(w) \quad (4)$$

To train the meta-learner, we need to sample a small meta-dataset with $\mathcal{Q}^{meta}$ and $\mathcal{D}^{meta}$. The meta-dataset represents the meta-knowledge of the true distribution of the protected group and the other group, where $|\mathcal{Q}^{meta}| = s \ll m$ and $|\mathcal{D}^{meta}| = r \ll n$. In the meta-dataset, we denote the feature vector of each item as $x^{(q_{meta})}$ and the relevance score as $y^{(q_{meta})}$ given a query q_{meta} from $\mathcal{Q}^{meta}$. Similar to $\mathcal{L}_i^{train}(w)$, we denote $\mathcal{L}_i^{meta}\big(w(\theta)\big)$ as the loss value for each meta-dataset sample. The goal of the meta-learner $g(\cdot; \theta)$ is to leverage the unbiased meta-dataset to learn how to re-weight the loss values to train the model $f(\cdot; w)$ on the biased dataset. Since

Algorithm 1 The MFR Learning Algorithm

Input: Training dataset $\mathcal{Q}^{train}$,$\mathcal{D}^{train}$, meta-dataset $\mathcal{Q}^{meta}$,$\mathcal{D}^{meta}$, batch size b, d, max iterations T.
Output: Classifier network parameter $w^{(T)}$

1: Initialize ranking model's parameter $w^{(0)}$ and the meta-learner's parameter $\theta^{(0)}$.
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\{x^{q_{meta}}, y^{q_{meta}}\} \leftarrow$ SampleMiniBatch($\mathcal{Q}^{meta}$, $\mathcal{D}^{meta}$, d).
4: $\{x^{q_{train}}, y^{q_{train}}\} \leftarrow$ SampleMiniBatch($\mathcal{Q}^{train}$, $\mathcal{D}^{train}$, b).
5: Update $\hat{w^{(t)}}(\theta)$ by Eq. (4) with $\{x^{q_{train}}, y^{q_{train}}\}$.
6: Update $\theta^{(t+1)}$ by Eq. (9) with $\{x^{q_{meta}}, y^{q_{meta}}\}$.
7: Update $w^{(t+1)}$ by Eq. (10) with $\{x^{q_{train}}, y^{q_{train}}\}$.
8: **end for**

	W3C Experts (gender)		Engineering Students (high school type)		Engineering Students (gender)		Law Students (gender)		Law Students (race)	
	P@10	Fairness	K. Tau	Fairness	K. Tau	Fairness	K. Tau	Fairness	K. Tau	Fairness
ListNet [5]	0.178	0.759	0.390	1.070	0.384	0.858	0.202	0.931	**0.184**	0.853
Lambdamart [4]	0.095	0.738	0.355	1.002	0.326	0.907	0.199	0.979	0.156	0.847
DELTR γ_{small} [21]	0.178	0.785	0.390	1.075	0.384	0.860	0.201	0.958	0.173	0.874
DELTR γ_{large} [21]	**0.180**	0.827	**0.391**	1.075	0.370	0.976	0.188	0.993	0.130	1.014
FA*IR post [20]	0.178	0.824	0.390	1.070	0.384	0.886	0.182	0.965	0.140	0.944
FA*IR pre [20]	**0.180**	0.770	0.374	1.020	0.360	0.942	0.203	0.931	0.161	0.895
MFR-ListNet	0.115	0.775	0.385	0.990	**0.385**	0.855	**0.225**	0.901	0.182	0.848
MFR	0.126	**0.830**	**0.391**	**1.086**	0.352	**1.052**	**0.225**	**1.015**	**0.184**	**1.654**

Table 1: Experimental results. To measure fairness, we compute the exposure ratio between the protected and the non-protected group, so the values greater than 1.0 indicate greater visibility for the protected group and vice versa. For the ranking metric, higher Kendall's Tau / Precision@10(P@10) scores indicate better performance. The bold text indicates the model with the best performance, and the results show that the MFR model is better on the fairness metrics with comparable performance on the ranking metrics against other state-of-the-art models.

w is a function of θ, we naturally formulate the proposed MFR as a bilevel optimization problem and give the objective function as

$$\begin{aligned} &\min_{\theta} \mathcal{L}^{meta}\big(w^*(\theta)\big) \\ &\text{s.t. } w^*(\theta) = \arg\min_{w} \mathcal{L}^{train}(w;\theta). \end{aligned} \tag{5}$$

Loss Functions. The proposed MFR jointly considers utility and fairness metrics by developing a listwise ranking loss with an exposure term following the DELTR loss [21], given by

$$\mathcal{L}(y^{(q)}, \hat{y}^{(q)}) = \ell(y^{(q)}, \hat{y}^{(q)}) + \gamma U(\hat{y}^{(q)}), \tag{6}$$

where $U(\hat{y}^{(q)})$ is a listwise fairness measurement, $\ell(y^{(q)}, \hat{y}^{(q)})$ is a listwise loss based on Cross Entropy [5], and $\gamma > 0$ is a balancing parameter. To obtain optimal parameters w^* and θ^*, we minimize the training loss by

$$w^*(\theta) = \arg\min_{w} \mathcal{L}^{train}(w;\theta) = \frac{1}{m}\sum_{i=1}^{m} \phi_i \mathcal{L}_i^{train}(w), \tag{7}$$

and the loss for the meta-learner by

$$\theta^* = \arg\min_{\theta} \mathcal{L}^{meta}\big(w^*(\theta)\big) = \frac{1}{s}\sum_{i=1}^{s} \mathcal{L}_i^{meta}\big(w^*(\theta)\big). \tag{8}$$

Parameters Update. At each step t, we compute the weighted loss values with θ^t and w^t, and update θ with the loss of the ranking model on the meta-dataset as the following:

$$\theta^{(t+1)} = \theta^{(t)} - \beta \frac{1}{d}\sum_{i=1}^{d} \nabla_\theta \mathcal{L}_i^{meta}\big(w^{(t)}(\theta)\big), \tag{9}$$

where β is the learning rate, d is the batch size of the meta-dataset. After we have the $\theta^{(t+1)}$, we update w as the following:

$$w^{(t+1)}(\theta) = w^{(t)} - \alpha \frac{1}{b}\sum_{i=1}^{b} \phi_i \nabla_w \mathcal{L}_i^{train}(w), \tag{10}$$

where α is the learning rate and b is the batch size of the training dataset. We adopt an alternating optimization strategy [13, 16, 19] to implement Eq. (9) and Eq. (10) instead of using nested optimization loops. The whole training process is summarized in Algorithm 1.

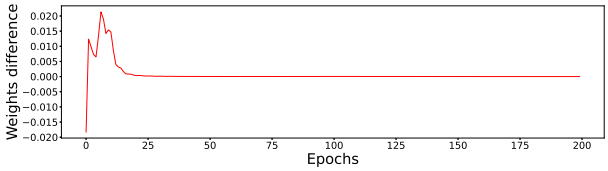
Although we consider the DELTR loss as the objective function of the ranking model, we could also use other fair ranking losses here. Besides the disparate exposure, there are other biases in the common ranking dataset such as selection bias and position bias. The model aims to provide a general meta-learning framework that can handle any fair ranking problems.

4 EXPERIMENTS

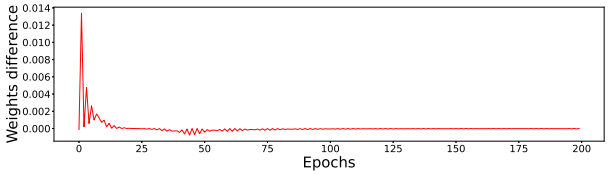
In the experiments, we train and evaluate the model on the three real-world datasets used in DELTR [21]. We study both the ranking and fairness metrics of our approach compared to other baseline models. The baseline models include the following: (i) ListNet [5]; (ii) Lambdamart [4]; (iii) the DELTR model with γ_{small} and γ_{large} which is the same setting as in [21]; (iv) the FA*IR [20] preprocessing approach that creates the fair dataset and trains on it; (v) the FA*IR post-processing approach that reorders the prediction results to ensure the fairness; (vi) MFR with different γ on a different dataset; (vii) MFR with the ListNet loss (MFR-ListNet). The code is available at https://github.com/ywang4/A-Meta-learning-Approach-to-Fair-Ranking.

For a fair comparison, we follow the same settings[1] as described in DELTR [21] to split the dataset and generate the item features. We use the following datasets: (i) W3C Experts (gender); (ii) Engineering Students (high school); (iii) Engineering Students (Gender); (iv) Law Students (gender); (v) Law Students (race). In the W3C Experts dataset, the task is the expert search originated from TREC 2005 Enterprise Track [8]. The protected attribute is female, and there are 200 items per query with an average of 21.5 items from the protected group. In the Engineering Students dataset, the task is the academic performance prediction, and the dataset contains anonymized historical information of college students. For the high school dataset, the protected attribute is public high school, and there are 480.6 items per query with 167.6 items from the protected group on average. For the gender dataset, the protected attribute is female, and there are 480.6 items per query with 97.6 items from the protected group on average. In the Law Students dataset, the task is also the academic performance prediction. For the gender dataset,

[1] https://github.com/MilkaLichtblau/DELTR-Experiments



(a) Engineering Students (high school)



(b) Law Students (gender)

Figure 3: The plot of the variation of learned weight over the two training datasets. The weight difference is computed as $\phi^t_{\text{diff}} = \frac{1}{m}\sum_{i=1}^{m}\phi^t_i - \phi^{t-1}_i$, and we plot the ϕ^t_{diff} over the training epochs. As shown in the plot, the weighting function is converging as the different values of weights between each epoch are decreasing to 0.0.

the protected attribute is female, and there is a total of 21791 items with 9537 items from the protected group. For the race dataset, the protected attribute is black, and there is a total of 19567 items with 1282 from the protected group. The queries are technical topics for the W3C dataset and academic years for the other datasets. For a fair comparison, we adapt the same evaluation metrics as [21]. To split the datasets, we have 50 queries for training and 10 queries for testing in the W3C dataset, 4 queries for training and 1 query for testing in the Engineering Students dataset, and 80% for training and 20% for testing in the Law Students dataset. We use Precision@10 (P@10) for the W3C dataset and Kendall's Tau for other datasets to evaluate the ranking performance. To measure fairness, we compute the exposure ratio between the protected and the non-protected group. Thus, in the fairness metric, values greater than 1.0 indicate greater visibility for the protected group and vice versa. As described in Sec. 3, the meta-dataset is required for our approach. Since the protected attribute in all datasets is binary, we perform random uniform sampling to collect the meta-dataset. Specifically, we randomly sample the same amount of data for the items from each query for each protected group and non-protected group.

Settings. In general, for the weighting function, we set the update frequency of the parameter $\theta$to be per 2 steps, the optimizer to be SGD, the momentum to be 0.98, the learning rate to be 0.02, the hidden layer dimension to be 30, and the number of hidden layers to be 3. For the ranking model, we set the learning rate for all datasets to be 0.005 except for W3C data to be 0.0005, the optimizer to be SGD, the momentum to be 0.95, and the weight decay to be 0.005. The values of γ and training epoch vary for different datasets: W3C dataset uses γ = 500 and 100 epochs, Engineering Students (high school) uses γ = 5000 and 500 epochs, Engineering Students (Gender) uses γ = 500 and 100 epochs, Law Students(gender) uses γ = 1200 and 3000 epochs, and Law Students (race) uses γ = 50000 and 100 epochs.

Results Analysis. As shown in Tab. 1, our approach performs better in terms of the fairness metrics on all datasets than both the DELTR γ_{small} and DELTR γ_{large}. The DELTR γ_{small} and DELTR γ_{large} models use different scales of γ values to weight the exposure measure in the loss function. With the meta learner, we can achieve higher fairness metrics by re-weighting the loss distribution during the training process. The intuition behind the observation is that the imbalanced pattern among the training data is observed and corrected by the meta learner. For the ranking metrics, we have similar or better results on all datasets except the W3C dataset. Since ListNet and Lambdamart do not consider any fairness measure during the training, the results are as expected that the fairness metrics are worse than the fairness ranking models. In addition, we train the MFR-ListNet that has the standard listwise ranking loss in the framework. The evaluation results show the worse performance on both the ranking and fairness metrics. As listwise loss does not consider the exposure measure, the meta-dataset that has a different data distribution as the training dataset has a negative effect on the meta-learner during the re-weighting process. Thus, we conclude that the meta-learning approach could help the model to further improve the fairness metrics compare to the model with only the DELTR loss function.

In Fig. 1, we plot the histogram of ranks on the protected attributes from the different models. From the plot, we can see the distribution of the predicted ranks shifts from right to left, which indicates the MFR model generally ranks the items from the protected group higher compared to ListNet. Note that at the plot, 1 means the top rank, so when more data samples fall in the bins at the left, the items receive higher ranks. The plot also agrees with the evaluation results. As we see that there is a large difference in Fig. 1b, the fairness metric of MFR on Law Students (race) dataset is about two times than that of ListNet.

In Fig. 3, we plot the variation of the learned weight for the training data. The plots show that the weighting function is converging as the different values of weights between each epoch are decreasing to 0. As suggested in Meta-Weight-Net [13], we use the multi-layer neural network as the weighting function because the multi-layer neural network is known as a universal approximator for the most continuous functions. The convergence shown in the plots indicates the successful learning process on the weighting function.

5 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a Meta-learning based Fair Ranking (MFR) model to improve the minority group's exposure. Our experiments on the real-world datasets demonstrate that our approach could achieve better fairness metrics compared to the fair ranking model without the meta-learning part. In the future, we will continue to study a better way to collect the meta-dataset as it is the key part to successfully training the weighting function.

6 ACKNOWLEDGMENTS

This work is supported in part by the Ciocca center research award at Santa Clara University.

REFERENCES

[1] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands). Association for Computing Machinery, New York, NY, USA, 1259–1276.

[2] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA). Association for Computing Machinery, New York, NY, USA, 2212–2220.

[3] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA). Association for Computing Machinery, New York, NY, USA, 405–414.

[4] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[5] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. 129–136.

[6] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming, 2018, July 9-13, 2018, Prague, Czech Republic*, Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella (Eds.), Vol. 107. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 28:1–28:15.

[7] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. AutoDebias: Learning to Debias for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 21–30.

[8] Nick Craswell, Arjen P De Vries, and Ian Soboroff. 2005. Overview of the TREC 2005 Enterprise Track.. In *Trec*, Vol. 5. 1–7.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1126–1135.

[10] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (jul 1996), 330–347.

[11] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference (Leibniz International Proceedings in Informatics, Vol. 94)*, Anna R. Karlin (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 33:1–33:17.

[12] Preethi Lahoti, Gerhard Weikum, and Krishna P. Gummadi. 2019. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. *2019 IEEE 35th International Conference on Data Engineering* (2019), 1334–1345.

[13] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems* 32 (2019).

[14] Dylan Slack, Sorelle A. Friedler, and Emile Givental. 2020. Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data. Association for Computing Machinery, New York, NY, USA, 200–209.

[15] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.

[16] Zhiqiang Tao, Yaliang Li, Bolin Ding, Ce Zhang, Jingren Zhou, and Yun Fu. 2020. Learning to Mutate with Hypergradient Guided Population. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 17641–17651.

[17] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 6035–6042.

[18] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA). Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages.

[19] Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Ruirui Li, and Zhenhui Li. 2020. Automated Relational Meta-learning. In *8th International Conference on Learning Representations*.

[20] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore). Association for Computing Machinery, New York, NY, USA, 1569–1578.

[21] Meike Zehlike and Carlos Castillo. 2020. *Reducing Disparate Exposure in Ranking: A Learning To Rank Approach.* Association for Computing Machinery, New York, NY, USA, 2849–2855.

[22] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching Code and Law: Achieving Algorithmic Fairness with Optimal Transport. *Data Min. Knowl. Discov.* 34, 1 (jan 2020), 163–200.

[23] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).

[24] Chen Zhao and Feng Chen. 2020. Unfairness Discovery and Prevention For Few-Shot Regression. In *2020 IEEE International Conference on Knowledge Graph*. 137–144.

[25] Chen Zhao, Feng Chen, and Bhavani Thuraisingham. 2021. Fairness-Aware Online Meta-Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, 2294–2304.

[26] Chen Zhao, Feng Chen, Zhuoyi Wang, and Latifur Khan. 2020. A primal-dual subgradient approach for fair meta learning. In *2020 IEEE International Conference on Data Mining*. IEEE, 821–830.