Michael Ekstrand
Boise State University

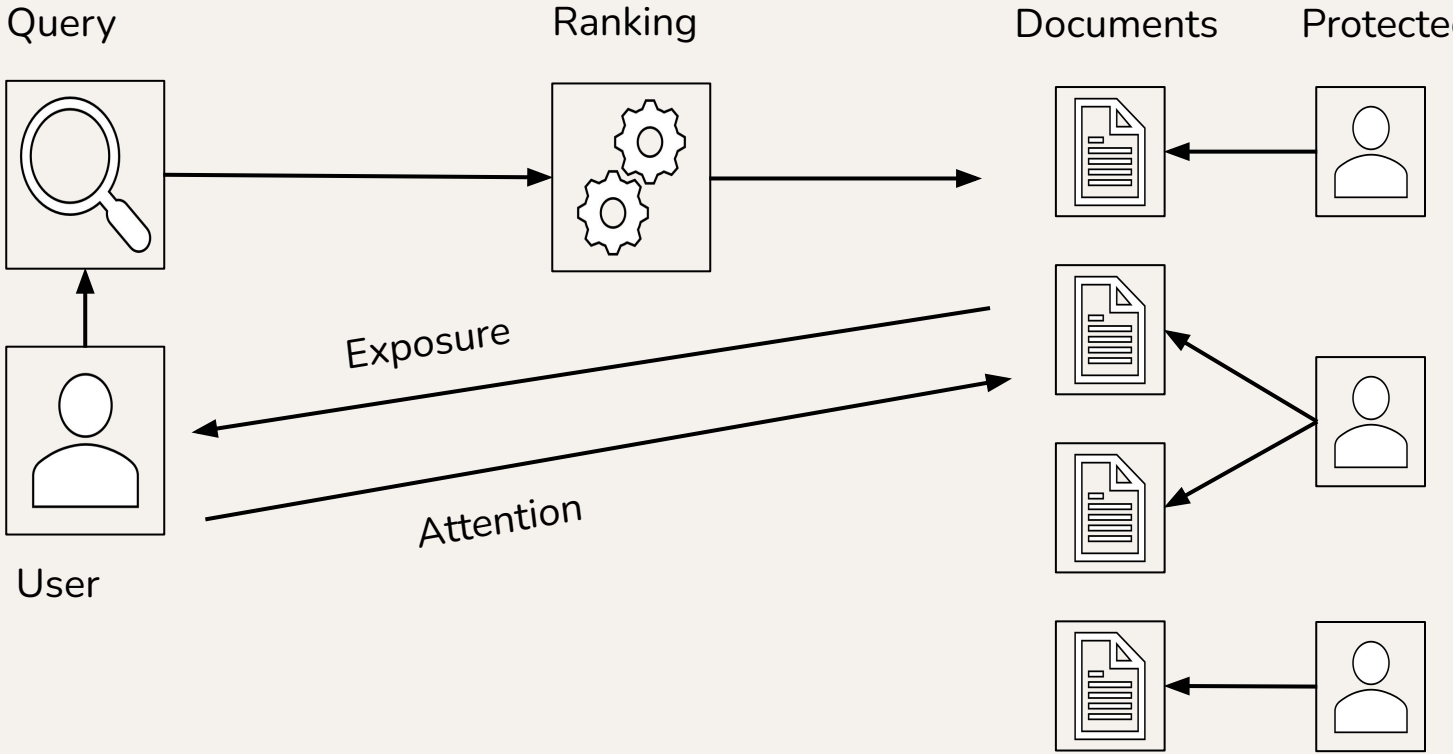
Graham McDonald
University of Glasgow

Amifa Raj
Boise State University

Isaac Johnson
Wikimedia Foundation

Morten Warncke-Wang
Wikimedia Foundation

TREC 2021 FAIR RANKING TRACK OVERVIEW



Do articles associated with protected attributes get fair exposure?

Fair Ranking Track 2021 vs Previous Years

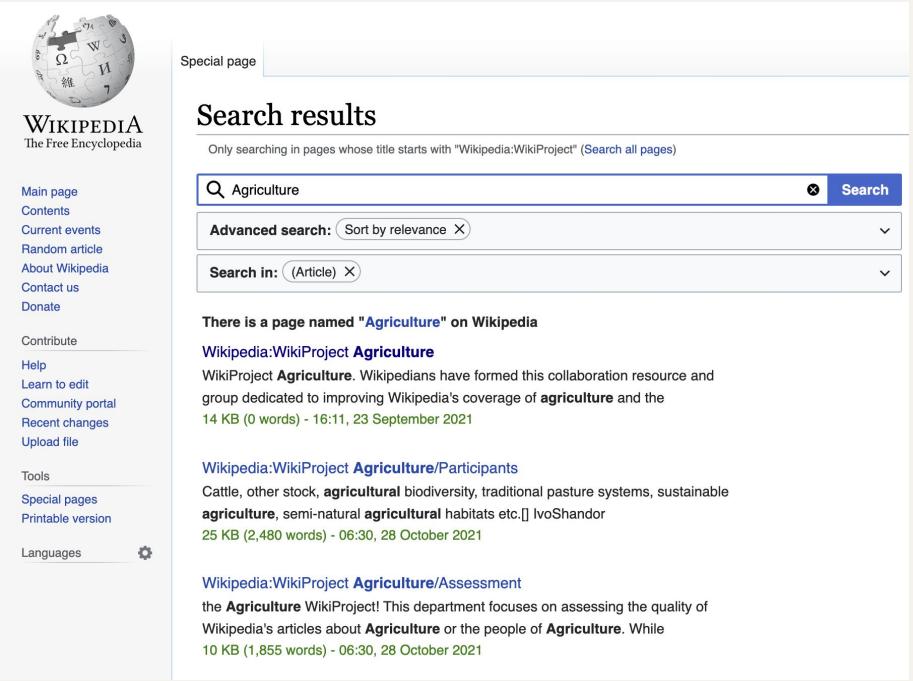
- 2019
  - Domain: Scholarly articles
  - Task: Re-ranking
- 2020
  - Domain: Scholarly articles
  - Tasks:
    - Task 1: Ad-hoc retrieval
    - Task 2: Re-ranking
- 2021
  - Domain: Wikimedia
  - Tasks: Ad-hoc retrieval
    - Task 1: Single ranking
    - Task 2: Stochastic ranking

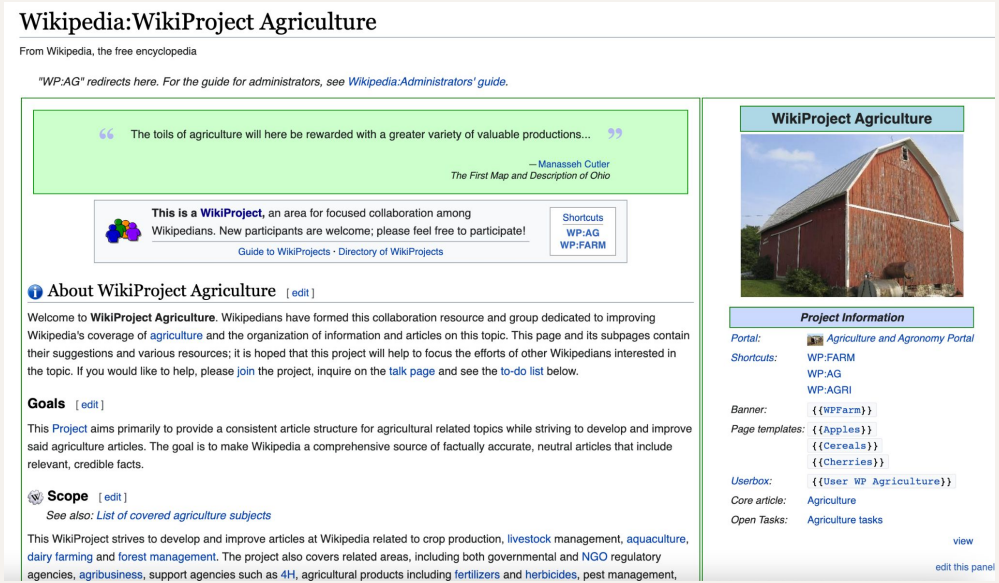
Equal Expected Exposure

“given a fixed information need, no item should have an expected exposure more or less than any other item of the same relevance.”

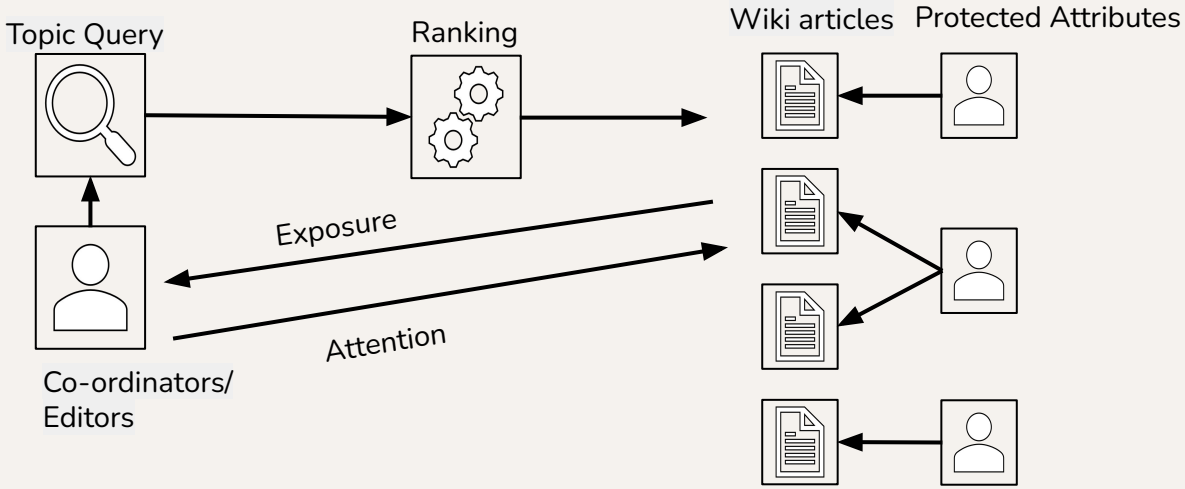
F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette. Evaluating stochastic rankings with expected exposure. CIKM, 2020.

WikiProject





WikiProject



Documents: A subset of English language Wikipedia article

Query: Topic form WikiProject

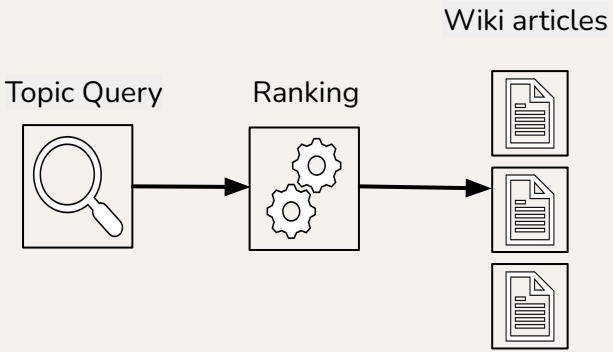
Fairness Objective: Ensure groups of articles associated with protected attributes get fair exposure

Task: ad-hoc retrieval

Task 1

Use Case: Help Wikiproject coordinators in finding articles for editors.

Single ranking per query



Evaluated based on

Geographic

Demographic

Relevance

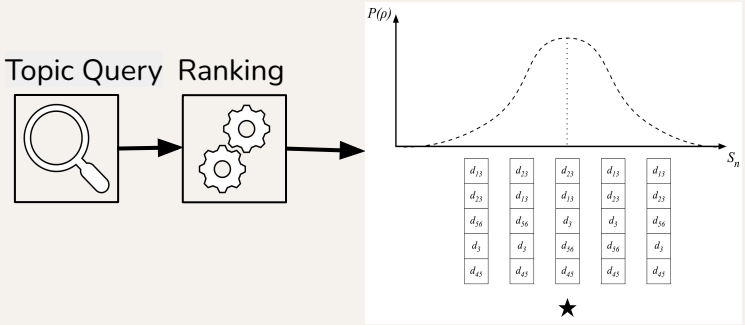
Fairness

Images are from https://unsplash.com/

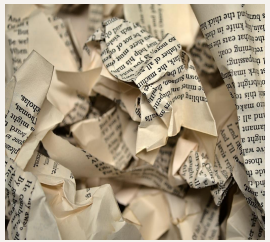
Task 2

Use Case: Help Wikiproject editors finding articles associated with a project (saved search)

Distribution over ranking



Evaluated based on



Geographic

Demographic

Relevance

Quality

Fairness

Images are from https://unsplash.com/

We Provided

WikiProject topic

English Wiki articles

Training data
- Geographic fairness ground truth

Evaluation data
- Hidden demographic attribute (gender)

Ranking Objectives

- Relevant documents come before irrelevant documents
- Fairness goal: Group exposure is fairly distributed according to the average of the distribution of relevant documents and the distribution of global population
- Relevant documents are sorted in nonincreasing order of work needed (Task 2)
    - Articles that need more editing will be on top
- For each query participants have to submit:
    - Task 1: single ranking consisting of 1000 articles
    - Task 2: 100 rankings each consisting of 50 articles

Annotations

NIST assessors annotated the articles with binary relevance score.
Annotation were incomplete because:

- Task 2 generates massive data (~700,000 article-topic pairs)
- Incomplete articles and not having enough information

We obtained assessment through tiered-pooling

- The first 20 items of all rankings for Task 1 (all queries)
- The first 5 items of the first 25 rankings from every submission to Task 2 (about 75% of the queries).

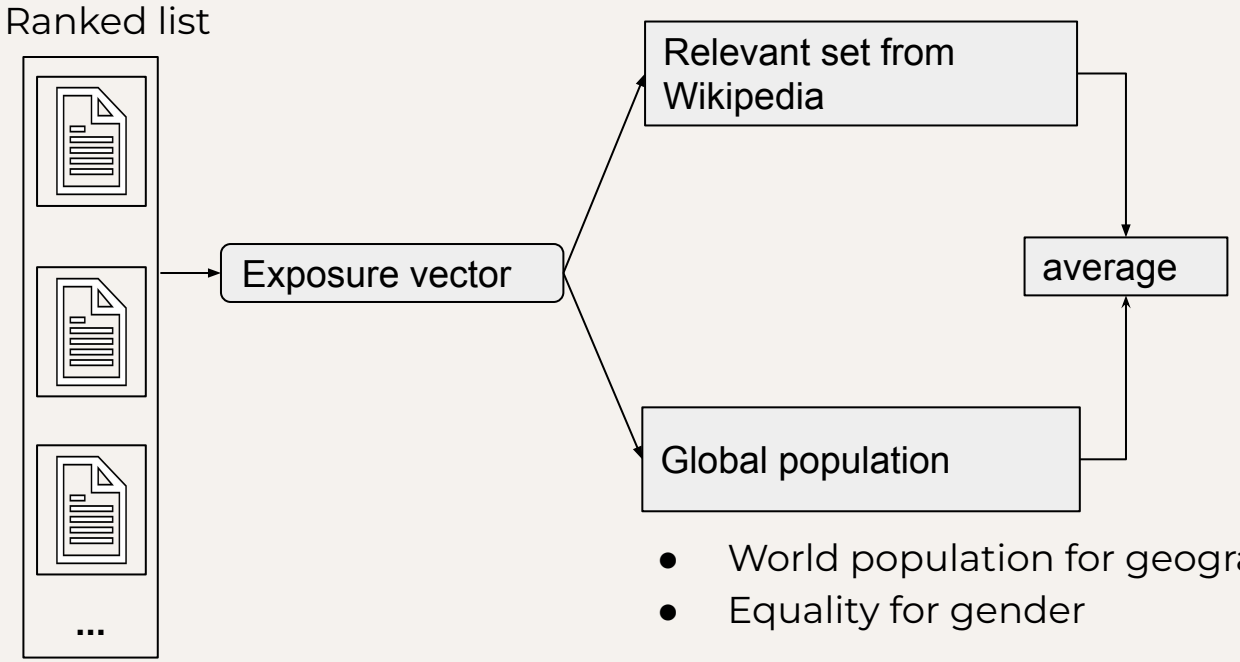
Evaluation Strategy

- Fairness with respect to *geographic* and *gender* information

$$G = G_{geo} \times G_{gender}$$

- “unknown” as a seperate group
- Log discounting for attention weight
- Measure exposure of group
- Compare with target exposure

Target Exposure



- Wikipedia has well-documented bias
- Imbalance in group distribution in topic relevance
- No data on ideal distribution for any particular topic

- World population for geography
- Equality for gender

Metrics for Single Ranking

- **Relevance Metric**: NDCG
- **Fair Ranking Metric:** Attention Weighted Rank Fairness (AWRF)
  - Jensen -shannon divergence between target and given exposure

$$AWRF(L) = 1 - d_{JS}(d_L, d^*_q)$$

d_L cumulated exposure a list gives to each group

d^*_q target exposure

- Combine fairness metric and relevance metric

$$M_1(L) = \mathrm{AWRF}(L) \times \mathrm{NDCG}(L)$$

Fairness Metric for Stochastic Rankings

Expected Exposure Metric

$\pi(\rho, q) = P(\rho, q)$

π ranking policy
q user query
ρ document ranking

$$EE(\Pi, q) = \|\varepsilon - \varepsilon^*\|_2^2$$

ε	expected exposure
ε^*	optimal expected exposure

F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette. Evaluating stochastic rankings with expected exposure. CIKM, 2020.

Fairness Metric for Stochastic Rankings

Expected Exposure Metric for Group Fairness

$\gamma_g = \sum_{d \in G_g} \varepsilon_d$

G_g set of documents associated with group g

γ $|G| \times 1$ group exposure vector

$EE_G(\pi, q) = \|\gamma - \gamma^*\|$

Expected-Exposure Disparity (EE-D)

$$EE_G(\pi, q) = \|\gamma\|^2_2 - 2\gamma^T\gamma^* + \|\gamma^*\|_2^2$$

Expected-Exposure Relevance (EE-R)

This is **Expected Exposure Loss (EE-L)**

F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette. Evaluating stochastic rankings with expected exposure. CIKM, 2020.

Submissions

- We received
  - Submissions from 4 teams for Task 1 (13 runs total)
  - Submissions from 3 teams for Task 2 (11 runs total)
- Approaches

Task 1

- RoBERTa for text fields
- BM25 for ranking
- Implicit diversification
- Tailored Diversification with Data Fusion
- Relevance-only

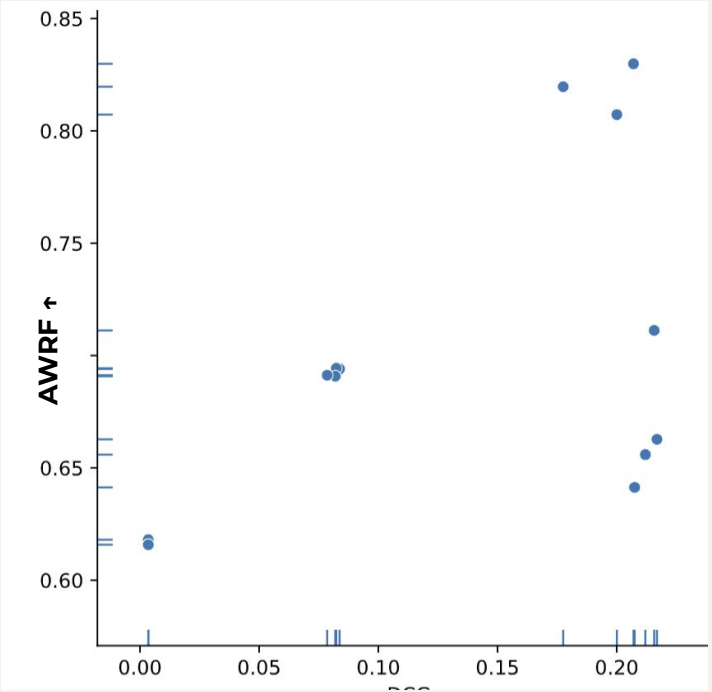
Task 2

- RoBERTa for text fields
- BM25 for ranking
- Tailored Diversification with Data Fusion
- Relevance-only

Result: Task 1

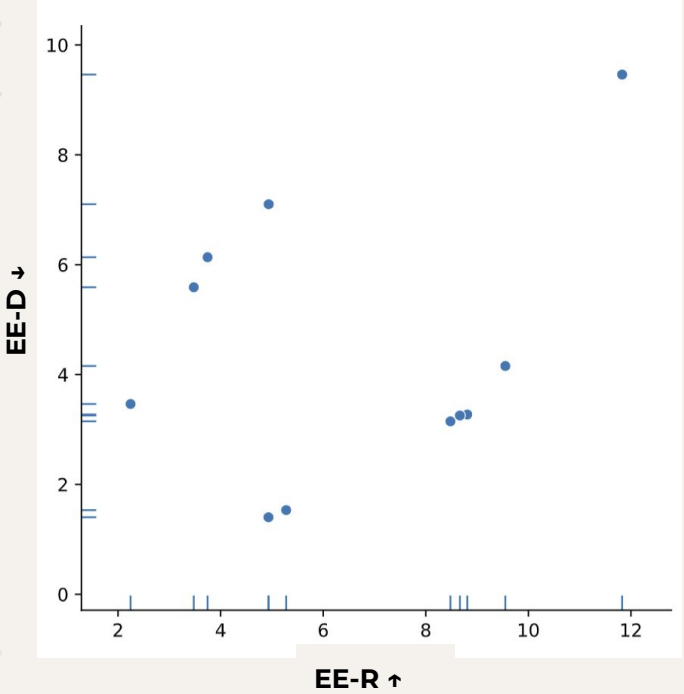
	nDCG	AWRF	Score (M_1)
UoGTrDExpDisT1	0.2071	0.8299	0.1761
UoGTrDRelDiT1	0.2001	0.8072	0.1639
UoGTrDivPropT1	0.2157	0.7112	0.1532
UoGTrDExpDisLT1	0.1776	0.8197	0.1459
RUN1	0.2169	0.6627	0.1425
UoGTrRelT1	0.2120	0.6559	0.1373
RMITRet	0.2075	0.6413	0.1317
1step_pair	0.0838	0.6940	0.0648
2step_pair	0.0824	0.6943	0.0638
1step_pair_list	0.0820	0.6908	0.0623
2step_pair_list	0.0786	0.6912	0.0607
RMITRetRerank_1	0.0035	0.6180	0.0026
RMITRetRerank_2	0.0035	0.6158	0.0026

Higher score is better



Result: Task 2

	EE-L (M_2)	EE-D	EE-R
RUN_task2	14.9007	4.1557	9.5508
pl_control_0.6	15.5017	3.2733	8.8091
UoGTrRelT2	15.6514	9.4609	11.8281
pl_control_0.8	15.7708	3.2550	8.6654
pl_control_0.92	16.0348	3.1486	8.4802
PL_IRLab_07	20.8213	1.5327	5.2790
PL_IRLab_05	21.3832	1.4029	4.9331
UoGTrDivPropT2	27.0726	7.1005	4.9372
UoGTrDRelDiT2	28.4816	5.5891	3.4770
UoGTrDExpDisT2	28.4903	6.1356	3.7459
UoGTrLambT2	28.8216	3.4644	2.2447



Lower EE-L is better

Lower EE-D is better; higher EE-R is better

Limitations

- Fairness Criteria
  - Geography: incomplete location information
  - Gender: possibility of misgendering
- Relevance Criteria
  - Missing relevance information
  - Coarse way to measure work needed
  - Incomplete assessment
- Task Definition
  - Doesn’t consider missing or deleted articles
  - There are more important protected attributes

Lessons Learned

- Our assessment budget was not enough for assessing the dataset generated by stochastic ranking
- Dealing with missing group labels was difficult in intersection

Fair Ranking Track Plenary Session is on November 16th (9am - 12pm)

Thanks

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**,and infographics & images by **Freepik**

grant IIS 17-51278