AMS42 at the NTCIR-18 FairWeb-2 Task

Clara Rus[(1)], Jasmin Kareem[(1)(3)], Chen Xu[(2)], Yuanna Liu[(1)], Zhirui Deng[(2)], Maria Heuss[(1)]

c.a.rus @uva.nl. (1)University of Amsterdam, (2)Renmin University of China, (3)Jheronimus Academy of Data Science

TASK OVERVIEW

ChuWeb 21D Web Page Collection

QUERY TOPICS

M (Movies) R (Researcher) Y (YouTube)

QUERY + DESCRIPTION ×

OVERVIEW OF RUNS (QUERY+DESCRIPTION)

RUN 1: BM25 + QUERY EXPANSION

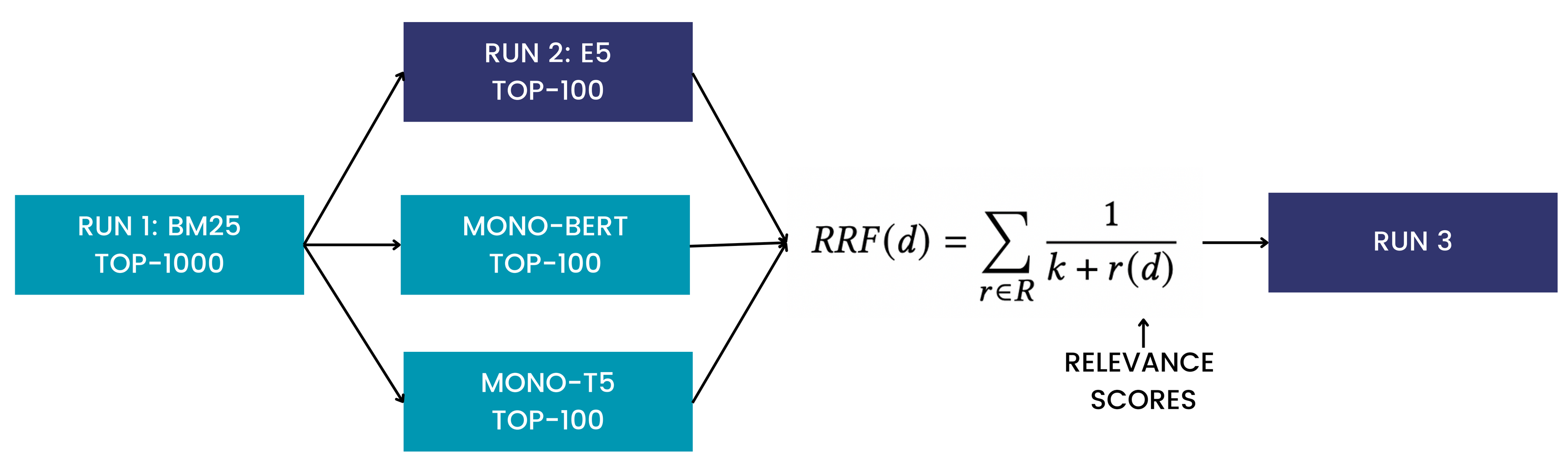
Query Expansion applied per topic:

Movies:
<movie/movies> on IMDb

Researcher:
<researchers/authors/coauthors> on Google Scholar

YouTube: <video/videos> - YouTube

RUN 2&3: FOCUSING ON RELEVANCE



RUN 4: IMPROVED MMR

Step 1: Estimation of Sensitive Attributes

Movies: Extracted movie name from title and searched on IMDb.

Researchers: Used Scholarly API to get first author via document title and their information (h-index, name); gender estimated from name.

YouTube: Extracted title before "- YouTube" and searched on YouTube.

Step 2: Apply MMR on RUN 2

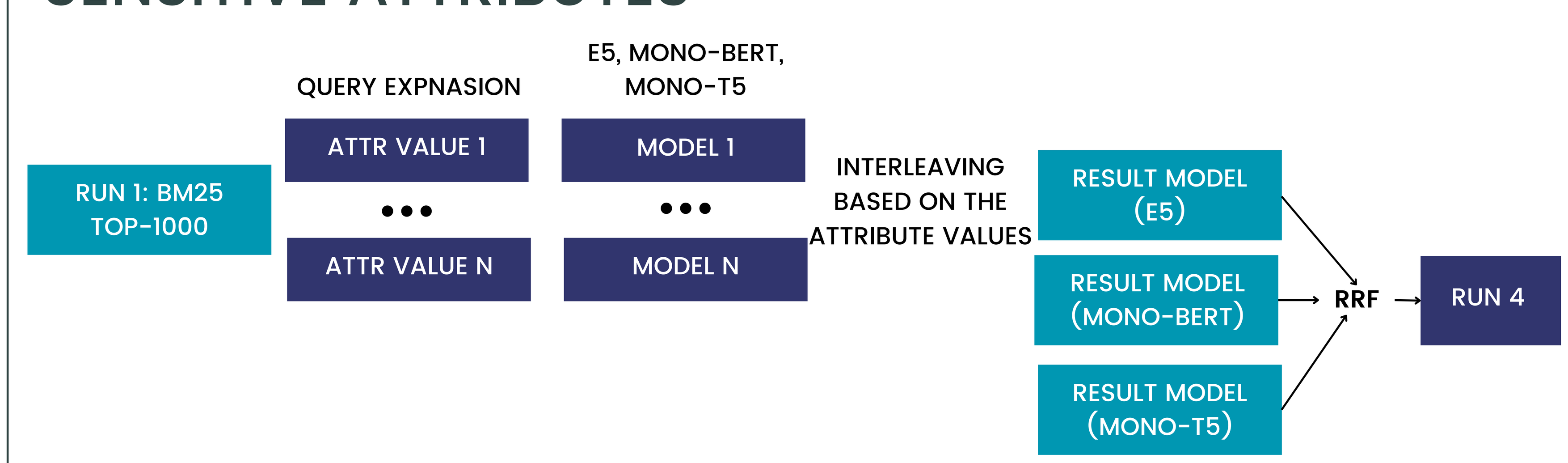
MMR selects a document d that maximizes the following objective function:

$$MMR(d) = \lambda \cdot r(q, d) + (1 - \lambda) \cdot f(S(d))$$

According to the evaluation metrics in the task requirements: Jensen-Shannon Divergence (JSD) and Root Normalized Order-aware Divergence (RNOD), MMR is adapted to optimize for:

$$f(S(d)) = \mathrm{JSD}(S(d), G) + \mathrm{RNOD}(s(d), G),$$

RUN 5: FUSION + QUERY EXPANSION WITH SENSITIVE ATTRIBUTES



M topic: <query> + " and these movies are from " + <attribute value>
R topic: <attribute value> + <researchers/authors/coauthors>
Y topic: None, as it is applied only for non-ordinal sensitive attributes

FAIRNESS RESULTS

Fairness Evaluation for Movie Topics

Run	Mean GF^{JSD} (ORIGIN)	Mean GF^{NMD} (RATINGS)	Mean GF^{RNOD} (RATINGS)	Mean GFR
AMS42-WS-QD-RG-1	0.2096	0.2329	0.2151	0.2235
AMS42-WS-QD-RG-2	0.4195 (>23)	0.4848 (>23)	0.4411 (>23)	0.4681 (>23)
AMS42-WS-QD-RG-3	0.3317 (>23)	0.3699 (>23)	0.3340 (>23)	0.3657 (>23)
AMS42-WS-QD-RG-4	**0.4491 (>23)**	**0.5029 (>22-23)**	**0.4644 (>22-23)**	**0.4877 (>22-23)**
AMS42-WS-QD-RG-5	0.3535 (>23)	0.3897 (>23)	0.3577 (>23)	0.3820 (>23)
Top 5 GFR of other teams runs				
RSLFW-WS-QD-RG-3	**0.4474** (>23)	**0.5207 (>22-23)**	**0.4496 (>23)**	**0.5101 (>22-23)**
RSLFW-WS-QD-RG-4	0.4465 (>23)	0.5110 (>22-23)	0.4464 (>23)	0.5044 (>22-23)
RSLFW-WS-QD-RR-2	0.4193 (>23)	0.5036 (>22-23)	0.4424 (>23)	0.4860 (>22-23)
RSLFW-WS-QD-RR-1	0.4176 (>23)	0.5006 (>22-23)	0.4409 (>23)	0.4835 (>22-23)
THUIR-WS-QD-REV-1	0.4034 (>23)	0.4973 (>22-23)	0.4437 (>23)	0.4758 (>22-23)

Fairness Evaluation for YouTube Topics

Run	Mean GF^{NMD} (SUBSCS)	Mean GF^{RNOD} (SUBSCS)	Mean GFR
AMS42-WS-QD-RG-1	0.0631	0.0560	0.0690
AMS42-WS-QD-RG-2	0.0645	0.0580	0.0715
AMS42-WS-QD-RG-3	0.0592	0.0502	0.0662
AMS42-WS-QD-RG-4	**0.0680**	**0.0635**	**0.0741**
AMS42-WS-QD-RG-5	0.0592	0.0502	0.0662
Top 5 GFR of other teams runs			
ORG-WS-run.qljm.Q	**0.2659 (>23)**	**0.2526 (>20-23)**	**0.2775 (>21-23)**
THUIR-WS-QD-RR-5	0.2484 (>23)	0.2401 (>23)	0.2531 (>23)
THUIR-WS-QD-RR-3	0.2407 (>23)	0.2322 (>23)	0.2437 (>23)
ORG-WS-run.bm25.Q	0.2367	0.2240	0.2368
THUIR-WS-QD-RR-1	0.2247	0.2153	0.2335

CONCLUSION

- Overall, higher performance leads to higher fairness.
- RUN 4&5 improved fairness across all topics in comparison with the other RUNS focused only on extracting relevant documents.
- Best results obtained on M (Movies) topic, while worst results are on Y (YouTube) topic. The dataset did not always contain clear links to actual YouTube videos, as opposed to IMDb pages.

RELEVANCE RESULTS

Run	M topics		Y topics		All topics	
	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU
AMS42-WS-QD-RG-1	0.0727	0.2460	0.0527	0.0821	0.0715	0.1817
AMS42-WS-QD-RG-2	**0.2027**	0.5437	0.0567	**0.0851**	0.1128 (>23)	0.2840 (>23)
AMS42-WS-QD-RG-3	0.1800	0.4314	0.0610	0.0823	**0.1142 (>23)**	0.2586 (>23)
AMS42-WS-QD-RG-4	0.1771	**0.5497**	0.0473	0.0847	0.0981 (>23)	**0.2897 (>23)**
AMS42-WS-QD-RG-5	0.1640	0.4349	**0.0610**	0.0823	0.0999 (>23)	0.2608 (>23)
Top 5 relevance of other teams runs						
RSLFW-WS-QD-RG-3	**0.2700**	**0.6332**	**0.1191**	**0.2540**	**0.2020 (>12-23)**	0.4556 (>15-23)
RSLFW-WS-QD-RG-4	0.2532	0.6204	0.1176	0.2340	0.1965(>12-23)	0.4434(>16-23)
RSLFW-WS-QD-RR-2	0.2523	0.5964	0.0939	0.2420	0.1942(>13-23)	**0.4613 (>15-23)**
RSLFW-WS-QD-RR-1	0.2506	0.5919	0.0912	0.2039	0.1931(>14-23)	0.4453 (>16-23)
THUIR-WS-QD-REV-1	0.2367	0.5804	0.0636	0.2327	0.1807 (>18-23)	0.4365 (>17-23)

CHALLENGES

- **Noisiness in the estimated sensitive attribute**
  → Difficulty in applying fairness interventions
  Solution: Query expansion with sensitive attribute values
  Limitation: It does not work for numerical attribute values (e.g. n of subscribers)

  → Difficulty in fairness evaluation
  – it is difficult to determine why certain approaches outperform others
  Solution: separating the evaluation of the estimation of attributes and the outcome of the fairness approach

- **Without first extracting relevant documents – estimating the sensitive attributes is not possible**
  E.g. if the document is somehow relevant to the query but it does not contain a YouTube video, one can't assess the sensitive attribute of this document

REFERENCES

[1] 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. 335–336.

[2] Gordon V Cormack, Charles LA 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. 758–759.

[3] Sijie Tao, Tetsuya Sakai, Junjie Wang, Hanpei Fang, Yuxiang Zhang, Haitao Li, Yiteng Tu, Nuo Chen, and Maria Maistro. 2025. Overview of the NTCIR-18 FairWeb-2 Task. In Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies.