Fair Group-Based Exposure of Documents in Ranked Search Results

Thomas Jänich

Submitted in fulfillment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

May 2025

Abstract

The main objective in the development and deployment of Information Retrieval (IR) systems has traditionally been to ensure that users receive relevant information, for example by ranking documents with respect to the documents' predicted relevance to the users' information needs, as represented by the users' queries. However, retrieval systems that only rank documents by predicted relevance can unintentionally introduce unfairness in the search results. In particular, search results can be unfair when groups of documents that are linked to specific entities such as individuals, organisations, or demographics are systematically ranked lower due to their associated attributes. Groups of documents that are ranked lower in the search results are less exposed to the user and therefore receive less attention. This lack of exposure can amplify biases and create unequal opportunities, thereby raising numerous ethical concerns and challenges.

Therefore, in this thesis, we address the task of ensuring fairness in the search results. In particular, we aim to ensure fairness of exposure over groups of documents. Such groups are formed based on shared characteristics, such as a common associated geographic location, language, or other given attributes.

Since maintaining relevance in the search results remains the main objective of every IR System, we aim to ensure fairness over the groups of documents without compromising the relevance or quality of the rankings. Indeed, search results that are perfectly fair according to a specific fairness definition but lack relevant documents would render an IR system ineffective and unusable. In this thesis, we provide insights into assessing exposure and demonstrate how IR systems using standard retrieval methods can be adapted to distribute exposure more fairly across groups of documents in search results while maintaining the quality of their results. We start our investigation by providing an overview of how exposure is distributed across document groups in search results. Based on our analysis, we argue that when adjustments to exposure distributions are necessary, modifying the ranked search results—by adding or removing documents from certain groups—is more effective than solely relying on re-ranking strategies. Moreover, we show that assessing the

expected exposure distribution before implementing fairness-enhancing modifications to IR systems is a critical step. By evaluating how exposure is likely to be allocated among document groups prior to retrieval, potential disparities can be detected and the necessity and type of required fairness interventions can be determined.

Specifically, we introduce a novel approach that predicts the exposure groups will receive prior to executing the retrieval process. Similar to Query Performance Prediction methods, which estimate the relevance of search results, our approach predicts how exposure will be distributed across groups of documents. Through experiments conducted over different document groups, we demonstrate that our proposed *Group Exposure Predictor* can assess the exposure distributions before traditional retrieval is performed.

To assess or predict the exposure received by document groups, it is necessary to know the labels that define these groups. However, in practice, such labels are often incomplete or entirely unavailable. To address this issue, we propose a robust method for assessing exposure in search results when document labels are missing, through the use of Quantification techniques. Unlike traditional classification methods that infer individual labels, Quantification focuses on estimating the distribution of labels across groups. We argue that this approach is more suitable for group exposure assessment, as it directly provides the aggregate information needed to evaluate group-level disparities, rather than relying on potentially inaccurate individual predictions.

To address unfair exposure distributions in search results, we propose adapting existing retrieval methods. Specifically, we argue that methods designed to increase the recall of search results are the most promising for adaptation, as adding more documents from underrepresented groups has a material impact on improving exposure distribution. To this end, we introduce a novel adaptation of the Graph-Based Adaptive Re-ranking (GAR) method to achieve a more fair exposure allocation over groups of documents in the search results. We propose several policies to adapt the GAR process and demonstrate through our experiments that each policy effectively improves the allocation of exposure across document groups in the search results, without compromising relevance.

Another effective approach to increasing the recall of relevant documents in search results is through query modification. In this thesis, we propose several query modification techniques that adapt and extend existing retrieval methods to improve the representation of underrepresented groups in the search results. These include techniques such as query expansion, where additional terms are added to the original query to retrieve more diverse and relevant documents and pseudo-relevance feedback, where the system analyses an initial set of retrieved documents to refine the query. By incorporating these approaches, we aim to retrieve more relevant documents for groups with a low representation in an initial retrieval, thereby enhancing recall and promoting a more balanced allocation of exposure in final rankings. To successfully modify a user's query, we propose to use pre-

trained large language models to generate query terms and full queries that help retrieve more documents from underrepresented groups. Through our empirical evaluation, we show that all of our techniques are suitable for improving the exposure distribution over groups of documents while maintaining the relevance and quality of the search results.

Overall, this thesis contributes to a growing body of research on fairness in IR systems. By addressing key challenges in balancing relevance and fairness, this work provides novel insights and practical methods for improving the equitable distribution of exposure in search results. Through the development of predictive models, robust measurement techniques, and adaptations of effective retrieval methods, including recent generative AI approaches, this thesis advances both the theoretical understanding and practical implementation of fairness-aware IR systems. These contributions not only enhance the design of more inclusive and socially responsible retrieval systems but also open new avenues for future research in this critical and evolving field.

Contents

List of Tables

List of Figures

List of Acronyms

ACC Adjusted Classify & Count. 58–61, 75

AvICTF Average Inverse Collection Term Frequency. 37, 47–49, 51, 52

AvIDF Average Inverse Document Frequency. 37, 47, 49, 51

AvPMI Averaged Pointwise Mutual Information. 37, 38, 47, 49, 51

AWRF Attention-Weighted Ranked Fairness. v, ix, xii, 31–33, 81, 92, 104, 118, 126, 129, 130, 132, 133, 138

CC Classify & Count. x, 57, 59, 71, 72, 74, 75

FGQE Fair Generative Query Expansion. vii, ix, 108–110, 112, 114–118, 126–133, 137, 138

GAR Graph-Based Adaptive Re-ranking. ii, vi–viii, xi, 76–82, 88, 92–98, 127–134, 136–139

GEP Group Exposure Prediction. viii, 35, 36, 42–44, 46, 48–53, 139

IID assumption Independent and Identically Distributed assumption. 58

IR Information Retrieval. i–iii, 1–4, 7–10, 17, 18, 20, 21, 34–36, 42, 47, 53, 72, 76, 77, 97, 107, 111, 126, 135, 137, 140, 141

JSD Jensen-Shannon Distance. 32, 47, 48

nDCG Normalised Discounted Cumulative Gain. v, ix, xii, 31, 33, 93, 104, 131–134

PACC Probabilistic Adjusted Classify & Count. x, 62, 68, 70–74

PLMs Pretrained Language Models. 10, 15–17, 19, 21, 108, 109, 112, 116, 118, 119, 125, 126, 128, 129, 135, 138

PMC Post-Metric Correction. 56, 57, 68–71

PMI Pointwise Mutual Information. 37, 38

PPS Prior Probability Shift. 58, 59, 61

PRF Pseudo-Relevance Feedback. vii, ix, 10, 13–15, 19–21, 42, 46, 50, 52, 53, 99–102, 105–109, 112–114, 116–119, 126, 137

QEP Query Exposure Prediction. 35, 36, 39, 44, 52–54, 136

QFE Query Fairness Estimation. 54, 55, 67, 75

QPP Query Performance Prediction. 35, 36, 38, 39

QSEE Query Similarity Exposure Enhancement. vii, xii, 119–121, 123, 125–134, 137, 138

RAE Relative Absolute Error. 64, 66, 72

SCS Simplified Clarity Score. 37, 47, 49, 51, 52

SSB Sample Selection Bias. 60, 68

List of Symbols

This list contains the most important symbols used in the chapters of this thesis.

β	Feedback weight controlling influence of expansion terms in ColBERT-PRF
γ	Smoothing weight for exposure-based re-ranking in QSEE
ζ	Binary variable indicating whether a document receives a favourable ranking outcome (e.g., appears in top-k)
η	Number of samples in the classifier training set L_{cls}
η'	Number of samples in the correction estimation set L_{corr}
λ_{RM3}	The smoothing parameter used in the RM3 query expansion model
ξ	Random variable indicating whether a document is relevant to a specific query ($\xi = 1$ for relevant, $\xi = 0$ for non-relevant) in quantification.
ρ	A position in a ranking
Υ	Candidate set of neighbours
ϕ	Instance-wise embedding function (used in quantification)
$\omega_{q_i}, \omega_{d_{ij}}, \omega_{x_{g,m}}$	Token embeddings for query, document, and expansion terms
$\mathcal{A}$	Sensitive attribute space in quantification
avgdl	Average document length across the collection
b_C	Initial batch of candidate documents passed to re-ranking in the GAR pipeline
b_F	Batch of frontier documents considered for re-ranking in the GAR pipeline
b_r	Scored batch of documents at a re-ranking step in the GAR pipeline
$CG = (VT, ED)$	Corpus graph representing the document collection, where VT represents the vertices and ED the set of directed similarity-based edges

d_ρ Document at rank position ρ

$D_{\text{top}k}$ The pseudo-relevant set of documents

dl Length (number of terms) in document d

$E(d_\rho)$ Exposure received by document d_ρ based on its rank

$E(g_l)$ Total exposure received by group g_l

$E_\tau(d)$ Exposure of document d at time τ

$E_{target}(g_l)$ Target exposure for group g_l

ED Set of directed edges in the corpus graph; each edge $(d_i, d_j) \in ED$ indicates d_j is among the k most similar documents to d_i

$\hat{E}(g_l)$ Predicted exposure for a document group g_l

F Frontier set of unexplored but potentially high-scoring documents in the GAR pipeline

G Set of groups of documents

g A single group of documents

$\text{IDF}(q_t, d)$ Inverse document frequency of term q_t

k Rank cutoff; the number of top-ranked items considered during evaluation

L_{cls} Labelled dataset used for training a classifier

L_{corr} Labelled dataset used for estimating correction factors

L_q Query-biased subset sampled from L_{corr} for estimating query-specific correction factors

$\mathcal{L}(n)$ First-stage retrieval function producing top-n candidates in the GAR pipeline

m Total number of groups considered in fairness evaluation

N Total number of documents in the corpus

N_t Number of documents containing term q_t

$\hat{p}_i, p_i$ Estimated and true prevalence of class i

$\mathbf{p}'$ Solution estimate for class prevalence

$\mathbf{p}, \mathbf{p}^k, \mathbf{p}^*$ Estimated class prevalences

q A query

q_t, q_{t_1}, q_{t_2} Individual terms within a query

q_μ A generated query

q_e An expanded query

Q A set of queries

r, r_f Original and re-scored relevance scores

rank_C Initial candidate ranking from first-stage retrieval in GAR

rank_R Final re-ranked document list after applying GAR

rel_ρ Relevance score of document at position ρ

$s(q, d), s(q_e, d)$ Similarity scores for document d with query q or expanded query q_e

$S(b)$ Re-ranking function applied to batch b of documents in the GAR pipeline

$\text{TF}(q_t, d)$ Term frequency of term q_t in document d

$\text{TF-IDF}(q_t, d)$ TF-IDF score of term q_t in document d

$\mathcal{V}$ Vocabulary

VT Set of vertices in the corpus graph; each vertex corresponds to a document

v_ρ Exposure value of document at position ρ, defined as $\frac{1}{\log_2(\rho+1)}$

V_q, V_d Sparse vector representations of query q and document d

$XT_{e,g}$ Expansion terms associated with group g

$\mathcal{X}$ Input label space in quantification

$\mathcal{Y}$ Class label space in quantifcation

$\hat{Y}, Y$ Predicted and true labels

Z_i Final set of k selected neighbours of document d_i in the corpus graph after fairness constraints are applied

Acknowledgements

I want to thank my supervisors, Graham McDonald and Iadh Ounis, for their unwavering support and direction over the course of my PhD. Their insight, encouragement, and expertise have been invaluable throughout this journey. I am especially grateful for their patience and critical feedback, which helped me grow both intellectually and personally. Moreover, I want to thank all of my colleagues from the Terrier Team, who made life in and outside the office both enjoyable and enriching, and who greatly contributed to the fantastic years I spent in Glasgow.

I feel especially fortunate to have had the friendship and support of Hitarth Narvala, Javier Sanz-Cruzado Puig, Jack McKechnie, Erlend Frayling, Ting Su and Sean MacAvaney. Moreover, I am thankful for all my colleagues from the different offices in the School of Computing Science. In particular, I want to express my gratitude to Haruna Umar Adoga, José Rodríguez Bacallado, Andreas Chari, Shen Dong, Sophie Fischer, Jinyuan Fang, Debasis Ganguly, Lubingzhi Guo, Sarawoot Kongyoung, Craig Macdonald, Richard McCreadie, Fabricio Mendoza, Zeyuan Meng, Aleksandr Petrov, Alexander Pugantsov, Eliyas Suleyman, Fanzheng Tian, Xiao Wang, Yaxiong Wu, Xinhao Yi, and Zixuan Yi, whose daily companionship and encouragement created a warm and supportive atmosphere that made even the busiest days enjoyable.

I am deeply grateful to my family for their constant support, patience, and encouragement throughout this journey. I am especially thankful to my loving parents, Ursula and Herbert, whose belief in me has never wavered and to my brother Sebastian and my sister Anna, for always being there with kindness, humour, and perspective.

Finally, I want to express my deepest gratitude to my partner, Toja, for her love, patience, and belief in me. Her strength, encouragement and steady presence through every high and low have been invaluable. Her support has meant more to me than words can express.

Declaration

With the exception of chapters 1, 2 and 3, which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

Traditionally, IR systems are deployed to find the most relevant documents to a user's query. In a standard IR system, users can issue a query, and the IR system retrieves and ranks documents based on their likelihood of being relevant to the users' request. In modern IR systems, deep neural rankers that use contextualised language representations, such as BERT (Devlin et al., 2019), have demonstrated high effectiveness in retrieving the most relevant documents for a user (Lin et al., 2022). However, the scoring functions of these neural rankers can typically be computationally intensive and impractical for use over large document collections (Hofstätter & Hanbury, 2019). Consequently, IR systems typically integrate neural rankers into a two-stage process (usually referred to as a re-ranking pipeline). In this process, an initial set of documents is first retrieved using an inexpensive retrieval model, such as BM25 (Robertson et al., 1994), and then this set of documents is re-scored using a neural re-ranker. This *re-ranking* process generally enhances the relevance of the final ranking for the user and often achieves state-of-the-art performance in retrieval tasks (Lin et al., 2022).

However, deploying IR systems that prioritise retrieving the most relevant documents can be beneficial for users but may have undesirable consequences for the creators of those documents or the entities mentioned within them. When documents relate to entities, for example, people, organisations, or geographic locations, the ranking decisions can significantly impact the lives and opportunities of those ranked entities (Pedreschi et al., 2008). Moreover, by disadvantaging certain entities, IR systems can be at odds with existing legislation such as the Equality Act 2010[1] or the EU AI Act[2].

[1] https://www.legislation.gov.uk/ukpga/2010/15/part/2/chapter/1

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

To prevent discrimination, it is important for an IR system to carefully manage the ranking positions of groups of documents in the search results. This consideration is crucial since users exhibit a *position bias*, i.e., a tendency to pay more attention to documents that appear higher up in the rankings. Hence, the further down a document is positioned in the ranking, the less likely it will be *exposed* to the user, i.e., the document receives less exposure to the user (Craswell et al., 2008).

The concept of *exposure* refers to the attention that documents receive in the search results. The exposure of a document can be modelled using different user browsing models. In this thesis, we use a logarithmic discount function in our user browsing model, which reflects the decreasing likelihood of users viewing documents as they move down the ranked list.

This user browsing model is one of the most prominent in the fairness literature for assessing exposure (Ekstrand et al., 2019; Raj et al., 2020) which is exemplified by its use in the evaluation of the 2021 & 2022 TREC Fair Ranking Tracks (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022).

When IR systems systematically rank relevant documents from certain groups lower than others, this results in reduced *exposure* to the user for these groups. This reduced exposure to the IR system's users can profoundly impact the economic, social, and professional opportunities of entities associated with these groups (Pedreschi et al., 2009).

For example, Wikipedia relies on ranking algorithms to ensure its articles receive fair exposure to editors and readers. If articles associated with certain demographics are consistently ranked lower, these articles have fewer opportunities for their work to be edited, leading to inequality (Raghavan et al., 2020).

To maintain the Wikipedia articles, a search engine is used by Wikipedia editors to find articles within WikiProjects. WikiProjects are collaborative efforts where groups of contributors work together to improve Wikipedia's coverage of specific topics, such as science, history, or biographies. This search engine ranks articles by predicted relevance scores. For example, the IR system might rank three articles about the male figures and three articles about the female figures within a particular WikiProject. Assume that the predicted relevance scores for the articles about the male figures are 0.82, 0.81, and 0.80, while the scores for the articles about the female figures are 0.79, 0.78, and 0.77.

If the articles about the male figures are ranked at positions 1, 2, & 3, their exposure values according to the user browsing model can be quantified as 1, 0.63, & 0.50 respectively, totalling 2.13. On the other hand, if the articles about the female figures are ranked at positions 4, 5 & 6, their exposure scores are 0.43, 0.39, & 0.36, totalling 1.18. The total exposure for all articles is $2.13+1.18 = 3.31$. The proportion of exposure for articles about the male figures is $\frac{2.13}{3.31} \approx 64.35\%$ and for articles about the female figures is $\frac{1.18}{3.31} \approx 35.65\%$. Despite the close relevance scores, there is a high exposure disparity.

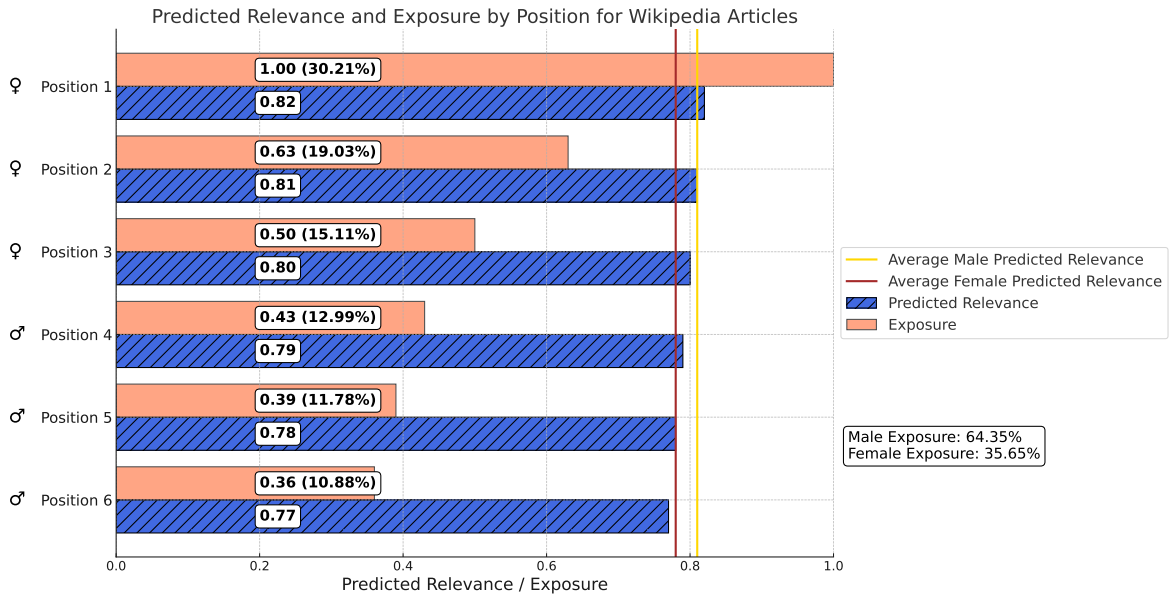


Figure 1.1: The figure shows an example using ranked Wikipedia articles, illustrating that small differences in relevance can lead to significant differences in exposure.

Figure 1.1 further illustrates this example, showing the predicted relevance scores and exposure levels for male and female articles across the six positions in the ranking. Blue bars with double slashes represent the predicted relevance scores, while solid orange bars show the exposure levels. Annotations on the bars indicate specific scores and exposure percentages. The plot highlights the average relevance values using distinct lines: solid yellow for average male predicted relevance and solid brown for average female predicted relevance. Observing the relevance and exposure values in the plot, the impact of the ranking on the two groups becomes apparent. Despite the small difference in average relevance scores (c.f. solid vertical lines), the impact on exposure is much more substantial, highlighting the importance of a fair exposure allocation in the search results.

As the example has shown, IR systems are increasingly deployed in contexts where ranking decisions can affect the exposure of people, organisations, or topics, hence it is crucial to account for retrieval effectiveness as well as exposure. Documents associated with marginalised or underrepresented groups may receive less exposure, not due to lower relevance, but as a result of the ranking process itself. This discrepancy has implications for fairness, particularly when exposure impacts downstream opportunities for the entities described in the retrieved documents. To address this, recent work in fair IR has introduced approaches for measuring and mitigating such disparities (Biega et al., 2020; Ekstrand et al., 2019, 2021; Ekstrand, McDonald et al., 2022; Singh & Joachims, 2018). However, much of this work remains limited to post hoc adjustments or settings where group labels are explicitly known. There is a need for methods that are both accurate and sensitive to fairness requirements, especially in practical systems. Therefore, in this

thesis, we contribute to the growing research on fairness in exposure by proposing different approaches to ensure a fair exposure in IR systems while maintaining satisfactory quality, i.e., utility in the search results. We demonstrate how existing state-of-the-art IR techniques can be adapted to achieve a fair allocation of exposure over multiple groups of documents while maintaining the relevance of the search results.We address crucial issues that can introduce unfairness in IR systems. For example, we show that if too few documents from a group are retrieved, achieving a fair exposure in a ranking becomes infeasible even if the IR system deploys a re-ranking pipeline that allows for an effective two-stage ranking process. To overcome this limitation, we introduce a novel method to predict the exposure distribution among document groups in a ranking to indicate when fairness interventions in an IR system are needed. Furthermore, we tackle another issue commonly found in practical applications regarding the availability of group labels. While most fairness interventions rely on labels of document groups, they are seldom available or complete. Therefore, we present a robust approach to measure exposure in rankings even when group membership labels are unknown, i.e., there is no explicit information about which documents belong to which group. In addition, we present various solutions to ensure fair exposure for underrepresented groups of documents in the final ranking. In this thesis, we focus on state-of-the-art IR methods that are particularly useful for fairness adaptation due to their ability to improve the effectiveness of a system by increasing recall. All of our fairness approaches offer distinct perspectives on the problem of fair exposure, enabling us to address fairness under different assumptions about group information, data distributions, and ranking pipeline configurations. Our proposed fair ranking methods include modern query modification techniques such as deep neural pseudo-relevance feedback, query generation, and adaptive re-ranking methods. Each approach is tailored to align with our fairness objectives, enhancing the exposure of underrepresented groups in the search results.

1.1 Scope of Thesis

To make a meaningful contribution to the literature on fairness in IR systems, it is essential to clearly define the scope of our research. Universal solutions to the fairness problem are challenging to achieve in practice since different value systems require distinct mechanisms for fair decision-making and fairness must typically be understood in context (Friedler et al., 2021). This involves acknowledging that fairness definitions are based on different, often conflicting world views and assumptions. As we will elaborate more in the upcoming chapters, the task and the context of fairness must be carefully defined to provide effective solutions. In the following paragraph, we define the chosen scope for this thesis in the context of IR systems, with more detailed background provided in Chapter 3.

In this thesis, we focus on the fairness of exposure for groups of documents in rankings. These groups are broadly defined and can, but do not necessarily, align with protected groups as defined by existing legislation. Examples of such categories include popularity, gender, race, and geographic location. Our research aims to ensure that each group within a category receives equitable exposure in search results, mitigating biases that could disadvantage particular groups. We adhere to the principle of statistical parity (Ekstrand, Das et al., 2022), aiming to ensure that every group has a fair chance of exposure to users. This principle requires that the likelihood of a document being displayed is independent of its group membership, preventing systematic disadvantages in the search results.

Under this principle, we investigate how the documents of all groups within a category can be positioned in the ranking to give them a fair, i.e., equal chance of exposure to the user. To illustrate, when focusing on the category of geographic locations, we ensure for example, that documents from each continent are fairly represented in the search results. This approach aims to prevent any single continent from dominating the top positions, thus giving documents from all continents a fair, i.e., equal chance to be seen by users. In this thesis, we focus on investigating multiple independent groups. This means we only consider the fairness of exposure within one single category.

1.2 Thesis Statement

The statement of this thesis is that it is possible to enhance the fair distribution of exposure over groups of documents in a ranking without significantly decreasing the relevance of the search results, i.e., the utility of the resulting information retrieval system for the user. We argue that modern ranking systems, that typically deploy re-ranking pipelines can be effectively adjusted to allocate the exposure in rankings more fairly. By predicting and assessing the number of documents per group in a ranking using Query Performance Prediction and Quantification, we can effectively tailor and expand existing state-of-the-art retrieval techniques towards the objective of a fair exposure distribution.

In particular, we posit that established retrieval methods aimed at increasing retrieval effectiveness by improving the recall using a two-stage process, such as Query Reformulation and Generation, as well as Adaptive Re-ranking and Diversification, can be successfully adapted to ensure both fair and relevant search results.

1.3 Contributions

The key contributions of this thesis can be summarised as follows:

1. We emphasise the importance of retrieving a sufficient number of documents per group (high recall per group) for a fair exposure allocation, highlighting the limitations of existing re-ranking approaches typically deployed in IR systems.

Existing works on fair exposure distribution (c.f. Chapter 3) primarily focus on re-ranking results from the first-stage retrieval. Our research, however, reveals the limitations of these re-ranking methods. When too few documents from certain groups of documents are retrieved initially, achieving a fair exposure allocation for all groups, through re-ranking alone is challenging and often infeasible. Therefore, we address the shortcomings of the first-stage retrieval (Chapters 6 & 7) by proposing methods that discover a sufficient number of documents per group to be able to ensure an equitable distribution of exposure.

2. We introduce a novel approach (Chapter 4) to predict the exposure groups of documents will receive in the search results after the first-stage retrieval.

To choose a suitable intervention that ensures a fair exposure allocation among groups in the rankings, it is crucial to estimate how the exposure is distributed over groups in the results of the first-stage retrieval. If too few documents from underrepresented groups are retrieved initially, fairness can be infeasible to achieve in the final ranking. To indicate when fairness interventions are needed, we propose the first exposure predictor, which can predict the exposure distribution in a single ranking produced by a lightweight retriever, i.e., in the first-stage retrieval of a re-ranking pipeline. This predictor helps identify potential imbalances early on, enabling us to potentially deploy fairness interventions (Chapters 6 &7) that aim to ensure that enough documents from every group are retrieved to achieve a fair representation in the final ranking.

3. In Chapter 5, we provide a reliable method to measure exposure when the group membership of documents is not explicitly known.

To measure the exposure groups of documents receive, the group membership of the documents needs to be known. While these labels are typically available, there are cases where they are not, for example, due to legal or privacy constraints (Bogen et al., 2020; Holstein et al., 2019). In such scenarios, exposure must be measured under "unawareness". We propose a robust quantification-based method that measures the exposure groups receive in a ranking even when the labels need to be inferred. This approach allows for an accurate assessment of exposure and, hence, the assessment of the need for fairness interventions without needing explicit group membership labels.

4. We introduce several adaptations of state-of-the-art information retrieval techniques (Chapters 6 & 7) to improve the fair allocation of exposure. Specifically, we adapt techniques that are widely used in modern re-ranking pipelines to improve recall in standard retrieval systems. Our adaptations of these techniques are designed to improve exposure allocation in the search results across different groups of documents while maintaining the utility of the system.

We show that retrieval techniques that are usually deployed to enhance the relevance of the search results, such as Pseudo-Relevance Feedback (Jaleel et al., 2004; X. Wang et al., 2021), Query Generation (Nogueira, Lin & Epistemic, 2019; Gospodinov et al., 2023), Generative Query expansion (Mackie et al., 2023) as well as adaptive re-ranking (MacAvaney et al., 2022) can ensure a sufficient number of documents in the first stage retrieval and hence can ensure fairness. All these techniques are typically deployed in re-ranking pipelines with the goal of discovering more relevant documents after an initial retrieval. In this thesis, we use their recall enhancing capabilities and tailor them towards fairness.

1.4 Origins of Material

Most of the material presented in this thesis has previously appeared in several conference papers published in the course of this PhD programme:

- In Chapter 4 we present our novel predictor for exposure called Group Exposure Predictor. This work has first been published in the proceedings of the ECIR 2024 conference (Jänich et al., 2024c).

- In Chapter 5, we propose a novel approach using quantification techniques to reduce the error in measuring exposure when the labels are unknown and need to be inferred. This work is currently under review at the Journal of Artificial Intelligence Research.

- In Chapter 6 we show how adaptive re-ranking techniques can be tailored to overcome limitations in the first-stage retrieval, such as too few retrieved documents per group. We show that our approach improves the fair allocation of exposure in the search results. This work first appeared in the proceedings of the SIGIR 2024 conference (Jänich et al., 2024a).

- In Chapter 7 we show how different Query Modification techniques can be used to improve the allocation of exposure over groups of documents while maintaining the utility of the IR system. Our work on Pseudo-Relevance Feedback for exposure fairness has been published in the proceedings of the ECIR 2023 conference (Jänich et al., 2023). Another work in this chapter based on how queries can be generated

has been published in the proceedings of the ECIR 2024 conference (Jänich et al., 2024b). Furthermore, in this chapter, we cover how prompting Pretrained Language Models can be used to generate query terms that help ensure fairness, which is work based on a full paper accepted at the ECIR 2025 conference.

1.5 Thesis Outline

The remainder of this thesis is structured as follows:

- Chapter 2 introduces the existing IR techniques that we adapt and evaluate in their ability to improve the fairness of the search results. Specifically, we introduce techniques, that increase recall of a system and highlight their limitations.
- Chapter 3 provides essential background on fairness in IR systems. Since there is no universal definition of the fairness problem in IR, we propose a brief taxonomy of how fairness definitions can be classified. After a short description of the existing fairness notions, the scope of the thesis is narrowed down to the fairness of exposure for groups of documents. Furthermore, we present our evaluation measures.
- Chapter 4 presents the novel task of *query exposure prediction* and a first solution for the task. We introduce a new predictor called Group Exposure Predictor, which can predict the exposure groups of documents receive in a first-stage retrieval.
- In Chapter 5, we investigate how exposure can be reliably measured when the group membership of documents is unknown, i.e., in the scenario of "unawareness." We show that naive approaches that deploy simple classification to infer group labels are insufficient for an accurate measurement of exposure. Therefore, we introduce an approach based on quantification that is more accurate in estimating group proportions in rankings and, hence, in measuring the exposure groups receive.
- In Chapter 6, we investigate how adaptive re-ranking pipelines, which are initially designed to increase the recall after a first-stage retrieval by using corpus graphs, can ensure a fair exposure allocation. Using customised versions of the adaptive re-ranking pipeline, we demonstrate that, when sufficient documents from each group are available, they can be effectively identified and incorporated into the ranking. This ensures that the resulting ranking is both fair and relevant.

- Chapter 7 further investigates how the distribution of exposure in the rankings can be improved, compared to standard systems that focus on relevance only. This chapter is specifically concerned with IR techniques that use modifications of the query to retrieve documents from every group of documents after a first-stage retrieval. We show that by using adaptations of existing query modification techniques such as Pseudo-Relevance Feedback, Query Generation, and Generative Query Expansion, exposure can be allocated more fairly amongst groups of documents.

- In Chapter 8, we compare our proposed query modification techniques with our modified adaptive re-ranking approach and conclude which methods are best used in which scenarios and where the limitations of the approaches apply.

- Chapter 9 concludes this thesis by summarising the contributions and conclusions presented throughout the chapters. Additionally, we explore potential directions for future work, highlighting areas where further research could expand upon our findings and discuss potential limitations of this thesis.

Chapter 2

Background

In this chapter, we provide the necessary background that is needed to conduct our research on fairness in the search results of IR systems. The primary goal in the development of fair IR systems is two-fold. First, the traditional objective of effectiveness must be maintained, i.e., the system should produce search results that are relevant to the user. Second, the IR system must ensure fairness in the search results according to a specified definition. Therefore, our primary objective in this thesis is to incorporate fairness into the search results. While fairness is our main focus, we also ensure that the retrieved documents are relevant. In other words, while our research is mainly on achieving fairness, relevance is also an important consideration that supports the primary goal of fairness. For the relevance of the search results, we build on existing effective retrieval techniques. The chapter is structured as follows:

- In Section 2.1 we introduce how documents can be represented using sparse models.
- In Section 2.2 , we introduce widely used retrieval models that rely on sparse document representations, i.e., TF-IDF (Section 2.2.2) or BM25 (Section 2.2.3).
- In Section 2.3 we cover sparse PRF techniques that can further increase the effectiveness of rankers in retrieving the most relevant items.
- In Section 2.4 we present Pretrained Language Models (PLMs).
- In Section 2.5 we show how PLMs can be deployed to retrieve documents. We cover different architectures and introduce the standard application of models in a re-ranking pipeline (Section 2.5.5).
- Finally, in Section 2.6 we give a summary of the presented content of this chapter.

2.1 Sparse Representations

In sparse retrieval models, queries V_q and documents V_d are represented as vectors in a high-dimensional space, with each dimension corresponding to a unique term in the vocabulary. These vectors are sparse because only a subset of terms typically appears in a document or query, resulting in many zero values. Consider the following documents:

- d_1: "The dog barks."
- d_2: "The cat meows."
- d_3: "The bird sings."

From these documents, a vocabulary can be formed: {*bird*, *barks*, *cat*, *dog*, *meows*, *sings*, *the*}. A document-term matrix can then be generated, where each row represents a document and each column a term. The resulting matrix is:

Document	bird	barks	cat	dog	meows	sings	the
d_1	0	1	0	1	0	0	1
d_2	0	0	1	0	1	0	1
d_3	1	0	0	0	0	1	1

From this matrix, we can obtain the vector representation for d_1 as $V_{d_1} = [0, 1, 0, 1, 0, 0, 1]$. The query "cat meows" would result in a sparse representation of: $V_q = [0, 0, 1, 0, 1, 0, 0]$. To calculate the values for each dimension in the vector, several methods can be used. For example, the term frequency (TF) (Salton & Buckley, 1988), which counts the number of times a term appears in a document, or the term frequency-inverse document frequency (TF-IDF) (Salton & Buckley, 1988), which adjusts the term frequency by considering how common or rare a term is across the entire document collection. Representing documents and queries in a sparse space enables efficient matching based on query terms (Singhal et al., 2001). This is facilitated by the use of inverted indices, which link terms to lists of documents where they appear, along with their frequency or position within each document. Preprocessing steps, such as stop-word removal (Kaur & Buttar, 2018) and stemming (Balakrishnan & Lloyd-Yemoh, 2014), further enhance the representation by reducing noise and consolidating different forms of the same word. Having demonstrated how sparse representations can be constructed for documents and queries, in the next section we explain how the sparse representations can be used for document retrieval.

2.2 Sparse Retrieval

2.2.1 Vector-Space Retrieval Models

A straightforward way to estimate the relevance between a query vector $\mathbf{V}_q$ and a document vector $\mathbf{V}_d$ is to compute the *dot product* of these vectors:

$$\text{Dot}(\mathbf{V}q, \mathbf{V}d) = \sum j = 1^{|\mathcal{V}|} Vq, j; \cdot; V_{d,j}, \tag{2.1}$$

In the equation $|\mathcal{V}|$ is the size of the vocabulary, and $V_{q,j}$ (or $V_{d,j}$) denotes the weight of the j-th term q_t in the query (or document). An alternative formulation is the *cosine similarity* (Salton et al., 1975), which normalises the dot product by the magnitudes of the query and document vectors thereby compensating for differences in vector lengths:

$$\text{CosineSim}(\mathbf{V}q, \mathbf{V}d) = \frac{\sum j = 1^{|\mathcal{V}|} Vq, j; \cdot; V_{d,j}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} (V_{q,j})^2}; \times; \sqrt{\sum_{j=1}^{|\mathcal{V}|} (V_{d,j})^2}}, \tag{2.2}$$

2.2.2 TF-IDF Retrieval Model

Although raw term counts can be used directly in the query and document vectors, additional weighting schemes often lead to more effective retrieval. One prominent approach is *TF-IDF* (Sparck Jones, 1972). TF-IDF combines term frequency (tf), which increases the weight of terms q_t occurring frequently in a document, and inverse document frequency (idf), which emphasises terms that occur infrequently in the corpus. A commonly adopted variant of idf is shown below:

$$\text{idf}(q_t, d) = \log\Big(\tfrac{N+1}{N_t+1}\Big), \tag{2.3}$$

In the equation N is the total number of documents in the collection, and N_t is the number of documents containing the term q_t. The $(+1)$ terms avoid division by zero. The TF-IDF weight for a term q_t in a document d can be expressed as:

$$\text{TF-IDF}(q_t, d) = \text{tf}(q_t, d) \times \text{idf}(q_t, d), \tag{2.4}$$

In the expression $\text{tf}(q_t, d)$ is the raw count of q_t in d. Terms that appear often in a specific document yet rarely across the corpus receive higher TF-IDF weights. A similar scheme can be applied to the query, after which the retrieval score may be computed using either the dot product (Equation 2.1) or cosine similarity (Equation 2.2).

2.2.3 Okapi BM25 Retrieval Model

TF-IDF does not explicitly account for document length normalisation or repeated term occurrences. Therefore, probabilistic retrieval models, such as *Okapi BM25* (Robertson et al., 1994), have been introduced to address these factors. The BM25 score for a document d given a query q is defined as:

$$\text{BM25}(q,d) = \sum_{q_t \in q \cap d} \log\left(\frac{N - N_t + 0.5}{N_t + 0.5}\right) \times \frac{\text{tf}(q_t, q)\,(k_2+1)}{\text{tf}(q_t, q) + k_2} \times \frac{\text{tf}(q_t, d)\,(k_1+1)}{\text{tf}(q_t, d) + k_1\left(1 - b + b\,\frac{dl}{\text{avgdl}}\right)}, \tag{2.5}$$

In this formulation, $\text{tf}(q_t, d)$ and $\text{tf}(q_t, q)$ denote the term frequency of the query term q_t in document d and in query q, respectively. N is the total number of documents in the corpus, and N_t is the number of documents that contain the term q_t. The variables dl and avgdl denote the length of document d and the average document length in the corpus. The parameters k_1, k_2, and b control the influence of term frequency saturation and document length normalisation. In this thesis, we deploy BM25 as our primary sparse retriever.

2.2.4 Limitations of Sparse Retrieval models

BM25 and TF-IDF are efficient in terms of computation and storage. However, their effectiveness is limited by their inability to predict the relevance of documents if there is no exact lexical match between the query and the document representations (Krovetz & Croft, 1992). Since the retrieval models based on sparse representations use individual terms as the basis for similarity prediction, they can typically only retrieve documents that contain the exact words or phrases specified in the query. This results in significant challenges when users use different terms than those present in the documents. For example, a query for *automobile* may not retrieve documents that use the term *car*. This issue is commonly referred to as the lexical mismatch problem (Krovetz & Croft, 1992).

A potential solution to the lexical mismatch problem involves modifying the user-issued query using PRF. Applying Pseudo-Relevance Feedback (PRF) mechanisms can improve retrieval effectiveness by bridging vocabulary gaps between queries and relevant documents (Keikha et al., 2018; Xu et al., 2009; Lv & Zhai, 2010). Therefore, in the next section, we introduce techniques that can further enhance retrieval performance.

2.3 Pseudo-Relevance Feedback

There are several ways to modify a user's query to improve the effectiveness of standard retrieval methods, including query expansion, query suggestions, or PRF (Carpineto & Romano, 2012). A commonly used strategy for modifying a user's query is PRF (Amati & van Rijsbergen, 2002; Krovetz & Croft, 1992). PRF assumes that the top-ranked documents from an initial retrieval, referred to as the *pseudo-relevant set* $D_{\text{top } k}$, contain relevant information. Therefore, PRF mechanisms extract terms from $D_{\text{top } k}$ and incorporate them into the original query, which is then resubmitted. This process can mitigate vocabulary mismatch by including terms in the query, that were not originally present but appear in the retrieved documents.

In this thesis, we deploy two well-known PRF methods: the RM3 (Jaleel et al., 2004) relevance language model and the KLQE model (Amati & van Rijsbergen, 2002). These methods differ in how they estimate the importance of terms in the pseudo-relevant set, yet both expand a query with terms that help overcome the vocabulary mismatch problem.

RM3 Relevance Language Model The RM3 model (Jaleel et al., 2004) expands a query by incorporating additional terms from a pseudo-relevant set of documents. Expansion terms from the documents are selected using a weighted combination of their importance in the original query and their presence in the pseudo-relevant set. Formally, this can be expressed as

$$P(t \mid q_e) = (1 - \lambda_{RM3})\, P(t \mid q) + \lambda_{RM3} \sum_{d \in D_{\text{top } k}} P(t \mid d)\, P(d \mid q) \tag{2.6}$$

The equation shows that for every potential expansion term t, its association with the original query terms q is measured. $P(t \mid q)$ represents the importance of t in the original query using the relative term frequency. The second component, $P(t \mid d)P(d \mid q)$, measures how often t appears in a document from the pseudo-relevant set ($P(t \mid d)$), weighted by the relevance of the document to the query (($P(d \mid q)$). ($P(d \mid q)$ can be derived from predicted relevance scores, e.g., using BM25. λ_{RM3} controls the degree of expansion. The terms with the highest expansion scores $P(t \mid q_e)$ are added to the query.

KL Query Expansion (KLQE) KL Query Expansion (KLQE) (Amati & van Rijsbergen, 2002) selects expansion terms by computing the Kullback-Leibler (KL) divergence between term distributions in the pseudo-relevant set and in the background collection. The KL divergence is used to quantify how much the term distribution in the pseudo-relevant documents deviates from that of the entire corpus. KLQE is defined as:

$$D_{KL}(P_{\text{feed.}}||P_{\text{corpus}}) = \sum_t P_{\text{feedback}}(t) \log \frac{P_{\text{feedback}}(t)}{P_{\text{corpus}}(t)} + (1 - P_{\text{feedback}}(t)) \log \frac{1 - P_{\text{feedback}}(t)}{1 - P_{\text{corpus}}(t)}$$

In the equation $P_{\text{feedback}}(t)$ represents the frequency of term t in the pseudo-relevant set, computed as:

$$P_{\text{feedback}}(t) = \frac{\text{TF}_{\text{feedback}}(t)}{\sum_{t'} \text{TF}_{\text{feedback}}(t')} \tag{2.7}$$

Moreover, $P_{\text{corpus}}(t)$ represents the probability of term t appearing in the entire corpus:

$$P_{\text{corpus}}(t) = \frac{\text{TF}_{\text{corpus}}(t)}{\sum_{t'} \text{TF}_{\text{corpus}}(t')} \tag{2.8}$$

KLQE selects expansion terms that have a higher frequency in the pseudo-relevant set compared to the corpus, thereby identifying terms that are characteristic of relevant documents. To account for statistical uncertainty in term frequency estimation, an additional correction factor is typically applied. This correction term stabilises the KL divergence computation by mitigating the impact of small sample sizes and extreme term probabilities in the pseudo-relevant set. We use the implementation of KLQ from PyTerrier.[1]

Limitations of PRF with Sparse Representations

While PRF methods such as RM3 and KLQE help mitigate the lexical mismatch problem by adding new terms to the query, they still rely on explicit keyword matching (H. Li, Xu et al., 2014). They do not inherently capture deeper contextual relationships among words. For instance, a query containing the ambiguous term *apple* may retrieve only documents referring to the fruit or only documents referring to the technology company, depending on which sense of *apple* can be found in the pseudo-relevant set (Levy, 2021). Without explicit semantic modelling, PRF expansions cannot easily resolve such ambiguities. Consequently, potentially relevant documents that use different terms may not be retrieved. To address these shortcomings, PLMs, such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) have been introduced.

2.4 Pretrained Language Models

PLMs are trained on large-scale text datasets to learn the statistical patterns of language. PLMs generate embeddings, i.e., numerical representations of words and phrases, that capture semantic meanings and relationships. The development of PLMs has been enabled by the transformer architecture (Vaswani et al., 2017). Unlike earlier methods, such as

[1] https://pyterrier.readthedocs.io/en/latest/rewrite.html

convolutional neural networks (Z. Li et al., 2021) and Long Short-Term Memory (LSTMs) networks (Hochreiter & Schmidhuber, 1997), which rely on sequential processing, transformers use self-attention mechanisms to process input in parallel. This capability allows transformers to model complex relationships between words in a sentence. Consequently, transformer-based models can be effectively deployed for several Natural Language Processing tasks, such as language translation (Lan et al., 2020), text classification (Devlin et al., 2019) or text generation (Raffel et al., 2020). The transformer architecture introduced by Vaswani et al. (2017) serves as the foundation for modern PLMs. The transformer consists of an encoder and a decoder. The encoder processes an input sequence to produce a representation that captures dependencies between words. This is done through self-attention and position-wise feed-forward neural networks. The decoder takes the encoded representation and generates an output sequence. It combines the information from the encoder with its own self-attention over the generated tokens to produce relevant outputs. The way the encoder and decoder are used defines different model architectures, which we will discuss in the next section.

Types of PLMs

PLMs can be grouped based on their use of the transformer architecture, which determines how they process and generate information. The three main categories are encoder-decoder, encoder-only, and decoder-only models.

- **Encoder-Decoder PLMs:** Encoder-decoder models such as T5 (Raffel et al., 2020) or BART (Lewis et al., 2020) apply both components of the transformer architecture. The encoder converts the input sequence into a representation and the decoder generates the output sequence. These models are commonly applied to tasks such as translation, summarisation and text generation. For instance, T5 reformulates NLP problems into text-to-text tasks. In Chapter 6, we deploy a re-ranker based on the T5 models within a re-ranking pipeline. In Chapter 7, we use a T5-based model to generate relevant queries for documents.

- **Encoder-Only PLMs:** Encoder-only models focus entirely on the encoder component of the transformer architecture. They process input sequences to generate embeddings that reflect relationships within the text, making them effective for tasks such as classification, question answering, and ranking. Examples include BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020). ELECTRA uses a pretraining method called replaced token detection, where the model predicts whether each token in a sequence has been substituted. This technique improves both computational efficiency and the quality of the learned embeddings. In Chapter 6, ELECTRA serves as the basis for our re-ranking model.

- **Decoder-Only PLMs:** Decoder-only models such as the GPT series (Radford et al., 2019) and LLaMA (Touvron et al., 2023) rely solely on the decoder component, generating text by predicting the next word in a sequence based on the preceding context. These models are particularly effective for text-generation tasks. In Chapter 7, we deploy decoder-based models to generate query expansion terms.

In the next paragraphs, we discuss how the PLMs are typically deployed in IR systems.

2.5 Dense Retrieval

PLMs have been effectively deployed in IR systems due to their ability to capture contextual information in their representation of text (Devlin et al., 2019; Khattab & Zaharia, 2020; Nogueira, Yang, Cho & Lin, 2019; Mitra, Craswell et al., 2018). To find relevant documents, PLMs represent queries and documents as dense vectors (embeddings) in a continuous space. Dense retrieval contrasts with the sparse retrieval methods discussed in Section 1.2, which rely on exact term matching. While sparse models represent queries and documents in high-dimensional spaces using the term presence or absence, dense retrieval captures semantic similarity through embeddings (Zhao et al., 2024).

Dense retrieval models can be categorised into two main architectural paradigms: cross-encoder and bi-encoder architectures (Zhao et al., 2024). Cross-encoders jointly encode the query and document, enabling detailed token-level interaction, but they are computationally intensive. In contrast, bi-encoders encode queries and documents separately, allowing for more efficient retrieval by enabling pre-computation of document embeddings. In the following subsections, we describe the two different architecture paradigms.

2.5.1 Cross-Encoder Architecture

Cross-encoder architectures (MacAvaney et al., 2019; Nogueira & Cho, 2019) process both the query and document as a single input sequence. This sequence typically includes a special separator token (`[SEP]`) and often begins with a classification token (`[CLS]`) (Zhao et al., 2024). By jointly encoding the query and document, cross-encoders capture detailed interactions between tokens from both the query and the document. The relevance score is computed from the output representation of the `[CLS]` token or an aggregated representation of the entire sequence.

In this thesis (Chapters 5-8), we use the monoT5 model (Nogueira et al., 2020), which employs a cross-encoder approach by feeding each query–document pair into T5's encoder-decoder framework. The term mono in monoT5 signifies that the model processes one query–document pair at a time, generating a relevance score.

We also use ELECTRA (Clark et al., 2020) in a cross-encoder configuration for re-ranking. ELECTRA concatenates each query with a candidate document into a single input sequence and encodes them to produce a relevance score. This design uses a replaced-token-detection pretraining objective, which yields robust contextual representations for more accurate query–document matching. Although cross-encoder re-ranking is more computationally expensive than bi-encoder retrieval, it remains tractable when applied only to a small set of documents initially retrieved by a lightweight ranker (e.g., BM25).

2.5.2 Bi-Encoder Architecture

Bi-encoder models (Devlin et al., 2019; Khattab & Zaharia, 2020; Karpukhin et al., 2020; Reimers & Gurevych, 2019) process queries and documents separately through their respective encoders, generating independent dense vector embeddings. A similarity function—such as cosine similarity—then calculates a relevance score (Zhao et al., 2024). Bi-encoders can be broadly categorised into single-representation and multi-representation types, based on how they encode queries and documents (Macdonald et al., 2021).

Single-Representation Bi-Encoder

This type of bi-encoder represents each query and document with a single dense vector, making it particularly efficient for large-scale retrieval tasks. Examples include models such as RoBERTa (Liu et al., 2019) and ANCE (Approximate Nearest Neighbour Negative Contrastive Estimation) (Xiong et al., 2021). However, representing an entire document with a single vector may limit the model's ability to capture fine-grained token-level interactions, which can be important when precise query–document alignment is required.

Multi-Representation Bi-Encoder

Multi-representation models generate multiple dense vectors per document, enabling detailed, token-level matching. For instance, ColBERT(Khattab & Zaharia, 2020) encodes queries and documents separately and produces multiple dense vectors for each. Instead of directly comparing aggregated embeddings, ColBERT uses a late interaction mechanism that computes relevance by finding the maximum cosine similarity between each query token embeddingω_{q_i}, and document token embeddings $\omega_{d_{ij}}$. This token-level interaction supports fine-grained matching and can be very effective. However, ColBERT's late interaction mechanism increases computational cost and query-time latency due to the need for real-time token-level similarity computations. Its multi-vector representation requires significantly more storage and complicates indexing compared to traditional dense retrieval models. Additionally, ColBERT does not integrate easily with classical IR systems and demands substantial resources for training and fine-tuning.

2.5.3 Neural Sparse Retrieval—SPLADE

SPLADE (Sparse Lexical and Expansion) (Formal et al., 2021) is a neural retrieval method that integrates PLMs while maintaining sparse representations. Instead of encoding queries and documents into dense embeddings, SPLADE predicts term importance in the vocabulary space, producing weighted sparse vectors that remain compatible with traditional inverted indexes. Unlike explicit query expansion, SPLADE learns contextualised term activations while applying self-regularisation techniques, such as L_1 regularisation or Kullback-Leibler (KL) divergence, to enforce sparsity. By balancing lexical precision with neural generalisation, SPLADE effectively bridges sparse and dense retrieval. We deploy SPARSE in Chapter 4 as a lightweight retrieval model. Next, we introduce how PRF mechanisms can be applied in the dense space.

2.5.4 Dense Pseudo-Relevance Feedback

While dense retrieval models reduce lexical mismatch by using contextualised representations, they can still fail to fully capture user intent (Devlin et al., 2019; Lin et al., 2022). This has led to the development of dense PRF methods, such as ColBERT-PRF (X. Wang et al., 2021), ANCE-PRF (Yu et al., 2021), and CEQE (Naseri et al., 2021). Among these, ColBERT-PRF, based on the ColBERT model (Khattab & Zaharia, 2020), has shown significant improvements over the original ColBERT ranker on several datasets (X. Wang et al., 2022).

ColBERT-PRF expands the original query q by incorporating dense token embeddings derived from the top-k pseudo-relevant documents $D_{\text{top } k}$ retrieved during an initial search. The expanded query q_e combines the embeddings of the original query ω_{q_i} with the embeddings of selected tokens from the pseudo-relevant documents $\omega_{d_{ij}}$. Formally, the expanded query is defined as:

$$q_e = \omega_{q_i} \cup \left\{ \omega_{d_{ij}} \mid d_j \in D_{\text{top } k} \right\} \tag{2.9}$$

Here, $\omega_{d_{ij}}$ denotes the dense embedding of token j in document d_j, and ω_{q_i} the embedding of token i in the original query. By integrating document-side embeddings, ColBERT-PRF enhances the contextual representation of the query, potentially improving retrieval performance. However, its effectiveness depends on the quality and diversity of the pseudo-relevant document set. In Chapter 7, we adapt ColBERT-PRF to promote fairness while maintaining relevance in search results.

2.5.5 Re-Ranking Pipelines

Dense retrieval models, despite their effectiveness, are computationally expensive to apply directly to large-scale document collections (Hofstätter & Hanbury, 2019; Nogueira & Cho, 2019). The high computational cost arises from the need to encode each query-document pair or perform complex similarity computations in real time. Therefore, the dense retrieval models are typically deployed in re-ranking pipelines to balance efficiency and effectiveness. A re-ranking pipeline operates in two stages:

1. **Initial Retrieval Stage:** A lightweight model, such as BM25 (Robertson et al., 1994), is used to efficiently retrieve a set of candidate documents from the full collection. This stage focuses on recall, ensuring that the retrieved candidate set contains most of the relevant documents.

2. **Re-Ranking Stage:** The candidate documents from the initial retrieval stage are re-ranked using a more computationally intensive dense retrieval model such as ELECTRA (Clark et al., 2020). These dense models use semantic and contextual embeddings to assess the relevance of each document to the query with greater precision. This re-ranking process ensures that the top-ranked documents in the final output are highly relevant.

This two-stage approach uses the strengths of both sparse and dense retrieval methods. Sparse models provide the efficiency needed to handle large-scale collections, while dense models refine the results to maximise relevance. By narrowing down the search space in the initial stage, the computational overhead of dense models is significantly reduced, as they are applied only to a manageable subset of documents (MacAvaney et al., 2019; Mitra, Craswell et al., 2018). In this thesis, re-ranking pipelines are deployed extensively in Chapters 6-8. For example, in Chapter 6, we use BM25 to generate candidate sets, which are then re-ranked using ELECTRA-based dense retrievers. Similarly, in Chapter 7, ColBERT is integrated into a pipeline to evaluate the effectiveness of dense PRF. These applications demonstrate how re-ranking pipelines enable the deployment of dense models in large-scale IR systems while maintaining computational efficiency.

2.6 Summary

In this chapter, we have presented the ranking methods that we deploy in this thesis to validate our thesis. In our description, we have traced the progression from sparse representations to dense retrieval models and their integration into re-ranking pipelines.

In Sections 2.1 and 2.2, we have introduced sparse representations, where queries and documents are represented as high-dimensional vectors with dimensions corresponding to terms in the vocabulary. We have highlighted sparse retrieval models, such as TF-IDF (c.f Section 2.2.2) and BM25 (c.f. Section 2.2.3) that provide an efficient way to compute the relevance. However, their reliance on exact lexical matches and inability to address synonymy and polysemy present significant challenges, particularly when semantically similar terms are treated as unrelated, limiting their applicability to more complex queries. Therefore in Section 2.3, we have introduced existing query modification techniques, specifically PRF methods such as RM3 (c.f. Section 2.3) and KLQ (c.f. Section 2.3). These approaches enhance retrieval by incorporating terms from pseudo-relevant documents into the query potentially increasing the recall of relevant documents. However, the inherent reliance on sparse representations restricts their ability to fully capture semantic relationships or resolve ambiguities.

In Section 2.4, we discussed PLMs and their ability to capture semantic and contextual relationships in text. Moreover, in Section 2.5 we explored dense retrieval methods that leverage the capabilities of PLMs. We categorised the dense retrieval methods into cross-encoder and bi-encoder architectures, highlighting their distinct approaches to relevance estimation.

Finally, Section 2.5.5 described re-ranking pipelines, which combine sparse retrieval for initial recall with dense re-ranking for precision. This two-stage process enables effective handling of large-scale collections while maintaining computational efficiency and is typically deployed in modern retrieval systems. According to our thesis statement (c.f. Section 1.2) we use the introduced retrieval methods and integrate fairness considerations into re-ranking pipelines, to enhance both the relevance and fairness of exposure in the search results. In the next chapter, we will provide the necessary background on fairness in IR and clarify how our thesis aligns with this research area.

Chapter 3

Fairness in IR

In this chapter, we provide an overview of existing fairness definitions from the literature and propose a taxonomy to categorise them. Furthermore, we present the definition of fairness we use in this thesis to validate our thesis statement. Moreover, we introduce the metrics that we deploy in our experiments to assess the fairness and relevance of the search results. The chapter is structured as follows:

- In Section 3.1 we introduce different fairness definitions based on individual fairness (c.f. Section 3.1.1) and group fairness (c.f. Section 3.1.2) and categorise them using our taxonomy.
- In Section 3.2 we introduce the fairness definition used in this thesis.
- In Section 3.4 we present the metrics we use to evaluate relevance and fairness.
- Finally, in Section 3.5 we give a summary of the presented content in this chapter.

3.1 Definitions of Fairness

Since there is no universally accepted definition of fairness (Friedler et al., 2021), we categorise fairness definitions along three key dimensions: *level*, *perspective*, and *output*. Figure 3.1 illustrates our proposed categorisation.

Level The leftmost branch of Figure 3.1 shows how fairness definitions can either focus on *individuals* or *groups*. Definitions that focus on individuals demand that similar entities, such as documents with comparable relevance scores, are treated similarly. In contrast, group-level fairness definitions demand a fair treatment over groups of documents.

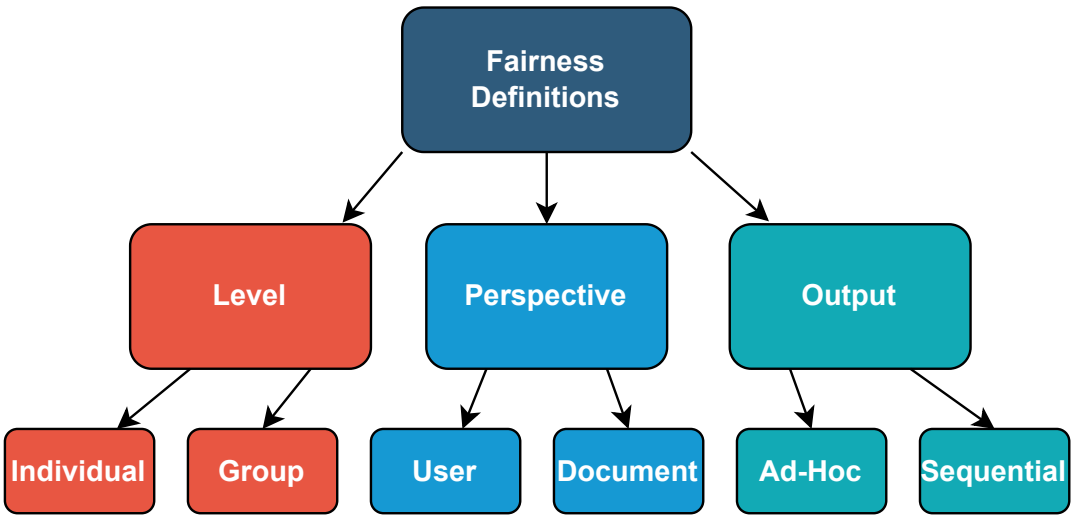


Figure 3.1: A taxonomy categorising fairness definitions by level, perspective, and output. Each branch highlights a distinct aspect of fairness in information retrieval.

Perspective The central branch of Figure 3.1 categorises fairness from two distinct viewpoints: *document-side* and *user-side*. Document-side fairness focuses on the fair treatment of entities represented by the documents. For example, authors from underrepresented demographics should not be disadvantaged in a ranking. Conversely, user-side fairness focuses on fairness from the user's perspective. For example, all users should have a fair access to relevant information in the search results.

Output The rightmost branch of Figure 3.1 distinguishes between *ad-hoc fairness* and *sequential fairness*. Ad-hoc fairness is measured in a single ranking, whereas sequential fairness is evaluated over multiple rankings. The next section outlines key fairness definitions and categorises them within our proposed taxonomy.

3.1.1 Individual Fairness Definitions

Individual fairness demands that entities with similar relevance are treated similarly within the search results (Dwork et al., 2012). Two prominent definitions of individual fairness are Counterfactual Fairness (Kusner et al., 2017) and Equity of Attention (Biega et al., 2018). These definitions are positioned within the taxonomy of fairness definitions (Figure 3.1), specifically within the *individual-level* category. Both definitions consider the *document* perspective. However, while Counterfactual fairness can in principle be applied for both *ad-hoc* and *sequential* outputs, *Equity of Attention* focuses specifically on *sequential* rankings.

Counterfactual Fairness

Counterfactual Fairness, proposed by Kusner et al. (2017), ensures that a document's sensitive attribute, such as gender or race, does not affect its ranking outcome. This definition is formally expressed as:

$$P(\zeta_{A=a} = 1 \mid X = x, A = a) = P(\zeta_{A=a'} = 1 \mid X = x, A = a'), \tag{3.1}$$

where $\zeta_{A=a}$ represents the ranking outcome for a document when its sensitive attribute A is set to a, and $\zeta_{A=a'}$ denotes the counterfactual outcome under an alternate value a'. The attribute A refers to a sensitive feature (e.g., gender or race), and a, a' are specific values of that feature. The variable X denotes the set of non-sensitive attributes of the document, and x is a specific value assignment to those attributes. Counterfactual Fairness requires that, for the same values of X, the probability of receiving a favourable ranking outcome must be the same regardless of the value of A. This ensures that ranking decisions are not influenced by sensitive attributes. This definition applies to both *ad-hoc* and *sequential* rankings, as illustrated in Figure 3.1.

Equity of Attention

Another fairness definition based on *individual fairness* is Equity of Attention, introduced by Biega et al. (2018). Specifically, Equity of Attention incorporates the concept of exposure into its fairness criterion. It ensures that documents with similar relevance receive comparable user attention across a sequence of rankings. The definition is formalised as:

$$\frac{\sum_{\tau=1}^{\mathcal{T}} E_\tau(d)}{\sum_{\tau=1}^{\mathcal{T}} R_\tau(d)} = \frac{\sum_{\tau=1}^{\mathcal{T}} E_\tau(d')}{\sum_{\tau=1}^{\mathcal{T}} R_\tau(d')}, \quad \forall d, d', \tag{3.2}$$

Here, $E_\tau(d)$ and $E_\tau(d')$ denote the exposure received by documents d and d' at ranking time step τ, and $R_\tau(d)$, $R_\tau(d')$ represent the respective relevance labels. The variable $\mathcal{T}$ denotes the total number of ranking iterations or time steps. The definition requires that the ratio of cumulative exposure to cumulative relevance be equal across documents with similar relevance. While this fairness concept explicitly considers exposure, it is tailored to sequential rankings. Since this thesis focuses on group fairness in ad-hoc rankings (see Section 1.2), Equity of Attention is not directly applied in our work. Nevertheless, it provides an important perspective on individual-level fairness.

3.1.2 Group Fairness Definitions

Group fairness definitions ensure that no group, defined by shared attributes such as race, gender, or age, is systematically disadvantaged in rankings. Unlike individual fairness, which focuses on the treatment of specific documents, group fairness evaluates how entire groups of documents are treated collectively. Below, we present several prominent definitions of group fairness.

Demographic Parity

Demographic parity (Feldman et al., 2015) requires statistical parity between groups defined by a sensitive attribute. In the context of rankings, it demands that the likelihood of receiving a favourable outcome (e.g., being ranked highly) is equal across all groups. Formally:

$$P(\zeta = 1 \mid A = a) = P(\zeta = 1 \mid A = a'), \quad \forall a, a'. \tag{3.3}$$

Here, $\zeta = 1$ indicates a positive ranking outcome (e.g., the document is ranked in a high position), and A is the sensitive attribute (e.g., gender or race) with values a and a' representing different groups. Demographic parity holds if documents from different groups have equal probability of achieving favourable ranks. This definition is typically applied in *ad-hoc* settings and adopts a *document-level* perspective.

Equality of Opportunity

Equality of Opportunity (Hardt et al., 2016) refines Demographic Parity by conditioning on document relevance. It ensures that relevant documents from all groups are equally likely to appear in favourable ranking positions:

$$P(\zeta = 1 \mid A = a, \xi = 1) = P(\zeta = 1 \mid A = a', \xi = 1), \quad \forall a, a'. \tag{3.4}$$

As before, A denotes the sensitive attribute, and a, a' are group identifiers. $\xi = 1$ indicates that a document is relevant, while $\zeta = 1$ denotes a positive ranking outcome. This definition focuses specifically on relevant documents, requiring that the probability of favourable ranking be equal across groups. Equality of Opportunity can be applied in both *ad-hoc* and *sequential* ranking contexts.

Fairness Constraints

Celis et al. (2018) introduced fairness constraints that impose quotas on protected groups in ranked results. These constraints ensure that a minimum fraction of the top-ranked documents belong to the protected group. Let g^+ denote the set of documents belonging to the protected group, and g^- those in the non-protected group. The constraint is formalised as:

$$\frac{|g^+_{1...k}|}{k} \geq \theta, \tag{3.5}$$

where $|g^+_{1...k}|$ denotes the number of protected documents appearing in the top-k positions of the ranking, and $\theta \in [0, 1]$ is the minimum acceptable proportion. This constraint ensures proportional representation in top results. It is typically applied to *ad-hoc* rankings from a *document-level* perspective.

Ranked Group Fairness Criterion

The Ranked Group Fairness Criterion, proposed by Zehlike et al. (2017), demands that the proportion of protected group members in the top-k positions is statistically consistent with expectations under random ranking. Formally, the observed proportion at rank k is given by:

$$P_k = \frac{|g^+_{1...k}|}{k}, \tag{3.6}$$

where P_k is the proportion of protected documents in the top-k, and $|g^+_{1...k}|$ is as defined above. This is compared to the expected proportion of protected items in the corpus:

$$\pi = \frac{|g^+|}{|g^+| + |g^-|}. \tag{3.7}$$

A one-sided binomial test is used to determine whether P_k is significantly lower than π, in which case the ranking is deemed unfair. This criterion is applied in *ad-hoc* settings and operates at the *document group* level.

Limitations of Group Fairness Definitions

The group fairness definitions discussed in the previous sections address important aspects of fairness but do not account for position bias (Craswell et al., 2008), meaning they overlook how document exposure varies based on ranking position. The previously introduced definitions focus on proportional representation in the ranking without considering exposure. However, users tend to exhibit a position bias, as highlighted in the literature (Craswell et al., 2008). Indeed, typically a user's probability of examining a document

decreases with rank in the search results, i.e., lower ranked documents are less exposed to a user. This can undermine fairness objectives even when proportional representation is maintained. To overcome these limitations, fairness definitions must explicitly incorporate exposure.

3.2 Demographic Parity in Exposure-Based Fairness

Since users are more likely to examine documents that are ranked higher in the search results, (Craswell et al., 2008) only focusing on the proportions in a ranking in many cases is not sufficient to ensure fairness. To address this limitation, Singh and Joachims (2018) introduced the concept of Fairness of Exposure, which explicitly incorporates exposure, which we will denote as E in this thesis.

3.2.1 Fairness of Exposure

Singh and Joachims (2018) introduced the concept of exposure into the definition of Demographic Parity, framing it as a statistical parity constraint that ignores relevance and requires equal exposure across groups. Their fairness of exposure definition can be positioned at the *group level*, adopts a *document perspective*, and applies to *ad-hoc outputs*. To incorporate exposure, they define the exposure of a group g_l as the average exposure received by documents in the group:

$$E(g_l) = \frac{1}{|g_l|} \sum_{d_\rho \in g_l} v_\rho \tag{3.8}$$

where $|g_l|$ denotes the number of documents in group g_l, and $v_\rho = \frac{1}{\log_2(\rho+1)}$ is the exposure of a document at rank position ρ. This logarithmic decay function reflects the assumption that user attention decreases with lower ranks, consistent with DCG (Järvelin & Kekäläinen, 2002).

The fairness constraint then requires that the average exposure across all groups is approximately equal:

$$E(g_l) \approx E(g_j), \quad \forall g_l, g_j. \tag{3.9}$$

While this definition incorporates exposure, it focuses on group-level averages rather than accounting for how exposure is distributed relative to the total available exposure. In contrast, our approach evaluates fairness in terms of proportional exposure allocation.

3.2.2 Proportional Fairness of Exposure

Similar to the approach of Singh and Joachims (2018), our fairness definition models exposure using an exposure decay function. However, we explicitly incorporate the total available exposure in the ranking and require that each group receives an equal share of it. Our definition operates at the *group level*, takes the *document perspective*, and is designed for *ad-hoc outputs*.The total available exposure in a ranking of length k is defined as:

$$\text{Total Available Exposure} = \sum_{\rho=1}^{k} v_\rho, \quad \text{with } v_\rho = \frac{1}{\log_2(\rho+1)} \tag{3.10}$$

The exposure received by a group g_l is the sum of exposure weights of its documents:

$$E(g_l) = \sum_{d_\rho \in g_l} v_\rho \tag{3.11}$$

The proportion of total exposure received by group g_l is computed as:

$$\text{Proportion of Total Exposure}(g_l) = \frac{E(g_l)}{\text{Total Available Exposure}} \tag{3.12}$$

According to our definition, a ranking is fair if all groups receive equal proportions of the total available exposure:

$$E(g_l) = E(g_j), \quad \forall g_l, g_j. \tag{3.13}$$

To illustrate the differences to the most related fairness definition introduced by Singh and Joachims (2018) we provide an example comparing the two fairness definitions.

Example of Our Proposed Fairness Definition

Consider a ranking scenario with two groups of documents, g_a and g_b. Table 3.1 shows a potential ranking with documents from each group and the corresponding exposure values. From the table, it is apparent that:

- Group g_a consists of three documents (A_1, A_2, A_3) with documents ranked first, fourth, and fifth.
- Group g_b contains two documents (B_1, B_2) occupying the second and third ranks.

We can calculate the exposure each group receives by summing the individual exposure values of its documents (c.f. Equation 3.11):

$$E(g_a) = v_1 + v_4 + v_5 = 1.000 + 0.390 + 0.250 = 1.640.$$

Table 3.1: Rank positions and exposure values for documents in Groups g_a and g_b. Higher ranks (lower ρ) receive greater exposure.

Rank (ρ)	**Group**	**Document**	**Exposure (v_j)**
1	g_a	A_1	1.000
2	g_b	B_1	0.630
3	g_b	B_2	0.430
4	g_a	A_2	0.390
5	g_a	A_3	0.250

$$E(g_b) = v_2 + v_3 = 0.630 + 0.430 = 1.060.$$

The total available exposure in the ranking can be obtained as follows:

$$\text{Total Available Exposure} = E(g_a) + E(g_b) = 1.640 + 1.060 = 2.700.$$

Using this setup, we assess the fairness according to both Singh and Joachims' (c.f. Section 3.2.1) and our proposed fairness definition. According to Singh and Joachims' proposed definition of fairness, the ranking is fair when the average exposure of both groups is the same. Therefore, we calculate the exposure per group.

$$\text{Average Exposure}(g_a) = \frac{E(g_a)}{|g_a|} = \frac{1.640}{3} \approx 0.547,$$

$$\text{Average Exposure}(g_b) = \frac{E(g_b)}{|g_b|} = \frac{1.060}{2} = 0.530.$$

According to our definition, the ranking is fair, when the proportion of total exposure allocated to each group is equal. Therefore, we measure the proportion per group.

$$\text{Proportion of Exposure}(g_a) = \frac{E(g_a)}{\text{Total Available Exposure}} = \frac{1.640}{2.700} \approx 0.607,$$

$$\text{Proportion of Exposure}(g_b) = \frac{E(g_b)}{\text{Total Available Exposure}} = \frac{1.060}{2.700} \approx 0.393.$$

Under Singh and Joachims' fairness definition, the ranking can be considered fair, since group g_a achieves a similar average exposure to group g_b. However, under our proportional exposure fairness assessment the ranking is not fair, since g_a receives 60.7% of the total available exposure, while g_b only receives 39.3%. This demonstrates how our fairness definition more effectively accounts for the total distribution of exposure and highlights exposure differences more explicitly.

3.3 Fair Ranking Approaches

Several works (Heuss et al., 2022; Kletti et al., 2022; Raj & Ekstrand, 2024; Singh & Joachims, 2018; Wu et al., 2022; T. Yang et al., 2023) have proposed methods to ensure fairness in the search results. However, the proposed approaches differ in the underlying fairness definition and the context they are deployed. In the following sections, we discuss the fair ranking approaches most related to our thesis.

Singh and Joachims (2018) introduced a framework along with their presented fairness definition (c.f. Section 3.2.1). Their framework integrates fairness constraints directly into the ranking process by formulating the problem of fair ranking as a linear program. Their probabilistic framework is able to maximise utility while meeting fairness constraints. However, their framework targets binary group scenarios, such as majority group versus minority group and does not generalise well to contexts involving non-binary groups. In this thesis, we explicitly address fairness for multiple groups.

Morik et al. (2020) proposed FairCo, a dynamic learning-to-rank approach to ensure fairness between groups in sequential rankings. FairCo measures and corrects imbalances in the exposure that groups receive over time ensuring the exposure per group remains proportional to their merit. However, FairCo relies on user interaction data to estimate merit and the exposure a group should receive which is often not available. Moreover, their approach is only suitable for sequential rankings and not for ad-hoc rankings which we investigate in this thesis.

There have been several works, that propose fair ranking approaches in ad-hoc rankings (Asudeh et al., 2019; Celis et al., 2018; Geyik et al., 2019; Zehlike et al., 2017; Zehlike & Castillo, 2020). Geyik et al. (2019), for example, developed post-processing algorithms to ensure proportional representation of protected groups within the top-k results. Similarly, Zehlike et al. (2017) introduced FAIR a fair top- k ranking approach, that uses a greedy strategy to select the most relevant item for the top-k positions, while ensuring that a protected group is not underrepresented (c.f Section 3.1.2). Unlike post-processing methods, FAIR integrates fairness directly into the ranking process. While these methods ensure proportional representation, they do not explicitly address exposure, which is the main focus of this thesis.

In principle, standard diversification approaches (Carbonell & Goldstein, 2017; Radlinski et al., 2008) such as xQuAD (Santos et al., 2010) or PM-2 (Dang & Croft, 2012) can also be deployed to ensure proportional fairness in the search results. For example, McDonald et al. (2022) have shown that diversification methods can indeed mitigate unfairness in the search results. However, standard diversification techniques can only ensure proportional fairness, rather than explicitly addressing the exposure the groups receive, which is the main focus of this thesis.

In addition to more general fair ranking methods, there has been work focusing on the specific context of gender fairness (Bigdeli et al., 2022, 2023; Krieg et al., 2022). For example, Rekabsaz and Schedl (2020) have shown that neural ranking models can significantly amplify gender biases in the search results. However, most of the works on gender fairness rely on text-based indicators (Rekabsaz et al., 2021) specifically tailored to gender fairness. In this thesis, we do not focus on a single fairness context such as gender fairness but instead investigate fairness for groups from different characteristics. In the next section, we introduce the metrics we use to evaluate fairness and relevance.

3.4 Evaluating Relevance and Fairness

Throughout this thesis, we use the Normalised Discounted Cumulative Gain (nDCG) for measuring relevance (Järvelin & Kekäläinen, 2002) and the attention-weighted ranked Fairness (AWRF) (Sapiezynski et al., 2019) for measuring fairness. Both metrics are chosen following the TREC Fair Ranking Track (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022), which has used them to assess the fairness of exposure in the search results of submitted systems.

3.4.1 Normalised Discounted Cumulative Gain (nDCG)

To assess the relevance of the search results, we use the Normalised Discounted Cumulative Gain (Järvelin & Kekäläinen, 2002). nDCG assigns higher scores to relevant documents appearing earlier in a ranking. It is calculated by normalising the Discounted Cumulative Gain (DCG) with the ideal DCG (iDCG). Specifically, the DCG can be formalised as:

$$\mathrm{DCG}_k = \sum_{\rho=1}^{k} \frac{2^{rel_\rho} - 1}{\log_2(\rho + 1)} \tag{3.14}$$

In this equation, rel_ρ is the relevance score assigned to the document at rank position ρ, and k is the cut-off rank. The logarithmic component models user attention, giving higher weight to documents ranked closer to the top. DCG is then normalised using the ideal DCG (iDCG), which represents the DCG of a perfectly ordered ranking. The normalised DCG (nDCG) is computed as:

$$\mathrm{nDCG}_k = \frac{\mathrm{DCG}_k}{\mathrm{iDCG}_k} \tag{3.15}$$

This normalisation ensures that nDCG values range between 0 and 1, with higher values indicating better-ranked results.

3.4.2 Attention Weighted Ranked Fairness (AWRF)

To evaluate the fairness of search results, we use Attention-Weighted Ranked Fairness (AWRF) (Sapiezynski et al., 2019). AWRF has been prominently used in the TREC Fair Ranking Track 2021 & 2022 (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022) to measure the fairness of exposure received by different groups in a ranking. AWRF quantifies how fairly exposure is distributed across groups by comparing the observed exposure distribution to a target distribution using the Jensen-Shannon Distance (JSD) (Fuglede & Topsøe, 2004).
According to our fairness definition (c.f. Section 3.2.2), the target exposure for each group reflects the ideal allocation:

$$E_{target}(g_l) = \frac{1}{m} \times \text{Total Available Exposure} \tag{3.16}$$

where m is the total number of groups, assuming each group should receive an equal share of the exposure. To evaluate deviation from the target exposure, we compute the *Jensen-Shannon distance* between the observed and target exposure distributions, denoted here as $\epsilon^{(o)}$ and $\epsilon^{(t)}$, respectively:

$$D_{\text{JS}}(\epsilon^{(o)} \| \epsilon^{(t)}) = \sqrt{\frac{1}{2} D_{\text{KL}}(\epsilon^{(o)} \| \mu) + \frac{1}{2} D_{\text{KL}}(\epsilon^{(t)} \| \mu)} \tag{3.17}$$

where $\mu = \frac{1}{2}(\epsilon^{(o)} + \epsilon^{(t)})$ is the average of the two distributions, and D_{KL} denotes the Kullback-Leibler divergence (Kullback & Leibler, 1951).To asses the AWRF we use the following equation:

$$\text{AWRF} = 1 - D_{JS\_dist}(E(g) \| \text{Target Exposure}(g)) \tag{3.18}$$

Here, $E(g)$ is the vector of observed exposure proportions over all groups, and Target Exposure(g) is the vector of their ideal exposure shares. A perfect AWRF score of 1 indicates complete fairness in exposure allocation.

3.5 Summary

In this section, we have discussed prominent fairness definitions relevant to this thesis and have presented our own fairness definition. Specifically, in Section 3.1, we introduced a taxonomy of fairness definitions which we used to categorise individual as well as group fairness definitions (c.f. Sections 3.1.1 and 3.1.2). Moreover, we have introduced the concept of fairness of exposure (c.f Section 3.2) and have presented our fairness definition (c.f. Section 3.2.2). In Section 3.3, we discussed different fair ranking approaches and clarified in which contexts they are suitable, highlighting the diverse range of approaches.

Finally, in Section 3.4, we presented the metrics that we use in this thesis to assess relevance and fairness. Specifically, we introduced the nDCG for relevance and the AWRF metric for fairness. This chapter has provided the necessary background on fairness which is needed for the validation of our thesis statement (c.f. Section 1.2) in the upcoming chapters.

Chapter 4

Query Exposure Prediction

4.1 Introduction

The main objective of an IR system is to provide the users with the documents that are the most relevant to their query. As previously discussed in Chapter 2, recently, deep neural ranking models that use contextualised language representations, such as BERT (Devlin et al., 2019), have been shown to effectively score documents and determine the documents' relevance to a user's query (Lin et al., 2022). However, while effective, applying these neural models is computationally expensive and is often infeasible over large collections of documents (Hofstätter & Hanbury, 2019). Therefore, deep neural ranking models are usually deployed in re-ranking pipelines (c.f. Section 2.5.5). In a re-ranking pipeline, an inexpensive ranking model, such as BM25 (Robertson et al., 1994) or the more recent SPLADE (Formal et al., 2021), is used to create an initial pool of documents in a first-pass retrieval process. The documents from this pool are then re-scored by deep neural ranking models such as ELECTRA (Clark et al., 2020). The re-scored documents are then presented to the user in decreasing order of predicted relevance. However, the amount of exposure to a user that a document is likely to receive is dependent on the document's position in the ranking, since according to the well-known position bias model (Craswell et al., 2008), lower-ranked documents have a lower probability of being exposed to, or examined by, a user. This can lead to an unintended bias against particular *groups* of documents. For example, if relevant documents that are associated with a demographic group, such as race or gender, are systematically ranked at lower positions, then the group will receive less exposure to the user than the groups that have documents ranked at higher positions (Singh & Joachims, 2018). In re-ranking pipelines, the exposure distribution over groups in the final ranking is dependent on the recall per group in the initial candidate pool, i.e., documents that are not retrieved by the first-pass retrieval process are not scored by the re-ranker. If the initial pool of documents contains only a few or no relevant documents from a group, a fair distribution of exposure in the final ranking might become infeasible.

For example, Bower et al. (2022) have shown that the existence of inequalities in the initial pool of documents in a multi-stage ranking process can hinder the effectiveness of existing fairness mitigation methods. Therefore, predicting the exposure that documents of a group receive for a given query is both important and useful. For example, predicting the exposure groups receive for a query enables the system to make adjustments during the initial retrieval stage, ensuring that enough documents from each group are retrieved. This ensures that every group has a fair chance of achieving the desired exposure in the final search results when a re-ranking process is applied. This predictive step aligns directly with our thesis statement (c.f. Section 1.2) in which we state, that modern ranking systems can be tailored to allocate exposure more fairly by leveraging information from the first-stage retrieval. Moreover, the ability to predict and control exposure early complements our thesis that IR systems can be adapted to produce both fair and relevant search results supporting our thesis statement.

In this chapter, we introduce the new task of Query Exposure Prediction (QEP), i.e., predicting how the available exposure in search results will be distributed among groups of documents in response to a user's query. We define QEP as the task of estimating how exposure to users will be distributed across predefined groups of documents in response to a query. This prediction is made before documents are ranked or shown, allowing interventions that can promote fair exposure. Predicting exposure for each query could help system designers select suitable rankers and apply fairness strategies, aligning with our thesis statement (c.f. Section 1.2). For example, when the predicted exposure distribution is not closely correlated with a desired or target distribution (c.f. Equation 3.13) fairness interventions can be deployed. In this Chapter, we propose a novel QEP approach, called Group Exposure Prediction (GEP). Our approach is inspired by pre-retrieval QPP (He & Ounis, 2006) approaches that aim to estimate how difficult it will be for an IR system to provide relevant documents in response to a particular query. Pre-retrieval QPP refers to estimating the effectiveness of a query before any documents are retrieved, often using lexical statistics such as IDF or term specificity. These predictors are efficient and do not rely on human relevance judgments. Our proposed GEP approach leverages statistics about how a query's terms are distributed in the indexed documents from each group to estimate how the available exposure will be distributed among the different groups in a ranking. The remainder of this chapter is structured as follows:

- In Section 4.2 we present the existing work that is related to exposure prediction.
- In Section 4.3, we show how exposure can be distributed over groups in a ranking.
- In Section 4.4, we propose a new predictor called **G**roup **E**xposure **P**redictor to predict the exposure distribution over groups of documents after a first-stage retrieval.

- In Section 4.5, we present our experimental setup to evaluate our predictor and the two test collections that we will use throughout the thesis.
- In Section 4.6 we present the results of our experiments that evaluate the performance of our GEP predictor.
- Finally, in Section 4.7 we summarise our contributions in this chapter.

4.2 Related Work

To the best of our knowledge, our work is the first to highlight the importance of predicting the exposure groups of documents receive in the search results by a user. Moreover, we propose the first query performance predictor for measuring the exposure of groups of documents. While there is no existing-related work to the task of QEP there are many works covering the related task of Query Performance Prediction (QPP) (Arabzadeh et al., 2024; He & Ounis, 2006; Hauff et al., 2008). QPP is traditionally applied to predict the effectiveness of the ranking produced by an IR system in response to a query, without relying on expensive and hard-to-obtain human-relevance judgment and shares fundamental concepts with QEP. For both QEP, as well as QPP, query and document features are used to predict the properties of the search results in relation to a query. In our experiments, we adapt the well-established concepts of QPP to the domain of QEP using similar methodologies to predict group-level exposure.

QPP predictors are usually separated into *post-retrieval* and *pre-retrieval* methods. Post-retrieval methods (Carmel & Yom-Tov, 2010) are applied after an initial ranking has been created. Therefore, post-retrieval QPP methods have access to the estimated relevance scores, which can be used for performance prediction. However, post-retrieval methods are not applicable for QEP, since after the initial retrieval stage, the exposure can be directly calculated, rendering the prediction redundant.

Pre-retrieval QPP methods (Hauff et al., 2008) typically rely on features derived from the underlying document corpus and query terms to make predictions. Similarly, our proposed predictor uses lexical features such as the term frequency and the inverse document frequency from the groups of documents to predict exposure distributions. Since pre-retrieval QPP methods are closely related to QEP, we include adaptations of traditional QPP methods in our experiments, specifically those that rely on term frequency and inverse document frequency features to make predictions. This ensures a fair comparison to our proposed predictor that uses the same features. In our experiments, we generate a prediction score for each group using a traditional QPP method based on lexical information and compare the scores between groups to predict the exposure distribution.

For example, we use the Average Inverse Document Frequency (AvIDF) (Cronen-Townsend et al., 2002) as one of these baselines. AvIDF approximates the average specificity of query terms by calculating the inverse document frequency (c.f. Section 2.2.2) for each query term. AvIDF can be formalised as follows:

$$\text{AvIDF} = \frac{1}{|q|} \sum_{q \in Q} \text{IDF}(q) \tag{4.1}$$

In the equation, $|q|$ is the number of query terms in q and $\text{IDF}(q)$ is the inverse document frequency of term q_t. Moreover, Q denotes the set of queries. We also use the Average Inverse Collection Term Frequency (AvICTF), introduced by He and Ounis (2004), as another baseline. AvICTF is a modification of AvIDF in which the document frequency is replaced by the collection term frequency. The collection term frequency $\text{cTF}(q_t)$ is the number of times term q_t appears across the entire corpus. The total number of term occurrences in the corpus is denoted as $\sum_{t \in \mathcal{V}} \text{cTF}(t)$, where $\mathcal{V}$ is the vocabulary. The AvICTF score is computed as:

$$\text{AvICTF} = \frac{1}{|q|} \sum_{q_t \in q} \log \left(\frac{\sum_{t \in \mathcal{V}} \text{cTF}(t)}{\text{cTF}(q_t)} \right) \tag{4.2}$$

Expanding on AvICTF, He and Ounis (2004) have proposed the Simplified Clarity Score (SCS). SCS measures query ambiguity by combining AvICTF with a maximum likelihood estimate of term frequency in the query. SCS can be obtained using the following formula:

$$\text{SCS} = \sum_{q \in Q} \left(\frac{\text{TF}(q_t, q}{|q|} \log \left(\frac{\text{TF}(q_t, q)}{\text{cTF}(q)} \right) \right) \tag{4.3}$$

In the formula, q_t represents an individual term in a query q and $\text{TF}(q_t, q)$ is the term frequency of the query term q_t in the query q. Moreover, we include the term-relatedness predictor Averaged Pointwise Mutual Information (AvPMI) (Hauff et al., 2010) as one of our baselines. AvPMI calculates the average association strength between query terms, based on the Pointwise Mutual Information (PMI) of term pairs. PMI quantifies the relationship between two terms q_{t_1} and q_{t_2} by comparing their observed co-occurrence probability $P(q_{t_1}, q_{t_2})$ with the expected probability assuming independence.

$$\text{PMI } (q_{t_1}, q_{t_2}) = \log \frac{P(q_{t_1}, q_{t_2})}{P(q_{t_1}) \cdot P(q_2)} \tag{4.4}$$

A higher PMI value indicates that the terms co-occur more often than expected, suggesting a stronger relationship. AvPMI extends this idea by aggregating the PMI scores for all distinct pairs of query terms q_{t_1}, q_{t_2} within a query q:

$$\text{AvPMI} = \frac{1}{|q|(|q|-1)} \sum_{\substack{q_{t_1}, q_{t_2} \in q \\ q_{t_1} \neq q_{t_2}}} \text{PMI}\ (q_{t_1}, q_{t_2}) \tag{4.5}$$

Here, $|q|$ represents the total number of terms in the query. By averaging PMI values, AvPMI provides an overall measure of how related the terms in a query are within the document collection. This predictor is particularly useful for evaluating queries where the relationships between terms play a significant role in determining their overall meaning, unlike other predictors that focus only on individual term properties.

Adapting QPP methods to assess the relevance of different groups of documents and predict their exposure has similarities to the concept of federated search. Federated search, also known as federated information retrieval, is a technique for searching multiple text collections simultaneously. Queries are sent to a subset of collections most likely to provide relevant results, and the returned results are integrated into a single ranked list (Shokouhi, Si et al., 2011). In federated search, resources—i.e., document collections—are selected based on their potential to provide the most relevant results, which requires predicting their relevance (Garba et al., 2023). Building on this concept, we adapt a resource selection method to identify which group of documents is most likely to contain the most relevant results, i.e., the group that will receive the highest exposure, rather than focusing on document collections. Specifically, we adapt the CORI algorithm (Callan, 2002), a resource selection method widely used in federated search. The CORI algorithm facilitates querying across multiple data sources and combines the results into a unified output (Shokouhi, Si et al., 2011), which we extend to support exposure prediction in our context. The CORI algorithm predicts the suitability of a data source for a given query by calculating a probability score known as the *belief* score. This score is derived from collection statistics, combining both term frequency and inverse document frequency components. For our experiments, we adapt the CORI algorithm to calculate the *belief* score for each group of documents, treating groups as analogous to collections in the original algorithm. This allows us to predict which groups are most likely to contain relevant documents.

In CORI, a term frequency component, TF_belief(q, g), evaluates the relevance of a term q within a group g based on its frequency and is computed as:

$$\text{TF\_belief}(q, g) = \frac{\text{TF}(q, g)}{\text{TF}(q, g) + o} \tag{4.6}$$

where $\mathrm{TF}(q, G)$ is the term frequency of q in the group, and o is a constant used to prevent overemphasis on high term frequencies, ensuring stability when term frequency values vary. A inverse document frequency component, $\mathrm{IDF\_belief}(q, g)$, evaluates the distinctiveness of a term q in the overall collection and is formalised as:

$$\mathrm{IDF\_belief}(q, g) = \frac{\mathrm{IDF}(q)}{\mathrm{IDF}(q) + o'} \tag{4.7}$$

where $\mathrm{IDF}(q)$ represents the inverse document frequency of q, and o' is a constant that prevents terms with extreme values from skewing the relevance calculation.

The final *belief* score for a group g is calculated by combining the term frequency and inverse document frequency components for all query terms $q \in Q$, weighted by a parameter w:

$$\mathrm{belief}(q, g) = \sum_{q \in Q} (w \cdot \mathrm{TF\_belief}(q, g) + (1 - w) \cdot \mathrm{IDF\_belief}(q, g)) \tag{4.8}$$

where w determines the relative importance of the term frequency and inverse document frequency components. By adapting the CORI algorithm in this way, we effectively translate its resource selection principles to the context of group-level exposure prediction, using the constants o and o' to stabilise calculations and improve robustness.

In our experiments, we focus on adapted QPP methods that rely on lexical information, as they offer clear and interpretable features, unlike neural approaches (Faggioli et al., 2023). This allows us to better understand the mechanisms behind predictions and to develop a new predictor tailored to our specific task of QEP.

4.3 Exposure in Rankings

In this chapter, we develop an approach to predict how much exposure to a user a group of documents will receive when the documents are ranked by their relevance to the user's query. The expected exposure of a given document depends on its position in the ranking since a user usually starts from the top and examines each document in order with a certain probability of stopping based on the number of relevant (or partially relevant) documents they have seen (c.f. Section 3.2). Therefore, the further down in the ranking the document is, the less likely it will be exposed to the user (often referred to as the position bias (Craswell et al., 2008) (c.f. Equation 3.14).

Numerous user browsing models have been proposed in the literature to estimate the likelihood that a user will examine a document at a given rank. For instance, the Cascade Model (Craswell et al., 2008) assumes that users examine documents sequentially from the top of the ranked list of results and stop once they encounter the first relevant

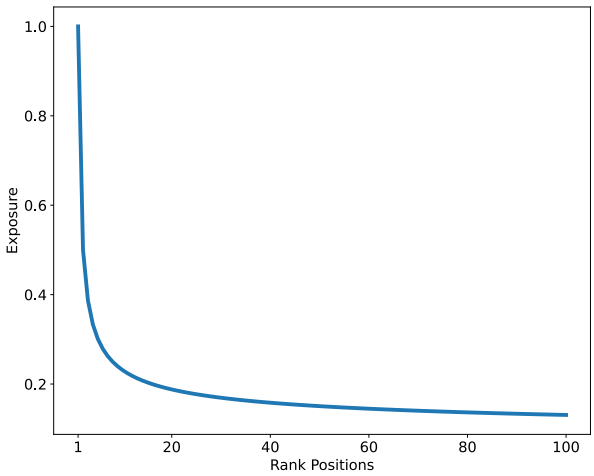


Figure 4.1: The plot shows the amount of available exposure for each position in a ranking of size k=100 according to the position bias model from the well-known nDCG metric (Järvelin & Kekäläinen, 2002) (c.f. Equation 3.14).

document, thus adopting an "all-or-nothing" approach to relevance. Another model, the Position-Based Model (PBM) (Agarwal et al., 2019), suggests that the likelihood of a user examining a document decreases as its rank increases, independent of the document's relevance. In Contrast, the Dependent Click Model (X. Wang et al., 2018) introduces the idea that users' decisions to stop browsing depend not only on the position of documents but also on the relevance of previously viewed documents.

In this thesis, we use the position bias model employed in the TREC Fair Ranking Track (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022), where the probability of a user examining a document decays logarithmically as the rank increases, i.e., we use PBM (Agarwal et al., 2019). This approach models exposure using the Discounted Cumulative Gain (DCG) framework (Järvelin & Kekäläinen, 2002) (c.f. Equation 3.14), which is commonly used to evaluate ranked retrieval systems. The exposure for a document at position ρ in the ranking is calculated as:

$$\text{Exposure}(\rho) = \frac{1}{\log_2(p+1)} \tag{4.9}$$

This formula reflects that users are more likely to view top-ranked documents, with exposure decreasing down the list (c.f. Section 3.2).

Figure 4.1 illustrates the exposure distribution available for a ranking of length k=100 under the position bias/exposure model. Figure 4.1 shows that the available exposure drops sharply in rank positions 1..10 and then decreases more gradually in positions 11..100. Figure 4.2 shows the number of possible orderings for rankings of size k=1..100. As can be seen from Figure 4.2, there are 10^{152} different possible orderings of documents in a ranking of size k=100. The amount of exposure that a group of documents receives is

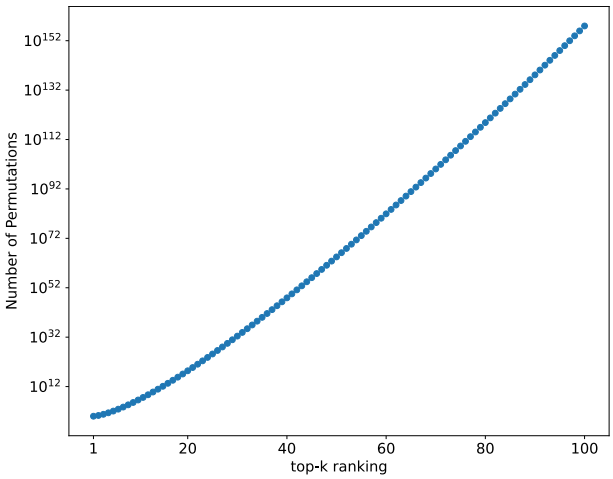


Figure 4.2: The plot shows the number of possible orderings of documents for rankings of size k=1...100. The y-axis is in log scale.

calculated by summing the exposure values for each position in the ranking where a document from the group appears. For example, if a group has two documents in a ranking, with the first document placed at position 1 (i.e., the top rank position) and the second document placed in the 5^{th} position, the first document receives an exposure of 1 and the second document receives an exposure of 0.39 (c.f. Equation 4.3). In such a scenario, the group will receive an exposure of $1 + 0.39 = 1.39$. In this case, the amount of exposure that a group of documents will receive is dependent on (1) the size of the ranking, (2) the number of documents from the group that are in the ranking, and (3) the positions in the ranking where the documents from the group appear (c.f. Section 3.2). On the other hand, there are multiple possible rankings (or orderings) that produce the same amount of exposure for a group. Returning to our previous example of two documents from a group in rank positions 1 and 5, if the positions of the two documents in the ranking are swapped, then the group's exposure is still 1.39.

For a ranking of size k=100, Figure 4.3 presents (1) the range of exposure values that are achievable for a single group depending on the number of documents from the group that are included in the ranking, and (2) the number of possible rankings that achieve that particular amount of exposure. From Figure 4.3, we can see that the range of achievable exposure is larger when there is a relatively small number of documents from a group in the ranking, but the amount of achievable exposure increases as more documents from the group are added to the ranking. Furthermore, the achievable exposure values depend on the number of documents included in the ranking. Hence, adding more documents from a group to the ranking can notably increase the group's exposure, much more than just moving the existing documents to higher ranks. This highlights the importance of the first-stage retrieval and the impact of low recall on the re-ranking pipeline since the exposure that a group receives will be substantially increased by adding more documents

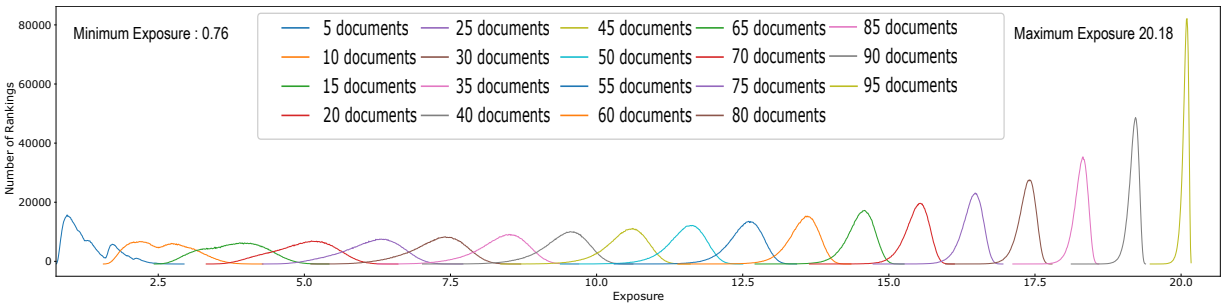


Figure 4.3: The plot shows the number of rankings that can achieve a particular amount of exposure for a given number of documents in a single group for a ranking of size k=100.

from a group in the first-pass retrieval. In addition, this supports our thesis statement (c.f Section 1.2) in which we posit that re-call enhancing retrieval techniques have the potential to notably impact the distribution of exposure. We investigate how the available exposure is distributed among groups of documents when different recall enhancing IR techniques such as PRF (c.f. Section 2.3) are applied in our experiments (RQ4.2) as well as in the upcoming Chapters 5-8.

4.4 Group Exposure Prediction

In this section, we outline the process of predicting group exposure, detailing each component. The GEP approach begins with representing document groups and queries using TF-IDF scores (c.f. Section 4.4.1) and then selecting the top-k highest scores (c.f. Section 4.4.2). GEP then progresses to calculating group-query similarity (c.f. Section 4.4.3) and concludes with predicting exposure distribution across groups (c.f. Section 4.4.4).

4.4.1 Group Representation and TF-IDF Scoring

To predict exposure, each group g is represented by a vector S_g, where each element corresponds to the group's relevance to a specific query term. Given a query q with terms $q_{t_1}, q_{t_2}, \ldots, q_{t_n}$, the vector S_g has a length of n, with each element s_i reflecting the group's relevance to term q_{t_i}. This relevance is calculated using TF-IDF scores, which measure the importance of query terms within the group's documents.

The relevance score s_i for a group g is based on the TF-IDF scores of the documents in that group. For a query term q_{t_i} in a document $d_{g,i}$, the TF-IDF score is calculated as:

$$\text{TF-IDF}(q_{t_i}, d_{g,i}) = \text{TF}(q_{t_i}, d_{g,i}) \cdot \text{IDF}(d_g), \tag{4.10}$$

where $\text{TF}(t_i, d_{g,i})$ represents the frequency of term q_{t_i} in document $d_{g,i}$, and $\text{IDF}(d_g)$ is computed as:

$$\text{IDF}(d_g) = \log\left(\frac{||d_g|| + 1}{df_g + 1}\right), \tag{4.11}$$

where $||d_g||$ is the total number of documents in group g, and df_g the number of documents containing term q_{t_i}. The "+1" in the IDF formula prevents division by zero.

4.4.2 Focusing on the Top-k Documents

Since only the top-k documents are shown to a user, we compute TF-IDF scores using only the most relevant documents from each group. For each query term q_{t_i}, we select the k highest TF-IDF scores within group g. If a query term appears in fewer than k documents in the group, the missing values are set to zero to prevent overestimation of relevance based on rare occurrences. The relevance score s_i for term q_{t_i} in group g is then computed as the average of these top-k scores:

$$s_i = \frac{1}{k}\sum_{j=1}^{k} \text{TF-IDF}(q_{t_i}, d_{g,j}), \tag{4.12}$$

where $d_{g,j}$ represents the document with the j-th highest TF-IDF score for term q_{t_i}. Averaging ensures that the relevance score reflects the overall importance of the term across multiple documents rather than being dominated by a single high-scoring document.

4.4.3 Calculating Group-Query Similarity

Once the group vector S_g is constructed, the query q is also represented as a vector V_q based on TF-IDF scores over the entire document collection. To estimate a group's alignment with the query, the similarity between the group vector S_g and the query vector V_q is calculated using the dot product (c.f. Section 2.1):

$$e_g = S_g \cdot V_q \tag{4.13}$$

A higher similarity score e_g indicates greater alignment with the query, suggesting the group is more likely to appear in the top-k positions.

4.4.4 Predicting Exposure and Fairness Evaluation

To predict the exposure of each group, the similarity scores are normalised across all groups:

$$GEP_g = \frac{e_g}{\sum_{g' \in G} e_{g'}}. \tag{4.14}$$

This normalisation ensures that exposure is distributed proportionally based on the group's relevance to the query. By providing a predicted exposure value for each group, this method allows us to evaluate the fairness of a ranking using our fairness definition, which seeks equal exposure across groups (c.f. Section 3.1).

GEP integrates group relevance scoring, query alignment, and exposure prediction into a unified approach. This enables a direct evaluation of fairness and supports the development of strategies to address exposure disparities in search results.

4.5 Experimental Setup

To evaluate our proposed GEP predictor for the task of QEP, we formulate the following two research questions:

- **RQ4.1:** Is our proposed GEP approach effective for predicting the exposure that groups will receive when an inexpensive ranking model is deployed?
- **RQ4.2:** Is GEP effective for predicting the exposure that groups receive in a ranking when the query is expanded using pseudo-relevance feedback?

To conduct our experiments and answer our research questions, we use the PyTerrier (Macdonald & Tonellotto, 2020) Information Retrieval Framework.

4.5.1 Test collections

In our experiments, we use the TREC 2021 and 2022 Fair Ranking Track test collections (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022). Both test collections consist of English-language Wikipedia articles. These test collections were created to encourage research on how editors of Wikipedia can be presented with a fair ranking of articles that need editing, in relation to a topic of their interest. In the TREC Fair Ranking Track, fairness is measured in terms of exposure. Hence, the corresponding test collections are well-suited for investigating our research questions. The test collections consist of a set of queries as well as fairness group characteristic labels for the documents in the underlying collection. The labels group the documents by different characteristics, such as their popularity or creation date. Both TREC collections were indexed using a Porter Stemmer and stop-word removal. Table 4.1 outlines the fairness characteristics and corresponding groups used to evaluate our GEP approach across both the TREC 2021 and TREC 2022 datasets. These fairness characteristics define the groups of documents for which we assess how exposure is distributed in the search results.

Table 4.1: Fairness characteristics and their groups.

Characteristic	Groups
Geographic location (TREC 21)	Unk, Europe, Africa, Northern America, Asia, Oceania, Latin America and the Caribbean, Antarctica
Continent (TREC 22)	Unk, Europe, Africa, America, Asia, Oceania, Antarctica
Age of the topic (TREC 22)	Unk, Pre-1900s, 20th century, 21st century
Languages (TREC 22)	5+ Languages, 2-5 Languages, Only English
Popularity (TREC 22)	Low, Medium-Low, Medium-High, High
Age of Article (TREC 22)	2001-2006, 2007-2011, 2012-2016, 2017-2022
Alphabetical (TREC 22)	a-d, e-k, l-r, s-z

In the TREC 2022 dataset, documents are grouped by several fairness characteristics. These include "Age of an Article", which measures how long an article has been available in Wikipedia, and "Age of Topic", which classifies documents by the time period of the subject matter they cover. We also use the "Alphabetical" fairness characteristic, which groups documents based on the first letter of their titles, as well as "Popularity" which refers to the number of views an article received in February 2022. Additionally, we include the "Languages" category, which reflects how many language versions an article has been translated into on Wikipedia. In both datasets, documents can be grouped by "Geographic Location". In the original TREC 2022 dataset, there were 24 fine-grained geographic labels, such as "Northern Europe" and "Western Asia". However, many of these smaller regions contained very few documents. To address this, we consolidate the geographic labels into broader categories based on continents. For instance, "Northern America", "Central America", "South America" and the "Caribbean" are all grouped under "America". Similarly, smaller regions in Oceania and Africa have been combined into broader continental categories. This ensures that each group has a sufficient number of relevant documents, allowing for more meaningful comparisons.

This consolidation of geographic labels also ensures consistency with the TREC 2021 dataset, where geographic locations were already grouped broadly by continent. For the TREC 2021 collection, we only use the "Geographic Location" category, which is the only characteristic that contains labels for all documents in the collection. By broadly aligning the TREC 2022 dataset to the groups of the TREC 2021 collection, we can better compare exposure patterns across the two datasets, ensuring that the analysis remains interpretable. Table 4.2 shows the exact grouping of the different locations.

Continent	Regions Included
America	Northern America, Central America, South America, Caribbean
Oceania	Australia and New Zealand, Polynesia, Micronesia, Melanesia
Europe	Northern Europe, Western Europe, Eastern Europe, Southern Europe
Africa	Eastern Africa, Western Africa, Southern Africa, Northern Africa, Middle Africa
Asia	Central Asia, Eastern Asia, Southern Asia, Southeastern Asia, Western Asia
Antarctica	Antarctica
Unknown	Documents without a clear geographic label

Table 4.2: Geographic Locations in the TREC 2022 Collection.

4.5.2 Queries

We use all of the evaluation queries from the two Fair Ranking Track collections. Specifically, we use the 48 evaluation queries in the TREC 2021 Fair Ranking Track collection and the 47 evaluation queries from the 2022 collection. All queries consist of keywords relating to a Wikipedia topic, e.g., for the topic "Architecture" the query is a collection of words such as "museum baroque librarian architecture library".

4.5.3 Retrieval Models

In our experiments, we use a selection of prominent first-pass ranking models commonly used in re-ranking pipelines. We evaluate our predictor on BM25 (Robertson et al., 1994) (with default parameters) (c.f. Section 2.2.3), TF-IDF (Salton & Buckley, 1988) (c.f. Section 2.2.2) and SPLADE (Formal et al., 2021) (c.f. Section 2.5.3). To evaluate the impact of PRF (c.f. Section 2.3) on the prediction performance of GEP, we use the RM3 relevance language model (Jaleel et al., 2004) (c.f. Equation 2.6) and KLQueryExpansion (Amati & van Rijsbergen, 2002) (KLQ) in our experiments, with 3 feedback documents and 10 expansion terms following the defaults of PyTerrier[1].

[1]https://pyterrier.readthedocs.io/en/latest/rewrite.html

4.5.4 Baselines

As baselines, we use the CORI (Callan, 2002) federated IR algorithm as well as several pre-retrieval Query Performance Predictors that we have presented in the related work Section 4.2. We use AvICTF (He & Ounis, 2006), AvIDF (Cronen-Townsend et al., 2002), AvPMI (Hauff et al., 2010)SCS) (He & Ounis, 2004) as baselines. To have a fair comparison with our proposed approach, we create a prediction score for each group and then normalise these group scores to be between 0 and 1. The normalised scores are used as a way to measure how much exposure a group will receive in the ranking.We assume that if a query is predicted to perform better on documents of one group, compared to the documents of another group, then the first group is likely to receive more exposure in the ranking. We normalise the prediction scores, since the predictions between groups are dependent i.e., if all prediction scores between groups are high, then the difference in exposure will likely be low. If the predicted performance of a query is high over one group of documents but low on the others, then it is likely that the high-performing group will receive most of the exposure in the ranking. As introduced in Section 4.2, the CORI algorithm contains a belief component, which depicts how likely a group of documents contains relevant information to a query. For our experiments, we choose to use an average of the belief values and set the belief parameter to 0.4 as suggested in the original paper (Callan, 2002).

4.5.5 Measures

To evaluate our predictions, we calculate the actual exposure the groups receive in a ranking and compare the exposure to our predicted values. We use the JSD for the calculation. The JSD is a symmetric and smoothed version of the Kullback-Leibler divergence used to measure the similarity between two probability distributions. In our case, it quantifies the difference between predicted and actual group-level exposure distributions. (c.f. Equation 3.17), i.e., we use it to calculate the distance between the distributions of our predicted exposure $\hat{E}$ and the actual Exposure E:

$$JSD(hatE\|E) = \sqrt{\frac{1}{2}\mathrm{KL}(\hat{E}\|\mu) + \frac{1}{2}\mathrm{KL}(E\|\mu)} \tag{4.15}$$

where $JSD(\hat{E}\|E)$ is the distance between the predicted exposure $\hat{E}$ and the actual exposure E. $KL(\hat{E}\|\mu)$ is the Kullback-Leibler (KL) divergence between $\hat{E}(d)$ and the average distribution μ. $KL(E\|\mu)$ is the Kullback-Leibler divergence between E and μ. μ denotes the midpoint distribution between $\hat{E}(d)$ and E. If there are no differences between the actual exposure and the predicted value, the distance is 0. In our experiments, we consider a predictor to be better when it produces values closer to 0 compared to another predictor.

In practice, when choosing a predictor for deployment, we usually set a target threshold to ascertain its usefulness. For example, a predictor should show a high degree of similarity (e.g., <0.2) with the real values of a test set before being deployed. However, in practice, the choice of threshold depends on the context of the predictor's application. We test the results of GEP for statistically significant differences to the baselines using a paired t-test ($p<0.01$) with Bonferroni correction.

4.6 Results

Results for RQ4.1: To answer RQ4.1, we evaluate the performance of our GEP predictor in terms of its ability to predict the exposure that groups of documents will receive in a ranking of size k=100. We generate rankings with different retrieval models namely BM25, TF-IDF and SPLADE. We calculate the actual exposure values the documents receive and compare these with the predictions made by our proposed approach and the baselines for each of the categories is a set of related groups, such as Europe, Asia, etc.

Table 4.3 shows the evaluation results when a first-pass retrieval model is used for ranking. The table reports the JSD between the prediction and the actual exposure of each group. The optimal prediction perfectly estimates the actual exposure and would result in a distance of 0. From the table, it is apparent that our GEP approach achieves the best results on all the groups when BM25 or TF-IDF is used as a ranker. GEP significantly outperforms (paired t-test, with Bonferroni correction, $p<0.01$) the baselines on five out of the seven categories. The differences are statistically significant for the fairness characteristics: Geographic Location, Continent, Age of Topic, Age of Article as well as Languages. Analysing the Popularity category, we note that our predictor is significantly better in predicting the exposure compared to the AvICTF and the CORI baselines. The biggest differences between the predictors can be observed in the geographic location category, on which our prediction outperforms the next best predictor by up to $\sim$40%. Geographic location is the category with the highest number of groups. When SPLADE is used as a retrieval model, we observe that our GEP predictor gives the best prediction on six out of seven categories. In all of these six categories, our predictor significantly outperforms the baselines. GEP is only outperformed in the Popularity category. However, there is no observed statistical significance. These results show that our approach is overall the most effective predictor compared to all other evaluated baselines.

Table 4.3: Evaluation of the predictors on common first-pass retrieval rankers, for rankings with k=100. * indicates significant differences (t-test, with Bonferroni correction, p<0.01) in prediction performance between GEP and the baselines.

Approach	Age of Topic	Popularity	Age of Article	Languages	Alphabetical	Continent	Geo-Location (TREC 21)
SPLADE							
GEP	**0.2624**	0.2013	**0.2129**	**0.1108**	**0.2041**	**0.3022**	**0.3543**
SCS	0.3738*	0.2243	0.2782*	0.1513*	0.2438*	0.4529*	0.5261*
AvIDF	0.3995*	0.2128	0.2889*	0.1545*	0.2471*	0.5081*	0.5820*
AvICTF	0.4088*	**0.1884**	0.2714*	0.1488*	0.2469*	0.5047*	0.5695*
AvPMI	0.3713*	0.2083	0.2758*	0.1424*	0.2414*	0.4614*	0.5474*
CORI	0.4062*	0.1888	0.2735*	0.1491*	0.2467*	0.4994*	0.5631*
BM25							
GEP	**0.2880**	**0.1567**	**0.1759**	**0.1585**	**0.1335**	**0.2810**	**0.2449**
SCS	0.3505*	0.1620	0.2336*	0.1803*	0.1407	0.4097*	0.4134*
AvIDF	0.3726*	0.1621	0.2443*	0.1854*	0.1436	0.4657*	0.4844*
AvICTF	0.3773*	0.1706*	0.2305*	0.1801*	0.1431	0.4601*	0.4668*
AvPMI	0.3562*	0.1645	0.2302*	0.1744*	0.1413	0.4226*	0.4388*
CORI	0.3735*	0.1703*	0.2338*	0.1794*	0.1422	0.4520*	0.4582*
TF-IDF							
GEP	**0.2886**	**0.1583**	**0.1759**	**0.1590**	**0.1336**	**0.2814**	**0.2454***
SCS	0.3512*	0.1628	0.2339*	0.1811*	0.1420	0.4094*	0.4138*
AvIDF	0.3734*	0.1629	0.2447*	0.1862*	0.1449	0.4655*	0.4847*
AvICTF	0.3780*	0.1716*	0.2307*	0.1808*	0.1445	0.4598*	0.4671*
AvPMI	0.3568*	0.1652	0.2304*	0.1751*	0.1423	0.4225*	0.4390*
CORI	0.3726*	0.1708*	0.2331*	0.1802*	0.1435	0.4518*	0.4586*

To further investigate the behaviour of the predictors, we consider how dispersed the exposure is distributed over the groups in the actual rankings. We have identified the coefficient of variation (CV) (Abdi, 2010), which is the ratio of the standard deviation of a distribution to the mean of the distribution, to be a suitable measure for this. A lower CV indicates a flatter distribution, e.g., if the actual exposure is distributed [0.25, 0.25, 0.25, 0.25] over four groups then $CV = 0$.

Figure 4.4 shows the distribution of the CV values over all queries per category for BM25. Analysing the figure, it is clear that the categories Popularity, Languages, and Alphabetical have the lowest median (orange line). The categories Geographic Location and Continent have the highest median and their interquartile range contains the highest CV values. Notably, the majority of categories on which our proposed GEP predictor significantly outperforms the baselines show a higher variability in the distribution of the actual exposure scores in the ranking than the other categories. This observation suggests

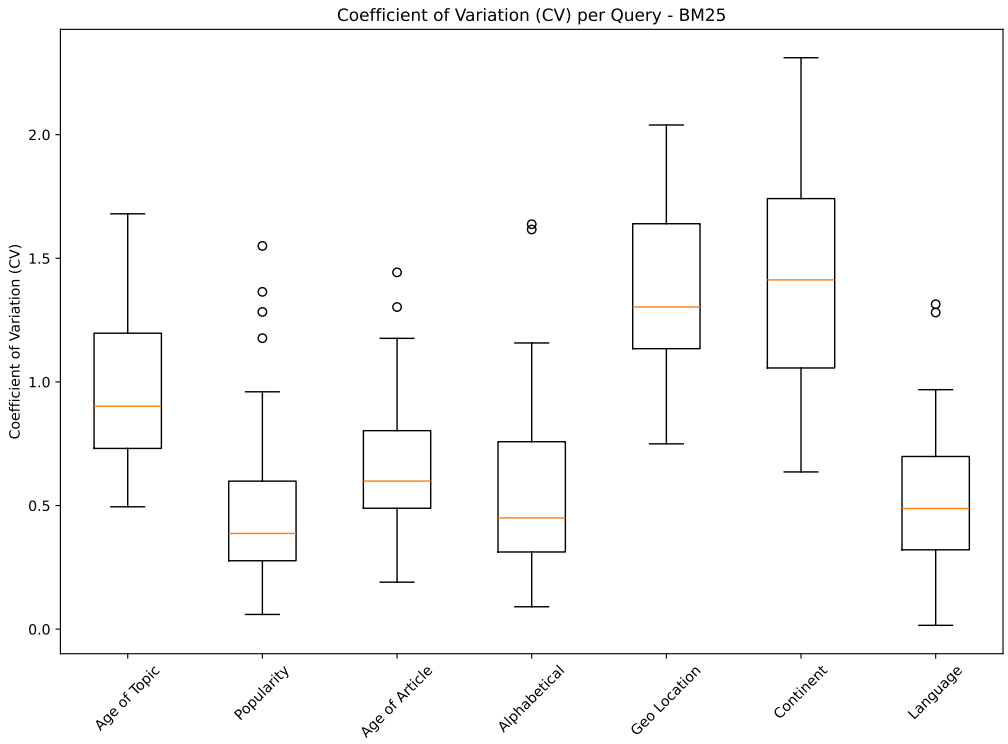


Figure 4.4: The figure shows the dispersion of the exposure in the categories per query measured by Coefficient of Variation. A lower CV indicates a flatter distribution, e.g., if the actual exposure is distributed [0.25, 0.25, 0.25, 0.25] over four groups then $CV = 0$.

that our predictor is better than the baselines when the exposure is not uniformly distributed. There are only two fairness characteristics on which our GEP predictor does not outperform the baselines with statistically significant differences: Popularity and Alphabetical. Notably, these two fairness characteristics have the lowest median CV over all fairness characteristics. This suggests that the coefficient of variation plays an important role in predicting the exposure over groups since all of our baselines provide better estimates for categories with more flat distributions. In answer to RQ4.1, we conclude that our proposed predictor can outperform all of the used baselines and provide promising results independent of the deviations in the exposure distribution.

Results RQ4.2: To answer RQ4.2, we investigate how the prediction performance is affected when a query is altered through Pseudo-Relevance Feedback. The first two sections of Table 4.4 show the results for the predictors when BM25 is combined with different PRF strategies (RM3, KLQ). We observe that for both PRF approaches over all categories, our predictor GEP outperforms the baselines. On five categories, our approach provides significant improvements over the baselines. This is the same trend we have observed when analysing the standard retrieval approaches. However, the results in Tables 4.3 and 4.4 show that for all of the categories and approaches, the prediction quality drops

Table 4.4: Evaluation of the predictors over common first-pass retrieval rankers with different Query Expansion models applied for rankings with $k = 100$. * indicates significant differences (t-test, with Bonferroni correction, $p < 0.01$) in prediction performance between GEP and the baselines.

Approach	Age of Topic	Popularity	Age of Article	Languages	Alphabetical	Continent	Geo-Location (TREC 21)
BM25 + RM3							
GEP	**0.3134**	**0.1722**	**0.2030**	**0.1684**	**0.1456**	0.**3156**	**0.2723**
SCS	0.3772*	0.1763	0.2507*	0.1943*	0.1548	0.4370*	0.4364*
AvIDF	0.3941*	0.1746	0.2603*	0.1996*	0.1571	0.4862*	0.5038*
AvICTF	0.3986*	0.1835	0.2493*	0.1940*	0.1567	0.4808*	0.4876*
AvPMI	0.3823*	0.1747	0.2481*	0.1874*	0.1561	0.4524*	0.4613*
CORI	0.3940*	0.1836	0.2514*	0.1933*	0.1557	0.4717*	0.4796*
BM25 + KLQ							
GEP	**0.2962**	**0.1625**	**0.1911**	**0.1658**	**0.1430**	**0.2963**	**0.2614**
SCS	0.3624*	0.1706	0.2491*	0.1905*	0.1489	0.4218*	0.4226*
AvIDF	0.3813*	0.1670	0.2598*	0.1957*	0.1513	0.4742*	0.4915*
AvICTF	0.3864*	0.1708	0.2467*	0.1902*	0.1510	0.4688*	0.4750*
AvPMI	0.3669*	0.1651	0.2461*	0.1836*	0.1505	0.4365*	0.4479*
CORI	0.3815*	0.1708	0.2487*	0.1894*	0.1500	0.4602*	0.4667*
SPLADE + RM3							
GEP	**0.1912**	**0.2870**	**0.2549**	**0.1516**	**0.1574**	**0.3262**	**0.3511**
SCS	0.2365*	0.3300*	0.2871*	0.1569	0.1670	0.4651*	0.5149*
AvIDF	0.2438*	0.4150*	0.2898*	0.1532	0.1637	0.5179*	0.5665*
AvICTF	0.2293*	0.4241*	0.2895*	0.1537	0.1576	0.5155*	0.5564*
AvPMI	0.2344*	0.3857*	0.2830*	0.1549	0.1610	0.4723*	0.5321*
CORI	0.2315*	0.4216*	0.2895*	0.1552	0.1572	0.5107*	0.5508*
SPLADE + KLQ							
GEP	0.2291	**0.2410**	**0.2662**	0.1780	0.1777	**0.3416**	**0.3251**
SCS	0.2200	0.3300*	0.2980*	0.1542	**0.1738**	0.4651*	0.4961*
AvIDF	0.2161	0.3538*	0.3009*	**0.1467**	0.1847*	0.5211*	0.5516*
AvICTF	**0.2120***	0.3613*	0.3009*	0.1535	0.2054*	0.5196*	0.5395*
AvPMI	0.2230	0.3245*	0.2933*	0.1659	0.1932*	0.4744*	0.5163*
CORI	**0.2120***	0.3615*	0.3010*	0.1541	0.2045*	0.5154*	0.5335*

compared to the standard retrieval rankings without pseudo-relevance feedback. This is to be expected since the predictors have no knowledge of how the query might be changed. Nevertheless, the drop in performances is marginal, indicating that the effect of PRF on the exposure is very limited and that our predictor is still useful even when the query is altered. When SPLADE is combined with RM3, our predictor outperforms the baselines on all categories with significant differences in five out of seven categories.

Analysing the SPLADE ranking with the KLQ expansion, we observe statistical differences to all baselines on four out of seven categories, with GEP providing the lowest numbers. On the Age of Topic category, the CORI and the AvICTF baseline produce the best prediction, outperforming our approach. In the Alphabetical category, however, SCS achieves the best scores. Overall, we observe that our predictor remains robust with only small drops in prediction quality even when PRF is applied in the pipeline.

A possible explanation for the robustness of our predictor when PRF is applied and unknown query terms are added can be found by revisiting Figure 4.3. First, let us assume a scenario where the PRF component leads to only a re-ranking of documents and does not add any new documents to the ranking. Assuming such a ranking, we can observe from Figure 4.3 how the exposure can change per group by looking at the individual bell curves. If the number of documents in a group stays the same, an individual bell curve shows the range of how much exposure a group can obtain. From the figure, it is clear that the range of possible exposure values is very restricted and narrows down even more for an increasing number of documents.

Explanations for a second scenario where the exposure changes with variations in the number of documents per group due to alterations of the query can also be found in Figure 4.3. The figure shows that substantial changes in the exposure distribution only occur if a relatively large number of documents (e.g., +5) from a group are added to the ranking. Moreover, as can be seen from Figure 4.1, if the documents are added in lower ranking positions, then relatively little additional exposure is accumulated, as previously discussed in Section 4.3. To answer RQ4.2, we can conclude that our proposed predictor is still effective even when PRF is applied. Moreover, we have given insights into why PRF interventions might only lead to limited effects on the exposure distribution.

4.7 Conclusions

In this chapter we have introduced the task of QEP, in which we predict how exposure will be distributed across groups of documents in response to a query. The task of QEP addresses the crucial issue of underrepresenting groups after the first-stage retrieval stage. In particular, when too few documents are retrieved from a group in the first-stage retrieval, fairness-aware re-ranking strategies may fail to correct the resulting imbalance.

These insights reinforce the importance of our thesis statement in which we claim that recall-enhancing techniques are essential to achieving fairer search results. To address this issue, we proposed the GEP model which is a first approach to solve the task of QEP. GEP is a pre-retrieval predictor designed to estimate the distribution of exposure across document groups before any ranking is generated. Our evaluation showed that the novel GEP model substantially outperforms standard query performance predictors and the CORI algorithm, reducing prediction error by up to 40%. The model was particularly effective in scenarios with high variability in exposure, demonstrating strong adaptability. Crucially, GEP maintained robust performance even in settings where queries were expanded via PRF, indicating its effectiveness across diverse IR systems. This robustness makes GEP a practical and generalisable tool for anticipating exposure imbalances before they appear in final rankings. In the subsequent chapters, we build on this insight by developing and evaluating methods that improve recall and exposure for underrepresented groups, thereby enhancing fairness without compromising relevance.

Chapter 5

Quantifying Query Fairness Under Unawareness

5.1 Introduction

In the previous chapter, we explained that a document's exposure to a user depends on its position in the ranking (c.f. Chapter 3, Section 4.3). In our thesis statement (c.f. Section 1.2), we posit that we can enhance the exposure allocation over groups of documents, i.e., improve the fairness in the search results. An important factor influencing the fairness of search results is the query that is issued by the user. The degree of unfairness in the search results can vary across different queries, depending on how a query is formulated. Such unfairness can be introduced either directly by the user, e.g., through the replication of existing biases when formulating the query (Kopeinik et al., 2023), or automatically, e.g., through the auto-completion features of a search system (L. Chen et al., 2018). Therefore, assessing the fairness of search results related to a specific query is crucial to determine whether fairness interventions are needed. In the previous chapter (c.f. Chapter 3), we have already introduced the task of QEP (c.f. Section 4.4) in which we predict the fairness of the search results before the retrieval process. In this chapter, we focus on the fairness of the search results after a retrieval has been executed. We introduce the terminology of Query Fairness Estimation (QFE) for assessing the fairness of search results for a given query. We define QFE as the task of assessing how equitably exposure is distributed across document groups in the ranked results for a given query.

To perform QFE, one typically needs access to the group labels of the ranked documents (Kuhlman et al., 2021; Raj & Ekstrand, 2022; Zehlike et al., 2022). Group labels refer to categorical attributes assigned to documents, such as region, gender, or publication date, that are used to evaluate fairness across document sets. These labels group

documents by different characteristics, such as geographic location or popularity. However, access to these group labels is often limited due to legal, ethical, or other data availability constraints (Bogen et al., 2020; Holstein et al., 2019). As a result, fairness evaluations can occur under "unawareness," where the actual labels must be inferred.

One way to obtain the labels under unawareness is to use human annotators. However, this is costly and impractical in most scenarios. An alternative is to deploy classifiers to infer the document labels automatically. A classification method deployed in practice is the *Bayesian improved surname geocoding* (BISG) which is used to infer race from surnames and ZIP codes using Bayesian statistics and U.S. Census data (Adjaye-Gbewonyo et al., 2014). While cost-efficient, deploying classifiers can introduce more unintended unfairness and bias by the classifiers themselves (Ghosh et al., 2021).

Moreover, even seemingly accurate classifiers can lead to unreliable results when performing QFE. For example, a good classifier may achieve high accuracy by focusing disproportionately on one class, thus minimising errors like false positives at the expense of increasing errors like false negatives (see §1.2 in Esuli et al., 2023). While technically accurate, this skewed performance fails to reflect the true distribution of groups, making it unreliable for assessing the proportions in a broader set of documents and can lead to significant errors and misjudgments in the measurements of fairness (J. Chen et al., 2019).

In this section, we propose using quantification techniques, i.e., machine learning models trained to estimate the relative frequencies of classes in unlabelled data (Esuli et al., 2023), to improve QFE. This aligns with our thesis statement (c.f. Section 1.2) since accurately measuring the search results' fairness is fundamental for developing methods that enhance the fair distribution of exposure across groups. The remainder of this section is structured as follows:

- In Section 5.2 we position our approach in relation to existing work.
- In Section 5.3 we first introduce the concept of quantification. We present the standard scenario (Prior Probability Shift) in which quantification methods are applied (Section 5.3.2) and introduce standard quantification methods (Section 5.3.3).
- In Section 5.4, we show that when assessing fairness in the search results, a sample selection bias is present and therefore, standard quantification methods fail. Therefore, in Section 5.4.1 we propose a method to make quantification methods robust to the sample selection bias in a binary group setting.
- In Section 5.4.2, we extend this method for the case of multiple groups within one fairness characteristic.
- In Section 5.5 we introduce our research questions and the experimental setup.

- In Section 5.6, we report our obtained results and answer our research questions.
- In Section 5.7, we briefly discuss the ethical implications of inferring group labels.
- Finally, in Section 5.8, we summarise the chapter and highlight our findings.

5.2 Related Work

To measure the fairness of search results for a given query, i.e., for the task of QFE, several measures have been introduced (Biega et al., 2018; Diaz et al., 2020; Kirnap et al., 2021; Kuhlman et al., 2021; Raj & Ekstrand, 2022; Sapiezynski et al., 2019; Singh & Joachims, 2018; E. Yang et al., 2024; K. Yang & Stoyanovich, 2017; Andrus & Villeneuve, 2022). While these measures cover different notions of fairness (c.f. Chapter 3, Section 3.1) they all depend on accurate knowledge of the document labels in a ranking. In this section, we consider an "unawareness" scenario, where group labels are unavailable, requiring alternative solutions to ensure accurate fairness assessments.

Related to this, Ghosh et al. (2021) have conducted an extensive study showing that inferring labels using standard classifiers can be problematic. Their work highlights the need for reliable methods to access accurate fairness labels. Although several studies have focused on enhancing the fairness of classifiers with noisy or incomplete group labels (Celis et al., 2021; Friedler et al., 2021; Ghosh et al., 2023; Mozannar et al., 2020; S. L. Wang et al., 2020), they primarily address fairness metrics for classification, not for ranking tasks.

In this thesis, we focus specifically on QFE for rankings under unawareness, an area that has received less attention compared to classification. In the absence of group labels, F. Chen and Fang (2023) proposed a distribution-based learning approach that leverages contextual features. Their approach uses a loss function that does not require explicit group labels but instead targets a fair distribution. Unlike their task of mitigating unfairness, our main objective is to obtain reliable estimates of group proportions in a ranking to accurately measure the fairness of search results.

Kirnap et al. (2021) proposed a method using a small query-dependent subset of data annotated by human assessors for QFE. Moreover, they only focused on binary group fairness, comparing protected versus non-protected groups. Our work addresses instead the characteristics of multiclass groups and does not require impractical human annotations for each query.

Related to our work on QFE in rankings, Ghazimatin et al. (2022) have proposed an approach that we will denote as Post-Metric Correction (PMC). Similar to our use of quantification, the PMC variants include a correction phase and rely on the outputs of an underlying classifier. However, there are significant differences between our quantification-based approaches and the PMC variants. First, the PMC variants are designed only for

a binary group fairness assessment; our method natively caters to multi-valued sensitive attributes. Moreover, each PMC variant is tied to a specific independence assumption, which may prove difficult to verify in general settings. Finally, the PMC methods are also tailored to one specific fairness metric, while our approach is more versatile. Our method applies a general pre-correction to the class prevalence estimates, which are then used to compute different fair ranking metrics. We will include the PMC method in our experiments as a baseline. In the next section, we will introduce quantification.

5.3 Learning to Quantify

Quantification is the task of estimating the class distribution (i.e., prevalence of each label) in a set of unlabelled data. It differs from classification, which predicts individual labels, by focusing on proportions over the whole set. Quantification relies on the observation that simply counting the labels predicted by a classifier tends to give inaccurate estimates of class prevalence, unless the classifier is perfect (see §1.2 in Esuli et al., 2023). This basic method, known as "Classify & Count" (CC) (Forman, 2005), is traditionally regarded as the baseline that any effective quantification method should outperform. Formally, a quantifier is a function $\lambda : \mathbb{N}^{\mathcal{X}} \to \Delta^{n-1}$ that maps bags (multi-sets) of instances from the input space $\mathcal{X} = \mathbb{R}^d$ to the probability simplex. This means $\lambda(\mathbf{X}) = \mathbf{p}$ lies on the $(n-1)$-simplex, $\Delta^{n-1} = \{p_1, \ldots, p_n : p_i \geq 0, \sum_i p_i = 1\}$, where n is the number of classes $\mathcal{Y} = \{1, \ldots, n\}$ and p_i represents the class prevalence of class i in bag $\mathbf{X}$. Given a classifier $h : \mathcal{X} \to \mathcal{Y}$ and a bag $\mathbf{X}$, CC is defined as:

$$\mathrm{CC}(\mathbf{X})_i = \hat{p}_i = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \mathbb{1}[h(\mathbf{x}) = i] \tag{5.1}$$

where $\hat{p}_i$ is the estimated prevalence of class i.

Throughout this chapter, we will interchangeably use the terms “classes” (here denoted by $\mathcal{Y}$) and “groups” as well as “labels” and “sensitive attributes” (denoted by $\mathcal{A}$). Moreover, we interchangeably use “bags” (or “multi-sets”) and “ranked lists of documents”, since our quantifiers regard the latter as unordered objects.

5.3.1 Practical Example

Consider an online hiring platform assisting recruiters to fill job openings with promising candidates. Recruiters query the platform with job descriptions and get ranked lists of candidates in return, in decreasing order of estimated job fitness. Platform developers want to ensure that their models are fair with respect to multiple attributes, including age, gender, race, and ethnicity. Given the sensitive nature of this information, most data

subjects will be hesitant to disclose it (Bogen et al., 2020). Developers, therefore, obtain sensitive attributes for a subset of users through voluntary data disclosure (Wilson et al., 2021; LinkedIn, 2024). This incomplete data can be used to estimate platform fairness across all queries and users. Finding the optimal way to perform this estimate is an open research problem tackled in the remainder of this section.

5.3.2 Dataset Shift

Quantification addresses situations where a difference (commonly referred to as a "shift" or "drift") occurs between the distribution P_{tr}, from which the training data is drawn, and the distribution P_{te}, from which the test data is drawn. This shift violates the Independent and Identically Distributed assumption (IID assumption). The IID assumption states that both the training and test instances are drawn from the same underlying distribution and that each instance is independent of the others (Bishop, 2006). When the IID assumption holds, estimating class prevalence in the test set is straightforward, as the training prevalence is a reliable estimate of the test prevalence.

However, when the IID assumption is violated due to a dataset shift, the training distribution $P_{tr}(X, Y)$ no longer mirrors the test distribution $P_{te}(X, Y)$. This mismatch complicates predictions, as the model trained on $P_{tr}(X, Y)$ cannot accurately estimate the class prevalence in the test data. An example of this in practice can be found in online hiring (c.f. Section 5.3.1), where the training set, consisting of candidates who disclose sensitive information (such as ethnicity y), may not be sampled IID from the same distribution as the test set. This situation, where $P_{tr}(X, Y) \neq P_{te}(X, Y)$, is commonly referred to as a *dataset shift* (Storkey, 2009).

Among the various types of dataset shifts described in the literature, quantification techniques are often deployed when a Prior Probability Shift (PPS) occurs. PPS is a type of shift seen in *anti-causal learning* (Schölkopf et al., 2012), where the covariates represent the predicted phenomenon's symptoms, rather than its causes. In such cases, the problem is often modelled with the factorisation $P(X, Y) = P(X|Y)P(Y)$. PPS occurs when the class distributions change, meaning $P_{tr}(Y) \neq P_{te}(Y)$, while the relationship between features and labels, $P_{tr}(X|Y) = P_{te}(X|Y)$, remains stable.

5.3.3 A Simple Quantifier

The Adjusted Classify & Count (ACC) method (Forman, 2005) is one of the simplest quantification techniques designed to address PPS. While ACC can be extended to multi-class settings, it is most straightforwardly illustrated in the binary case where $\mathcal{Y} = \{0, 1\}$. The key principle of ACC is captured by the following equation:

$$P_{te}(\hat{Y}=1) = P_{te}(\hat{Y}=1|Y=1)P_{te}(Y=1) + P_{te}(\hat{Y}=1|Y=0)P_{te}(Y=0) \tag{5.2}$$

Here, $P_{te}(\hat{Y}=1)$ represents the *Classify & Count* (CC) estimate of the prevalence of the positive class in the test set. Class prevalence refers to the proportion of instances belonging to each class in a dataset, and is the key quantity estimated by quantification methods. $P_{te}(\hat{Y}=1|Y=1)$ and $P_{te}(\hat{Y}=1|Y=0)$ denote the classifier's *true positive rate* (tpr) and *false positive rate* (fpr), respectively, which can be estimated using the training data. Under the assumption that the class-conditional distributions, $P(X|Y)$, remain consistent between the training and test sets, ACC adjusts the CC estimate by incorporating the classifier's error rates, as follows:

$$\text{ACC}(\mathbf{X})_1 = \frac{\text{CC}(\mathbf{X})_1 - \hat{\text{fpr}}}{\hat{\text{tpr}} - \hat{\text{fpr}}} \tag{5.3}$$

ACC is more reliable than CC under PPS by accounting for the classifier error rates.

5.4 Countering Sample Selection Bias

Quantification methods such as the just introduced ACC (Equation 5.3) are designed to estimate group distributions more accurately than traditional classification methods, particularly in the presence of PPS. Consequently, it is reasonable to assume that they produce more reliable evaluations of fairness in search results compared to simpler techniques such as CC . However, quantification methods can perform worse than CC when the underlying data shift is not PPS—the scenario for which they are typically designed. For instance, let ξ represent a random variable indicating whether a document is relevant to a specific query ($\xi = 1$) or non-relevant ($\xi = 0$). The class-conditional distribution in this case is a mixture of relevant and non-relevant documents, expressed as:

$$P(X,\xi|Y) = P(X|Y,\xi=1)P(\xi=1) + P(X|Y,\xi=0)P(\xi=0) \tag{5.4}$$

In the training data, it is expected that $P_{tr}(\xi=1) \ll P_{te}(\xi=1)$, as the majority of training examples are unrelated to specific queries, whereas most test examples are query-relevant. This discrepancy leads to differences in class-conditional distributions between the training and test sets, violating the assumptions of PPS.

Here, ξ acts as a *selection variable*, introducing a form of dataset shift known as Sample Selection Bias (SSB). SSB occurs when the samples used for training or evaluation are not representative of the target population due to the selection mechanism, often leading to biased estimations. While SSB traditionally refers to biases in the training data (Storkey, 2009), in this case, it affects the test set.

Despite the effectiveness of quantification methods in addressing other types of dataset shift (Gonzalez et al., 2024), no existing method is robust against SSB. In the following section, we present the first approach to adapt quantification methods for robustness against SSB, specifically for use in fairness evaluation (QFE).

5.4.1 Proposed Approach for QFE

The failure of traditional quantifiers in the presence of SSB arises from the violation of the assumption that $P_{tr}(X|Y) = P_{te}(X|Y)$, which underpins the concept of PPS.

Mitigation

To address this, we revisit the ACC method (Equation 5.3). ACC assumes that class-conditional distributions are consistent between the training and test sets, meaning $P_{tr}(\hat{Y}|Y) = P_{te}(\hat{Y}|Y)$. This assumption is valid as long as the classifier, h, is a measurable function (Lipton et al., 2018).In general, $P_{te}(\hat{Y} = i|Y = j)$ represents the classifier's *classification rates* in the test set and can be expressed as:

$$P_{te}(\hat{Y} = i|Y = j) = \mathbb{E}_{\mathbf{x}\sim P_{te}(X|Y=j)}[\mathbb{1}[h(\mathbf{x}) = i]] \tag{5.5}$$

Since the true distribution $P_{te}(X|Y = j)$ is unavailable, it is typically approximated using the training data, assuming $P_{te}(X|Y) = P_{tr}(X|Y)$:

$$P_{te}(\hat{Y} = i|Y = j) \approx \frac{1}{m}\sum_{k=1}^{m}\mathbb{1}[h(\mathbf{x}_k) = i] \tag{5.6}$$

However, as discussed in Section 5.4 , this assumption does not hold under SSB. The problem is not with the classifier itself but with the biased empirical distribution used to estimate the necessary parameters.

To address this, we propose separating the training of the classifier from the correction phase. This requires two labelled datasets: $L_{cls} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\eta}$, used for the classifier training, and $L_{corr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\eta'}$, used to estimate corrections. By isolating L_{corr}, we address the selection bias affecting the correction process.

Main Idea

Since sampling bias in the test data is inherent due to query-specific retrieval, we propose retrieving a biased subset from an auxiliary pool L_{corr} using the same retriever and query. This subset, $L_q \subset L_{corr}$, mirrors the query bias present in the test data.

By treating L_q as a sample from the query-biased distribution P_q, and noting that $P_q(\xi) \approx P_{te}(\xi)$, we can reasonably assume $P_q(X, \xi|Y) \approx P_{te}(X, \xi|Y)$. This restores the PPS assumption, neutralising the impact of selection bias. This approach uses L_q to compute query-specific corrections for estimating group prevalence in the top-k prefixes of unlabelled rankings U_q. For example, for our online hiring example (c.f. Section 5.3.1), ACC could be applied to estimate the prevalence of ethnicities in U_q by applying classification rates from the query-specific subset L_q instead of the entire labelled set L_{corr}.

5.4.2 Quantifying Query Fairness

In the previous section, we introduced the intuition behind our approach using the binary ACC method, a simple quantification technique. In this section, we extend this foundation to advanced multiclass quantification methods.

In multiclass quantification (Bunse, 2022b), approaches are typically formulated as solving for $\mathbf{p} \in \Delta^{n-1}$ (the unknown prevalences) in a system of linear equations:

$$\mathbf{t} = \mathbf{M}\mathbf{p} \tag{5.7}$$

where $\mathbf{t} = \Phi(U)$ represents the test set U, and $\mathbf{M} = [\Phi(L_1), \ldots, \Phi(L_n)]$ is a matrix containing the class-specific representations of the training subsets $L_i = \{\mathbf{x}_k : (\mathbf{x}_k, y_k) \in L, y_k = i\}$. The representation function $\Phi : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^z$ maps sets of instances to z-dimensional vectors. Typically, Φ takes the form:

$$\Phi(\mathbf{X}) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \phi(\mathbf{x}) \tag{5.8}$$

where $\phi : \mathcal{X} \to \mathbb{R}^z$ is an instance-wise embedding function, and Φ computes the mean embedding over the set. Different definitions of Φ and ϕ lead to various quantification methods. For the ACC method, ϕ represents a one-hot encoding of the predictions made by a classifier h. The matrix $\mathbf{M}$ is constructed from the classification rates of h, estimated using the training data. The class proportions in the test set are then computed as follows:

$$\mathbf{p} = \mathbf{M}^{-1}\mathbf{t} \tag{5.9}$$

More advanced methods refine this approach. For instance, Probabilistic Adjusted Classify & Count (PACC) (Bella et al., 2010), used in our experiments, employs a probabilistic classifier to define ϕ, which outputs posterior probabilities for each class. When $\mathbf{M}$ is non-invertible (Bunse, 2022a), PACC reformulates Equation (5.4.2) as the following optimisation problem:

$$\mathbf{p}' = \underset{\mathbf{p}\in\Delta^{n-1}}{\arg\min} |\mathbf{t} - \mathbf{M}\mathbf{p}|^2 \tag{5.10}$$

We also use KDEy (Moreo et al., 2024), a state-of-the-art multiclass quantification method. KDEy defines Φ as a Gaussian mixture model derived through kernel density estimation on the posterior probabilities produced by a probabilistic classifier. The maximum-likelihood variant of KDEy solves Equation (5.4.2) by minimising the Kullback-Leibler divergence:

$$\mathbf{p}' = \underset{\mathbf{p}\in\Delta^{n-1}}{\arg\min} \mathcal{D}_{\mathrm{KL}}(\mathbf{t}\|\mathbf{M}\mathbf{p}) \tag{5.11}$$

Both PACC and KDEy depend on a probabilistic classifier pre-trained on L_{cls}. Solving Equations (5.10) and (5.11) involves converting the documents in L_q and the test documents in U_q—retrieved for the same query but originating from separate pools, L_{corr} and U—into posterior probabilities and solving the respective optimisation problems. The efficiency of these computations is maintained through modern optimisation methods and the limited size of the retrieved sets, capped at 1,000 items (see Section 5.6.4).

5.5 Experimental Setup

In this section, we address the following research questions:

- **RQ5.1**: Does our proposed quantification method improve upon existing baselines?
- **RQ5.2**: Do quantification techniques allow for accurate QFE in a multiclass setting?
- **RQ5.3**: How does our algorithm's performance depend on the training pool's size and the rank exposure?

Table 5.1: Accuracy of Standard Classifiers over different fairness characteristics in the TREC 2022 collection.

Attribute	Groups	LR	SVM
Geo. Location	{Unk., Europe, Africa, Oceania, North America, Asia, Lt. America}	.800	**.807**
Gender	{Unknown, Male, Female}	**.968**	**.968**
Age of Topic	{Unk., Pre-1900s, 20th century, 21st century}	**.769**	.762

5.5.1 Dataset & Retrieval

To answer our research questions, we use the TREC 2022 Fair Ranking Track collection (Ekstrand, McDonald et al., 2022) as introduced in Section 4.5.1. Moreover, we use the queries of the TREC 2021 Fair Ranking Track collection to select a suitable model as we detail in the coming section 5.5.4.

To evaluate the generalisation of our approach, we use three fairness characteristics from the TREC 2022 collection on which traditional classifiers can be successfully applied. On geographic location, gender, and age of topic, standard classifiers such as logistic regression and support vector machines (SVM) achieve reasonable accuracy (> 0.75) as can be seen in Table 5.1. Moreover, the table shows the different groups per fairness characteristics. On all other available characteristics, the accuracy is below 0.75.

For our experiments, we use all available 97 queries from the collection, including the training and the evaluation queries. We employ the BM25 retrieval model (Robertson et al., 1994) for document retrieval, using its standard parameters (c.f. Section 2.2.3).

5.5.2 Evaluation Measures

In our experiments, we evaluate the accuracy of QFE when our protocol is applied using metrics that implement an exposure drop-off based on rank position. In Section 5.6.1, we evaluate our method against baseline approaches that are specifically designed for the normalised discounted difference (rND) (Ghazimatin et al., 2022), a specific binary-only fairness metric. Therefore, we include rND in our evaluation, defined as:

$$\mathrm{rND}(U_q) = \frac{1}{Z} \sum_{k \in K} \frac{1}{\log_2 k} \left| p_1^k - p_1^* \right| \tag{5.12}$$

Note that rND only considers the prevalence of the positive class, which is typically the disadvantaged or protected group. To evaluate fairness estimation in the multiclass setting, we use the normalised discounted Kullback-Leibler divergence (rKL) (Kullback & Leibler, 1951). rKL extends the binary fairness metric rND to the multiclass setting. Unlike rND,

which only quantifies fairness as the difference in representation between a single protected group and the rest, rKL captures distributional shifts across multiple groups. Since our approach relies on quantification, rKL provides a structured way to compare predicted and actual group proportions across ranking levels, ensuring that discrepancies in the estimated distributions are systematically assessed. For a given set of retrieved documents U_q and different ranking levels k (where $K = \{50, 100, 500, 1000\}$), rKL is defined as:

$$\mathrm{rKL}(U_q) = \frac{1}{Z} \sum_{k \in K} \frac{1}{\log_2 k} \mathcal{D}_{\mathrm{KL}}(\mathbf{p}^k || \mathbf{p}^*) \tag{5.13}$$

where $\mathbf{p}^k$ represents the group distribution for the top-k documents, and $\mathbf{p}^*$ is the group distribution across all judged-relevant documents in the test collection U from which the ranked list U_q is retrieved. Z is the normalisation factor, defined as $Z = \sum_{k \in K} \log_2 k$.

Since we work under the unawareness assumption, the true distribution $\mathbf{p}^k$ is unknown and must be estimated. Therefore, we use $\hat{\mathrm{rKL}}(U_q)$ (and $\hat{\mathrm{rND}}(U_q)$) to denote the score calculated using the predicted distributions $\hat{\mathbf{p}}^k$ instead of the true ones.

To evaluate the accuracy of our fairness predictions, we report the *absolute error* (AE), averaged across all ranked lists U_q retrieved for all queries. AE is defined as:

$$\mathrm{AE}(Queries, M) = \frac{1}{|Queries|} \sum_{U_q \in Queries} \left| M(U_q) - \hat{M}(U_q) \right| \tag{5.14}$$

where M represents the fairness metric (either rKL or rND).

Since our methods are based on quantification, we also report the Relative Absolute Error (RAE), a quantification-specific measure that compares a predicted distribution with the true distribution. RAE is defined as follows:

$$\mathrm{RAE}(\mathbf{p}, \hat{\mathbf{p}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{p}_i - p_i|}{p_i} \tag{5.15}$$

We select RAE because it is particularly sensitive to minority classes, where estimation errors may be small in absolute terms but proportionally large (Sebastiani, 2020).

Algorithm 1: Experimental protocol.

Input : • data
• Classifier learner CLS
• Quantification method QUANT
• Group labels

Output: • rKL of the fairness estimates
• RAE of the group prevalence estimates

1 $L, U \leftarrow$ split(data; 50%, 50%)
2 $L_{cls} \leftarrow$ draw(L; 500 documents per group)
3 $L \leftarrow L - L_{cls}$
4 $h \leftarrow \text{CLS}(L_{cls})$
5 **for** $size \in \{10\text{K}, 50\text{K}, 100\text{K}, 500\text{K}, 1\text{M}, 3.25\text{M}\}$ **do**
6 $\quad L_{size} \leftarrow$ undersample(L; $size$)
7 $\quad L_{corr} \coloneqq L_{size}$ # alias
8 $\quad$ **for** query $\in$ *Queries* **do**
9 $\quad\quad U_q \leftarrow \text{Retrieve}(U, \text{query})@1000$
10 $\quad\quad L_q \leftarrow \text{Retrieve}(L_{corr}, \text{query})@1000$
11 $\quad\quad L_q \leftarrow$ keep(L_q; max 200 most relevant documents per group)
12 $\quad\quad \lambda_h \leftarrow \text{QUANT}(L_q, h)$
13 $\quad\quad$ **for** $k \in \{50, 100, 500, 1000\}$ **do**
14 $\quad\quad\quad \hat{\mathbf{p}}^k \leftarrow \lambda_h(U_q^k)$
15 $\quad\quad\quad \mathbf{p}^k \leftarrow \text{prevalence}(U_q^k)$
16 $\quad\quad\quad$ compute RAE($\mathbf{p}^k, \hat{\mathbf{p}}^k$)
17 $\quad$ compute AE(*Queries*, rKL)

5.5.3 QFE Protocol

This section details the protocol we propose for QFE, designed to address our research questions (RQs). The pseudo-code for the protocol is provided in Algorithm 1. We use CLS as a variable for the classifier and Quant as a variable for a quantification method. Below, we describe the key steps and variables in detail.

Protocol Steps

1. **Data Splitting (Line 1)**: The dataset is split into two equal parts: a training pool L and a test pool U, each containing 50% of the documents. This ensures a balanced dataset for training and evaluation.

2. **Classifier Training (Lines 2–4)**: A subset of 500 documents per group is sampled from L to create a dedicated training set L_{cls} for the classifier. We select 500 documents per group as this represents a realistic number of labels that could feasibly be obtained in practical scenarios. Choosing a lower number might not provide enough data to make reliable predictions, while a higher number could impose significant resource and time constraints, making it impractical for many real-world applications. These documents are removed from L to prevent overlap. The classifier h is then trained using L_{cls}.

3. **Training Pool Resampling (Lines 5–7)**: The size of the remaining training pool L is varied by undersampling to create subsets of specific sizes ($size$). These subsets, referred to as L_{corr}, are used for retrieving query-relevant training documents later in the protocol. The sizes explored are $size \in \{10K, 50K, 100K, 500K, 1M, 3.25M\}$, balancing labelling cost and performance.

4. **Query Processing (Lines 8–15)**:

   (a) **Retrieving Documents (Lines 9–10)**: Relevant documents for the query are retrieved from the test pool U, producing a ranked list U_q of up to 1,000 documents. Similarly, query-relevant documents are retrieved from L_{corr} to form L_q, a training pool subset biased towards the query.

   (b) **Relevance Filtering (Line 11)**: To ensure fairness, L_q is filtered to include at most 200 of the most relevant documents per group. This limit helps mitigate potential group imbalances in the retrieved training data while also ensuring that the selected documents maintain high relevance. Including a larger number of documents could result in diminishing relevance, which might negatively impact the overall quality of the training data

   (c) **Quantification Method Application (Line 12)**: The quantification method QUANT is applied to L_q using the trained classifier h, producing a query-specific quantification model λ_h.

5. **Rank Cutoff Evaluation (Lines 13–15)**:

   (a) **Prevalence Estimation (Line 14)**: The quantification model λ_h is used to estimate the group prevalence $\hat{\mathbf{p}}^k$ in the top-k results of U_q. The true group prevalence $\mathbf{p}^k$ is also computed for evaluation.

   (b) **Accuracy Metrics (Line 15)**: The RAE between the estimated and true prevalence values is computed for each rank cutoff.

6. **Aggregate Metrics (Line 16)**: The mean relative Kullback-Leibler divergence (rKL) is calculated across all queries to evaluate the fairness estimation performance.

Key Experimental Variables

- **Training Pool Size ($size$)**: The size of L_{corr} affects the quality of L_q. Larger training pools reduce the distributional mismatch between L_q and U_q but incur higher labelling costs. Sizes range from 10K documents (realistic) to 3.25M documents (optimistic). This is controlled in Line 5. This is important since there is an evident trade-off between cost and performance: a larger pool size implies higher labelling cost, while at the same time helps in reducing the discrepancy

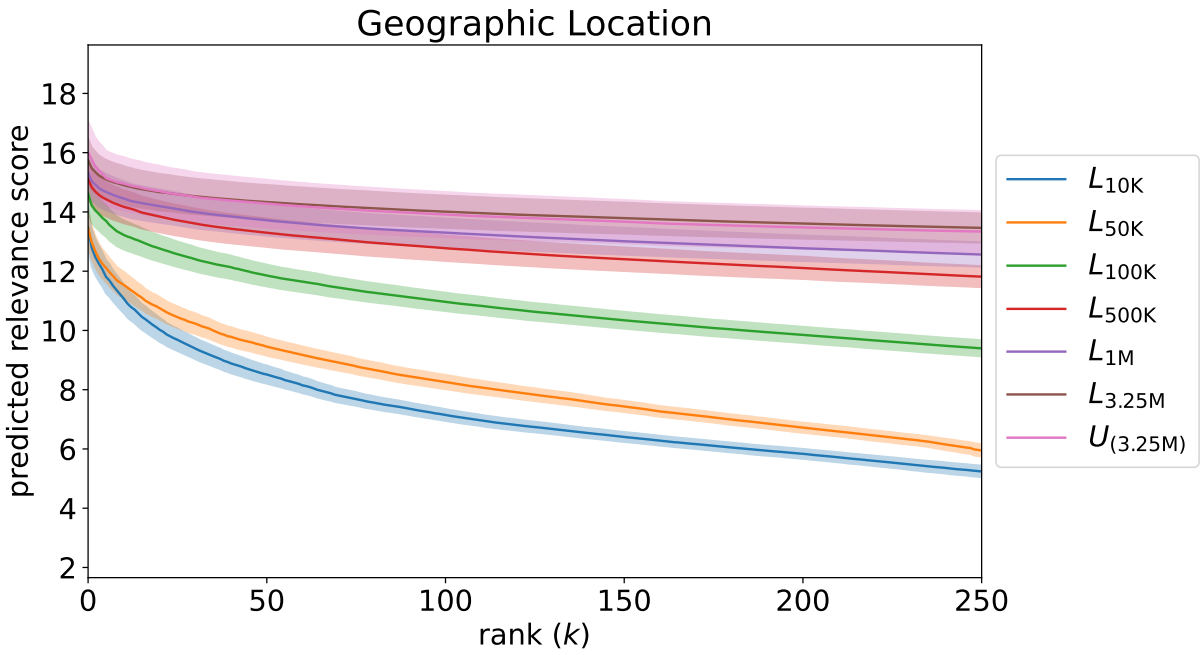


Figure 5.1: The plot shows the distribution of predicted relevance score per rank across all queries. As the size of L_{corr} increases, its distribution becomes more aligned with U.

between the distribution of the documents retrieved from the training and test pool. Figure 5.1 illustrates the predicted relevance scores across all queries at different rank positions (k) for Geographic Location, comparing different correction pool sizes ($L_{10K}, L_{50K}, L_{100K}, L_{500K}, L_{1M}, L_{3.25M}$) and the test pool ($U_{3.25M}$). Each curve represents the average predicted relevance score, while the shaded regions indicate variability or confidence intervals.

As the figure shows, larger correction pools (e.g., $L_{3.25M}$) tend to maintain higher predicted relevance scores across ranks compared to smaller pools (e.g., L_{10K}), particularly for lower ranks where the most relevant documents are concentrated.

This highlights that a larger correction pool is more likely to contain relevant documents, contributing to higher predicted relevance scores throughout the ranking. This demonstrates the importance of using sufficiently large training pools to improve the effectiveness of QFE.

- **Rank Cutoff (k)**: The protocol evaluates how different rank cutoffs influence the accuracy of QFE. The tested values are $k \in \{50, 100, 500, 1000\}$, as specified in Line 13.

- **Queries**: All 97 queries from the TREC 2022 collection are processed to ensure comprehensive evaluation (Line 8).

Limitations of the Protocol

In quantification research, experimental evaluations typically simulate class distribution shifts to stress-test a quantifier's performance. These shifts are typically applied to the test set. However, it is unclear how to simulate such shifts in our case without interfering with the presence of SSB.

Nevertheless, it is important to introduce a shift in the prior since both the training and test pools are derived from the same preexisting collection, resulting in nearly identical true distributions P_{tr} and P_{te}. This is an oversimplification of the problem.

To address this, we artificially limit the training data to a fixed number of documents per class. Specifically, we limit the training set for the classifier to 500 documents per class (L_{cls} in Line 2) and restrict the training documents retrieved for each query to 200 per class (L_q in Line 11), to test our quantifiers under more realistic conditions.

5.5.4 Methods

We evaluate the following methods using our proposed protocol for QFE:

1. **PACC** (Bella et al., 2010): The Probabilistic Adjusted Classify & Count method, as described in Section 5.4.2. In our implementation, the classification rates matrix for the classifier h is estimated using L_q.

2. **KDEy** (Moreo et al., 2024): The KDE-based quantification method, also detailed in Section 5.4.2. In this variant, the class-wise densities of the posterior probabilities produced by h are modelled using L_q.

We compare these methods against the following baselines:

1. **Naive@k**: This baseline does not analyse the search results to estimate the prevalence in U_q^k. Instead, it reports the prevalence from the top-k training documents retrieved for each query from L_{corr}.

2. **CC** (Forman, 2005): The Classify & Count method, as defined in Section 5.3 and Equation (5.1). This approach relies solely on the predictions from h and does not use any corrections based on L_q.

3. **PMC**$_b$ (Ghazimatin et al., 2022): A binary-only method that adjusts an initial estimate $\hat{\text{rND}}_h$, obtained using proxy labels from h. The adjustment is applied using:

$$\text{rND}(U_q) = \frac{\hat{\text{rND}}_h}{(1-p)-w} \tag{5.16}$$

where $p = P_{tr}(\hat{Y} = 1|Y = 0)$ and $w = P_{tr}(\hat{Y} = 0|Y = 1)$.

4. **PMC$_b^+$**: An enhanced version of PMC$_b$, where the correction factors p and w are derived from L_q rather than L_{cls}.

5. **PMC$_d$** (Ghazimatin et al., 2022): Another binary-only method that applies the following adjustment to an initial estimate $\hat{\text{rND}}_h$:

$$\text{rND}(U_q) = \hat{\text{rND}}_h \cdot \left(\frac{(1-w)\cdot\beta}{x} - \frac{w\cdot\beta}{y} \right) \tag{5.17}$$

Here, $x = (1-w)\cdot\beta + p\cdot(1-\beta)$ and $y = w\cdot\beta + (1-p)\cdot(1-\beta)$, with p and w as defined for PMC$_b$.

6. **PMC$_d^+$**: An improved version of PMC$_d$, where p, w, β, x, and y are calculated using L_q instead of L_{cls}.

Classifier Training

All methods (except Naive@k) rely on the outputs of a classifier. We use Logistic Regression (LR) as our classifier. LR is chosen because it provides well-calibrated probabilistic outputs, which are required by many quantification methods. LR is trained offline on L_{cls}.

Model Selection

We perform model selection for LR by tuning the regularisation strength $C \in \{10^i\}$ for $-4 \leq i \leq 4$, and the CLASS_WEIGHT parameter in $\{Balanced, None\}$. For KDEy, which depends on the bandwidth of the kernel, we select the bandwidth using 99 queries (training and evaluation) from the TREC 2021 Fair Ranking Track (Ekstrand et al., 2021) collection (c.f. Section 4.5.1). We explore bandwidth values in the range $\{0.01, 0.02, \ldots, 0.10\}$.

5.6 Results

In the next sections, we report and discuss the experimental results obtained. Section 5.6.1 presents a comparison against the PMC variants focusing on the binary-only setting (RQ5.1). Section 5.6.2 presents our main experiments, where we assess the effectiveness of QFE in multiclass group settings (RQ5.2). Section 5.6.3 analyses the impact of rank k and the training pool size on the performance of our methods (RQ5.3). Lastly, Section 5.6.4 provides average time measurements for our methods.

Table 5.2: Results of QFE in terms of AE (lower is better) for rND prediction for L_{10K} (binary case). KDEy outperforms all methods.

Method	Geographic Location	Gender	Age of Topic
Naive@k	$0.014^{\dagger}\pm.023$	$0.047\pm.048$	$0.040\pm.040$
CC	$0.013^{\ddagger}\pm.029$	$0.014^{\ddagger}\pm.042$	$0.025^{\ddagger}\pm.040$
PMC$_b$	$0.020^{\dagger}\pm.049$	$0.014^{\ddagger}\pm.041$	$0.028\pm.040$
PMC$_b^+$	$0.030\pm.087$	$0.015^{\ddagger}\pm.037$	$0.042\pm.068$
PMC$_d$	$0.020^{\dagger}\pm.049$	$0.014^{\ddagger}\pm.041$	$0.028\pm.040$
PMC$_d^+$	$0.031\pm.083$	$0.015^{\ddagger}\pm.037$	$0.044\pm.063$
PACC (ours)	$0.043^{\ddagger}\pm.143$	$0.016^{\ddagger}\pm.040$	$0.019^{\dagger}\pm.023$
KDEy (ours)	**0.012±.027**	**0.011±.026**	**0.017±.021**

5.6.1 Binary Document Groups

We compare PACC and KDEy deployed within our protocol to previous work, specifically the PMC variants (Ghazimatin et al., 2022) (RQ5.1). The results discussed here are based on the realistic scenario where $size = 10K$.

The PMC variants, like our methods, include a correction phase and rely on underlying classifiers. However, while our methods apply a pre-correction to class prevalence estimates before calculating a fairness metric, the PMC methods apply a post-correction to the fairness metric itself, making them metric-specific (rND) and binary-only.

To enable a fair comparison, we binarise our datasets, with the positive class ($Y = 1$) representing the disadvantaged group. Specifically, we consider the following groups: "Africa" for Geographic Location, "Female" for Gender, and "Pre-1900s" for Age of Topic. All other groups are merged into the negative class ($Y = 0$).

Table 5.2 shows the absolute error (AE) for estimating rND. The values are averaged across all 97 queries, and a lower AE indicates a better performance. Boldface highlights the best method for each category, and superscripts † and ‡ indicate methods not statistically different from the best one at varying confidence levels ($0.001 < p\text{-value} < 0.01$ for †, and $p\text{-value} \geq 0.01$ for ‡), based on the non-parametric Wilcoxon signed-rank test. The colour coding (green for best, red for worst) further aids interpretation.

Table 5.2 summarises the absolute error (AE) for rND prediction across three fairness characteristics: Geographic Location, Gender, and Age of Topic. The results demonstrate that KDEy consistently performs best, with the lowest AE values across all categories. For Geographic Location, KDEy outperforms all other methods with an AE of 0.012 ± 0.027, highlighted as the best method. Similarly, for Gender and Age of Topic, KDEy maintains its superior performance with AE values of 0.011 ± 0.026 and 0.017 ± 0.021, respectively.

Table 5.3: Results of QFE in terms of AE (lower is better) for rKL prediction for L_{10K} (multiclass case). KDEy outperforms all methods.

	Naive@k	CC	PACC (ours)	KDEy (ours)
Geo. Location	.188†± .255	.212 ± .246	.298 ± .241	**.132 ± .181**
Gender	.305 ± .356	.068†± .143	.064 ± .108	**.037 ± .060**
Age of Topic	.213†± .265	.173†± .219	.285 ± .321	**.129 ± .158**

The CC method displays a competitive performance, particularly for Geographic Location and Gender, with AE values of 0.013 ± 0.029 and 0.014 ± 0.042. Although CC performs consistently worse than KDEy, the statistical test reveals that these differences are not significant, as indicated by the superscripts ‡. This may be due to a high classifier accuracy in the binary case, leaving a small room for improvement.

The PMC-based methods generally perform worse than KDEy, with notable variability across the fairness characteristics. For example, PMC_b demonstrates a reasonable performance for Geographic Location and Gender but exhibits noticeably weaker results for the Age of Topic. The PMC_b^+ and PMC_d^+ variants, which model corrections on L_q, show lower performance than their original counterparts, particularly for Geographic Location and Age of Topic, suggesting potential misalignment with the QFE's requirements.

Naive@k performs poorly overall, particularly for Gender and Age of Topic, highlighting the importance of correcting for class prevalence. PACC underperforms for Geographic Location and Gender, likely due to the limited training data available for modelling classification rates. However, achieves the second-best result in Age of Topic.

5.6.2 Multiple Groups

To answer RQ5.2, we investigate the performance of the different quantification methods over multiple groups of documents.

The results for the multiclass setting, presented in Table 5.3, are based on a scenario where $size = 10K$. As this is the first multiclass method developed for QFE, we compare our approaches against standard quantification baselines. Lower AE values indicate a better performance, and boldface highlights the best-performing methods. Superscripts † indicate methods not statistically different from the best-performing method ($0.001 < p$-value < 0.01), as determined by the non-parametric Wilcoxon signed-rank test.

KDEy emerges as the best-performing method across all three fairness characteristics: Geographic Location, Gender, and Age of Topic. For Geographic Location, KDEy achieves an AE of 0.132 ± 0.181, outperforming Naive@k (0.188 ± 0.255), CC (0.212 ± 0.246), and PACC (0.298 ± 0.241). The statistically significant difference between KDEy and the baselines highlights its robustness.

In the Gender category, KDEy also leads with an AE of 0.037 ± 0.060. While CC and PACC produce comparable results (0.068 ± 0.143 and 0.064 ± 0.108, respectively), Naive@k performs notably worse, with an AE of 0.305 ± 0.356. This underscores the limitations of methods that do not account for class prevalence in the test set.

For Age of Topic, KDEy achieves the lowest AE at 0.129 ± 0.158, with CC and Naive@k following at 0.173 ± 0.219 and 0.213 ± 0.265, respectively. PACC again underperforms, with an AE of 0.285 ± 0.321, further illustrating its lack of competitiveness in this context.

These results demonstrate the consistent superiority of KDEy in the multiclass setting, with statistically significant improvements over standard baselines in most cases. The findings also highlight the variability and limitations of other methods, particularly Naive@k and PACC, when applied to QFE across diverse fairness characteristics.

5.6.3 Impact of Rank k and Training Pool Size

To answer RQ5.3, we analyse variations in quantification performance at different rank levels k and training pool sizes. The results are presented in terms of RAE between the true and predicted distributions.

Figure 5.3 shows variations due to changes in the training pool size at $k = 100$. Figure 5.2 shows the performance at different ranks ($k \in \{50, 100, 500, 1000\}$) for $L_{corr} = L_{10K}$.

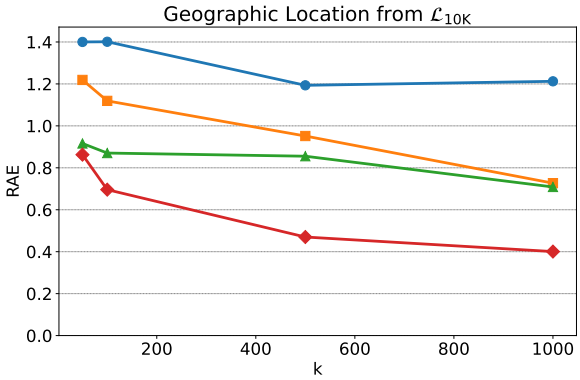
It is notable that KDEy consistently outperforms all other methods across different rank cutoffs (k) and training pool sizes, demonstrating its robustness. While most methods show improved performance as k increases, Naive@k does not exhibit this trend, indicating limitations in its approach to rank-specific prevalence estimation.

PACC and KDEy maintain a stable performance as the training pool size increases, suggesting that these methods are less reliant on the quantity of training data. In contrast, Naive@k is more sensitive to the training pool size, highlighting its dependency on the availability of sufficient labelled data.

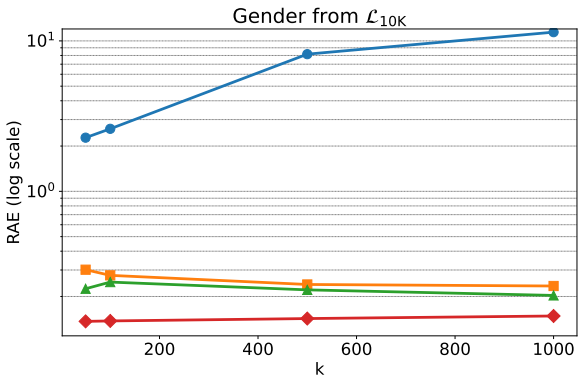
The CC method, however, shows no improvement with larger training pools. This is likely because CC does not use additional training data effectively, relying solely on the classifier's predictions without incorporating corrections or modelling enhancements.

5.6.4 Time Measurements

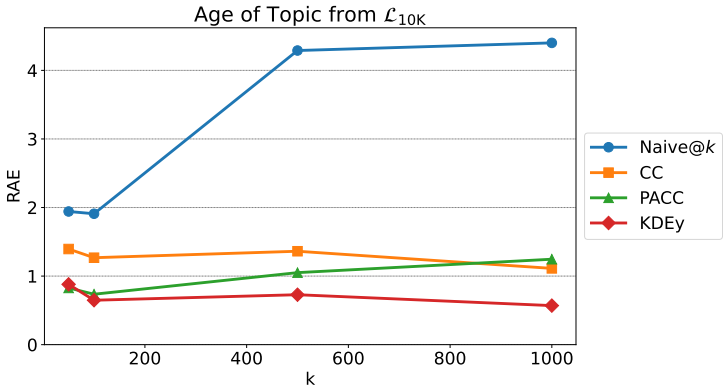
We measured the training and testing times on a desktop with a 12th Gen Intel(R) i9-12900K processor and 64GB of RAM, running Ubuntu 22. Our methods consist of a per-query correction learning phase followed by prevalence predictions. On average, PACC required 0.907 ms for learning and 3.316 ms per prediction, while KDEy took 1.586 ms for learning and 6.395 ms per prediction. Despite the query-time learning phase, our methods demonstrate high efficiency and are well-suited for integration into standard IR pipelines.



(a) Continent (10K)



(b) Gender (10K)

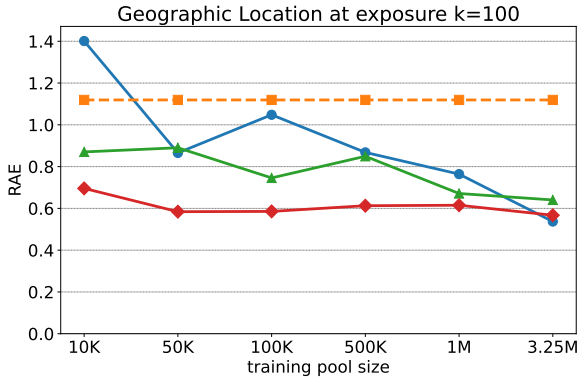


(c) Years Category (10K)

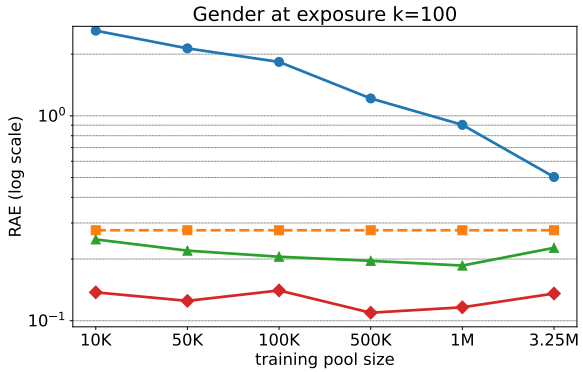
Figure 5.2: The plots show variations in quantification performance (measured in RAE – lower is better) at different values of k. KDEy and PACC outperform all baselines.

5.7 Ethical Considerations

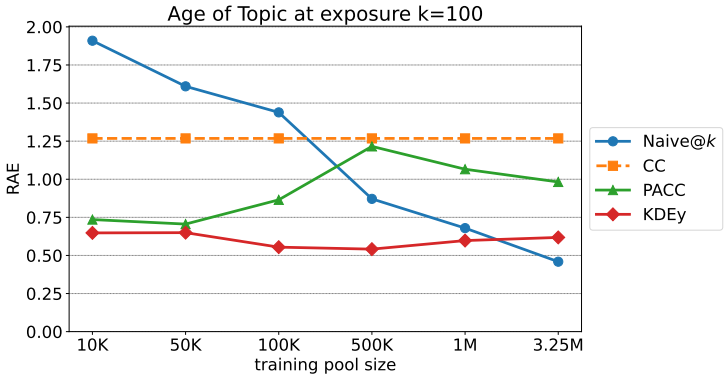
When addressing QFE, especially when dealing with sensitive attributes like demographics, it is crucial to consider both accuracy and bias when inferring group membership. When labels for individuals are inferred automatically, i.e., by machine-learned models, errors can occur, particularly if these models are not well-trained or if they are used on populations different from those they were designed for. These errors can lead to misleading fairness assessments and might even worsen existing biases. Moreover, if the data used to train these models is biased, the predictions made by the models can unfairly disadvantage certain groups, further reinforcing those biases. However, with our proposed technique, we argue that we are taking a step forward in addressing this issue by providing more accurate fairness assessments and by being mindful and explicitly addressing the challenges of inferring group labels. While we aim to improve the situation, we acknowledge that some level of error will always be present and can potentially be harmful.



(a) Continent (100)



(b) Gender (100)



(c) Years Category (100)

Figure 5.3: The plots show variations in quantification performance (measured in RAE – lower is better) at different training pool sizes. KDEy and PACC demonstrate stability and outperform CC and Naive@k.

Another concern when inferring labels is the issue of privacy and confidentiality, particularly when dealing with sensitive attributes. The process of inferring group membership involves handling sensitive demographic data, which carries an increased risk of compromising individuals' privacy. It is crucial to ensure that this data is kept confidential and that individuals' identities remain protected, avoiding any possibility of re-identification. To be considerate of this, we have used the TREC Fair Ranking dataset in our experiments, where private & confidential data is limited, making it more suitable for this purpose. The dataset primarily relies on publicly available information, which is already covered by Wikipedia, thereby reducing the risks associated with privacy concerns. This makes the TREC Fair Ranking dataset a more appropriate choice, as it allows us to focus on improving fairness while minimising potential privacy issues.

5.8 Conclusions

In this chapter we have introduced the task of QFE under the constraint of unawareness of group labels. While previous approaches that measure fairness of the search results, assume that sensitive attributes of documents are available, we addressed the open problem of measuring group exposure when labels are unavailable. We showed that standard classification-based methods, such as CC, are insufficient for QFE. CC based methods produce unreliable group prevalence estimates in ranked results, as they fail to account for shifts between training and test distributions. We demonstrated that more advanced approaches from the domain of Quantifiaction, such as ACC, offer improved estimates by correcting for prior probability differences. However, existing methods remain vulnerable to the *sample selection bias*. We identified the *sample selection bias* as a key challenge when applying quantification, since retrieved documents are inherently non-random and highly relevant. To resolve this, we proposed a novel quantification protocol that decouples classifier training from the correction phase. By introducing an auxiliary dataset that mirrors the selection bias in the test rankings, our method enables more accurate corrections. Our proposed protocol provides robust estimates of group prevalence and thereby enabling successful and reliable QFE fairness evaluation under conditions of unawareness. We show that our method can be applied in the multiclass settings, addressing QFE beyond the binary group scenario. This advancement is particularly important in real word scenarios. Our evaluations confirmed the effectiveness of our approach. Especially the KDEy variant consistently outperformed baseline methods, producing more reliable group prevalence estimates across both binary and multiclass settings. Moreover, we have shown that our approach is robust to variations in training pool sizes and ranking cutoffs. In summary we can conclude, that our proposed quantification protocol is an effective method to address the challenge of QFE under unawareness, validating our thesis statement.

Chapter 6

Fairness-Aware Adaptive Re-ranking

As we have previously discussed in Section 2.5.5, deep neural rankers such as BERT (Devlin et al., 2019) are highly effective (Lin et al., 2022) but computationally expensive. Consequently, IR systems deploy them within re-ranking pipelines (c.f. Section 2.5.5), where an initial retrieval model, such as BM25 (Robertson et al., 1994), retrieves documents that are then re-scored by a neural ranker to enhance the relevance of the search results (Nogueira, Yang, Cho & Lin, 2019; Lin et al., 2022). However, as noted in Section 4.3, the fairness of exposure allocation in such pipelines depends on the recall of the initial retrieval stage. Fair exposure in the final ranking becomes unfeasible when insufficient relevant documents are retrieved for certain groups (c.f. Section 4.3).

This recall limitation not only affects fairness but can also constrain a system's overall effectiveness. Relevant documents omitted during the initial retrieval stage are excluded from subsequent re-ranking, preventing their inclusion in the final ranked results. In line with our thesis statement (c.f. Section 1.2), we posit that improving recall, especially for underrepresented groups, can serve as a mechanism to mitigate exposure disparities while maintaining or improving system utility. In this chapter, we explore strategies for improving the recall of underrepresented groups in re-ranking pipelines, focusing on GAR (MacAvaney et al., 2022). GAR extends the standard re-ranking pipeline by employing a graph-based structure to introduce additional documents dynamically during the re-ranking process. The adaptive re-ranking approach of GAR addresses the recall limitations of traditional pipelines. Adaptive re-ranking refers to a retrieval process that dynamically augments the initial candidate set with additional documents during re-ranking, based on structural or semantic relationships. GAR improves this process by identifying and incorporating contextually relevant documents that may have been overlooked in the initial retrieval.

GAR's dynamic character offers a promising solution for tackling the challenges of insufficient recall and fairness. By expanding the pool of re-ranking candidates beyond the initial retrieval set, GAR has the potential to enhance the recall of underrepresented groups, thereby addressing exposure disparities. While GAR has demonstrated improvements in relevance compared to standard re-ranking pipelines (Frayling et al., 2024; MacAvaney et al., 2022), its impact on fairness has not yet been examined. In this chapter, we aim to fill this gap by evaluating GAR's effectiveness in improving the fairness of exposure allocation across document groups. Furthermore, we investigate whether the GAR process can be adapted and tailored to specifically enhance fairness, thereby providing a comprehensive evaluation of its alignment with the goals outlined in our thesis statement.

The remainder of this chapter is structured as follows:

- In Section 6.1 we present the related work to our investigation on GAR.
- In Section 6.2 we introduce the standard GAR re-ranking pipeline.
- In Section 6.3, we present our experimental evaluation of the standard GAR pipeline regarding the fairness of the search results.
- In Section 6.4, we introduce six policies for improving the fairness of the search results using the adaptive re-ranking method of GAR. In particular, we present In-Process policies and Pre-retrieval policies.
- In Section 6.5, we present our experimental evaluation of our proposed policies. We evaluate our policies in terms of fairness (Section 6.5.2) as well as relevance (Section 6.5.3). Moreover, we compare the Pre-retrieval and In-Process policies in Section 6.5.4.
- Finally, Section 6.6 summarises our contributions and presents the conclusions.

6.1 Related Work

To enhance the recall in a re-ranking pipeline, Macdonald et al. (2013) have investigated how many documents should be retrieved in the first stage to ensure a sufficient tradeoff between effectiveness and efficiency in an IR system. Their findings indicate that the smallest effective size of the initial set of retrieved documents depends on various factors, such as the information need and the evaluation measures. To reduce the impact of the recall effectiveness of the first stage retrieval, MacAvaney et al. (2022) proposed the graph-based adaptive re-ranking process, called GAR. In this chapter, we evaluate the potential of GAR for fair ranking, i.e., distributing the exposure equally amongst groups, aligning with our thesis statement (see Section 1.2). While Section 3.3 presented existing fair ranking

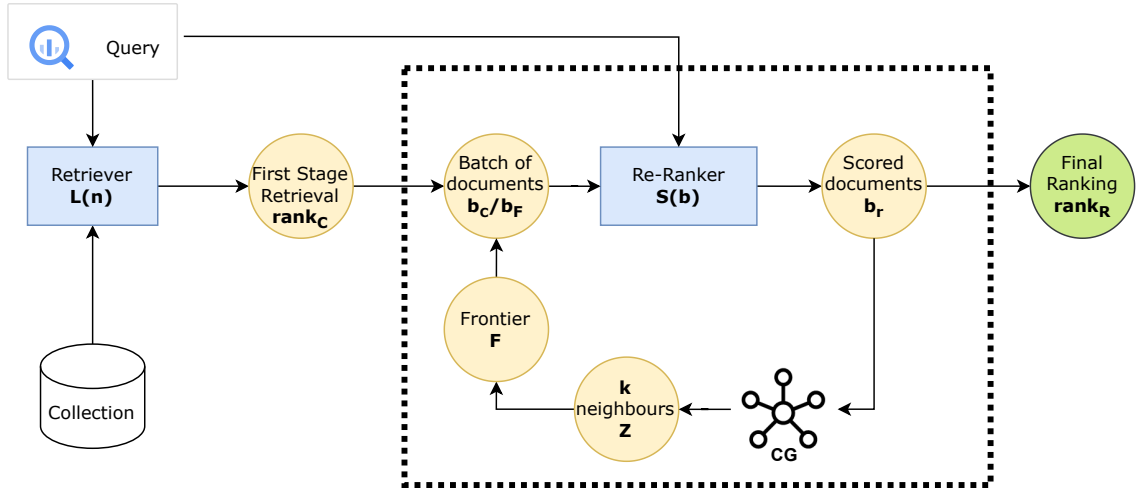


Figure 6.1: The figure shows the standard Adaptive Re-Ranking process. Starting on the left, the figure shows the standard re-ranking process extended by adaptive re-ranking (dotted rectangle). Blue rectangles represent ranking functions, whereas squares denote sets.

methods, these approaches do not align with our fairness definition (c.f. Equation 3.13) and do not explicitly focus on recall enhancement. Therefore, we compare our approach to systems using the standard GAR architecture and traditional re-ranking pipelines to demonstrate how our method addresses these fairness gaps.

6.2 Standard GAR

GAR extends the traditional re-ranking pipeline by identifying additional documents related to those retrieved during the initial stage and scoring them in the subsequent re-ranking phase. Specifically, GAR employs a corpus graph. The term corpus graph refers to a graph-based representation of the document collection, where documents are nodes and edges encode similarity relationships between them. This structure enables discovery of related documents beyond initial retrieval results. Figure 6.1 illustrates the GAR-enhanced re-ranking pipeline. In the figure, circles represent sets of documents, and rectangles depict rankers.

In a standard re-ranking pipeline, an initial retrieval model $L(n)$ (e.g., BM25) retrieves a set of n documents denoted as $rank_C$ and a neural re-ranker re-scores these documents without introducing new ones. However, GAR extends this by allowing new, related documents to be identified through the corpus graph and scored alongside the originally retrieved set. GAR, as illustrated in Figure 6.1, operates through an iterative cycle. It alternates between re-ranking the documents retrieved in the first stage and retrieving additional documents related to those initially retrieved. The steps are as follows:

1. **Step 1: Retrieving Initial Documents (Figure: Left-hand side)** The process begins when a *Query* is submitted. The first-stage retrieval model, $L(n)$ (e.g., BM25), retrieves an initial set of documents, $rank_C$, from the *Collection*. This initial retrieval aims to efficiently identify a pool of candidate documents that are potentially relevant to the query. These documents are ranked based on their relevance scores provided by $L(n)$. $rank_C$ forms the foundation for the subsequent stages.

2. **Step 2: Standard Re-Ranking (Figure: Top section)** From the initial retrieval $rank_C$, a *batch of documents*, b_C, is selected for re-ranking. This batch consists of the highest-scoring documents in $rank_C$ that have not yet been re-ranked. The selected documents are sent to the *Re-Ranker*, $S(b)$, which typically uses a neural model (e.g., ELECTRA (Clark et al., 2020)). The re-ranker assigns updated relevance scores to the batch, producing a set of re-scored documents, b_r. These re-scored documents are added to the *Final Ranking*, $rank_R$ (depicted as the green circle in the figure). This step ensures that documents with the highest predicted relevance in $rank_C$ are re-ranked using a more accurate relevance model.

3. **Step 3: Discovery of Related Documents (Figure: Lower portion)** For each document d_i in the re-scored batch b_r, the *Corpus Graph (CG)* is used to identify k-most similar neighbouring documents, Z_i. The corpus graph represents the document collection as a graph, where nodes correspond to documents, and edges represent their similarity to other documents. Each document in Z_i inherits the relevance score of its corresponding source document, d_i. Any documents in Z_i that are already present in $rank_C$ or the final ranking $rank_R$ are removed to avoid duplication. The remaining new documents are added to the frontier F, which contains a pool of newly identified but unscored documents.

4. **Step 4: Re-Ranking Documents from the Frontier (Figure: Cyclic loop)** From the frontier F, a new batch of documents, b_f, is selected. This batch includes the documents with the highest inherited relevance scores in F. The selected frontier documents are sent to the re-Ranker, $S(b_f)$ to be scored. The re-ranker assigns updated relevance scores to the documents and adds them to the $rank_R$.

5. **Step 5: Overall Process (Figure: Dashed rectangle)** The GAR process operates iteratively, alternating between re-ranking documents from the initial set $rank_C$ and discovering and re-ranking new documents through the corpus graph. During each iteration, newly identified documents from F are re-ranked and incorporated into the final ranking $rank_R$. This iterative cycle ensures a balance between refining the initial retrieval and expanding the pool of potentially relevant documents. The process continues until a predefined re-ranking budget is depleted.

In summary, GAR systematically alternates between re-ranking documents from the initial retrieval set and exploring additional documents through graph-based similarity in the corpus. This iterative process helps discover additional relevant documents that may have been missed in the first-stage retrieval, leading to improved final rankings.

6.3 Evaluation of the Standard GAR

To assess the impact of adaptive re-ranking on fairness, we compare the fairness of search results produced by standard re-ranking pipelines against those generated by the GAR adaptive re-ranking pipeline. In particular, we answer the following research question:

> **RQ6.1:** Does adaptive re-ranking lead to a fairer distribution of exposure compared to a standard re-ranking pipeline?

To answer our research question, we evaluate GAR using the TREC 2021 & 2022 Fair Ranking Track (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022) test collections (c.f. Sections 4.5.1). We use the same document groups as in Section 4.5, Table 4.1.

6.3.1 Baselines

To evaluate the effect of GAR on exposure allocation in search results, we compare its performance against two standard re-ranking baselines. To ensure a fair comparison, the retrieval model and re-ranker remain consistent across all pipelines, including the GAR process. All pipelines use BM25 for the first-stage retrieval and a neural ranking model for re-ranking. Specifically, we use monoT5 (Nogueira et al., 2020) and ELECTRA (Clark et al., 2020) as re-rankers. Document content is truncated to 512 tokens to align with the neural models' standard settings. We use the >> operator, from PyTerrier to describe our pipelines. Our baselines are denoted as BM25>>monoT5 and BM25>>ELECTRA, indicating that the output of BM25 is re-ranked by either monoT5 or ELECTRA. This setup ensures that differences in exposure allocation can be directly attributed to the application of GAR, as all other aspects of the pipelines are identical.

6.3.2 GAR Approaches

For our GAR approaches, we use the same ranker (BM25) and re-rankers (monoT5 & ELECTRA). The key difference to the standard re-ranking pipelines is the incorporation of graph-based exploration into the ranking process. To enable graph-based exploration, we construct a corpus graph for each test collection. Due to the size of the test collections (over 6 million articles each), we use the efficient TasB model (Hofstätter et al., 2021) to

index the collection and generate the corpus graph. TasB (Transformer-agnostic sparse bi-encoder) is a retrieval model designed to encode document and query representations for approximate nearest-neighbour search efficiently. It combines sparse retrieval techniques with the effectiveness of transformer-based encoders, allowing large-scale collections to be processed efficiently. In our setup, TasB is used to compute embeddings for each document in the collection and to identify neighbours based on vector similarity. By default, TasB truncates documents to 512 tokens during indexing aligning with other transformer based models. Following the setup in the original work proposed by MacAvaney et al. (2022), we configure the graph with 16 neighbours per document.

6.3.3 Metric

To measure the fairness of the search results, we use the AWRF (Sapiezynski et al., 2019), following the TREC Fair Ranking Track evaluation methodology. AWRF measures how similar the exposure distribution in the search results is to a given target distribution (c.f. Section 3.4.2). In our experiments, our target exposure consists of ensuring an equal distribution of exposure over all groups, i.e., if there are four considered groups, each group should receive 25% of the total available exposure (c.f. Section 3.13). If there are no differences between the actual exposure and the target exposure, the value will be 1. To measure the exposure a document receives in the final ranking, we use the exposure drop-off from the well-known Discounted Cumulative Gain (DCG) (Järvelin & Kekäläinen, 2002) metric (c.f. Section 3.12).

6.3.4 Results RQ6.1

To answer RQ6.1, we evaluate the fairness of exposure allocation achieved by GAR compared to standard re-ranking pipelines. Figure 6.2 presents the AWRF values across the seven evaluation categories across both collections for the baseline pipelines and GAR.

Our results reveal that applying GAR in its standard configuration reduces fairness compared to baseline pipelines. This decrease in AWRF is observed across all categories when ELECTRA is used as the neural ranker and in six out of seven categories when monoT5 is used. For instance, in the "Alphabetical" category, AWRF drops from 0.7756 to 0.7361 for ELECTRA and from 0.7565 to 0.7102 for monoT5. Similar reductions are observed in other categories, such as "Age of Article" (0.6790 to 0.6553 for ELECTRA and 0.7028 to 0.6907 for monoT5) and "Popularity" (0.6931 to 0.6733 for ELECTRA and 0.6760 to 0.6616 for monoT5). These results highlight the limitations of GAR in im-

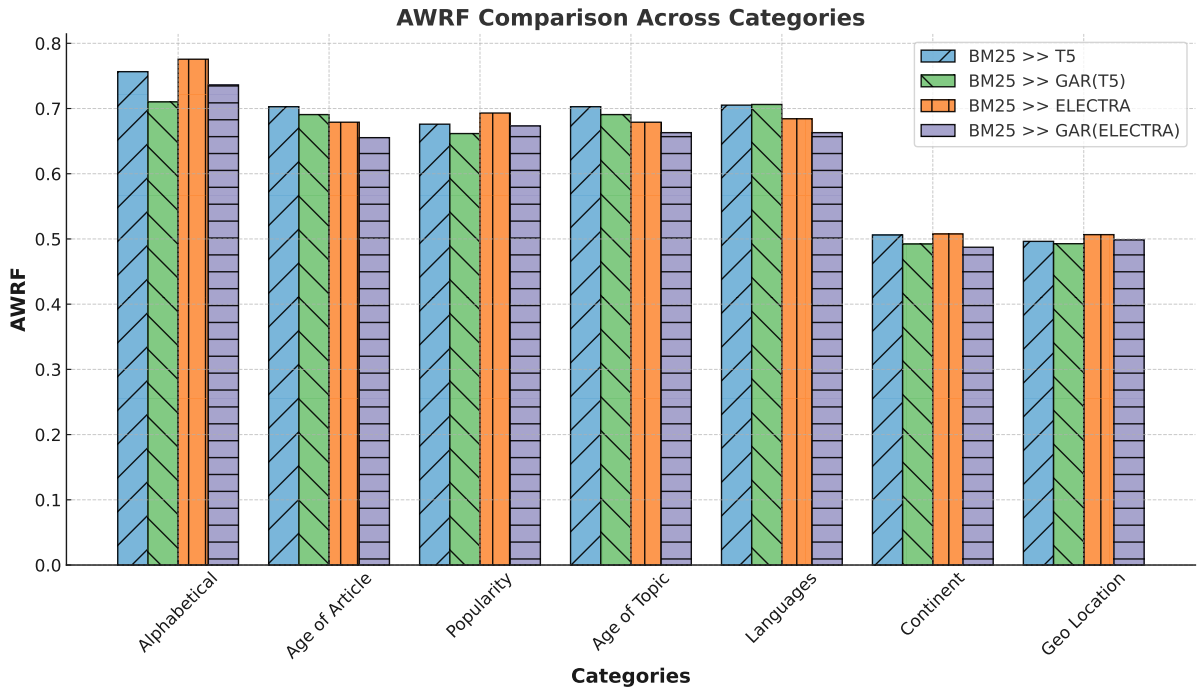


Figure 6.2: The plot shows a comparison of GAR to standard re-ranking pipelines in terms of AWRF across the TREC Fair Ranking Collections. Higher scores indicate a fairer distribution of exposure in the search results.

proving fairness within the standard re-ranking setup. The highest AWRF value across all approaches is achieved by BM25>>ELECTRA with 0.7756 in the "Alphabetical" category. However, all approaches fall short of the ideal AWRF score of 1, underscoring the need for better exposure distribution across groups.

In summary, our findings indicate that GAR, originally designed to optimise relevance, is ineffective in improving fairness in search results under its standard configuration. These results underline the need for adaptations to the GAR framework to address fairness.

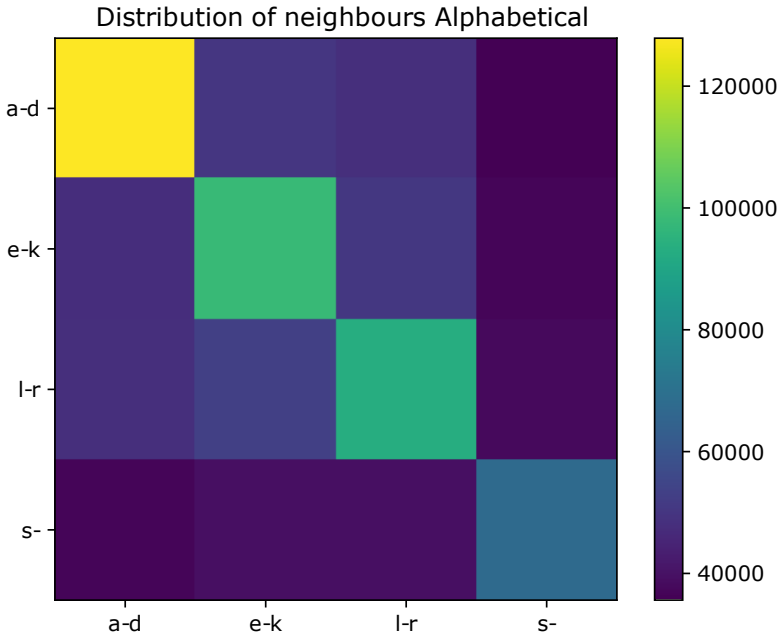
6.3.5 Analysis of Neighbour Distribution Across Groups

To better understand the reduction in fairness observed when applying GAR in the standard setting, we analyse how neighbours are distributed within the corpus graph used by GAR for both TREC track collections. Specifically, we focus on the neighbours of all relevant documents for each query in the graph, using the ground truth relevance labels from the collections. This analysis allows us to determine whether the graph structure amplifies the unfairness present in the initial retrieval.

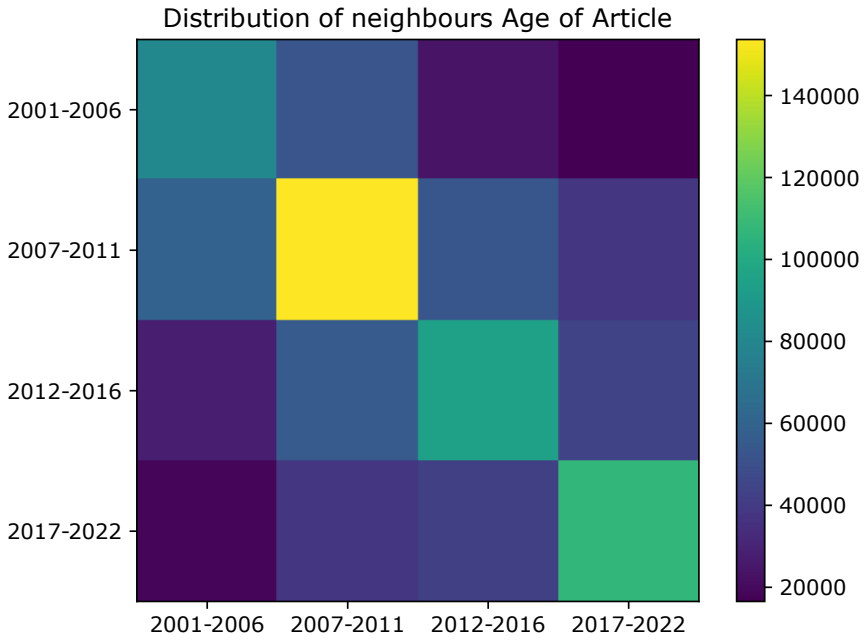
Figures 6.3-6.6 visualise the distribution of neighbours across multiple fairness characteristics and their respective groups. Each subplot represents a fairness characteristic (e.g., "Alphabetical", "Age of Topic", "Geo Category"), and the matrix shows how frequently neighbours from one group are linked to another. A brighter cell in the matrix indicates a higher number of neighbours between the corresponding group pair.

The plots reveal that, for most fairness characteristics, the majority of neighbours for a group are from the group itself. For example, in the "Alphabetical" characteristic (Figure 6.3), most neighbours for documents in the "a-d" group also belong to "a-d", i.e., the same group. This pattern is consistent across other characteristics, such as "Geo Category" and "Popularity". These findings indicate that the graph structure tends to reinforce the group composition of the initial retrieval, leading to a concentration of neighbours within the same group. Moreover, this trend suggests that GAR, in its standard configuration, inherits and amplifies biases present in the initial retrieval. If a query's initial retrieval is skewed toward a particular group, the corpus graph is likely to reinforce this imbalance by identifying additional neighbours from the same group.

To address this issue, we propose adjustments to the re-ranking process to promote a more balanced exposure distribution. In the next section, we introduce six policies aimed at improving fairness by modifying the adaptive re-ranking process. These policies are designed to ensure both relevance and a more fair distribution of exposure across groups.



Alphabetical



Age of Article

Figure 6.3: Each plot shows the number of neighbours between group combinations, with brighter cells indicating higher counts (Part 1-Alphabetical & Age of Article).

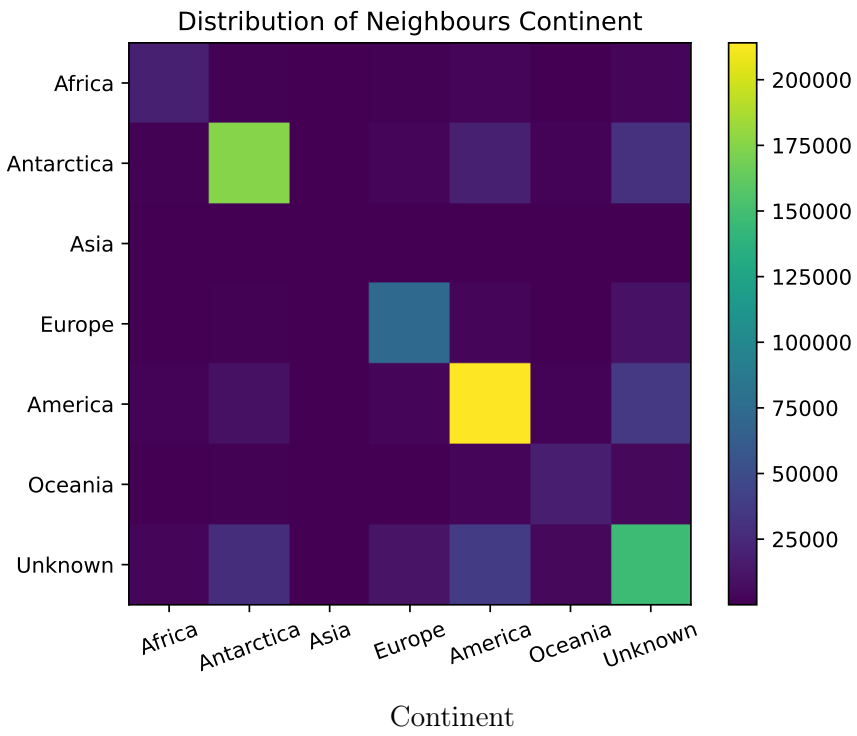
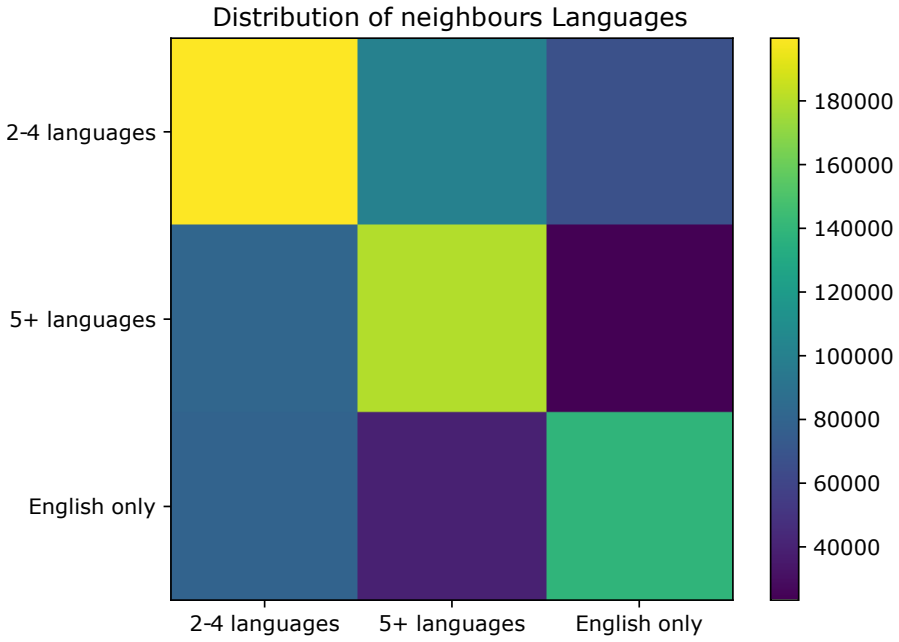
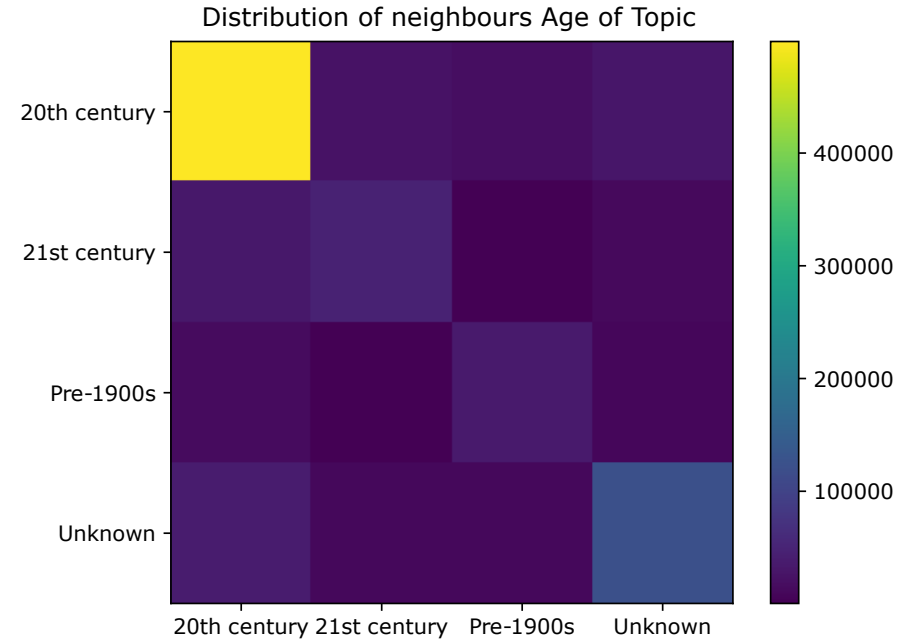


Figure 6.4: The plot shows the number of neighbours between group combinations, with brighter cells indicating higher counts (Part 2-Continent).

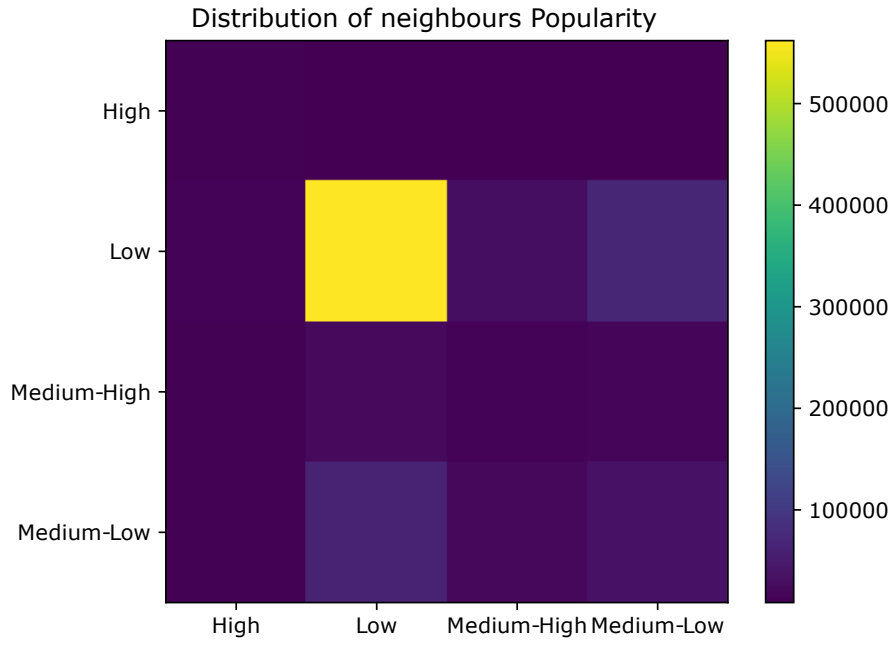


Languages



Age of Topic

Figure 6.5: Each plot shows the number of neighbours between group combinations, with brighter cells indicating higher counts (Part 3-Languages & Age of Topic).



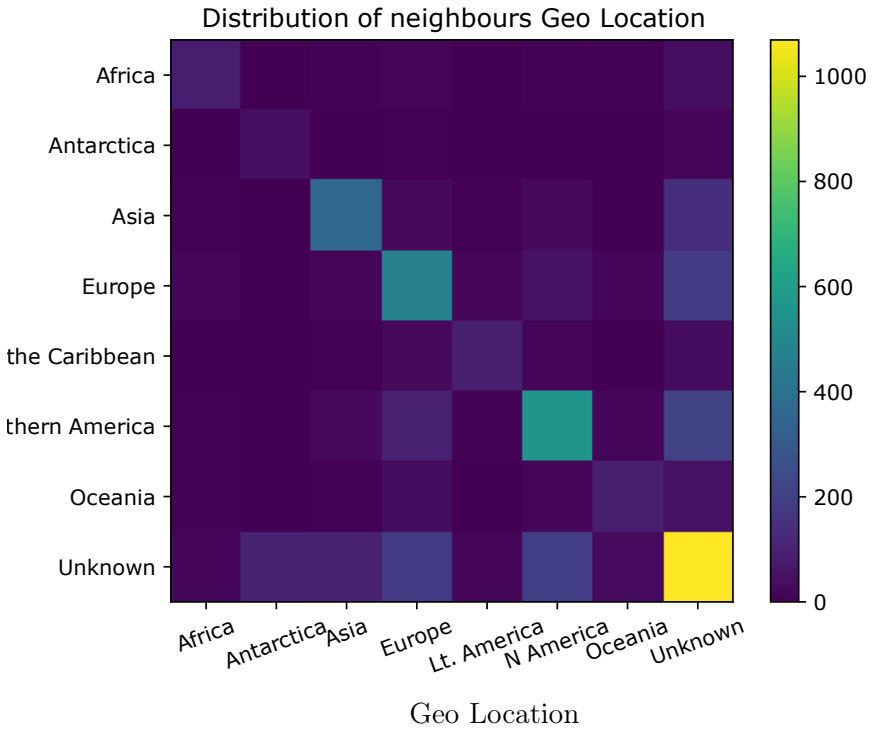


Figure 6.6: Each plot shows the number of neighbours between group combinations, with brighter cells indicating higher counts (Part 4-Popularity & Geo Location).

6.4 Optimising GAR for Fair Exposure Distribution

To address the limitations of the standard GAR process described in Section 6.2, we propose six fairness policies aimed at improving the distribution of exposure across document groups while maintaining relevance. These policies adjust different components of the GAR pipeline to ensure fair exposure across groups.

Figure 6.7 provides an overview of the GAR pipeline, highlighting where each policy is applied. Policies 1 and 2 modify the corpus graph during its construction in the Pre-Retrieval phase (red double-lined box around the Corpus Graph). Policies 3 to 6 are applied dynamically during the adaptive re-ranking process (red double-lined boxes around b_I, b_F, and Z_i). The corpus graph $CG = (VT, ED)$ represents the document collection

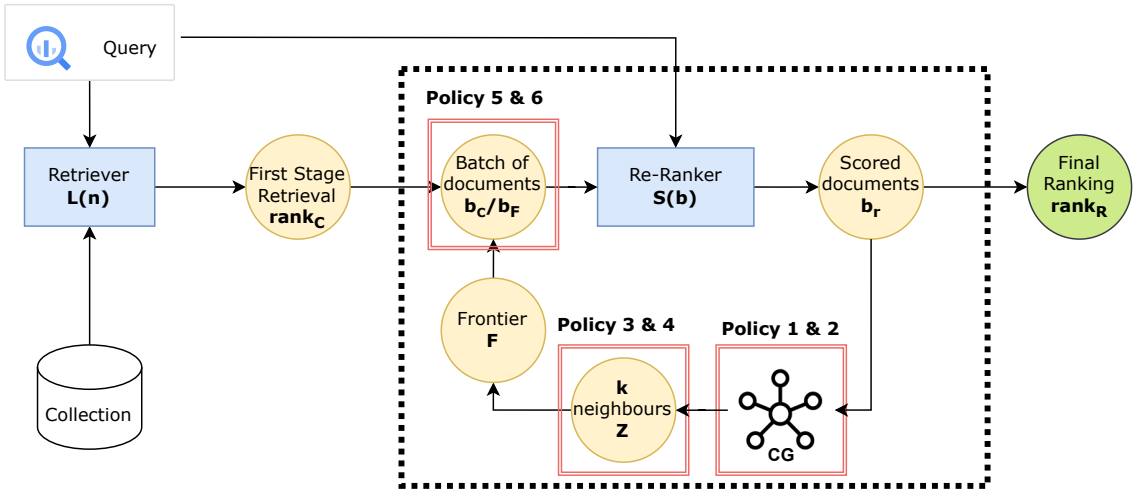


Figure 6.7: The plot shows the standard GAR process with highlighted areas where policies are applied. The Pre-Retrieval Policies 1 and 2 modify the corpus graph (CG) during construction. Policies 3 and 4 influence Z_i, the set of k neighbours, while Policies 5 and 6 adjust batch selection during re-ranking.

as a directed graph, where each vertex in VT corresponds to a document d_i, and each directed edge $(d_i, d_j) \in ED$ signifies that d_j is among the k most similar documents to d_i based on a predefined similarity metric. Typically, these documents are selected solely based on similarity. However, the selection process is modified to incorporate fairness constraints alongside similarity. To achieve this in our pre-retrieval policies, each document d_i is initially associated with a candidate set Υ of size $|\Upsilon|$, where $|\Upsilon| > k$. From this candidate set, a subset Z_i containing the k documents is selected. The larger size of Υ provides flexibility to balance relevance and fairness constraints during graph construction, ensuring that the final selection of documents in Z_i is not only similar to d_i but also aligned with fairness considerations. Our in-process policies operate only over the original set of neighbours.

6.4.1 Policy 1 - Diverse Group Linking

Policy 1 ensures that the set of neighbours Z_i for a document d_i contains only documents from groups other than the group of d_i. This is achieved by excluding documents from the same group, thereby promoting diversity in the selection of related documents.

Figure 6.8 shows how Policy 1 constructs a corpus graph by excluding documents from the same group. For instance, if a document belongs to Group A, only documents from Groups B and C are included in Z_i. If fewer than k documents from other groups are available in Υ, the size of Z_i is reduced to avoid including irrelevant documents.

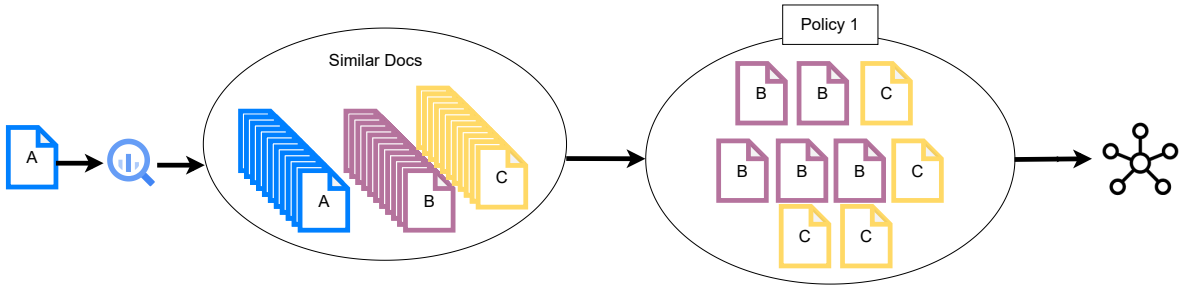


Figure 6.8: The figure shows the selection strategy of Policy 1. Policy 1 constructs a corpus graph ($k = 9$) by only including neighbours from groups other than the source document (e.g., Document from Group A).

6.4.2 Policy 2 - Balanced Group Quota

Policy 2 enforces proportional representation of groups in the set of neighbours Z_i by applying a quota that defines the maximum number of documents each group can contribute. The quota is calculated as $k/|G|$, where k is the total number of neighbours to be selected and $|G|$ is the number of groups. This ensures that each group is allocated an equal share of the neighbour set, up to the defined quota. For example, if $k = 9$ and there are three groups, the quota allows each group to contribute up to three documents.

If some groups lack sufficient candidates in Υ, the set of most similar documents, the policy redistributes the unfulfilled quota among the other groups. Redistribution ensures that Z_i remains balanced by reallocating unused slots to groups with additional candidates. This mechanism prevents dominant groups from occupying all available slots while also maximising the use of the available candidate pool. By enforcing these quotas, Policy 2 ensures that Z_i contains a fair distribution of groups, intending to enhance exposure fairness in the final ranking. Figure 6.9 demonstrates this process, showing how the policy balances group representation during the graph construction phase.

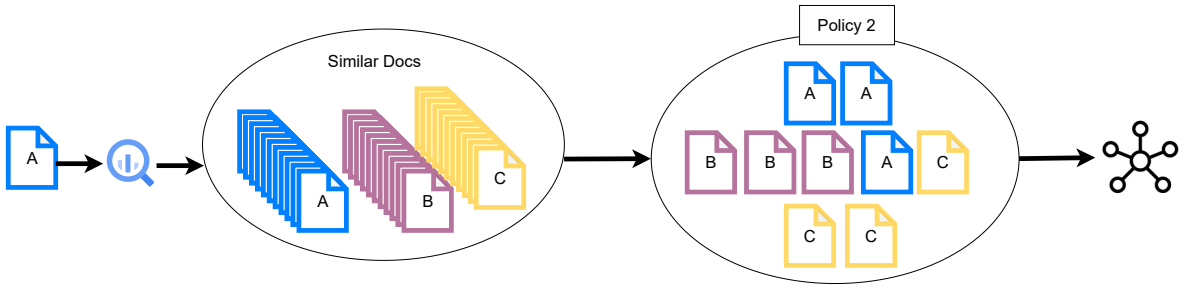


Figure 6.9: The figure shows the selection strategy of Policy 2. Policy 2 enforces equal proportions in the set of neighbours when building the corpus graph.

6.4.3 Policy 3 - Removing Documents from the Same Group

Policy 3 modifies Z_i dynamically during re-ranking. After identifying the k most similar documents for a source document d_i, all documents belonging to the same group as d_i are removed. This ensures that only documents from other groups are added to the frontier (F). Figure 6.10 illustrates how Policy 3 excludes documents from the same group. For instance, if Z_i contains five documents from Group A, two from Group B, and two from Group C, only the documents from Groups B and C are added to F.

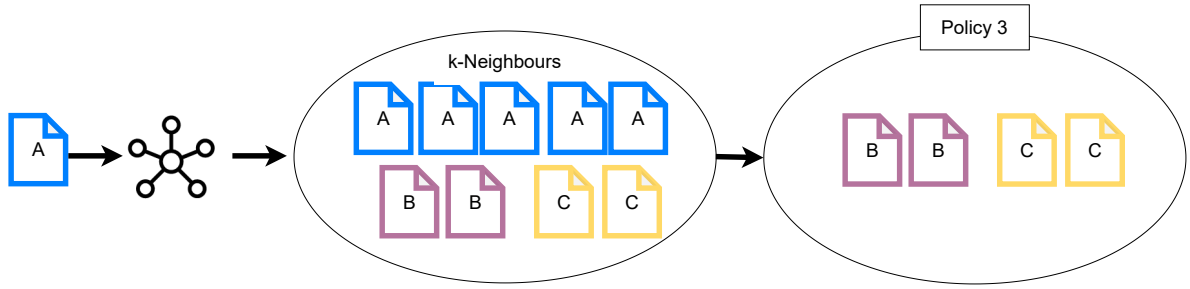


Figure 6.10: The figure shows the selection strategy of Policy 3. Policy 3 dynamically removes neighbours from the same group during re-ranking to increase diversity in the frontier.

6.4.4 Policy 4 - Equal Proportions in the Selected Documents

Policy 4 enforces equal representation of groups in Z_i by selecting a fixed number of documents from each group, ensuring that every group has a fair opportunity for exposure in the ranking process. For example, if $k = 9$ and there are three groups, the ideal distribution would allocate three documents to each group. However, when some groups lack sufficient candidates to meet this quota, the policy adjusts the selection proportionally across the remaining groups. This proportional adjustment means that the total number of selected documents is recalculated based on the group with the fewest available candidates.

For instance, if one group can only provide two documents instead of three, the total number of documents selected for Z_i is reduced from nine to six (i.e., 2×3, as each group is now limited to two documents). This ensures that the relative representation of groups remains balanced, even when certain groups have fewer candidates.

By maintaining this proportional distribution, Policy 4 avoids over-representation of larger or more dominant groups. Figure 6.11 illustrates this process, demonstrating how the policy dynamically adjusts the selection to balance group proportions while adhering to the constraints of the available candidates.

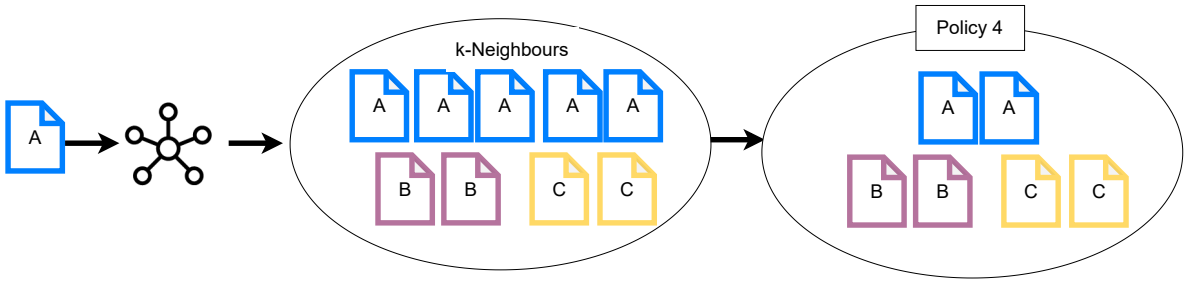


Figure 6.11: The figure shows the selection strategy of Policy 4. Policy 4 enforces equal proportions in neighbours ensuring a balanced representation of groups.

6.4.5 Policy 5 - Group-Based Selection of Top Documents

Policy 5 modifies the batch selection process by ensuring that high-scoring documents from every group are included in b_I or b_F. This prevents dominance by a single group during re-ranking. Figure 6.12 illustrates where Policy 5 selects high-scoring documents from each group, ensuring a balanced distribution of relevance scores in the frontier.

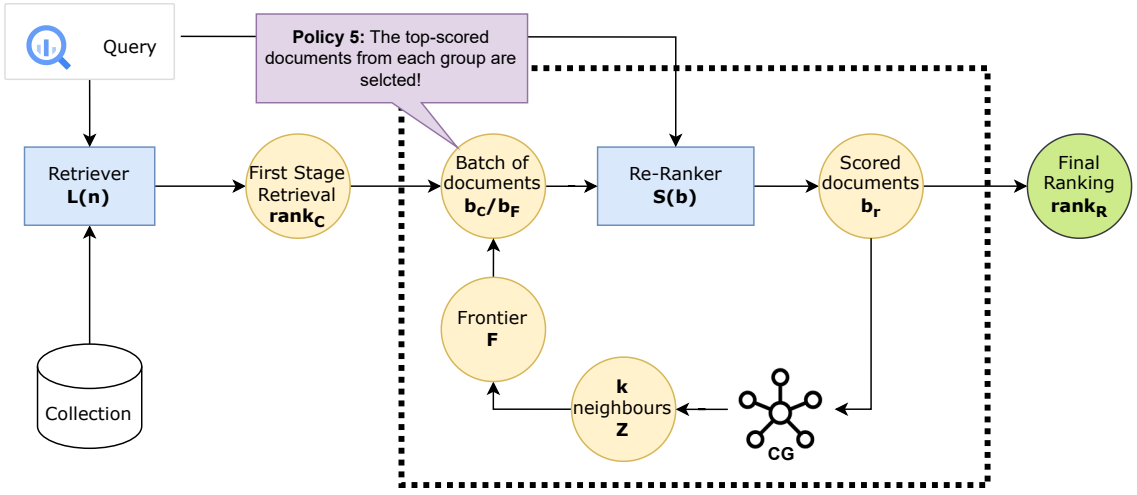


Figure 6.12: The figure shows the selection strategy of Policy 5. Policy 5 intervenes in the batch selection by ensuring high-scoring documents are equally distributed among groups.

6.4.6 Policy 6 - Prioritising Documents from Early Iterations

We extend Policy 5 by modifying the document selection strategy from the frontier b_F. Instead of selecting the highest-scoring documents from each group in the frontier, we follow the order in which the set of neighbours Z_i was added to the frontier. During selection, we iterate through each Z_i in F and choose the highest-scoring document from each group until the predefined number of documents per group is reached. This approach prioritises the neighbours of documents scored in the early iterations of the process. For selection from the initial retrieval $rank_C$, we continue to select the highest-scoring documents from each group, as in Policy 5.

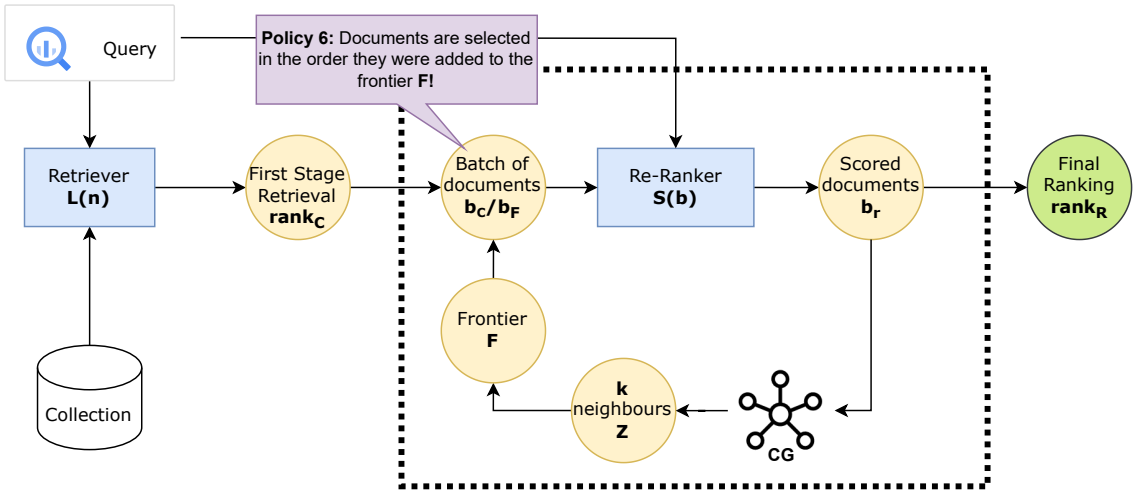


Figure 6.13: The figure shows the selection strategy of Policy 6. Policy 6 prioritises neighbours from earlier iterations in the re-ranking process to ensure fair consideration of documents over time.

The six proposed policies address limitations in the standard GAR process by introducing fairness constraints during graph construction (Pre-Retrieval Policies) and by dynamically adjusting the document selection process (In-Process Policies). In the following section, we present an evaluation of these policies to determine their impact on fairness in exposure distribution across groups.

6.5 Experimental Evaluation of the Proposed Policies

To evaluate our proposed policies, we use the experimental setup described in Section 6.3. The evaluation compares our policies against the standard re-ranking baselines and the standard GAR process. Performance is assessed using two metrics:

1. **Fairness**: Measured using the AWRF, where higher values indicate a more equal exposure distribution among document groups.

2. **Relevance**: Measured using nDCG (Järvelin & Kekäläinen, 2002), which quantifies the utility of the search results

We answer the following research questions:

RQ6.2: Can our proposed policies for adaptive re-ranking distribute exposure more fairly than a standard re-ranking pipeline?

RQ6.3: How does improving fairness affect the relevance of the search results?

RQ6.4: How do pre-retrieval policies (which create a Corpus Graph) compare to in-process policies (which modify re-ranking during retrieval) in terms of fairness and relevance?

6.5.1 Statistical Testing

To determine whether the observed differences between our policies and the baselines are statistically significant, we use the following tests:

- A paired t-test (Student, 1908) is used to compare the AWRF scores of our policies against the baselines. We apply a significance threshold of $p < 0.05$ and use Bonferroni corrections (Dunn, 1961) to account for multiple comparisons.
- To test whether the nDCG scores of our policies are statistically equivalent to those of the baselines, we use the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987). The TOST procedure evaluates whether the differences in nDCG scores fall within a predefined equivalence margin of 0.1, with a significance level of $\alpha = 0.05$.

We use the paired t-test to provide evidence of improvements in fairness (AWRF), while the TOST procedure ensures that relevance nDCG is not compromised by the application of our policies. We start our analysis by answering RQ6.2.

6.5.2 Impact on Fairness

To address RQ6.2, we investigate whether the proposed policies can distribute exposure more equitably than the standard re-ranking pipelines and the GAR process. Table 6.1 presents the AWRF scores for each fairness characteristic. The table is structured to show the results for the proposed policies at the top, with pre-retrieval policies (Policy 1 and Policy 2) followed by in-process policies (Policies 3–6). Baseline results are shown at the bottom of the table. Each policy is evaluated against the corresponding baseline using the same re-ranking model (e.g., ELECTRA-based policies are compared to ELECTRA-based baselines). In the table, statistically significant differences from the standard re-ranking baseline are marked with an asterisk (*), while differences from GAR are denoted by (†).

Table 6.1: Comparison of our policies regarding fairness (AWRF) to the baselines over the categories. We use * to indicate a statistically significant difference to the standard re-ranking baseline and † to denote statistically significant differences to the GAR baseline.

	Approaches	Alphabetical	Age of Article	Popularity	Age of Topic	Languages	Continent	Geo Location (TREC 21)
Pre-Retr.	P1-GAR(T5)	0.7468†	0.7150†	0.6943†	0.5856†*	0.7310†*	**0.5150†**	0.5113
	P1-GAR(ELECTRA)	0.7752†	0.6867†	0.7122†	0.5824†*	0.7023†	0.5128†*	0.5339
	P2-GAR(T5)	0.7535†	0.7269†	0.7038†*	0.5968†*	0.7438†*	0.4920*	0.4965
	P2-GAR(ELECTRA)	0.7835†	0.7100†	**0.7411†***	0.6002†*	0.7105†	0.4920	0.5103
In-Proc.	P3-GAR(T5)	0.7417†	0.7158†	0.6691	0.5821†*	0.7340†*	0.5040	0.4962
	P3-GAR(ELECTRA)	0.7717†	0.6844†	0.7088†	0.5748†*	0.7105†	0.5089†	0.5026
	P4-GAR(T5)	0.7497†	0.7226†	0.7056†*	0.6000†*	0.7477†*	0.5075	**0.5574†***
	P4-GAR(ELECTRA)	0.7785†	0.7084†*	0.7368†*	**0.7084†***	0.7235†*	0.5086†	0.5388†*
	P5-GAR(T5)	0.7458†	0.7312†*	0.7261†*	0.6041†*	**0.7626†***	0.4903	0.5001
	P5-GAR(ELECTRA)	0.7810†	0.7249†*	0.7400†*	0.6088†*	0.7286†*	0.4878*	0.5137†
	P6-GAR(T5)	0.7739†*	0.7415†*	0.7206†*	0.5987†*	0.7566†*	0.4905	0.4996
	P6-GAR(ELECTRA)	**0.7983†***	**0.7418†***	0.7407†*	0.6046†*	0.7286†*	0.4882	0.5095
Base	BM25>>T5	0.7565	0.7028	0.6760	0.5622	0.7052	0.5061	0.4963
	BM25>>GAR(T5)	0.7102	0.6907	0.6616	0.5589	0.7062	0.4923	0.4927
	BM25>>ELECTRA	0.7756	0.6790	0.6931	0.5478	0.6843	0.5077	0.5066
	BM25>>GAR(ELECTRA)	0.7361	0.6553	0.6733	0.5402	0.6630	0.4873	0.4985

Performance of Pre-Retrieval Policies

Policy 1 deployed with ELECTRA demonstrates consistent improvement in AWRF over all categories. In the "Age of Topic" and the "Languages" category we observe statistically significant differences to the standard re-ranking and the GAR baseline. Policy 2 deployed with ELECTRA outperforms the GAR baseline in 5 out of 7 categories with statistically significant differences. In the "Popularity" category, Policy 2 paired with ELECTRA reaches an AWRF score of 0.7411, outperforming both baselines and achieving the highest score out of all policies. However, both Policy 1 and Policy 2, when deployed with ELECTRA show only small improvements in the "Geographic Location" category. This suggests that the effectiveness of pre-retrieval policies can vary depending on the structural characteristics of specific categories. Analysing the performance of the policies when monoT5 is deployed we can observe a similar trend. Policy 1 shows statistically significant differences to the GAR pipeline improving fairness in the majority of categories. Moreover, in the categories "Age of Topic" and "Languages" Policy 1 shows statistically significant differences to the standard re-ranking pipeline as well. Notably, in the "Continent" category Policy 1 is able to outperform all other approaches including our policies, achieving the highest overall AWRF score of 0.5150 in this category. Policy 2 with monoT5 shows improvements with significant differences to both baselines in three categories, namely "Popularity", "Age of Topic" and "Languages".

Performance of In-Process Policies

Analysing the results of our in-process policies, it is notable, that all of our policies, i.e., Policies 3–6 enhance fairness in the search results across the majority of categories. In particular, it is notable, that in the "Alphabetical" all policies are able to outperform the GAR baselines with statistically significant differences, independent of the underlying ranking model. Policy 6 shows statistically significant differences to the standard re-ranking baselines as well. Analysing the results in the "Age of Article" category we can observe that both versions of Policy 3 and Policy 4 deployed with monoT5 show improvements over the GAR baselines, with statistically significant differences. All other *in-process* policies show improvements over both baselines, also with statistically significant differences. In the "Popularity" category, only Policy 3 is not able to improve over both baselines. Moreover, in "Age of Topic" all policies show improvements in AWRF with statistically significant differences to both baselines. A similar trend can be observed in the "Languages" category. The only outlier is Policy 3 deployed with ELECTRA which is not able to outperform the standard re-ranking pipeline. It is notable, that in the "Continent" category only marginal improvements can be observed over all the policies. Similarly, in the "Geographic Location" category only Policy 4 achieves improvements with statistically significant differences to the GAR and the re-ranking baseline.

From the different strategies, Policy 4 stands out for its strong performance in the "Geo Location" and "Age of Topic" categories. In particular, in the "Age of Topic" category Policy 4 achieves the highest score out of all policies (0.7084) outperforming the baseline score of 0.5478. The differences to both baselines are statistically significant. Policy 5 achieves the highest score in the "Languages" category, showing improvements of up to 8% over the baselines. However, Policy 6 emerges as the most effective policy overall, particularly when deployed with ELECTRA. It achieves significant improvements across 5 out of 7 categories to both baselines and records the highest overall AWRF score (0.7983) in the "Alphabetical" category. This demonstrates its robustness in addressing fairness over different categories of groups. Furthermore, Policy 6 achieves a notable improvement in the "Age of Article" category, increasing the AWRF from 0.6553 (GAR baseline) to 0.7418, representing a 13% improvement. Overall, it is notable, that all of our policies can be successfully deployed to increase the fairness of the search results.

Table 6.2: nDCG Results Comparison for Policies and Baselines Across Categories. We use ‡ to indicate equivalence to the GAR baselines, and • to indicate equivalence to the standard pipeline for the nDCG values. Bold values indicate the best results.

	Approaches	Alphabetical	Age of Article	Popularity	Age of Topic	Languages	Continent	Geo Location (TREC 21)
Pre-Retr.	P1-GAR(T5)	0.5763•‡	0.5816•‡	0.6006•‡	0.5801•‡	0.5983•‡	0.5768•‡	0.3937•‡
	P1-GAR(ELECTRA)	0.5985•‡	0.5759•‡	0.5914•‡	0.5688•	0.5872•‡	0.5776•‡	0.3440•‡
	P2-GAR(T5)	0.5802•‡	0.5864•‡	0.5944•‡	0.5930•‡	0.5813•‡	0.5875•‡	0.3915•‡
	P2-GAR(ELECTRA)	0.5826•‡	0.5867•‡	0.5832•‡	0.5741•‡	0.5920•‡	0.5875•‡	0.3426•‡
In-Proc.	P3-GAR(T5)	0.5797‡	0.5803•‡	0.5856•‡	0.5856•‡	0.5983•‡	0.5854•‡	0.3849•‡
	P3-GAR(ELECTRA)	0.6029•‡	0.5867•‡	0.5954•‡	0.5776•‡	0.5865•‡	0.5833•‡	0.3274•‡
	P4-GAR(T5)	0.5748•‡	0.5691‡	0.6011•‡	0.5909•‡	0.6117•‡	0.5922•‡	**0.3933•‡**
	P4-GAR(ELECTRA)	0.5841•‡	0.5717•‡	0.5717•‡	0.5717•‡	0.6105•‡	0.5873•‡	0.3346•‡
	P5-GAR(T5)	0.5700‡	0.5629‡	0.5974•‡	0.5707•‡	0.5682‡	0.5730•‡	0.3816•‡
	P5-GAR(ELECTRA)	0.5892•‡	0.5848•‡	0.5810•‡	0.5790•‡	0.5901•‡	0.5821•‡	0.3347•‡
	P6-GAR(T5)	**0.6206•‡**	**0.6074•‡**	**0.6308•**	**0.6022•‡**	**0.6141•‡**	0.5716•‡	0.3836•‡
	P6-GAR(ELECTRA)	0.6111•‡	0.5938•‡	0.6075•‡	0.5743•‡	0.6000•‡	0.5827•‡	0.3331•‡
Base	BM25>>T5	0.5956	0.5956	0.5956	0.5956	0.5956	**0.5956**	0.3900
	BM25>>GAR(T5)	0.5774	0.5774	0.5774	0.5774	0.5774	0.5774	0.3801
	BM25>>ELECTRA	0.5795	0.5795	0.5795	0.5795	0.5795	0.5795	0.3316
	BM25>>GAR(ELECTRA)	0.5825	0.5825	0.5825	0.5825	0.5825	0.5825	0.3299

6.5.3 Impact of the Policies on Relevance

To address RQ6.3, we examine the extent to which the proposed policies affect the relevance of search results. Table 6.2 summarises the nDCG scores for each policy across the different fairness characteristics, comparing these scores against the baseline approaches. Statistical equivalence is denoted by ‡ for GAR baselines and • for standard re-ranking baselines. Higher nDCG values indicate improved relevance.

In the "Alphabetical" category, the majority of policies (10 out of 12) achieve nDCG scores statistically equivalent to all the baselines. However, Policy 3 and Policy 5 when deployed with monoT5 show statistical equivalence to the GAR baseline only. Notably, Policy 6 paired with monoT5 outperforms the standard re-ranking baseline, achieving an nDCG score of 0.6206 compared to the baseline value of 0.5956.

For the "Age of Article" category, 10 out of 12 policies maintain equivalence to all the baselines. Policy 6 with monoT5 again demonstrates strong performance with an nDCG score of 0.6074, exceeding the baseline score of 0.5956. Policies 4 and 5 with monoT5, however, remain statistically equivalent only to the GAR baselines. Moreover, in the "Popularity" category, 11 out of 12 policies achieve statistical equivalence to the baselines. It is notable, that Policy 6 with monoT5 achieves a nDCG score of 0.6308, substantially higher than the baseline score of 0.5956. Furthermore, when analysing the results for

the "Age of Topic" category, it is notable, that 11 out of 12 policies achieve statistical equivalence to re-ranking and GAR baselines. Policy 1 shows equivalence to the standard re-ranking pipeline only. Moreover, Policy 6 deployed with monoT5, again emerges as the most effective strategy. In the "Languages" category, all policies except Policy 5 deployed with monoT5 achieve equivalence to both baselines. However, many policies demonstrate higher nDCG scores than the baselines. For instance, both versions of Policy 4 achieve an nDCG of approximately 0.6100, exceeding the baseline score of 0.5956. These results suggest that our policies can maintain and sometimes even improve the relevance.

The "Continent" category results show that BM25>>T5 achieves the highest nDCG score with 0.5956. However, all of our policies achieve nDCG scores statistically equivalent to the baselines. This result indicates that the policies do not adversely impact relevance for this category, even if significant improvements are not observed.

For the "Geographic Location" category, all policies maintain statistical equivalence to the baselines. From our analysis it became apparent, that Policy 6 consistently achieves the highest relevance scores, particularly when paired with monoT5, demonstrating its effectiveness in enhancing relevance across categories such as "Alphabetical", "Age of Article", and "Popularity". In summary, our proposed fairness-enhancing policies largely preserve the relevance of the search results, with Policy 6 emerging as the most effective strategy. The policies' ability to achieve equivalence with baseline relevance scores underscores their potential for integrating fairness into IR systems without compromising utility.

6.5.4 Comparison of In-Process and Pre-Retrieval policies

To answer RQ6.4, we evaluate the performance of the *pre-retrieval* policies (Policies 1 & 2) and the *in-process* policies (Policies 3–6) to understand their effectiveness across different categories. While both sets of policies aim to improve fairness, the *pre-retrieval* policies modify the corpus graph before retrieval, whereas the *in-process* policies intervene during the adaptive re-ranking process. Conceptually, Policies 1 and 2 are related to Policies 3 and 4, as both enforce similar constraints on neighbour selection. However, *in-process* policies operate on a smaller set of neighbours and dynamically adjust fairness interventions during retrieval. The results show that both policy types perform similarly across most categories. For instance, in the "Languages" category, Policy 1 with T5 achieves an AWRF score of 0.7310, closely matching Policy 3's score of 0.7340. A similar trend is observed in the "Age of Topic", "Alphabetical", and "Age of Article" categories, indicating that both *pre-retrieval* and *in-process* policies can achieve comparable fairness improvements. However, differences emerge when comparing Policies 1 & 2 with Policies 5 & 6. Policy 5 outperforms the *pre-retrieval* policies in "Age of Topic" and "Languages", while Policy 6 achieves the highest fairness scores in "Alphabetical" and "Age of Article". In contrast, Policies 1 and

2 perform better than Policies 5 and 6 in the "Popularity" and "Continent" categories, highlighting that the category influences policy effectiveness. Overall, it is notable, that both types of policies can be successfully deployed to improve the fairness in the search results while maintaining the utility of a ranking.

6.6 Conclusions

In this chapter, we explored the impact of GAR on the fair distribution of exposure in rankings. We have shown that while the standard GAR approach enhances retrieval effectiveness, it can inadvertently exacerbate exposure disparities across groups. To address this, we introduced a set of six fairness-enhancing policies that operate either before retrieval or during re-ranking. Our evaluation showed that all proposed policies significantly improve fairness as measured by AWRF, with Policy 6 standing out as the most effective achieving up to 13% improvement. Importantly, these gains in fairness did not come at the expense of relevance. Most policies maintained and in several cases improved, nDCG scores relative to baseline methods. This confirms that fairness-aware adaptations can preserve, and in some cases even enhance, the relevance of the search results. In particular, Policy 6 paired with monoT5 yielded improved relevance in the majority of evaluated categories, supporting the broader claim that re-ranking pipelines can be effectively tailored to promote both fairness and relevance. Comparing the two types of policy, we found that in-process interventions generally yielded stronger trade-offs between fairness and relevance, with Policy 6 again emerging as the most robust policy. Overall, our findings validate our claim that re-ranking strategies can be adapted to more equitably allocate exposure in rankings, without decreasing relevance. Moreover, with our investigation we have shown, that state-of-the art ranking systems originally designed to increase recall can be successfully tailored towards fairness goals. However, we have also revealed that components originally designed to increase retrieval effectiveness such as a corpus graph can inadvertently introduce unfairness and even amplify it. This highlights the need for the explicit fairness solutions which we present over the technical chapters in this thesis.

Chapter 7

Fair Query Expansion & Generation

In the previous chapters, we have outlined that a low recall of underrepresented groups in the first-stage retrieval can lead to an unfair exposure distribution over groups of documents in the search results (c.f. Chapters 4 & 5). In Chapter 6, we have presented an approach that mitigates this unfairness by increasing recall after the first-stage retrieval using graph-based adaptive re-ranking. The adaptive re-ranking approach uses the set of documents from the first-stage retrieval to discover additional relevant documents from underrepresented groups using the graph component, which effectively enhances recall from disadvantaged groups while maintaining the utility of the system. As postulated in the thesis statement (c.f. Chapter 1, Section 1.2) retrieval methods aimed at increasing recall effectiveness of the first-stage retrieval can also help improve fairness. As outlined in Chapter 2 (Section 2.3), there are several approaches to increase a standard retrieval's recall by modifying and resubmitting the originally issued query, for example, PRF. Therefore, in this chapter, we examine how modifying queries can help identify relevant documents from underrepresented groups that traditional retrieval methods may not retrieve, thereby also improving the search results fairness. The chapter is outlined as follows:

- First, in Section 7.1 , we analyse how adapting PRF methods in dense retrieval can improve the fairness of exposure in a ranking.
- In Section 7.2, we extend the traditional PRF concept by using pre-trained large language models to generate query expansion terms to enhance the fairness of the search results.
- In Section 7.3, we explore the potential of generation techniques to automatically create queries that identify more relevant documents from an underrepresented group.
- Finally, in Section 7.4 we conclude and summarise the chapter.

7.1 PRF for Fair Dense Retrieval

As outlined in Chapter 2, Section 2.3, extending a user's initial query with additional informative terms, i.e., query expansion, has been shown to be a useful mechanism for improving the effectiveness of search systems. Query expansion refers to the process of reformulating a user's query by adding additional terms or embeddings in order to retrieve more relevant documents. It is typically used to increase recall and reduce vocabulary mismatch in retrieval. Many approaches from the literature, for example (Yan et al., 2003; Keikha et al., 2018; Xu et al., 2009; Lv & Zhai, 2010), use PRF for identifying such informative terms and expanding the user's query. Essentially, PRF approaches expand the initial query by appending it with the terms that are expected to be the most informative from the top-ranked documents (often referred to as the pseudo-relevant set) in an initial ranking (c.f. Section 2.3). Recently, novel PRF approaches for dense retrieval, such as ColBERT-PRF (X. Wang et al., 2021), have been shown to be successful in improving the relevance of search results (c.f. Section 2.5). However, there has been little work investigating the relationship between PRF in dense retrieval and fairness in ranking. Indeed, using representative and discriminative embeddings from an initial retrieval to revise a query can, in principle, allow the optimisation of fairness in the search results, supporting the postulated thesis statement (c.f. Section 1.2). To validate our statement-that retrieval methods modifying the query can be effectively adapted to ensure both fair and relevant search results-we investigate the effects of PRF in dense retrieval on fairness. In this section, we specifically examine how PRF in dense retrieval impacts the exposure that different groups of documents receive in the search results. For example, we analyse whether documents relating to different geographic locations can be fairly exposed to users when PRF is applied. We propose a novel fair PRF mechanism for multiple representation dense retrieval, named ColBERT-FairPRF, that is effective for enhancing the fairness of the distribution of exposure over groups of documents in a ranking. We focus on multiple representation models (c.f. Section 2.5), which represent each document as a set of embeddings rather than a single vector. The rest of the section is structured as follows:

- In Section 7.1.1 we present the existing related work on fairness and PRF.
- In Section 7.1.2 we present a novel PRF approach to ensure a fair exposure allocation in the search results.
- In Section 7.1.3 we present the experimental setup we use to evaluate our approach.
- In Section 7.1.4 we discuss the results of our experiments.
- Finally, in Section 7.1.5 we summarise and conclude the section.

7.1.1 Related Work

PRF has been shown to be effective in increasing the retrieval performance in terms of relevance in many previous works, for example (Yan et al., 2003; Keikha et al., 2018; Xu et al., 2009; Lv & Zhai, 2010). Traditional PRF models, such as the RM3 relevance language model (Jaleel et al., 2004), or the Bo1 model from the Divergence from Randomness framework (Amati & van Rijsbergen, 2002) use statistical information about the distributions of terms in the pseudo-relevant set and the larger collection of documents (that the pseudo-relevant set is retrieved from) for selecting informative terms (c.f. Chapter 2, Section 2.3). However, with the emergence of large pre-trained language models such as BERT (Devlin et al., 2019), contextualised embeddings, known as *dense* representations, have become widely used for representing queries and documents (c.f. Chapter 2, Section 2.4). As such, the emergence of BERT-like retrieval approaches has sparked the development of new PRF methods for dense retrieval. In particular, ColBERT-PRF (X. Wang et al., 2021; X. Wang, Macdonald, Tonellotto & Ounis, 2023; X. Wang et al., 2022; S. L. Wang et al., 2020), which builds upon ColBERT (Khattab & Zaharia, 2020), has been shown to be effective in improving the retrieval performance of the original model. ColBERT-PRF enhances retrieval by expanding the user's initial query with additional contextual embeddings from a pseudo-relevant set of documents.

Aligned with our thesis statement (c.f. Section 1.2)-that it is possible to enhance the fair distribution of exposure over groups of documents in a ranking without significantly decreasing the relevance of the search results-we investigate in this section how multi-representation dense retrieval methods based on ColBERT-PRF can be adapted to create search results that are both relevant to the query and provide fair exposure to multiple *groups of documents*. We focus on multiple representation models (c.f. Section 2.5.2), which represent each document as a set of embeddings rather than a single vector.

There is little previous work that investigated the relationship between fairness and PRF. Bashir and Rauber (2009) introduced a cluster-based PRF mechanism for decreasing bias in the search results set by increasing the findability of documents. Related, Wilkie and Azzopardi (2014) investigated how revising a user's query can impact the retrievability of documents. Bigdeli et al. (2021) introduced a bias-aware PRF framework for selecting neutral (i.e., non-biased) documents to form the pseudo-relevant set of documents that the informative terms are selected from. Different from the works of Bashir and Rauber (2009), Wilkie and Azzopardi (2014) and Bigdeli et al. (2021) that mainly focused on the investigation and mitigation of bias with PRF, our focus is on providing groups of documents a fair exposure to the user by using PRF. This aligns with the fairness definition introduced in Section 3.2.2. Next, we introduce our novel PRF approach.

7.1.2 ColBERT-FairPRF

We extend the PRF mechanism of ColBERT-PRF (X. Wang et al., 2021) to provide fair exposure in search results for multiple groups of documents. In ColBERT (Khattab & Zaharia, 2020) and ColBERT-PRF (X. Wang et al., 2021), scoring documents with respect to a query q involves finding the nearest document token embeddings for each query token embedding using an approximate nearest neighbour search. The token similarities are then aggregated to calculate the overall similarity between a document and the query. ColBERT-PRF enhances this process by selecting useful embeddings from a pseudo-relevant set and appending them to the initial query representation.

To provide fair exposure to documents from a set of fairness groups G, our ColBERT-FairPRF approach modifies the PRF mechanism by selecting feedback embeddings from the highest-ranked documents of *each* individual group in the pseudo-relevant set. By selecting embeddings from each group individually, we ensure that all groups we aim to be fair to are represented in the expanded query. ColBERT-FairPRF aims to generate a fair ranking rank_R where each group $g \in G$ receives a fair exposure. To achieve this, we follow several steps: First, for each group $g \in G$, we select the top k documents belonging to the group g from an initial ranking rank_C to form the pseudo-relevant set $D_{\text{top } k_g}$. Next, following the method described in ColBERT-PRF (X. Wang et al., 2021), we apply a simple KMeans clustering to the embeddings of the documents in each $D_{\text{top } k_g}$ to obtain representative centroid embeddings for each group g. For each centroid, we identify the closest term embedding from $D_{\text{top } k_g}$. These term embeddings are expected to be the most useful for representing group g. We then append the selected term embeddings from all groups to the original query q to form the expanded query q_e :

$$q_e = q \cup \bigcup_{g \in G} XT_{e,g} \tag{7.1}$$

In the equation $XT_{e,g}$ denotes the expansion embeddings for group g. The expanded query q_e is then used to generate a refined ranking of documents rank_R, following the methodology of ColBERT-PRF (X. Wang et al., 2021), where documents are scored based on their similarity to the expanded query. The relevance score $s(q_e, d_i)$ for a document d_i given the expanded query q_e is defined as:

$$s(q_e, d_i) = \sum_{i=1}^{|q|} \max_{j=1}^{|d_i|} \omega_{q_i}^{\top} \omega_{d_{ij}} + \beta \sum_{g \in G} \sum_{e_{g,m} \in XT_{e,g}} \max_{j=1}^{|d_i|} \omega_{x_{g,m}}^{\top} \omega_{d_{ij}} \tag{7.2}$$

In Equation (7.2), $|q|$ is the number of embeddings in the original query q, and $|d_i|$ is the number of embeddings in document d_i. The term ω_{q_i} represents the embedding for the i-th token in the query q, and $\omega_{d_{ij}}$ represents the embedding for the j-th token in document d_i. The first term in the equation:

$$\sum_{i=1}^{|q|} \max_{j=1}^{|d_i|} \omega_{q_i}^{\top} \omega_{d_{ij}}, \tag{7.3}$$

captures the relevance between the original query and the document by summing the maximum similarities between each query token embedding ω_{q_i} and the document token embeddings $\omega_{d_{ij}}$. The second term:

$$\beta \sum_{g \in G} \sum_{e_{g,m} \in XT_{e,g}} \max_{j=1}^{|d_i|} \omega_{x_{g,m}}^{\top} \omega_{d_{ij}}, \tag{7.4}$$

integrates fairness-aware expansion by adding the weighted sum of the maximum similarities between expansion embeddings from each group g and the document token embeddings. Here, $\omega_{x_{g,m}}$ is the m-th expansion embedding from group g, and β controls the contribution of these expansion embeddings to the final score. By incorporating expansion embeddings from each group $g \in G$, we improve the likelihood that all groups are fairly represented in the search results. The parameter β has a range of $1 > \beta > 0$ and allows us to adjust the influence of the expansion embeddings on the final score, balancing relevance and fairness. For further details about the centroid creation and selection of term embeddings, we refer the reader to the original ColBERT-PRF work (X. Wang et al., 2021).

7.1.3 Experimental Setup

To evaluate the effectiveness of the ColBERT-FairPRF approach, we answer the following two research questions:

- **RQ7.1**: Can our ColBERT-FairPRF improve the fairness of multiple representation dense retrieval?
- **RQ7.2**: What is the effect of improving fairness using ColBERT-FairPRF on the relevance of the search results?

Test Collection

Analogous to the experiments we have presented in Section 4.5 we use the test collection of the 2021 & 2022 TREC Fair Ranking Tracks (Ekstrand et al., 2021) and seven different fairness characteristics (c.f. Table 4.1).

Table 7.1: Comparison of our policies to the baselines across different groupings. $*$ indicates statistical equivalence to both of the baselines in terms of nDCG. † indicates statistically significant differences from both of the baselines in AWRF. Bold values indicate the best AWRF (fairness) scores for each grouping.

Grouping	ColBERT		ColBERT-PRF		ColBERT-FairPRF	
	AWRF	nDCG	AWRF	nDCG	AWRF	nDCG
Alphabetical	0.6764	0.6413	0.6941	**0.6566**	**0.7298**†	0.6390*
Age of Article	0.6953	0.6413	0.6746	**0.6566**	**0.7110**†	0.6227*
Popularity	**0.6852**	0.6413	0.6780	**0.6566**	0.6577	0.6520*
Age of Topic	0.5552	0.6413	0.5595	**0.6566**	**0.5797**†	0.6120
Languages	0.7278	0.6413	0.7284	**0.6566**	**0.7329**†	0.6392*
Continent	0.4850	0.6413	0.4811	**0.6566**	**0.5156**†	**0.6566***
Geo Locations (TREC 21)	0.4803	**0.2777**	0.4700	0.2499	**0.5065**†	0.2490$*$

Baselines

We compare our proposed ColBERT-FairPRF approach against the ColBERT (Khattab & Zaharia, 2020) and ColBERT-PRF (X. Wang et al., 2021) approaches. We deploy the default parameters for ColBERT-PRF (X. Wang et al., 2021), i.e. we select 3 feedback documents and identify 24 representative centroids, before appending 10 embeddings to the query. We set ColBERT-PRF's β value to 1, thereby giving the expansion embeddings the maximum possible influence. For our proposed ColBERT-FairPRF approach, we use the same parameter values as ColBERT-PRF, however they are applied per group.

Metrics

We use AWRF (Sapiezynski et al., 2019) as presented in 6.3.3 to measure the fairness of search results. For relevance, we report the *nDCG* (Järvelin & Kekäläinen, 2002). Both measures are computed over the top 100 documents within a ranking. We use the same statistical tests as presented in 6.5.1, i.e., we evaluate the statistical significance of our results by comparing the AWRF scores between our proposed policies and the baselines using a paired t-test (Student, 1908). Additionally, we test for statistical equivalence between the nDCG scores of the baselines and our policies using the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987).

7.1.4 Results

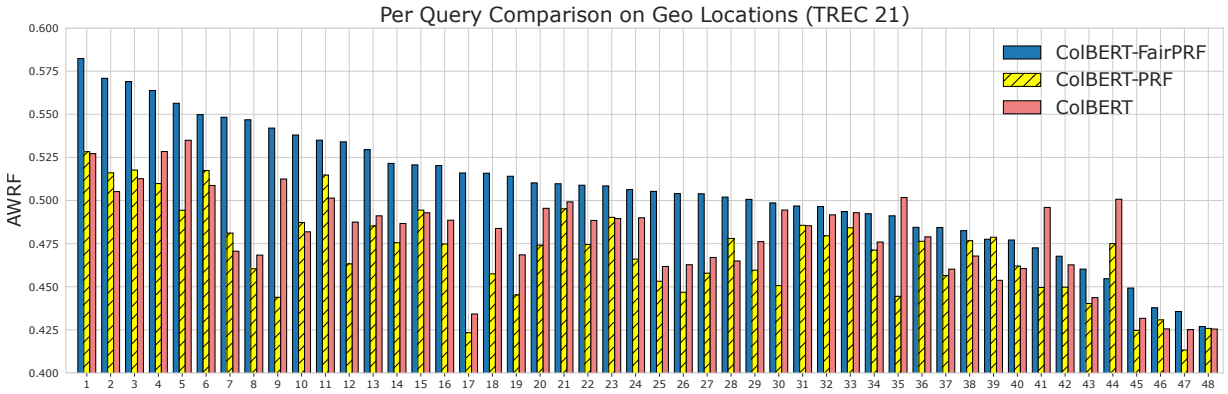
Table 7.1 shows the results over the seven different fairness characteristics, comparing the ColBERT-FairPRF approach to the baseline methods using AWRF and nDCG as metrics. In the "Alphabetical" category, ColBERT-FairPRF achieves the highest AWRF score of 0.7298, showing statistically significant differences (t-test, $p<0.05$) to ColBERT

(0.6764) and ColBERT-PRF (0.6941). Although ColBERT-PRF has the highest nDCG score (0.6566), there are only small differences in nDCG scores between ColBERT-FairPRF (0.6390), ColBERT-PRF, and ColBERT (0.641). A Two One-Sided Test (TOST) confirms statistical equivalence among the three methods, indicating that the improvements in fairness do not negatively impact relevance. In the "Age of Article" category, ColBERT-FairPRF also outperforms the baselines with an AWRF score of 0.7110. The TOST test indicates statistical equivalence to the ColBERT baseline in terms of nDCG. Within the "Popularity" category, Colbert-FairPRF leads to a decrease in fairness. This is the only category in which our approach does not increase fairness. However, the nDCG results are statistically equivalent, with Colbert-FairPRF achieving higher scores than the ColBERT baseline. For the "Age of Topic" category, ColBERT-FairPRF achieves the highest AWRF score of 0.5797 with statistical significance differences to both baselines. The nDCG scores are not statistically equivalent. In the "Languages" category, ColBERT-FairPRF achieves the highest AWRF score of 0.7329 outperforming the baselines, while the nDCG scores of all evaluated approaches are statistically equivalent. In the "Continent" category, ColBERT-FairPRF scores 0.5156 in AWRF, outperforming ColBERT (0.4850) and ColBERT-PRF (0.4811). The differences between the approaches are statistically significant for the AWRF metric and statistically equivalent for the nDCG metric.

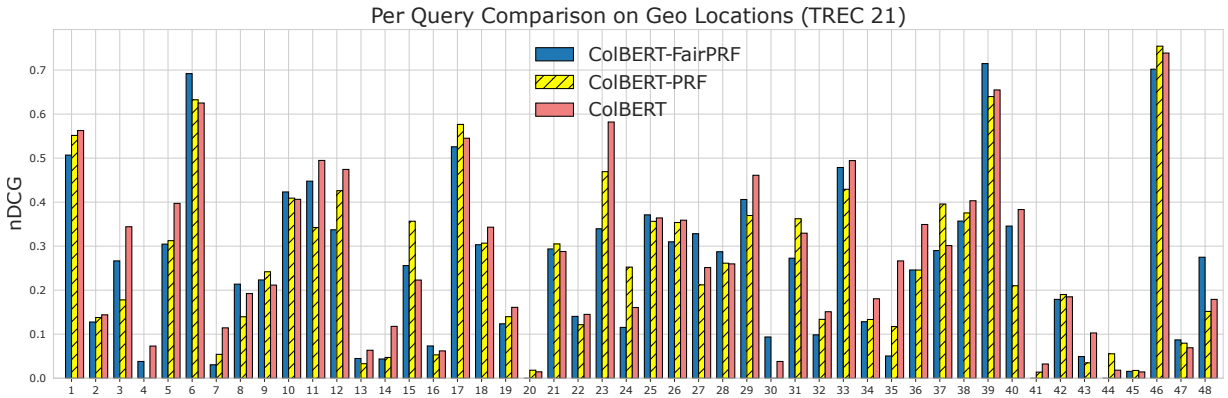
Finally, in the "Geographic Location" category (TREC 21), ColBERT-FairPRF achieves the highest AWRF score of 0.5065 with statistically significant differences to the baselines. Moreover, the nDCG scores of our approach are statistically equivalent compared to the baselines. Overall, the results show that ColBERT-FairPRF improves fairness in six out of seven categories, as indicated by higher AWRF scores, while still maintaining similar nDCG scores to the baselines. This suggests that ColBERT-FairPRF is able to successfully balance fairness and relevance.

To further understand the behaviour of the approaches, we examine the per-query performances on both AWRF and nDCG. Figure 7.1a shows the AWRF scores of the approaches for every query in the "Geographic Location" category of the TREC 2021 test collection. The queries are ordered by decreasing AWRF of ColBERT-FairPRF. Accordingly, the query on which ColBERT-FairPRF performs the best is at the left side of the plot. From Figure 7.1a we can see that ColBERT-FairPRF has the highest scores on the majority of the queries compared to the baselines. In particular, for 44 out of 48 queries, our approach achieves the highest AWRF score. On one query ColBERT-PRF is the highest scoring approach. ColBERT has the fairest ranking for three queries.

Figure 7.1b shows the nDCG scores per query on the "Geographic Location" category of the TREC 2021 test collection. To allow an easy comparison with Figure 7.1b the queries in both plots are aligned in the same order. From comparing the plots, we note that there is not a clear relationship between the amount of gain in fairness (Figure 7.1a)



(a) The plot shows the per query effectiveness in terms of AWRF for the category geographic locations. In the plot, the queries are ordered by decreasing AWRF score for ColBERT-FairPRF.



(b) The plot shows the per query effectiveness in terms of nDCG for the category geographic locations. In the plot, for comparison to Figure 7.1a, the queries are ordered by decreasing AWRF score for ColBERT-FairPRF.

Figure 7.1: The plots show the per query effectiveness for the category geographic locations. Subfigure (a) shows the AWRF scores, while subfigure (b) shows the nDCG scores, both ordered by decreasing AWRF score for ColBERT-FairPRF.

and the increase/decrease or overall relevance for a query (Figure 7.1b). For example, for queries 6 and 16 ColBERT-FairPRF manages to increase both fairness and relevance compared to both of the baselines. This shows that it is possible to increase fairness while also improving the relevance of the results.

To further clarify how Colbert-FairPRF expands a query, we present a detailed investigation of the "Disco Music" query for the Geo Locations category. Table 7.2 presents an overview of how the query is enriched by ColBERT-PRF and ColBERT-FairPRF. ColBERT-PRF creates 10 new query tokens that are used to expand the original query. Colbert-FairPRF applies the same approach for every geographic group and creates 10 query tokens per group. The lower portion of Table 7.2 shows the identified expansion tokens per group. It is noticeable that for some groups, the expansion embeddings are related to the geographic location. For example, the group Antarctica contains the token

Table 7.2: Example expansions for ColBERT PRF and ColBERT-FairPRF.

Original Query	'disco', 'discotheque', 'nightclub', 'deejay', 'dj', 'remix', 'dance music
ColBERT-PRF	'gibbons','proposition','stereo', 'mix', #mat', 'disco', '1975', '8', 'new', 'which'

Fairness Group	Colbert-FairPRF
Unknown	'thieves', 'collin', 'remixes', 'mix', 'disco', 'dj', 'got', 'de', 'united', 'which'
Africa	'bays', 'gee', 'vision', 'mt', '#oa', 'financial', 'dance', 'secondary', 'active', 'fr'
Antarctica	'#ater', '#dur', 'antarctica', 'antarctica', 'disco', 'ci', 'm', 'named', '#r', 'south'
Asia	'honda', 'apartment', 'resignation', 'creating', 'dj', 'hop', '#shi', '#shi', '#2', 'age'
Europe	'djs', 'afro', 'disco', '#ty', 'together', 'late', '#es', 'used', 'music', '#k'
Latin America	'flores', 'reggae', '#itt', 'rican', '#mus', 'duo', 'dj', '#ja', '#de', 'music'
North America	'gibbons', 'bronx', 'leonard', '#rra', 'disco', 'dj', 'hip', '36', 'video', 'york'
Oceania	'botany', 'edwin', 'dj', 'hip', '#ren', 'dance', 'zealand', '23', 'station', '-'

"antarctica", while the group Northern America contains "bronx" and "york" that point to distinct relevant geographic locations. By adding these tokens to the query embeddings, our proposed approach boosts the initially underrepresented groups and brings their exposure closer to the desired target exposure.

7.1.5 Conclusions

We have demonstrated that PRF, while traditionally employed to enhance relevance, can be adapted to promote fairness. We have proposed ColBERT-FairPRF, a novel mechanism that modifies the PRF process so that expanded queries reflect a broader spectrum of document groups. Our evaluation on the TREC Fair Ranking Tracks shows that ColBERT-FairPRF consistently improves fairness across six out of seven evaluated group categories. These gains, measured using AWRF, were statistically significant in most cases. Importantly, the relevance of the retrieved results remained largely unaffected, i.e., the nDCG scores statistically equivalent in six out of seven categories. This indicates that enhancing fairness through query expansion is achievable without compromising retrieval performance. Moreover, our work has been the first to show that fairness-aware query modification is both feasible and effective in the context of multi-representation dense retrieval. Our findings, represent a significant step towards retrieval pipelines that are not only relevant but also fair, validating our claims from the thesis statement that fairness can be integrated into IR systems without decreasing their utility.

7.2 Fair Generative Relevance Feedback

As we have shown in the previous section, dense PRF can be successfully adapted to improve the fairness of search results. In Section 7.1.2, the query was modified using terms from an initial retrieval, i.e., the pseudo-relevant set. However, as discussed in Section 2.4, large language models (PLMs) such as PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and the GPT models (Brown et al., 2020; OpenAI, 2023) have introduced refined language modelling capabilities, making them adaptable to both retrieval (Devlin et al., 2019) and generation tasks (Raffel et al., 2020). These generative capabilities of PLMs offer an alternative approach to the traditional PRF, allowing queries to be modified with terms generated beyond the pseudo-relevant set obtained from an initial retrieval (Mackie et al., 2023; X. Wang, MacAvaney et al., 2023). For example, methods such as Generative Relevance Feedback (GRF) (Mackie et al., 2023), or GenQR and GenPRF (X. Wang, MacAvaney et al., 2023) expand a user's query by generating terms using PLMs (Touvron et al., 2023; Brown et al., 2020; OpenAI, 2023) with various techniques, including fine-tuning and direct prompting. Generative relevance feedback extends traditional PRF by using large language models to generate new query terms that are expected to retrieve relevant documents. These terms are not restricted to those found in retrieved documents, enabling broader and more informed query reformulations.

Generative PRF has been shown to improve the effectiveness of IR systems. According to our thesis statement (Section 1.2), the two-stage ranking process in generative relevance feedback can, in principle, be adapted to enhance fairness in search results while maintaining the effectiveness of the IR system. In this section, to validate our thesis statement, we propose a generative PRF approach that uses PLM prompting to generate expansion terms, to increase fairness in the search results.

The section is structured as follows:

- In Section 7.2.1 we present the existing related work on generative relevance feedback.
- In Section 7.2.2 we propose a novel generative relevance approach called FGQE. FGQE employs zero-shot generation through PLM prompting using different strategies to generate expansion terms.
- In Section 7.2.3 we introduce the experimental setup we use to evaluate FGQE.
- In Section 7.2.5, we report the results of our evaluation of FGQE in section 7.2.5.
- Finally, in Section 7.2.6 we present our conclusions.

7.2.1 Related Work

As we have discussed in Section 2.3 as well as in Section 7.1.1, expanding a user's query has been shown to increase the effectiveness of a search system in retrieving relevant documents for a user's query. (Yan et al., 2003; Keikha et al., 2018; Xu et al., 2009; Lv & Zhai, 2010). PRF approaches expand an initial query by extracting additional keywords from the top-k documents in the initial retrieval (the pseudo-relevant set), which are presumed to contain relevant information. Traditional models like Bo1 (Amati & van Rijsbergen, 2002) and RM3 (Jaleel et al., 2004) used term statistics for expansion, while recent approaches such as ColBERT-PRF (X. Wang et al., 2021) use contextualised embeddings.

However, recent advancements have shifted away from the traditional principle of PRF. For example, X. Wang, MacAvaney et al. (2023) and Mackie et al. (2023) have introduced relevance feedback approaches that rely on PLMs to generate relevant expansion terms directly, without using a pseudo-relevant set of documents.

While PLM-based query expansion and traditional PRF methods were primarily developed to enhance the retrieval effectiveness of rankers, our focus is on improving fairness in search results. We have shown that PRF methods can be adapted to support fairness objectives (c.f. Section 7.1.5). However, to the best of our knowledge, no investigation has been conducted into the potential of generative relevance feedback to improve the fairness of exposure in the search results. Therefore, we propose Fair Generative Query Expansion (FGQE) to fill this gap. FGQE is designed to improve the allocation of exposure over groups of documents. Our FGQE method is closely related to GRF (Mackie et al., 2023) and GenQR (X. Wang, MacAvaney et al., 2023), as it also takes the original user query as input to generate expansion terms without relying on a pseudo-relevant set.

7.2.2 The FGQE Method

We propose FGQE to improve the fairness of the exposure distribution across groups of documents in search results. FGQE expands the initial query, q_0, by adding generated query terms related to specific groups $g \in G$ that were underrepresented in the initial retrieval, $rank_C$.

We quantify the exposure a group receives using the user model from DCG as introduced in Section 3.2 in Equation (3.12). This allows us to identify the groups with the lowest exposure for which we generate expansion terms by prompting an PLM. FGQE uses prompting strategies that instruct an PLM to produce terms that are directly relevant to each underrepresented group. Our four distinct prompting strategies each involve the characteristic of a group and the initial query terms.

Entities	For the [query], generate relevant entities that promote a diverse range of results, considering various aspects for the geographic locations [groups].
Keywords	For the [query], generate relevant keywords that promote a diverse range of results,considering various aspects for the geographic locations [groups].
CoT-Keywords	For the [query], generate relevant keywords that promote a diverse range of results, considering various aspects for the geographic locations [groups]. For every keyword, briefly describe why you chose it.
Document	For the [query], generate a relevant document that promotes a diverse range of results, considering various aspects for the geographic locations [groups].

Figure 7.2: The figure gives an overview of our four prompting strategies.

Prompting Strategies

Figure 7.2 shows four different prompting strategies that can be deployed within FGQE to identify query terms for underrepresented groups. The strategies differ in their approach, with some focusing on specific entities or keywords, others incorporating reasoning, and one generating a more detailed response. In the following, we describe each strategy.

1. **Entities** (blue, first row in Figure 7.2): The Entities strategy prompts the model to generate key concepts or named entities related to both the initial query and the context of the underrepresented group. For example, if the group is ”Europe” under the characteristic “Geographic location” the generated entities might include “European cultural heritage” or “regional landmarks”. In this strategy, the prompt specifically requests “relevant entities”, instructing the model to produce focused, group-specific terms.

2. **Keywords** (yellow, second row in Figure 7.2): The Keywords strategy instructs the model to generate terms that capture core themes relevant to each group and the initial query. For example, for the group “Asia”, the model might generate keywords such as “traditional Asian art” or “regional cuisine”. Here, the prompt specifically requests “relevant keywords”, i.e., terms that cover essential themes unique to each group.

Query	"Fishes"
Instruction	For the query 'Fishes', generate relevant keywords that promote a diverse range of results, considering various aspects for the geographic locations Oceania AND Antarctica AND Africa.
Response	Oceania marine life, Antarctic aquatic species, African freshwater fishes, Oceania fishing spots, Antarctic fish research, African fish cuisine, Oceania fish conservation, Antarctic ice fishes, African fish diversity.
New Query	"Fishes, fish, oceania, antarctic, african, marine, life, aquatic, species, freshwater, spot"

Table 7.3: Instruction and response for generating keywords with the GPT-4 model for the query 'Fishes'.

3. **CoT Keywords** (purple, third row in Figure 7.2): The CoT (chain-of-thought) Keywords strategy uses a zero-shot CoT prompting approach, guiding the model to generate keywords with brief explanations of their relevance. For example, if a keyword relates to a cultural practice, the model explains the relevance of the cultural practice to the group, ensuring the expanded query remains aligned with the group's perspective. In this strategy, the prompt explicitly requests both terms and their explanations.

4. **Document** (orange, fourth row in Figure 7.2): The Document strategy prompts the model to generate a longer response that includes both the query and relevant group information. For instance, for the group "North America", the generated document might cover regional themes, events, or cultural topics central to that group. The prompt in this strategy requests a "relevant document" in the prompt.

All of the proposed strategies generate content for the n most underrepresented groups in the retrieved document set, $rank_C$. The generated content is tokenised and split into a set of potential expansion terms, XT. From XT, stop-words, duplicates, and any irregular characters are removed. In the next step, we select the f terms from XT with the highest probability of relevance to the query. To estimate the relevance of different terms in the generated content, we use the frequency of each term in XT, selecting the most frequent terms in the generated content. If multiple terms have the same frequency, we use the order of their generation to break ties. For this work, we set $n = 3$ and $f = 10$, aligning these values with the default settings of prior approaches such as ColBERT-FairPRF and ColBERT-PRF (X. Wang et al., 2021; Khattab & Zaharia, 2020), which used three feedback documents and appended 10 expansion terms or embeddings to ensure comparability across experiments. The f selected terms are then appended to the original query q, producing the expanded query q_e, which is resubmitted to theIRsystem.

Example Query "Fishes"

To better illustrate how our approach works, we use the query "Fishes" from the TREC 2022 Ranking collection and show how it is deployed with GPT-4 using the "Keyword" Prompt (Figure 7.2, yellow row 2). Table 7.3 shows how the expansion terms are identified. First, an initial retrieval is executed and the three groups with the least amount of exposure are identified. In our example, these groups are "Oceania", "Antarctica" and "Africa". Following the prompting strategy for keywords we have presented in Section 7.2.2 we then prompt the PLM with the Instruction seen in the second row of Table 7.3. The model returns potential query expansion terms (third row of the table) as a response. We tokenise the response and remove the duplicates and any irregular characters. We then select the 10 most frequent terms (i.e., f=10) and append them to the query (fourth row). If there are multiple terms with the same frequency, we select terms based on the order of their generation. We then resubmit the query using the original retrieval model, i.e., BM25. In the following section, we detail our experimental setup which we use to evaluate FGQE.

7.2.3 Experimental Setup

To evaluate FGQE, we answer the following three research questions.

- **RQ7.3:** Does our FGQE approach lead to a more fair exposure allocation over groups of documents in the search results compared to traditional retrieval methods?
- **RQ7.4:** What is the effect of FGQE on the relevance of the search results?
- **RQ7.5:** How does our FGQE approach compare to dense PRF variants ColBERT-PRF and ColBERT-FairPRF (c.f. Section 7.1.2)?

Test Collections

We use the same collections as presented in 4.5 and used in 7.1.3. We evaluate FGQE over the groups "Geo Location" and "Continent" as presented in previous sections (see Table 4.1). These groups were chosen because they represent distinct categories that align well with external knowledge accessible to PLMs, providing useful context for generating relevant expansion terms. In contrast, categories such as "Popularity" or "Age of Article" lack specific thematic content that could guide an PLM in producing meaningful expansions. For example, without having access to the specific user statistics of Wikipedia, the PLMs have no knowledge of which articles might have low or high popularity.

7.2.4 Baselines

We compare our FGQE approach to the following baselines:

- **BM25** (Robertson et al., 1994): A sparse retrieval model (c.f. Section 2.2.3), deployed with the default parameters used in the PyTerrier implementation $k_1 = 1.2$ and $b = 0.75$.
- **BM25>>RM3** (Jaleel et al., 2004): The RM3 relevance language model applied after a first-pass retrieval using BM25, with PyTerrier's standard expansion setting: 3 feedback documents & 10 expansion terms (c.f. Section 2.3, 2.6). We use the >> operator from the declarative PyTerrier framework to define the retrieval pipeline: first, BM25 is applied for the initial retrieval, and then RM3 to expand the query.
- **ColBERT** (Khattab & Zaharia, 2020): A neural ranking model (c.f. Section 2.5) that uses contextualised late interaction for efficient and effective document retrieval.
- **ColBERT-PRF** (X. Wang et al., 2021): A recent dense PRF model developed on top of ColBERT (c.f. Section 2.5.4). We use the standard settings for ColBERT-PRF with 3 feedback documents, 24 representative centroids and 10 expansion embeddings. This also matches the RM3 baseline.
- **ColBERT-FairPRF**: Our fairness-aware dense PRF approach, introduced in Section 7.1.2, designed to improve fair exposure in search results.

FGQE Settings

FGQE can be applied with any PLM that is capable of long-text generation. In our experiments, we use FlanT5 (Chung et al., 2024), GPT-4 (OpenAI, 2023) and Llama2 (Touvron et al., 2023). We choose FlanT5 due to its previous success in Query Relevance Reformulation (X. Wang, MacAvaney et al., 2023). We also deploy a GPT model following the experiments of Mackie et al. (2023). Moreover, we include Llama 2 (Touvron et al., 2023) as an open-weight baseline in our comparison.

- **Llama2**: We give our prompt to *meta-llama/Llama-2-7b-chat-hf*[1] and use the following tuned parameters when generating the expansion terms: *frequency_penalty*: 0.8 and *temperature*: 0.7. These parameters were specifically chosen to balance the diversity and relevance of the generated terms, ensuring a robust query expansion process.
- **FlanT5**: We use the *google/flan-t5-large*[2] model with the following parameters: *do_sample*: False, *top_k*: 200, *top_p*: 0.92, *repetition_penalty*: 2.1, and *early_stopping*: True. These settings were selected to ensure deterministic output, promote term diversity, and minimise repetitive terms, aligning with the requirements of our query expansion strategy.

[1] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[2] https://huggingface.co/google/flan-t5-large

- **GPT-4**: We use the GPT-4-API to access the *gpt-4o*[3] model with the default parameters provided by the API.

All of our models use BM25 as their ranking model following the experimental setup of Mackie et al. (2023). For consistency in this thesis, we set the number of underrepresented groups included in the prompt to n=3 and the number of expansion terms to 10 to align with Section 7.1. We use the same measures and statistical testing as in Section 7.1.3.

7.2.5 Results

In the following paragraphs, we answer our three research questions.

RQ7.3: Does FGQE improve the allocation of exposure over groups of documents in the search results?

Table 7.4 shows the results of our FGQE approaches across various prompting strategies, compared to the baselines BM25 and BM25>>RM3 in terms of fairness using the AWRF metric. On the TREC 2022 Fair Ranking Track test collection, BM25 yields an AWRF score of 0.5028, while applying PRF using BM25>>RM3 leads to a small improvement of fairness at 0.5030. In contrast, on the TREC 2021 collection BM25>>RM3 leads to a slight decrease in AWRF from 0.4845 to 0.4835. Moreover, the results show that each of our FGQE strategies achieves notably higher fairness improvements.

The *Entities* prompting strategy consistently achieves the highest AWRF scores. FGQE (GPT-4)-Entities achieves an AWRF score of 0.5419 on the TREC 2022 Fair Ranking Track test collection, a 7.8% increase over BM25. On the TREC 2021 Fair Ranking Track test collection, the Entities strategy has an AWRF score of 0.5080, making it the best-performing strategy for fairness across both datasets. This highlights the effectiveness of expanding queries with entities related to the most underrepresented groups of documents. The *Keywords* strategy also performs strongly in terms of fairness. FGQE (GPT-4)-Keywords achieves an AWRF score of 0.5408 on the TREC 2022 test collection, only slightly lower than the *Entities* strategy, and 0.5067 on the TREC 2021 test collection. This indicates that including group-relevant keywords as additional query terms effectively promotes fairness, although the gains in fairness are smaller than the entity-focused approach in maximising fair exposure. The *CoT-Keywords* strategy, which combines keywords with brief rationales, shows competitive AWRF scores across both datasets but does not consistently outperform the simpler *Keywords* strategy. For instance, FGQE (GPT-4)-CoT-Keywords achieves an AWRF of 0.5376 on the TREC 2022 test collection and 0.5060 on the TREC 2021 test collection, indicating that the added rationales may not consistently enhance fairness over the direct keyword expansions alone.

[3]https://platform.openai.com/docs/models

Table 7.4: Comparison of baselines and FGQE variants in the TREC 2021 and TREC 2022 test collections. ‡ indicates statistically significant differences (t-test, $p < 0.05$, using Bonferroni correction) compared to BM25 and BM25>>RM3. * indicates statistical equivalence (TOST, α<0.05).

Approach	**TREC 2021**		**TREC 2022**	
	nDCG	**AWRF**	**nDCG**	**AWRF**
BM25	0.1742	0.4845	0.5227	0.5028
BM25>>RM3	0.1959	0.4835	0.5233	0.5030
FGQE (Keywords)				
Llama-2	0.1939*	0.5038‡	0.5038*	0.5241‡
Flan-T5	0.1598*	0.4982‡	0.4689*	0.5352‡
GPT-4	0.1973*	0.5067‡	**0.5703**	0.5408‡
FGQE (Entities)				
Llama-2	0.1775*	0.5038‡	0.5125*	0.5248‡
Flan-T5	0.1585*	0.4997‡	0.4960*	0.5276‡
GPT-4	0.2083*	**0.5080**‡	0.5419*	**0.5419**‡
FGQE (CoT-Keywords)				
Llama-2	0.2068*	0.5014‡	0.5340*	0.5238‡
Flan-T5	0.1610*	0.4994‡	0.5074*	0.5359‡
GPT-4	0.1959*	0.5060‡	0.5154*	0.5376‡
FGQE (Document)				
Llama-2	0.1939*	0.5070‡	0.5497*	0.5390‡
Flan-T5	0.1598*	0.4982‡	0.5305*	0.5261‡
GPT-4	**0.2226**	0.5069‡	0.5506*	0.5401‡

Finally, the *Document* strategy yields substantial fairness improvements over the baselines but does not reach the same levels as *Entities* or *Keywords*. FGQE (GPT-4)-Document achieves an AWRF of 0.5401 on the TREC 2022 test collection and 0.5069 on the TREC 2021 test collection, showing that it is effective for balanced exposure, though its more generalised content generation approach may not target group-specific terms as precisely as the *Entities* strategy. Across both datasets, the FGQE strategies improve fairness over BM25 and BM25>>RM3 in all cases independent of the underlying PLM, with the *Entities* strategy emerging as the most effective for maximising AWRF. The t-test reveals statistically significant differences to the baselines for all FGQE prompting strategies.

RQ7.4: What is the effect of FGQE on the relevance of the search results?

For RQ7.4, in which we examine the impact of a fairer exposure distribution on the relevance of search results, the findings from Table 7.4 indicate that the FGQE approaches improve fairness and maintain or enhance relevance compared to traditional baselines. The *Keywords* and *Document* strategies perform particularly well, often outperforming the baselines and other prompting strategies.

On the TREC 2022 test collection, FGQE (GPT-4)-Keywords achieves the highest relevance score with an nDCG of 0.5703, representing a 9.1% improvement over BM25 and an 8.9% improvement over BM25>>RM3. The *Document* strategy, also demonstrates a strong relevance performance. FGQE (GPT-4)-Document reaches an nDCG score of 0.5506 on the TREC 2022 test collection, representing a 5.3% improvement over BM25, making it the second-best strategy for effectiveness on this dataset. This result suggests that detailed content generation provides contextual depth that complements group-specific keywords, thereby enhancing retrieval relevance.

On the TREC 2021 test collection, FGQE (GPT-4)-Document outperforms all other strategies, achieving an nDCG score of 0.2226, a substantial 27.9% improvement over BM25 and a 13.6% improvement over BM25>>RM3. FGQE (GPT-4)-Entities also performs well on the TREC 2021 test collection, with an nDCG of 0.2083, representing a 19.7% improvement over BM25.

The *CoT-Keywords* strategy is only able to outperform the baselines when LLama-2 is deployed. With all other PLMs, we can observe small decreases in nDCG over both datasets compared to the baselines. Notably, deploying Flan-T5 leads to a small loss in relevance over the majority of strategies.

The *Entities* strategy is able to outperform the baselines on the TREC 2022 test collection when GPT-4 is applied. Indeed, when GPT-4 is deployed as the PLM in FGQE, we can observe increases in nDCG in the majority of cases.

Overall, the FGQE methods outperform the BM25 baseline in nDCG in 6 out of 12 cases on the TREC 2022 test collection dataset and in 8 out of 12 cases on the TREC 2021 test collection, indicating the general relevance benefits of FGQE. The Two One-Sided Tests (TOST) confirm statistical equivalence across all FGQE prompting strategies, except for FGQE (GPT-4)-Keywords on the TREC 2022 test collection and FGQE (GPT-4)-Document on the TREC 2021 test collection, both of which show improvements.

RQ7.5: How does FGQE compare to existing dense PRF baselines?

Table 7.5 presents a comparison of the ColBERT-based approaches—ColBERT, ColBERT-PRF, and ColBERT-FairPRF—with all FGQE prompting strategies using GPT-4. The GPT-4 strategies are included in this comparison because they have been demonstrated to be the most successful among the FGQE strategies. Among the ColBERT-based meth-

Table 7.5: Comparison of ColBERT variants and the best FGQE performers on the TREC 2021 and TREC 2022 datasets.

Approach	**TREC 2021**		**TREC 2022**	
	nDCG	**AWRF**	**nDCG**	**AWRF**
ColBERT	**0.2777**	0.4803	0.6413	0.4850
ColBERT-PRF	0.2499	0.4700	**0.6566**	0.4811
ColBERT-FairPRF	0.2490	0.5065	**0.6566**	0.5156
FGQE (Keywords)				
GPT-4 (Keywords)	0.1973	0.5067	0.5703	0.5408
FGQE (Entities)				
GPT-4 (Entities)	0.2083	**0.5080**	0.5419	**0.5419**
FGQE (CoT-Keywords)				
GPT-4 (CoT-Keywords)	0.1959	0.5060	0.5154	0.5376
FGQE (Document)				
GPT-4 (Document)	0.2226	0.5069	0.5506	0.5401

ods, ColBERT-FairPRF performs best in terms of fairness, with AWRF scores of 0.5065 on TREC 2021 and 0.5156 on TREC 2022. These results highlight the effectiveness of ColBERT-FairPRF's approach in enhancing fair exposure. However, our FGQE methods consistently also outperform all ColBERT variants.

On the TREC 2021 test collection, FGQE (GPT-4)-Entities achieves the highest fairness score with an AWRF score of 0.5080, surpassing ColBERT-FairPRF's 0.5065. On the TREC 2022 collection, both FGQE (GPT-4)-Entities and FGQE (GPT-4)-Document achieve the highest AWRF scores with 0.5419 and 0.5401, respectively, compared to ColBERT-FairPRF's 0.5156. The *Keywords* strategy demonstrates strong performance in terms of fairness, achieving an AWRF score of 0.5408 on the TREC 2022 collection. Although the *CoT-Keywords* strategy performs slightly below other FGQE strategies, it achieves comparable results to the baselines with an AWRF of 0.5060 on the TREC 2021 test collection and surpasses all baselines on the TREC 2022 test collection with a score of 0.5376. These findings indicate that the FGQE approaches offer a robust alternative to fairness-enhancing PRF methods that improve the exposure allocation over groups of documents, such as ColBERT-FairPRF.

When examining relevance, ColBERT achieves the highest nDCG scores among the ColBERT-based methods, with 0.2777 on the TREC 2021 test collection. Moreover, ColBERT-FairPRF and ColBERT-PRF achieve the highest nDCG with 0.6566 on the TREC 2022 test collection, outperforming the FGQE methods. While FGQE generally yields lower nDCG scores, FGQE (GPT-4)-Document reaches an nDCG of 0.2226 on the TREC 2021 test collection, achieving similar scores to ColBERT-FairPRF. However, it is notable that FGQE relies on BM25 as its underlying ranking model, a sparse approach, rather than a neural model, indicating that while less competitive in relevance FGQE still provides reasonable utility. We leave the investigation of substituting the BM25 initial retrieval with ColBERT to future work.

7.2.6 Conclusions

In this section we have demonstrated that generative PRF can be effectively adapted to promote fairness in rankings without reducing the relevance of the search results. By introducing FGQE, a novel method that uses PLMs to generate expansion terms tailored to underrepresented document groups, we have shown that it is possible to use generative capabilities for fairness-aware retrieval. FGQE can be deployed with four distinct prompting strategies Keywords", Entities", CoT-Keywords", and Document" that all offer distinct advantages and allow to align the query expansion process to marginalised and underexposed groups. Among our strategies, the "Entities" strategy proved particularly effective, achieving the highest AWRF scores across both TREC 2021 and TREC 2022 datasets, and consistently outperforming traditional baselines. Moreover, the Keywords" and Document" strategies improved nDCG relative to BM25, confirming that fairness-aware generative expansion does not inherently reduce result quality. FGQE achieved favourable trade-offs when compared to existing dense PRF models, such as ColBERT-based baselines, demonstrating that zero-shot generative expansion can compete with existing state-of-the-art techniques. With the introduction of FGQE we provide an alternative to the more traditional PRF approaches. The use of FGQE allows to access external information from the PLMs not explicitly inherent in the documents allowing for a more flexible fairness improvements. Our findings from this chapter validate our thesis statement that fairness can be integrated into modern retrieval systems using query expansion. FGQE represents a scalable and adaptable mechanism for achieving balanced exposure in search results, offering a compelling direction for future fairness-aware IR research using generative models.

7.3 Fairness through Query Generation

In the previous two sections, namely 7.1.2 and 7.2.2, we provided evidence supporting our thesis statement that established retrieval methods, which improve recall by modifying the query, can be effectively tailored to achieve a fairer exposure distribution. In Section 7.2.2, we have shown that we can use the generative capabilities of PLMs to effectively improve the relevance and fairness of the search results. Specifically, we have focused on generating query expansion terms (c.f. Section 7.2.2) independently without using a pseudo-relevant set from an initial retrieval. Related to using generative relevance feedback, generating relevant queries and appending them to documents has been shown to increase the effectiveness of a standard retrieval pipeline (Gospodinov et al., 2023; Nogueira, Yang, Cho & Lin, 2019). Following our thesis statement, in this section, we investigate whether the generation of full queries can also be adapted to improve the fairness of the search results. We introduce a new re-ranking method to study whether generating full queries based on document content is suitable for addressing the uneven distribution of attention towards underrepresented groups in the search results. The rest of the section is structured as follows:

- In Section 7.3.1 we discuss the related work to query generation for fairness.
- In Section 7.3.2, we introduce our approach, QSEE, which employs various policies to generate queries and re-rank documents based on the generated queries.
- In Section 7.3.3, we present the experimental setup we use to evaluate our QSEE approach and analyse the results in section 7.3.4.
- Finally, in section 7.3.5 we conclude this section.

7.3.1 Related Work

In Section 2.3 as well as in Section 7.1.1 and Section 7.2.1, we have presented different methods that help overcome the vocabulary mismatch problem and increase the recall of an IR system. The presented methods involve modifying the using traditional query expansion methods, i.e., RM3 (Jaleel et al., 2004) as well as neural PRF methods such as ColBERT-PRF (X. Wang, Macdonald, Tonellotto & Ounis, 2023).

Alternatively to expanding the user's query, augmenting the documents' contents with terms that are expected to be relevant and representative of the documents' contents has also been shown to be an effective approach for improving retrieval effectiveness (MacAvaney et al., 2020). Methods such as Doc2Query (Nogueira, Lin & Epistemic, 2019) and Doc2Query– (Gospodinov et al., 2023) generate queries that are likely to be answered by a document, and append the generated queries to the original document's text. Appending such generated query terms can increase the chances of a relevant docu-

ment being retrieved in response to a user's query. However, in addition to the traditional objective of providing the most relevant search results to a user, following our thesis statement (c.f. Section 1.2) we focus on the fair allocation of exposure over groups of documents.

Related to this, Abolghasemi et al. (2023) have successfully used doc2T5query to generate synthetic queries to estimate the retrievability bias of pre-trained language models. Moreover, Penha et al. (2023) apply controlled query generation to improve the retrievability of single documents. However, in this work, we are not focusing on the retrievability of documents but on the exposure groups of documents obtain in a ranking.

7.3.2 Query Similarity Exposure Enhancement

In this section, we introduce a novel re-ranking approach called Query Similarity Exposure Enhancement (QSEE). QSEE is designed to re-score the relevance of documents from an underexposed group by leveraging queries generated from the documents' text. Query generation refers to the task of producing plausible user queries from the content of a document. In the context of re-ranking, such generated queries can be used to assess how likely a document would be retrieved for a given user query, enabling better estimation of its relevance. For each document in an underexposed group, QSEE generates a set of representative queries that capture the document's content and uses these queries to better predict the document's relevance to the user's query.

By representing documents as queries, QSEE mitigates existing biases in the document text, providing a fairer opportunity for these documents to receive appropriate exposure. We propose three policies that use the similarity between a generated query and the user's query as an additional relevance signal. This similarity score is used to re-score the documents before re-ranking them.

Figure 7.3 illustrates the retrieval and re-ranking process that integrates our proposed QSEE approach and policies. From left to right, the standard retrieval pipeline, denoted as $P(n)$, begins by retrieving documents for a user query q_0 using an efficient ranking model, such as BM25 (Robertson et al., 1994), to retrieve an initial ranked list, rank_C. This is followed by a more computationally intensive neural re-ranking model, such as ELECTRA (Clark et al., 2020), producing a ranked list, rank_U, where documents are ordered by their predicted relevance scores, $r \in \mathbb{R}$.

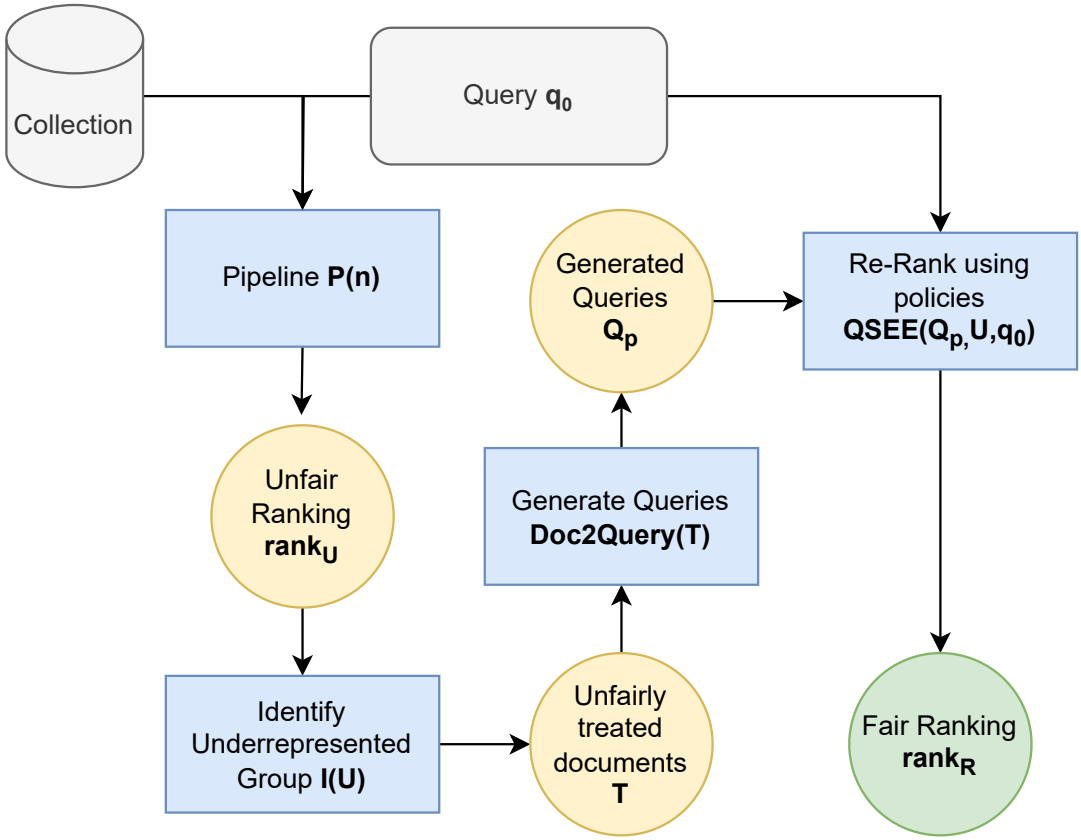


Figure 7.3: The figure shows the retrieval process integrating our QSEE approach and policies. Squares represent retrieval functions, while circles represent document sets and rankings.

Once rank_U is obtained, we assess the exposure of each group g to identify the most underexposed group, T, which receives the least exposure. We use the same process to identify the most underrepresented group as described in Equation 3.12, which has also been applied in our ColBERT-FairPRF 7.1.2 and FGQE 7.2.2 approaches. For each document within T, we generate a set of representative queries, Q_μ, using a sequence-to-sequence model, such as T5 (Raffel et al., 2020). This approach produces queries that reflect the document's content from multiple perspectives (Nogueira et al., 2020).

To use these generated queries in re-ranking, QSEE selects, for each document in T, the query $q_\mu \in Q_\mu$ that is most relevant to the user query q_0. The re-ranking model from $P(n)$ calculates a similarity score $s(q_0, q_\mu)$ for each $q_\mu \in Q_\mu$, and we select the query with the maximum score, $s_{\max}$. This score is then applied in each of our policies to re-rank the documents in rank_U, aiming to create a fairer ranking, rank_R, based on adjusted fair relevance scores, r_f. Focusing on a single underexposed group allows for efficient query generation and targeted fairness adjustments.

Policy 1: Replacing Original Relevance Score with Query Similarity Score

In this policy, to re-rank rank_U, we replace each document's original relevance score, r, produced by the standard retrieval pipeline $P(n)$, with a new score if the document belongs to an underrepresented group T. Specifically, we use the score of the document's generated query that is most similar to the user's query q_0, denoted as $s_{\max}$. The final relevance score is then given by:

$$r_f = s_{\max} \quad \text{if} \quad s_{\max} > r \tag{7.5}$$

Our rationale is that if a generated query is more similar to the user's query than the document itself, the document might be more relevant than indicated by its original score r. However, to avoid penalising documents due to potential inaccuracies in query generation, we only replace r with $s_{\max}$ when $s_{\max}$ exceeds r. Replacing the entire score with $s_{\max}$ reduces the risk of underestimation but may lead to some overestimation.

Policy 2: Averaging Original and Query Similarity Scores

In this policy, if $s_{\max} > r$, we set the final relevance score as the mean of the original score and the maximum query similarity score. The final relevance score is thus defined as:

$$r_f = \frac{r + s_{\max}}{2} \tag{7.6}$$

By taking the mean, we aim to balance relevance estimation, reducing the risk of overestimating a document's relevance. However, exposure within the ranking is finite. Increasing exposure for an underrepresented group can reduce exposure for other groups. Therefore, if the most underrepresented group in rank_U is already close to its ideal exposure, the final score r_f should lean closer to the original relevance score r rather than to $s_{\max}$, to prevent overexposure.

Policy 3: Mixture Model of Relevance Scores

In Policy 3, to determine the final score r_f, we incorporate a factor γ, which controls the influence of the maximum query similarity score $s_{\max}$. This factor γ is calculated based on the ratio of the target exposure E_{target} to the actual exposure E_{Actual}:

$$\gamma = \frac{E_{target}}{E_{Actual}} \tag{7.7}$$

If the discrepancy between E_{Actual} and E_{target} is large, indicating significant underexposure, we reduce reliance on the original relevance score and place more weight on $s_{\max}$. The final relevance score is given by:

$$r_f = \frac{r + \gamma \cdot s_{\max}}{2} \tag{7.8}$$

This weighting allows the model to adjust relevance proportionately to the group's exposure needs, improving the accuracy of exposure distribution in the final ranking. By prioritising the most underexposed group, this policy prevents conflicting adjustments that could arise from modifying multiple groups simultaneously. Since ranking positions are limited, adjusting several groups at once could lead to competing fairness corrections.

7.3.3 Experimental Setup

We aim to answer the following two research questions:

- **RQ7.6:** Can our QSEE policies distribute the available exposure to users more fairly among groups than a standard retrieval pipeline?
- **RQ7.7:** How do our re-ranking policies affect the relevance of the ranking?

To evaluate our policies, we again use the TREC 2021 and 2022 Fair Ranking Track test collections (Ekstrand, McDonald et al., 2022, 2022). This evaluation employs the same categories used in our ColBERT-FairPRF experiments, as outlined in Section 7.1 and detailed in Table 4.1. As a baseline, we implement a re-ranking pipeline that uses BM25 (Robertson et al., 1994) for initial retrieval, followed by ELECTRA (Clark et al., 2020) as the re-ranker (i.e., BM25>>ELECTRA). For query generation, we use the PyTerrier implementation[4] of the T5 Doc2Query model from Nogueira and Lin Nogueira et al. (2020), generating 20 queries for each document. We apply the same metrics as in the previous sections 7.1.3 and 7.3.3. Specifically, we use AWRF (Sapiezynski et al., 2019), with an equal exposure target to measure fairness and nDCG to assess relevance.

7.3.4 Results

To address our research questions, we evaluate the performance of our policies over rankings of size $k = 100$. Table 7.6 shows the results for our approach QSEE as measured by AWRF and nDCG, comparing each policy against the baseline BM25>>ELECTRA.

[4]https://github.com/terrierteam/pyterrier_doc2query

Table 7.6: Comparison of our policies to the baseline BM25>>ELECTRA. We use † to indicate a significant difference from the re-ranking baseline in terms of AWRF (t-test, p<0.05), and * to indicate statistical equivalence (TOST, $\alpha < 0.05$) in terms of nDCG.

Category	Policy-1		Policy-2		Policy-3		BM25>>ELECTRA	
	AWRF	nDCG	AWRF	nDCG	AWRF	nDCG	AWRF	nDCG
Alphabetical	0.7664	0.5762*	0.8021†	0.5792*	**0.8222**†	**0.5863***	0.7756	0.5795
Age of Article	0.7128†	0.5783*	0.7072†	0.5787*	**0.7363**†	0.5779*	0.6790	**0.5795**
Popularity	0.7125	0.5818*	0.7180†	0.5795*	**0.7446**†	**0.5940***	0.6931	0.5795
Age of Topic	0.5744†	0.5739*	0.5595†	0.5769*	**0.5967**†	**0.5811***	0.5478	0.5795
Languages	0.7118	0.5749*	0.7060†	0.5776*	**0.7540**†	0.5763*	0.6843	**0.5795**
Continent	0.5169	0.5769*	0.5116	0.5783*	**0.5285**†	**0.5820***	0.5077	0.5795
Geo Locations	0.5046	**0.3344***	0.5068	**0.3344***	**0.5089**	**0.3344***	0.5066	0.3316

RQ7.6: Do our policies improve the fairness in the search results?

In response to RQ1, we observe that Policy 1 improves fairness in five of the seven categories, with statistically significant gains in "Age of Article" and "Age of Topic" (†, t-test, p<0.05). By replacing the relevance score with the maximum query similarity score, this policy effectively enhances exposure for the underrepresented group, i.e., the exposure is distributed more equally in the search results.

Policy 2, which averages the original and maximum query similarity scores, consistently produces stable improvements across all categories, with statistically significant differences in five out of seven categories. Though the gains are generally smaller than those of Policy 1, the stability of Policy 2 over the different categories is notable.

Policy 3, using a weighted adjustment with factor γ based on the actual exposure and the target exposure, achieves the highest AWRF scores across all categories. The improvements over the baseline are statistically significant in 6 out of the 7 evaluated categories, with AWRF gains of up to approximately 10%. The highest AWRF score is achieved in the "Alphabetical" category with 0.8222. This result underscores the effectiveness of adaptive weighting, which adjusts relevance based on the group's exposure discrepancy.

RQ7.7: How do our re-ranking policies affect the relevance of the ranking?

To answer RQ7.7, we examine the relevance of search results using nDCG as reported in Table 7.6. Policy 1 shows slight reductions in nDCG across five categories, but statistical equivalence (TOST) is maintained, suggesting these changes are minor. Policy 1 even achieves a small nDCG improvement in the "Popularity" category. Policy 2 consistently maintains equivalence in nDCG with the baseline, with only slight decreases observed across categories. On the 2021 TREC Fair Ranking Track test collection, i.e., the Geographic Location category, Policy 2 even yields marginal nDCG improvements, indicating that the averaging approach retains relevance effectively. Policy 3 shows minor

nDCG reductions in three categories but achieves increases in all others, maintaining overall equivalence with the baseline in terms of relevance. Moreover, it is notable that all policies outperform the baseline in terms of nDCG in the Geo Location category. Overall, the results show, that Policy 3 is the most successful in improving the exposure allocation of the search results, without compromising the utility of the system.

Example for Query "Catholicism"

To further clarify the operation of our policies, we provide an example of a re-ranked document for the query *Catholicism*, which includes keywords such as "catholicism", "papacy", "jesuit", "pope", "papal", "catholic", and "bishop". Initially, the "a-d" group from the Alphabetical category was underrepresented. To address this imbalance, our policy generated additional queries specifically for documents in this group.

As shown in Table 7.7, after applying Policy 3, the document titled "Catholic Church in Germany" experienced the most significant rank improvement. Initially positioned at rank 945 with an ELECTRA relevance score of -6.566262, the document was reassessed through a generated query, "who is the catholic bishop of Germany?". This generated query achieved a similarity score of -0.16, which indicated greater alignment with the original query *Catholicism*. Policy 3 then adjusted the document's relevance score to -3.330994, which raised its rank to 33.

Table 7.7: The table shows an example of a re-ranked Document using Policy 3 for the query *Catholicism*.

Title	**Query**	**Original Rank**	**New Rank**	**Original Score**	**New Score**
Catholic Church in Germany	Catholicism	945	33	-6.566262	-3.330994
Abstract				**Query Generation**	**Sim Score**
The Catholic Church in Germany, or Roman Catholic Church in Germany, is part of the worldwide Roman Catholic Church, in communion with the Pope, assisted by the Roman Curia, and with the German bishops.				who is the catholic bishop of Germany	-0.16

7.3.5 Conclusion

Our introduced QSEE re-ranking approach is able to balance group exposure by generating queries that reflect underrepresented perspectives. By incorporating generated group-specific queries into the re-ranking process, QSEE enables fairer distribution of exposure in search results. In particular, we have proposed three policies that can be applied within QSEE. Our evaluation reveals that all of our policies are successful in improving the fairness over the majority of categories. This shows, that using generative capabilities of PLMs to create full queries is another successful method for fair ranking. In particular,

Policy 3 which adaptively re-weights scores based on exposure disparities, was the most effective. It consistently outperformed baseline methods across all categories, achieving fairness improvements of up to 10% (as measured by AWRF) without degrading relevance. While Policy 1 and Policy 2 are also able to outperform the baselines, their performance is more dependent on specific categories over which they are applied. In summary we can conclude, that our thesis statement in which we posit, that IR systems can be successfully adapted to promote fairness alongside relevance by using query generation can be validated.

7.4 Conclusion

All of our approaches contain distinct strategies to enhance the fairness of the search results by using query modification techniques. We show that traditional IR concepts such as PRF can be as successful as more modern IR variants that use the generative capabilities of PLMs. The success of all three methods demonstrates that applying fairness interventions at the query level is an effective strategy for improving fairness in search results. While the PRF variant can be incorporated in any IR system that uses a PRF mechanism, approaches such as FGQE and QSEE can be deployed with any system that has access to PLMs. The systems with the PLMs can then access external information to avoid existing unfairness. However, this needs to be carefully considered, since existing PLMs can introduce other unintended unfairness while the PRF approaches operate on the closed system of the existing retrieval.

In summary, in this chapter, we have contributed three novel approaches aligned with our thesis statements, that enhance the fairness of a ranking without significantly decreasing the quality of results. In the next section, we compare the performance of our query modification methods from Chapters 6 with our adaptive re-ranking approach from Chapters 7.

Chapter 8

Comparison of Fairness Interventions

In Chapters 6 and 7, we introduced multiple fairness-aware adaptations of retrieval methods, designed to improve the distribution of exposure across different groups of documents in search results. In Chapter 6, we introduced six distinct policies for improving fairness in the graph-based adaptive re-ranking process (c.f. Chapter 6.4), which we refer to as Fair GAR in this chapter. In Chapter 7, we presented three approaches that modify a user's query to improve the exposure distribution in the search results. Specifically, in Section 7.1, we described ColBERT-FairPRF, an approach that uses dense pseudo-relevance feedback mechanisms to improve the allocation of exposure across document groups. Moreover, in Section 7.2, we proposed FGQE, an approach that prompts pre-trained large language models to generate query expansion terms to increase the fairness in the search results. Finally, in Section 7.3, we introduced QSEE, a re-ranking approach that uses the similarity between generated queries for an underrepresented group and the user's original query to improve the exposure allocation. All of our proposed methods deploy different approaches to improve the fairness in the search results while maintaining satisfactory relevance. Therefore, in this chapter, we provide a comparison of the performance of the different techniques. The chapter is structured as follows:

- In Section 8.1, we summarise the core concepts of the methods Fair GAR, ColBERT-FairPRF, FGQE, and QSEE.
- In Section 8.2, we compare these methods in terms of fairness metrics, focusing on their ability to achieve balanced exposure across document groups.
- In Section 8.3, we analyse the impact of our techniques on the relevance of the search results.
- In Section 8.4, we analyse the trade-off between relevance and fairness when our proposed methods are deployed.

- Finally, in Section 8.5, we conclude by reflecting on the comparative strengths and limitations of each method.

8.1 Methods Overview

The following paragraphs provide a summary of four approaches that we previously introduced and discussed in Chapters 6 and 7. Specifically, we cover Fair GAR which was introduced in Section 6, where we outlined its adaptive fairness-enhancing re-ranking mechanism. Moreover, we give a brief summary of ColBERT-FairPRF our proposed approach focusing on fairness-aware pseudo-relevance feedback, which we have previously presented in Section 7.1. Furthermore, we discuss FGQE which was detailed in Section 7.2, where we explored its query expansion strategies using PLMs. Finally, we return to QSEE, introduced in Section 7.3, which generates queries to re-rank documents from an underrepresented group to improve fairness.

8.1.1 Fair Graph Based Adaptive Re-ranking (Fair GAR)

Our proposed fairness-aware adaptive re-ranking (Fair GAR) (c.f. Section 6) method introduces six policies that can be deployed in the original GAR process to enhance the fairness of the search results (c.f. Section 6.5.2). Our proposed policies (c.f. Section 6.4) intervene either during the construction of the corpus graph (*pre-retrieval*, Policies 1 & 2) or during the re-ranking process (*in-process*, Policies 3–6). From our evaluation it became apparent, that Policy 6 is the best-performing out of our compared policies, therefore we use it for comparison with the other fairness-aware ranking methods. Policy 6 adjusts the batch selection in the adaptive re-ranking process by prioritising neighbours from earlier iterations (c.f. Section 6.4).

8.1.2 ColBERT-FairPRF

ColBERT-FairPRF (c.f. Section 7.1) extends the ColBERT-PRF (X. Wang et al., 2021) approach by selecting feedback embeddings from the highest-ranked documents of each group in the pseudo-relevant set. This ensures, that all the groups are represented in the expanded query, improving fairness in the search results (c.f. Section 7.1.4).

8.1.3 FGQE

FGQE (c.f. Section 7.2) prompts PLMs to generate query expansion terms. FGQE is designed to be deployed with a single prompting strategy at a time, each tailored to retrieve documents from underrepresented groups. By incorporating models like GPT-4, FGQE generates terms that expand the query to help the underrepresented groups receive a fair exposure. The most successful of our proposed strategies is FGQE (GPT-4)-Entities, which uses GPT-4 as the base language model and generates terms focusing on entities related to the underrepresented groups (c.f. Section 7.2.2).

8.1.4 QSEE

QSEE (c.f. Section 7.3) uses generated queries to re-rank documents from an underrepresented group to enhance exposure. We have proposed three policies that QSEE can be deployed with (c.f Section 7.3.2). The most effective policy (Policy 3) for improving exposure allocation in search results dynamically adjusts rankings using a mixture model. Policy 3 accounts for discrepancies between the target exposure and the actual exposure received by the underrepresented group in the initial retrieval (c.f. Section 7.3.4).

In the next sections, we compare our proposed approaches in terms of fairness in the search results (Section 8.2) and their impact on relevance (Section 8.3). Moreover, we investigate the trade-offs between the fairness and relevance of each method (Section 8.4).

8.2 Analysis of Fairness Performance

To ensure a fair and consistent comparison, all methods were evaluated using the same experimental setup as described in the previous chapters. Specifically, we used the TREC 2021 and 2022 Fair Ranking Track collections (Ekstrand et al., 2021; Ekstrand, McDonald et al., 2022) and the AWRF (Sapiezynski et al., 2019) (c.f. Section 3.4.2) as our evaluation metric to assess fairness. Table 8.1 presents a summary of the results from all approaches across seven categories (c.f. Table 4.1). Moreover, we include BM25>>ELECTRA as a baseline representing the standard re-ranking pipeline in our comparison. From our proposed approaches, we report the results of the best-performing policies and strategies to ensure a competitive evaluation. Specifically, for Fair GAR, we use Policy 6, which was shown in Chapter 6 to provide the most significant fairness improvements across multiple categories. Policy 6 optimises exposure by balancing relevance and fairness scores while targeting underrepresented groups. For QSEE, we deploy Policy 3, which adjusts relevance scores using a mixture model that incorporates query similarity and exposure needs,

as discussed in Chapter 7.3. Similarly, the ColBERT-FairPRF and FGQE methods use their most effective configurations, with pseudo-relevance feedback in ColBERT-FairPRF (Section 7.1) and GPT-4's entity-driven prompting strategy in FGQE (c.f. Chapter 7.2), to yield the best performance across the evaluated categories.

Table 8.1: AWRF Scores Across Categories and Methods. Bold values indicate the highest AWRF score for each category. FGQE results are limited to Continent and Geographic Location, reflecting the experimental scope outlined in Chapter 7.2.

Method	Alphabetical	Age of Article	Popularity	Age of Topic	Languages	Continent	Geo. Location
BM25 $\gg$ ELECTRA	0.7756	0.6790	0.6931	0.5478	0.6843	0.5077	0.5066
ColBERT-FairPRF	0.7298	0.7110	0.6577	0.5797	0.7329	0.5156	0.5065
FGQE (GPT-4)	–	–	–	–	–	**0.5419**	0.5080
QSEE (Policy 3)	**0.8222**	0.7363	**0.7446**	0.5967	**0.7540**	0.5285	0.5089
Fair GAR (Policy 6)	0.7983	**0.7418**	0.7407	**0.6046**	0.7286	0.4882	**0.5095**

Analysing the results in Table 8.1 it becomes apparent the baseline BM25>>ELECTRA produces the most unfair search results, i.e., it receives the lowest AWRF scores on the majority of categories. Moreover, it is notable that QSEE is the best-performing approach in three out of the seven categories. Specifically, QSEE achieves the highest scores in the categories Alphabetical (0.8222), Popularity (0.7446), and Languages (0.7540). In the Alphabetical category, QSEE shows improvements over the baseline BM25>>ELECTRA of up to approximately 6.5% (0.8222 vs. 0.7756). In Languages, QSEE shows a 10.1% improvement over the baseline. While QSEE achieves the highest scores in these categories, it also remains competitive in the others. For example, QSEE is the second-best performing approach in Age of Topic, Age of Article and Geographic Location.

The Fair GAR method achieves the highest AWRF scores out of all systems on the Age of Article, Age of Topic and Geo Location categories. In all of these categories, Fair GAR is able to outperform the baseline and other approaches. Moreover, Fair GAR achieves competitive performance in all remaining categories providing the second-best result in the Alphabetical category (0.7983) and the Popularity category in which we can only observe small differences to the best-performing QSEE approach (0.7407 vs 0.7446).

Analysing the results of FGQE, we can observe that in both evaluated categories (Continent and Geo Location), FGQE achieves competitive AWRF scores. Specifically, in the Continent category FGQE achieves the highest score with 0.5419. Moreover, in the Geo Location category, FGQE outperforms the baseline and ColBERT-FairPRF.

Observing the results for ColBERT-FairPRF it is notable that its fairness improvements are only moderate compared to the other approaches. While ColBERT-FairPRF is able to outperform the baseline in four out of seven categories, namely Age of Article, Age of Topic, Languages, and Continent. It fails to improve over the baseline on the remaining three categories, making it the proposed approach with the least amount of fairness gains.

In summary, our analysis highlights that QSEE and Fair GAR are the best-performing approaches, showing competitive fairness performance across all categories. Moreover, we note that FGQE also provides strong fairness improvements over the baselines. Furthermore, it is apparent that while ColBERT-FairPRF is able to improve fairness in many cases it provides only moderate fairness gains. In the next section, we will analyse the relevance of the search results when our approaches are applied.

8.3 Analysis of Relevance Performance

Table 8.2 shows the nDCG scores for each method across the different categories.

Table 8.2: nDCG Scores Across Categories and Methods. Bold values indicate the highest nDCG score for each category. FGQE is evaluated only in the Continent and Geographic Location categories.

Method	Alphabetical	Age of Article	Popularity	Age of Topic	Languages	Continent	Geo. Location
BM25 ≫ ELECTRA	0.5795	0.5795	0.5795	0.5795	0.5795	0.5795	0.3316
ColBERT-FairPRF	**0.6390**	**0.6227**	**0.6520**	**0.6120**	**0.6392**	**0.6566**	0.2490
FGQE (GPT-4)	–	–	–	–	–	0.5419	0.2083
QSEE (Policy 3)	0.5863	0.5779	0.5940	0.5811	0.5763	0.5820	**0.3344**
Fair GAR (Policy 6)	0.6111	0.5938	0.6075	0.5743	0.6000	0.5827	0.3331

From the table, it becomes evident, that ColBERT-FairPRF achieves the highest relevance scores on the majority of categories. Only in the Geographic Location category, QSEE achieves higher relevance scores than ColBERT-FairPRF. However, the differences between the approaches are only marginal. ColBERT-FairPRF achieves the highest overall nDCG score (0.6566) on the Continent category showing a 13% increase over the BM25>>ELECTRA baseline. Analysing our other proposed methods, it is notable, that Fair GAR is the second best-performing approach in terms of relevance. Fair GAR achieves the second highest nDCG scores in the majority of categories showing improvements of

up to 6% over the BM25>>ELECTRA baseline. QSEE only shows moderate relevance improvements compared to our other proposed approaches. However, it is able to outperform the BM25>>ELECTRA baseline on the majority of categories and is the best performing approach in the Geographic Location category with an nDCG score of 0.3344.

QSEE demonstrates moderate relevance improvements, with its strongest performance observed in the Geographic Location category, where it achieves 0.3344 compared to the baseline's 0.3316. While these gains are less strong compared to Fair GAR, QSEE remains a competitive approach for integrating fairness interventions. Finally, FGQE shows the lowest nDCG scores out of all proposed approaches. However, this is to be expected, since FGQE only uses BM25 as its underlying ranking model. In summary, our results show, that ColBERT-FairPRF is the best-performing system in terms of relevance. ColBERT-FairPRF uses the ColBERT dense retrieval model (Khattab & Zaharia, 2020) as a base, which typically shows strong retrieval performance. However, using ColBERT is generally computationally demanding and requires a large embedding index (Nogueira, Yang, Lin & Cho, 2019; Thakur et al., 2021) making it impractical for our other proposed approaches as the underlying model.

8.4 Trade-off Between Fairness and Relevance

After we have analysed fairness and relevance separately, we will now analyse both measures in combination. Figure 8.1 shows a stacked bar chart which combines the nDCG and AWRF scores per category. The perfect combined score is 2.0. The lower half of a bar shows the relevance score (nDCG), while the upper half shows the fairness score, i.e., AWRF. In the chart, the lower segment of each bar corresponds to relevance (nDCG), while the upper segment reflects fairness (AWRF). The total height of a bar illustrates the combined score (nDCG+AWRF). The bar chart further strengthens the insights from the previous sections, in which we could observe that all of our approaches are able to outperform the BM25>>ELECTRA baseline on the majority of categories. Moreover, it is notable, that ColBERT-FairPRF achieves the highest combined score on the categories Age of Topic, Languages and Continent. The chart shows, that Fair GAR and QSEE achieve a similar combined score of approximately 1.4 in the Alphabetical category. However, it is notable, that Fair GAR has a more even balance between relevance and fairness, while QSEE is more influenced by a high fairness score. We can observe the same pattern in the Languages category in which both approaches achieve a similar score of approximately 1.3. Overall, it is notable, that both Fair GAR and QSEE achieve very similar scores on the majority of categories. However, Fair GAR more often leads to a higher combined

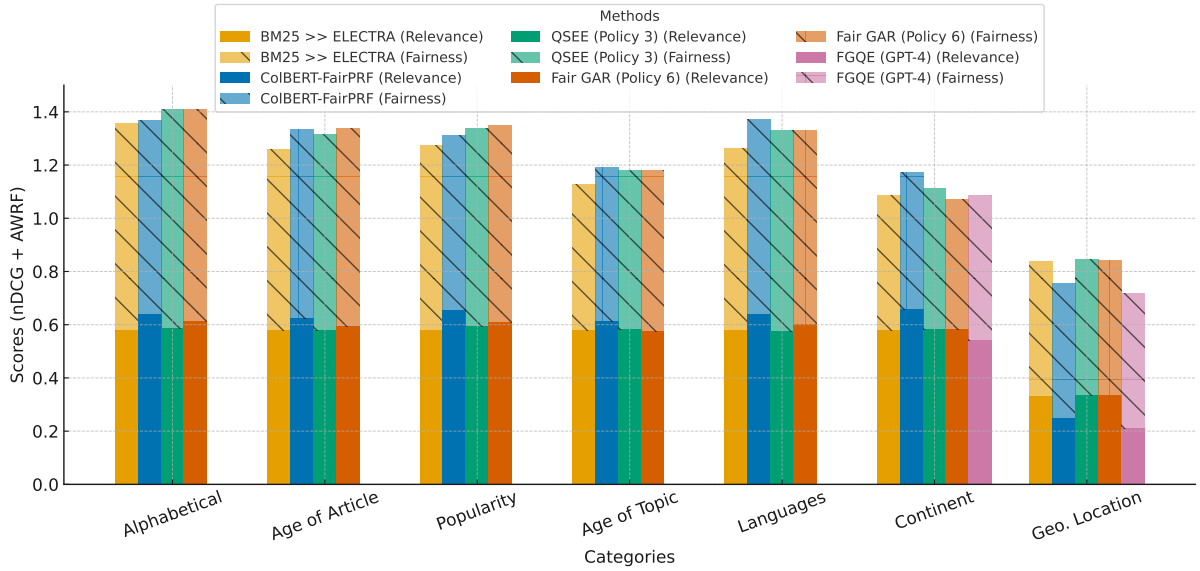


Figure 8.1: The stacked bar chart illustrates the trade-offs between fairness (AWRF) and relevance (nDCG) across methods and categories. Each bar is divided into two segments: the lower segment represents relevance, while the upper segment represents fairness. The total height of each bar reflects the combined scores.

score (Age of Article and Popularity category). Observing the FGQE approach we note, that while it has the lowest relevance scores, the improvements in fairness are substantial. However, the big improvements in fairness cannot always mitigate the effect of the low relevance scores as can be seen in the Geographic Location category.

In summary, we have observed, that all of our approaches are able to maintain a reasonable trade-off between relevance and fairness further supporting our thesis statement (c.f. Section 1.2). Our results show, that ColBERT-FairPRF achieves high combined scores with a focus on relevance and moderate fairness improvements. In contrast, our QSEE approach achieves high fairness improvements with small decreases in nDCG. Moreover, Fair GAR is the most robust version when balancing fairness and relevance.

8.5 Conclusion

In this chapter, we have demonstrated that all of our proposed fairness-aware approaches can substantially improve the allocation of exposure across document groups when evaluated on the TREC 2021 and 2022 Fair Ranking Track collections. Each method outperformed a standard re-ranking baseline in terms of fairness, as measured by AWRF, thereby reinforcing the claim that fairness can be systematically integrated into modern retrieval systems. Among the evaluated approaches, Fair GAR and QSEE achieved the most apparent gains in fairness. These results underscore the effectiveness of both graph-based and generative query-based re-ranking mechanisms in mitigating exposure disparities. With

respect to relevance, ColBERT-FairPRF emerged as the strongest strategy in terms of nDCG, indicating that dense PRF methods can be adapted for fairness without decreasing retrieval quality. While other approaches did not reach the same level of relevance, they nevertheless maintained satisfactory performance, confirming that fairness improvements do not inherently compromise utility. When jointly considering fairness and relevance, Fair GAR proved to be the most robust overall, offering the most balanced trade-off across the two objectives. ColBERT-FairPRF provided the most consistent relevance performance, while QSEE delivered the largest fairness improvements. These findings substantiate the thesis claim that fairness-aware adaptations can achieve improvements in exposure equity without diminishing the effectiveness of retrieval. They also highlight the importance of selecting the appropriate method depending on the optimisation priorities of the system in question.

Chapter 9

Conclusions and Future Work

While the primary objective of an IR system is to retrieve the most relevant documents to a user's query, ensuring fairness in search results has been increasingly recognised as an important additional objective (Ekstrand et al., 2019). Advancements in retrieval techniques (c.f. Chapter 2.5), such as the use of PLMs, can enhance the system's ability to retrieve relevant documents effectively. However, these improvements in effectiveness must be paired with efforts to ensure fairness in search results. Without such considerations, imbalances of exposure across document groups can introduce or even amplify economic and social disadvantages (Pedreschi et al., 2008). In this thesis, we have shown how exposure can be fairly distributed in the search results of IR Systems without affecting the quality or relevance of results. In particular, we have shown how exposure imbalances can be predicted (Chapter 4) and how exposure can be measured when document labels are unavailable (Chapter 5). Moreover, we have shown how modern ranking pipelines (c.f. Section 2.5.5) based on adaptive re-ranking can be tailored towards ensuring fairness in the search results. Furthermore, we demonstrated how other re-ranking approaches that modify a user's query can also enhance the exposure distribution (Chapter 7). In Chapter 7 we have adapted established retrieval methods such as Pseudo-Relevance Feedback (c.f. Section 7.1) as well as modern ranking methods that use Query Generation (c.f Section 7.3) and Prompting (c.f. Section 7.2) to ensure fairness in the search results. In all adaptations of these retrieval techniques, we ensured that not only is fairness introduced as an objective alongside traditional relevance-focused metrics, but that the quality of ranking is not markedly or significantly reduced. The remainder of this chapter is structured as follows:

- In Section 9.1 we summarise the main contributions of this thesis.
- In Section 9.2 we validate the claims from our thesis statement (c.f. Section 1.2).
- In Section 9.3 we present the limitations of our work.
- In Section 9.4 we discuss potential future research directions.

- Finally, in Section 9.5 we provide concluding remarks for this thesis.

9.1 Contributions

The main contributions of this thesis are as follows:

- In Chapter 4, we gave a comprehensive introduction into how exposure can be distributed in the search results across groups of documents (c.f. Section 4.3). We have highlighted that the number of documents associated with a single group present in a ranking is a crucial factor in the exposure that a particular group will receive (c.f. Figure 4.3). In our analysis of exposure properties (c.f Section 4.3), we showed, that the exposure will change more substantially when documents are added or removed compared to simply re-ranking the already present documents. Moreover, we have presented the novel task of QEP (c.f. Section 4.2), in which the exposure groups receive in a ranking is predicted prior to retrieval. We have provided a first approach to tackle QEP, called Group Exposure Predictor (c.f. Section 4.4). The GEP uses lexical features to construct a vector that represents a group of documents which can be used to predict the relevance to a query.

- In Chapter 5, we have shown how the exposure distribution of groups of documents can be measured when document labels are unavailable. We have proposed the use of quantification (c.f. Section 5.3) to reliably assess the document labels under unawareness and in turn measure the exposure distributions in a ranking. In Section 5.4.1, we have proposed a novel quantification protocol, that is specifically tailored towards the task of *Query Fairness Estimation*, i.e., the assessment of fairness in relation to a specific query. Our proposed protocol is the first that is able to alleviate the Sample Selection Bias (c.f. Section 5.4) in which traditional quantification methods typically fail. We have demonstrated, that quantification can be successfully used to measure the exposure distribution in rankings, i.e., the fairness of the search results (c.f. Section 5.6).

- In Chapter 6, we highlighted that the original GAR approach does not succeed in improving the fairness of exposure distribution across document groups. Although GAR was initially developed to enhance recall in search results, which might intuitively support fairer rankings, our analysis shows that it does not achieve this goal (c.f Section 6.3). From our analysis, it has been notable, that the graph-based ranking approach can even increase imbalances in the exposure distribution (c.f.

Section 6.3). To be able to use GAR for improving fairness, we have proposed six policies (c.f. Section 6.4) that can be deployed pre-retrieval or in-process, which all help to improve the exposure allocation over groups while maintaining the quality of a ranking (c.f. Sections 6.5.2 and 6.5.3).

- In Chapter 7, we proposed several approaches that adapt existing retrieval methods-originally designed to improve recall by modifying a user's query-to enhance the allocation of exposure across groups of documents. In particular, we have adapted the concept ofPRF in dense retrieval to improve the fairness of the search results (c.f. Section 7.1). Moreover, we have extended the traditional PRF paradigm and presented different strategies to prompt pre-trained large language models to generate query expansion terms, that expand queries and help to create fairer search results (c.f. Section 7.2). Finally, in Section 7.3 we used generated queries to identify and re-rank relevant documents from an underrepresented group of documents. We show that all of our proposed approaches increase the fairness in the search results while maintaining the quality of a ranking (c.f. Sections 7.1.4, 7.2.5 and 7.3.4).

- In Chapter 8 we compared our proposed methods from Chapters 6 and 7. We compared our methods in terms of fairness (c.f. Section 8.2) and relevance (c.f. Section 8.3). Moreover, we analysed the tradeoff of both (c.f. Section 8.4).

All of our contributions were developed with practical deployment in mind, aiming to provide fairness interventions that are feasible to integrate into real world IR systems and adaptable to a range of operational constraints. In this thesis we offer a variety of possible fairness interventions that, depending on context, can be chosen. For example, when traditional re-ranking pipelines are deployed, our prediction method from Chapter 4 can easily be deployed to assess the need for fairness interventions in the development of IR systems. Moreover, the contributions from Chapter 5 can help practitioners to make fairness assessments possible and replace existing problematic processes that rely on classification-based methods. However, while our approach improves on existing methods, it is important to consider the specific use case, since the inference of labels can still introduce discrimination as we have previously discussed in Section 5.7. With Chapter 6 we show how a state-of-the-art re-ranking pipeline can be deployed with fairness and relevance as objectives. In particular, the GAR pipeline offers practitioners a dynamic system that can be deployed with different rankers and re-rankers. Our contributions from Chapter 7 cover a wide range of query modification methods from which practitioners can choose the most suitable. For example, with ColBERT-FairPRF we have proposed a modern PRF variant that allows the effective ColBERT model to be deployed and used for fair query expansion within the IR system. If more lightweight rankers are deployed in the retrieval pipeline, our approaches QSEE and FGQE offer a viable solution which uses the

knowledge of PLMs for successful and fair query expansion. The choice of model depends on the constraints of where the ranker is deployed, i.e., whether access to PLMs is feasible or the memory constraints of the ColBERT model during retrieval can be met. Overall, we have demonstrated a variety of approaches that generalise well, are easy to integrate, and can be further adapted. In the next section, we validate our thesis statement.

9.2 Validation of Thesis Statement

Our thesis statement presented in Section 1.2 contains three main claims, which we will validate in the following section. The first claim we validate is that modern ranking systems that use a two-stage process can be adapted to improve the exposure allocation over groups of documents in the search results. Moreover, in our thesis statement, we posit that established methods such as Query Performance Predictors and quantification methods can be modified to assess and predict the exposure groups receive in an initial ranking. Finally, we state, that is possible to enhance the fair distribution of exposure across groups of documents in a ranking without significantly compromising the relevance of the search results.

- **Claim 1:** ***Modern ranking systems, which typically employ re-ranking pipelines, can be effectively adjusted to allocate the exposure in rankings more fairly.... In particular, established retrieval methods aimed at increasing retrieval effectiveness by improving recall using a two-stage process can be successfully adapted...*** In Chapter 6 we have proposed an adaptation of the graph-based adaptive re-ranking process and in Chapter 7 we have introduced three different modifications of established query modification techniques, namely ColBERT-FairPRF (c.f Section 7.1), FGQE (c.f. Section 7.2) and QSEE (c.f. Section 7.3). To validate our claim, we have investigated whether our adaptations can improve the exposure allocation over groups in the search results. Our experiments using the AWRF metric (Section 3.4.2) indicate that all of our methods are able to enhance the fairness of the search results, validating our first claim. Specifically, we demonstrated, that all of our proposed adaptions of the original GAR process significantly enhance fairness compared to standard re-ranking baselines and the original GAR process. Our best-performing policy achieves improvements of up to 13% compared to the baselines (Table 6.1). Moreover, our analysis of the results in Section 7.1.4 revealed that ColBERT-FairPRF also achieved higher AWRF scores than the baselines in the majority of our evaluated categories (Table 7.1). Furthermore, our experimental evaluation in Section 7.2.5 as well as in Section 7.3.4, showed that our proposed approaches FGQE and QSEE consistently lead to fairness improvements in the search results (c.f. Tables 7.4 & 7.6).

- **Claim 2:** ***By predicting and assessing the number of documents per group in a ranking using Query Performance Prediction and quantification, we can tailor and expand existing state-of-the-art retrieval techniques towards the objective of a fair exposure distribution.*** In Chapters 4, we demonstrated that the exposure allocation can be effectively predicted in the search results. Moreover, in Chapter 5 we showed that exposure can be measured even in the absence of document labels. To validate Claim 2, we introduced GEP, the first method for Query Exposure Prediction (Section 4.4), and showed its effectiveness in predicting the exposure received by groups in search results. Our evaluation shows, that GEP can achieve up to a 40% reduction in prediction error (Table 4.6). Additionally, we validated that quantification is a suitable technique for measuring exposure. In Section 5.6, we showed that our quantification-based methods consistently outperformed standard baselines, surpassing both traditional classification approaches and earlier correction methods (c.f. Tables 5.3 and 5.2). These advancements enabled us to assess unfairness in search results and helped in the development of fair retrieval adaptations introduced in Chapters 6 and 7.

- **Claim 3:** ***It is possible to enhance the fair distribution of exposure across groups of documents in a ranking without significantly decreasing the relevance of the search results, i.e., the utility of the resulting information retrieval system for the user.*** To validate this claim, we have analysed the quality of the search results created by our proposed approaches from Chapters 6 and 7 using the nDCG metric (c.f. Section 3.4.1). Moreover, we have deployed the TOST procedure (c.f. Section 6.5.1) to investigate if the relevance of the search results remains statistically relevant to the baselines. Specifically, we have investigated whether any differences in nDCG scores between our methods and the baselines are marginal enough to be considered negligible. For the adaptations of the GAR methods (c.f. Section 6.5.3), our results showed, that the best-performing policies achieved significant fairness improvements while maintaining or even exceeding the relevance scores of the adapted baselines (c.f. Table 6.2). Furthermore, our evaluation of the QSEE approach (c.f. Section 7.3.4) showed substantial improvements in fairness, without any significant reductions in relevance (c.f. Table 7.6). The same trend can be observed when investigating the results of the FGQE strategies (c.f. Section 7.2.5). Our proposed FGQE approach was able to improve fairness while maintaining statistically equivalent nDCG scores to the baselines (c.f. Table 7.4). Our analysis of ColBERT-FairPRF (c.f. Section 7.1.4) also revealed that the improvements in fairness do not negatively affect the relevance of the search results and remain statistically equivalent to the baselines (c.f. Table 7.1). Our analysis of

the trade-off between relevance and fairness in Section 8.4 further strengthens these insights. In summary, it is notable, that all of our methods can effectively balance fairness and relevance supporting the claim that fairness can be improved without diminishing the utility of the information retrieval system.

In the next section, we discuss some of the limitations of this thesis:

9.3 Limitations of this Work

One limitation of this thesis is the lack of real-world user studies. In the evaluation of our experiments, we rely on offline metrics and existing user browsing models to calculate the exposure groups of documents received. Therefore, it would have been beneficial, to evaluate our proposed approaches in real-world scenarios to analyse whether our fairness interventions are effective. However, user studies are complex and require substantial resources. The scale and scope of effort required for effective user studies are highlighted by initiatives such as the Participatory Harm Auditing Workbenches and Methodologies (PHAWM) project[1] which aims to develop standards and evaluation tools for such scenarios. However, the development of standardised evaluation benchmarks goes beyond the scope of our thesis. Another limitation is the potential for missing groups in the evaluation of fairness. In this thesis, we rely on the available labels in the TREC Fair Ranking Track collections to group documents and evaluate fairness. However, this means that certain groups of documents, that need fairness in real-life ranking scenarios, may be overlooked and can not be included in the evaluation. Moreover, the evaluation relies on the characteristics of the TREC Fair Ranking Track collections. For example, the nature of the documents and queries as well as the relevance labels are all collection-dependent. Furthermore, another limitation is the computational complexity of our fairness interventions. While efforts were made to optimise the efficiency of our fairness interventions, their application in real-world retrieval systems where latency and resource constraints are critical has yet to be investigated. However, adding efficiency as a third objective next to fairness and relevance would add substantial complexity and further require appropriate evaluation measures.

9.4 Directions of Future Work

While this work addresses key aspects of fairness-aware IR, several areas remain for further exploration:

[1] https://phawm.org/

- **Additional Fairness Definitions:** In our thesis we focus on a single group fairness definition. However, in future work, different or multiple fairness definitions (c.f. Section 3.1) could be incorporated in the adaptation of existing retrieval approaches. In particular, fairness definitions covering different fairness perspectives could be combined. For example, approaches that ensure individual fairness under group fairness constraints could be developed.

- **Multiple Intersectional Groups:** Related to the investigation of different fairness definitions, future work could investigate the potential of our proposed approaches to cover different group concepts. In particular, the compatibility of our proposed approaches with multiple intersectional groups, i.e., groups that combine multiple aspects such as geographic location and popularity.

- **Real-World User Studies:** As discussed in the previous section our evaluation of fairness relied on offline metrics. Therefore, in future work, it would be beneficial to conduct real-world user studies.

9.5 Closing Remarks

In this thesis, we have addressed the important task of ensuring fairness in search results. In particular, we have contributed several adaptations of established IR techniques that can be deployed to make the outputs of IR systems fairer, Moreover, all of our proposed methods are able to maintain the quality of the search results. We believe that research on developing IR systems, that are fair and useful at the same time, will further grow. Especially, since the societal significance of ensuring fair information and unbiased access to information is getting increasing recognition. As we have outlined in Section 9.4, there are still many exciting topics and complex challenges that future research can address.

Bibliography

Abdi, H. (2010). Coefficient of variation. *Encyclopedia of research design*, *1*(5).

Abolghasemi, A., Verberne, S., Askari, A., & Azzopardi, L. (2023). Retrievability bias estimation using synthetically generated queries. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3712–3716.

Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., & Omer, S. B. (2014). Using the Bayesian improved surname geocoding method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: A validation study. *Health Services Research*, *49*(1), 268–283.

Agarwal, A., Zaitsev, I., Wang, X., Li, C., Najork, M., & Joachims, T. (2019). Estimating position bias without intrusive interventions. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 474–482.

Amati, G., & van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *Transactions on Information Systems*, *20*(4), 357–389.

Andrus, M., & Villeneuve, S. (2022). Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. *Proceedings of FAccT '22: Conference on Fairness, Accountability, and Transparency*, 1709–1721.

Arabzadeh, N., Meng, C., Aliannejadi, M., & Bagheri, E. (2024). Query performance prediction: From fundamentals to advanced techniques. *Proceedings of the 46th European Conference on Information Retrieval*, 381–388.

Asudeh, A., Jagadish, H. V., Stoyanovich, J., & Das, G. (2019). Designing fair ranking schemes. *Proceedings of the 2019 International Conference on Management of Data*, 1259–1276.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*, *2*(3).

Bashir, S., & Rauber, A. (2009). Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. *Proceedings of the 18th ACM conference on Information and knowledge management*, 1863–1866.

Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2010). Quantification via probability estimators. *Proceedings of the 10th International Conference on Data Mining*, 737–742.

Biega, A. J., Diaz, F., Ekstrand, M. D., & Kohlmeier, S. (2020). Overview of the TREC 2019 fair ranking track. *In Proceedings of The Twenty-Eigth Text REtrieval Conference*.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 405–414.

Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Mitra, B., Zihayat, M., & Bagheri, E. (2023). De-biasing relevance judgements for fair ranking. *Proceedings of the 45th European Conference on Information Retrieval*, 350–358.

Bigdeli, A., Arabzadeh, N., SeyedSalehi, S., Zihayat, M., & Bagheri, E. (2022). Gender fairness in information retrieval systems. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3436–3439.

Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., & Bagheri, E. (2021). On the orthogonality of bias and utility in ad hoc retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1748–1752.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bogen, M., Rieke, A., & Ahmed, S. (2020). Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. *Proceedings of FAccT '20: Conference on Fairness, Accountability, and Transparency*, 492–500.

Bower, A., Lum, K., Lazovich, T., Yee, K., & Belli, L. (2022). Random isn't always fair: Candidate set imbalance and exposure inequality in recommender systems. *arXiv preprint arXiv:2209.05000*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Proceedings of Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Bunse, M. (2022a). On multi-class extensions of adjusted classify and count. *Proceedings of the 2nd International Workshop on Learning to Quantify (LQ 2022)*, 43–50.

Bunse, M. (2022b). Unification of algorithms for quantification and unfolding. *52. Jahrestagung der Gesellschaft für Informatik, Informatik in den Naturwissenschaften*, *P-326*, 459–468.

Callan, J. (2002). Distributed information retrieval. In *Advances in information retrieval: Recent research from the center for intelligent information retrieval* (pp. 127–150). Springer.

Carbonell, J. G., & Goldstein, J. (2017). The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum*, *51*(2), 209–210.

Carmel, D., & Yom-Tov, E. (2010). *Estimating the query difficulty for information retrieval*. Morgan & Claypool Publishers.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, *44*(1), 1–50.

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2021). Fair classification with noisy protected attributes: A framework with provable guarantees. *Proceedings of the 38th International Conference on Machine Learning*, 1349–1361.

Celis, L. E., Straszak, D., & Vishnoi, N. K. (2018). Ranking with fairness constraints. *Proceedings of 45th International Colloquium on Automata, Languages, and Programming.*, 28:1–28:15.

Chen, F., & Fang, H. (2023). Learn to be fair without labels: A distribution-based learning framework for fair ranking. *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 23–32.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proceedings of FAT*19 the Conference on Fairness, Accountability, and Transparency*, 339–348.

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. *Proceedings of the 2018 Conference on Human Factors in Computing Systems*, 651.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1–113.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., ... Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*, 70:1–70:53.

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *Proceedings of the 8th International Conference on Learning Representations*.

Craswell, N., Zoeter, O., Taylor, M. J., & Ramsey, B. (2008). An experimental comparison of click position-bias models. *Proceedings of the International Conference on Web Search and Web Data Mining*, 87–94.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299–306.

Dang, V., & Croft, W. B. (2012). Diversity by proportionality: An election-based approach to search result diversification. *Proceedings of the 35th International ACM SIGIR conference on research and development in Information Retrieval*, 65–74.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., & Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 275–284.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, *56*(293), 52–64.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, 214–226.

Ekstrand, M. D., Burke, R., & Diaz, F. (2019). Fairness and discrimination in retrieval and recommendation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1403–1404.

Ekstrand, M. D., Das, A., Burke, R., Diaz, F., et al. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, *16*(1-2), 1–177.

Ekstrand, M. D., McDonald, G., Raj, A., & Johnson, I. (2022). Overview of the trec 2022 fair ranking track. *Proceedings of the Thirty-First Text REtrieval Conference*.

Ekstrand, M. D., Raj, A., McDonald, G., & Johnson, I. (2021). Overview of the TREC 2021 fair ranking track. *Proceedings of the Thirtieth Text REtrieval Conference*.

Esuli, A., Fabris, A., Moreo, A., & Sebastiani, F. (2023). *Learning to quantify* (Vol. 47). Springer.

Faggioli, G., Formal, T., Marchesin, S., Clinchant, S., Ferro, N., & Piwowarski, B. (2023). Query performance prediction for neural IR: are we there yet? *Proceedings of the 45th European Conference on Information Retrieval*, 232–248.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Formal, T., Piwowarski, B., & Clinchant, S. (2021). SPLADE: sparse lexical and expansion model for first stage ranking. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2288–2292.

Forman, G. (2005). Counting positives accurately despite inaccurate classification. *Proceedings of the 16th European Conference on Machine Learning*, 564–575.

Frayling, E., MacAvaney, S., Macdonald, C., & Ounis, I. (2024). Effective adhoc retrieval through traversal of a query-document graph. *Proceedings of the 46th European Conference on Information Retrieval*, 89–104.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136–143.

Fuglede, B., & Topsøe, F. (2004). Jensen-shannon divergence and hilbert space embedding. *Proceedings of the 2004 IEEE International Symposium on Information Theory*, 31.

Garba, A., Wu, S., & Khalid, S. (2023). Federated search techniques: An overview of the trends and state of the art. *Knowledge and Information Systems*, *65*(12), 5065–5095.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data mining*, 2221–2231.

Ghazimatin, A., Kleindessner, M., Russell, C., Abedjan, Z., & Golebiowski, J. (2022). Measuring fairness of rankings under noisy sensitive information. *Proceedings og FAccT '22: Conference on Fairness, Accountability, and Transparency*, 2263–2279.

Ghosh, A., Dutt, R., & Wilson, C. (2021). When fair ranking meets uncertain inference. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1033–1043.

Ghosh, A., Kvitca, P., & Wilson, C. (2023). When fair classification meets noisy protected attributes. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 679–690.

Gonzalez, P., Moreo, A., & Sebastiani, F. (2024). Binary quantification and dataset shift: An experimental investigation. *Data Mining and Knowledge Discovery*, 1–43.

Gospodinov, M., MacAvaney, S., & Macdonald, C. (2023). Doc2query–: When less is more. *Proceedings of the 45th European Conference on Information Retrieval*, 414–422.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the Annual Conference on Neural Information Processing Systems 2016*, 3315–3323.

Hauff, C., Azzopardi, L., Hiemstra, D., & de Jong, F. (2010). Query performance prediction: Evaluation contrasted with effectiveness. *Proceedings of the 32nd European Conference on IR Research*, *5993*, 204–216.

Hauff, C., Hiemstra, D., & de Jong, F. (2008). A survey of pre-retrieval query performance predictors. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 1419–1420.

He, B., & Ounis, I. (2004). Inferring query performance using pre-retrieval predictors. *Proceddings of String Processing and Information Retrieval*, *3246*, 43–54.

He, B., & Ounis, I. (2006). Query performance prediction. *Information Systems*, *31*(7), 585–594.

Heuss, M., Sarvi, F., & de Rijke, M. (2022). Fairness of exposure in light of incomplete exposure estimation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 759–769.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hofstätter, S., & Hanbury, A. (2019). Let's measure run time! extending the ir replicability infrastructure to include performance aspects. *Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 12–16.

Hofstätter, S., Lin, S., Yang, J., Lin, J., & Hanbury, A. (2021). Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 113–122.

Holstein, K., Vaughan, J. W., III, H. D., Dudík, M., & Wallach, H. M. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 600.

Jaleel, N. A., Allan, J., Croft, W. B., Diaz, F., Larkey, L. S., Li, X., Smucker, M. D., & Wade, C. (2004). Umass at TREC 2004: Novelty and HARD. *Proceedings of the Thirteenth Text REtrieval Conference*.

Jänich, T., McDonald, G., & Ounis, I. (2023). Colbert-fairprf: Towards fair pseudo-relevance feedback in dense retrieval. *Proceedings of the 45th European Conference on Information Retrieval*, 457–465.

Jänich, T., McDonald, G., & Ounis, I. (2024a). Fairness-aware exposure allocation via adaptive reranking. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1504–1513.

Jänich, T., McDonald, G., & Ounis, I. (2024b). Improving exposure allocation in rankings by query generation. *Proceedings of the 46th European Conference on Information Retrieval*, 121–129.

Jänich, T., McDonald, G., & Ounis, I. (2024c). Query exposure prediction for groups of documents in rankings. *Proceedings of the 46th European Conference on Information Retrieval*, 143–158.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S. H., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 6769–6781.

Kaur, J., & Buttar, P. K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *4*(4), 207–210.

Keikha, A., Ensan, F., & Bagheri, E. (2018). Query expansion using pseudo relevance feedback on wikipedia. *Journal of Intelligent Information Systems*, *50*, 455–478.

Khattab, O., & Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48.

Kirnap, Ö., Diaz, F., Biega, A., Ekstrand, M. D., Carterette, B., & Yilmaz, E. (2021). Estimation of fair ranking metrics with incomplete judgments. *Proceedings of the Web Conference*, 1065–1075.

Kletti, T., Renders, J., & Loiseau, P. (2022). Pareto-optimal fairness-utility amortizations in rankings with a DBN exposure model. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 748–758.

Kopeinik, S., Mara, M., Ratz, L., Krieg, K., Schedl, M., & Rekabsaz, N. (2023). Show me a" male nurse"! how gender bias is reflected in the query formulation of search engine users. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15.

Krieg, K., Parada-Cabaleiro, E., Schedl, M., & Rekabsaz, N. (2022). Do perceived gender biases in retrieval results affect relevance judgements? *Proceedings of advances in Bias and Fairness in Information Retrieval - Third International Workshop*, 104–116.

Krovetz, R., & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, *10*(2), 115–141.

Kuhlman, C., Gerych, W., & Rundensteiner, E. A. (2021). Measuring group advantage: A comparative study of fair ranking metrics. *Proceedings of the Conference on AI, Ethics, and Society*, 674–682.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *Proceedings of the 8th International Conference on Learning Representations*.

Levy, S. (2021). *In the plex: How google thinks, works, and shapes our lives*. Simon & Schuster.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Li, H., Xu, J., et al. (2014). Semantic matching in search. *Foundations and Trends® in Information Retrieval*, *7*(5), 343–469.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, *33*(12), 6999–7019.

Lin, J., Nogueira, R., & Yates, A. (2022). *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

LinkedIn. (2024). Linkedin settings [Accessed: 2024-08-06].

Lipton, Z. C., Wang, Y., & Smola, A. J. (2018). Detecting and correcting for label shift with black box predictors. *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 3128–3136.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*.

Lv, Y., & Zhai, C. (2010). Positional relevance model for pseudo-relevance feedback. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 579–586.

MacAvaney, S., Nardini, F. M., Perego, R., Tonellotto, N., Goharian, N., & Frieder, O. (2020). Expansion via prediction of importance with contextualization. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 1573–1576.

MacAvaney, S., Tonellotto, N., & Macdonald, C. (2022). Adaptive re-ranking with a corpus graph. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1491–1500.

MacAvaney, S., Yates, A., Cohan, A., & Goharian, N. (2019). CEDR: contextualized embeddings for document ranking. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1101–1104.

Macdonald, C., Santos, R. L., & Ounis, I. (2013). The whens and hows of learning to rank for web search. *Information Retrieval Journal*, *16*, 584–628.

Macdonald, C., & Tonellotto, N. (2020). Declarative experimentation in information retrieval using pyterrier. *Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval*, 161–168.

Macdonald, C., Tonellotto, N., & Ounis, I. (2021). On single and multiple representations in dense passage retrieval. *Proceedings of the 11th Italian Information Retrieval Workshop 2021*, *2947*.

Mackie, I., Chatterjee, S., & Dalton, J. (2023). Generative relevance feedback with large language models. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026–2031.

McDonald, G., Macdonald, C., & Ounis, I. (2022). Search results diversification for effective fair ranking in academic search. *Information Retrieval Journal*, *25*(1), 1–26.

Mitra, B., Craswell, N., et al. (2018). An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, *13*(1), 1–126.

Moreo, A., González, P., & del Coz, J. J. (2024). Kernel density estimation for multiclass quantification. *CoRR*, *abs/2401.00490*.

Morik, M., Singh, A., Hong, J., & Joachims, T. (2020). Controlling fairness and bias in dynamic learning-to-rank. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 429–438.

Mozannar, H., Ohannessian, M. I., & Srebro, N. (2020). Fair learning with private demographic data. *Proceedings of the 37th International Conference on Machine Learning*, 7066–7075.

Naseri, S., Dalton, J., Yates, A., & Allan, J. (2021). CEQE: contextualized embeddings for query expansion. *Proceedings of the 43rd European Conference on IR Research*, 467–482.

Nogueira, R. F., & Cho, K. (2019). Passage re-ranking with BERT. *CoRR*, *abs/1901.04085*.

Nogueira, R. F., Jiang, Z., Pradeep, R., & Lin, J. (2020). Document ranking with a pre-trained sequence-to-sequence model. *Findings of the Association for Computational Linguistics: EMNLP*, 708–718.

Nogueira, R. F., Lin, J., & Epistemic, A. (2019). From doc2query to doctttttquery. *Online preprint*, *6*, 2.

Nogueira, R. F., Yang, W., Cho, K., & Lin, J. (2019). Multi-stage document ranking with BERT. *CoRR*, *abs/1910.14424*.

Nogueira, R. F., Yang, W., Lin, J., & Cho, K. (2019). Document expansion by query prediction. *CoRR*, *abs/1904.08375*.

OpenAI. (2023). GPT-4 technical report. *CoRR*, *abs/2303.08774*.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. *Proceedings of the SIAM International Conference on Data Mining*, 581–592.

Penha, G., Palumbo, E., Aziz, M., Wang, A., & Bouchard, H. (2023). Improving content retrievability in search with controllable query generation. *Proceedings of the ACM Web Conference 2023*, 3182–3192.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Radlinski, F., Kleinberg, R., & Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. *Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, 784–791.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Raghavan, M., Barocas, S., Kleinberg, J. M., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of FAT*'20: Conference on Fairness, Accountability, and Transparency*, 469–481.

Raj, A., & Ekstrand, M. D. (2022). Measuring fairness in ranked results: An analytical and empirical comparison. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 726–736.

Raj, A., & Ekstrand, M. D. (2024). Towards optimizing ranking in grid-layout for provider-side fairness. *Proceedings of the 46th European Conference on Information Retrieval*, 90–105.

Raj, A., Wood, C., Montoly, A., & Ekstrand, M. D. (2020). Comparing fair ranking metrics. *CoRR*, *abs/2009.01311*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3980–3990.

Rekabsaz, N., Kopeinik, S., & Schedl, M. (2021). Societal biases in retrieved contents: Measurement framework and adversarial mitigation of BERT rankers. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 306–316.

Rekabsaz, N., & Schedl, M. (2020). Do neural ranking models intensify gender bias? *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2065–2068.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. *Proceedings of The Third Text REtrieval Conference*, 109–126.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Santos, R. L. T., Peng, J., Macdonald, C., & Ounis, I. (2010). Explicit search result diversification through sub-queries. *Proceedings of the 32nd European Conference on IR Research*, 87–99.

Sapiezynski, P., Zeng, W., Robertson, R. E., Mislove, A., & Wilson, C. (2019). Quantifying the impact of user attentionon fair group representation in ranked lists. *Companion Proceedings of The 2019 World Wide Web Conference*, 553–562.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. M. (2012). On causal and anticausal learning. *Proceedings of the 29th International Conference on Machine Learning*.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*, 657–680.

Sebastiani, F. (2020). Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, *23*(3), 255–288.

Shokouhi, M., Si, L., et al. (2011). Federated search. *Foundations and trends® in information retrieval*, *5*(1), 1–102.

Singh, A., & Joachims, T. (2018). Fairness of exposure in rankings. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2219–2228.

Singhal, A., et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Engeneering Bulletin*, *24*(4), 35–43.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11–21.

Storkey, A. (2009). When training and test sets are different: Characterizing learning transfer. In *Dataset shift in machine learning* (pp. 3–28). The MIT Press.

Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. *CoRR*, *abs/2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S. L., Guo, W., Narasimhan, H., Cotter, A., Gupta, M. R., & Jordan, M. I. (2020). Robust optimization for fairness with noisy protected groups. *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Wang, X., MacAvaney, S., Macdonald, C., & Ounis, I. (2023). Generative query reformulation for effective adhoc search. *Proceedings of Gen-IR @ SIGIR 2023 Workshop*.

Wang, X., Macdonald, C., & Ounis, I. (2022). Improving zero-shot retrieval using dense external expansion. *Information Processing & Management*, *59*(5), 103026.

Wang, X., Macdonald, C., Tonellotto, N., & Ounis, I. (2021). Pseudo-relevance feedback for multiple representation dense retrieval. *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, 297–306.

Wang, X., Macdonald, C., Tonellotto, N., & Ounis, I. (2023). Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, *17*(1), 1–39.

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., & Najork, M. (2018). Position bias estimation for unbiased learning to rank in personal search. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 610–618.

Wilkie, C., & Azzopardi, L. (2014). Best and fairest: An empirical analysis of retrieval system bias. *Proceedings of the 36th European Conference on IR Research*, 13–25.

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L. J., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. *Proceedings of FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677.

Wu, H., Mitra, B., Ma, C., Diaz, F., & Liu, X. (2022). Joint multisided exposure fairness for recommendation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 703–714.

Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P. N., Ahmed, J., & Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Proceedings of the 9th International Conference on Learning Representations*.

Xu, Y., Jones, G. J. F., & Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 59–66.

Yan, R., Hauptmann, A. G., & Jin, R. (2003). Multimedia search with pseudo-relevance feedback. *Proceedings of Image and Video Retrieval, Second International Conference, CIVR 2003, Urbana-Champaign, IL, USA, July 24-25, 2003, Proceedings*, 238–247.

Yang, E., Jänich, T., Mayfield, J., & Lawrie, D. J. (2024). Language fairness in multilingual information retrieval. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2487–2491.

Yang, K., & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 22:1–22:6.

Yang, T., Xu, Z., Wang, Z., & Ai, Q. (2023). FARA: future-aware ranking algorithm for fairness optimization. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2906–2916.

Yu, H., Xiong, C., & Callan, J. (2021). Improving query representations for dense retrieval with pseudo relevance feedback. *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 3592–3596.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). Fa* ir: A fair top-k ranking algorithm. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569–1578.

Zehlike, M., & Castillo, C. (2020). Reducing disparate exposure in ranking: A learning to rank approach. *Proceedings of the Web Conference 2020*, 2849–2855.

Zehlike, M., Yang, K., & Stoyanovich, J. (2022). Fairness in ranking, part i: Score-based ranking. *ACM Computing Surveys*, *55*(6), 1–36.

Zhao, W. X., Liu, J., Ren, R., & Wen, J.-R. (2024). Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, *42*(4), 1–60.