University of
Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization:
Applied Data Science

Spring semester, 2023

Open access

Writer: Maria Eriksen

Maria Eriksen
(Writer's signature)

Faculty supervisor: Krisztian Balog, Nolwenn Marie Emilie Bernard

External supervisor(s):

Thesis title:
Evaluating Fairness in Information Retrieval Systems:
A Study on the Performance of the FAIR Metric

Credits (ECTS): 30

Key words:
Information Retrieval
Fairness
Utility

Pages: 57

Stavanger, 15.06.2023

Front page for master thesis
Faculty of Science and Technology
Decision made by the Dean October 30th 2009

Faculty of Science and Technology
Department of Electrical Engineering and Computer Science

Evaluating Fairness in Information Retrieval Systems: A Study on the Performance of the FAIR Metric

Master's Thesis in Computer Science
by

Maria Eriksen

Internal Supervisors

Krisztian Balog

Nolwenn Marie Emilie Bernard

June 15, 2023

Declaration of Authorship

I, Maria Eriksen, declare that this thesis and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master's degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:
Maria Eriksen

Date:
June 15, 2023

“Learning is the only thing the mind never exhausts, never fears, and never regrets.”

Leonardo da Vinci, inventor and polymath.

Abstract

In the digital era, the use of information retrieval (IR) technologies has surged, enabling users to access vast amounts of data quickly. However, concerns have arisen regarding bias and unfairness in these systems, leading to unequal treatment and outcomes for different user groups, such as racial, gender, or socioeconomic bias. To ensure equitable access to information, it is crucial to establish a fair IR system.

This thesis focuses on the FAIR metric proposed by Gao et al. [1] and investigates its performance by replacing standard information retrieval (IR) utility metrics. It explores the impact of different metrics on the overall performance of FAIR, providing a comprehensive analysis of its effectiveness in evaluating fairness in IR systems.

Acknowledgements

I would like to thank my supervisors for their fantastic enthusiasm and help with writing this thesis. Words cannot express my gratitude to my professor Krisztian Balog, Professor of Computer Science at the University of Stavanger, who generously provided knowledge and expertise. I would also like to thank Nolwenn Marie Emilie Bernard for her invaluable patience and feedback, without which I could not have embarked on this voyage.

Thanks should also go to the classmates, librarians, research assistants, and study participants from the university, who impacted and inspired me.

It would be improper of me not to mention my family. Their belief in me has kept my spirits and motivation high during this process.

Abbreviations

Acronym	Description
AP	Average Precision
BM25	Best Match 25
DCG	Discounted Cumulative Gain
ERR	Expected Reciprocal Rank
ERR-IA	Intent-Aware Expected Reciprocal Rank
IA	Intent-Aware
IDF	Inverse Document Frequency
IR	Information Retrieval
IRM	Information Retrieval Metric
KL	Kullback-Leibler
MAP	Mean Average Precision
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
NRBP	Novelty- and Rank-Biased Precision
RBP	Rank-Biased Precision
RPI	Recidivism Prediction Instrument
S2ORC	Semantic Scholar Open Corpus
TF-IDF	Term Frequency-Inverse Document Frequency
TREC	Text REtrieval Conference
UDC	Universal Decimal Classification

Contents

Chapter 1

Introduction

In the current digital era, the use of information retrieval (IR) technologies has increased significantly. These systems allow users to rapidly and simply find enormous amounts of data to identify pertinent information [2]. The possibility of bias and unfairness in IR systems, that can result in unequal treatment and outcomes for various user groups (e.g., racial, gender or socioeconomic bias [3–5]), is causing significant concern.

Therefore, it is important to establish a fair IR system in order to guarantee that all users have fair access to information. Similar to Gao et al. [1], this thesis defines fairness as a subjective moderation of the ratio between different groups. Fairness entails creating a system that does not discriminate against specific users because of sensitive attributes, like gender or age. Bias in data collection creates distortions that prevent the information from accurately representing the situation under investigation [6].Fairness may then be defined as lack of bias [7].

The following subsections give the reader a general overview on the topic from different perspectives.

1.1 Historical perspectives of the information retrieval

The history of information retrieval (IR) systems can be traced back to the early twentieth century, long before the digital era, when the development of new information technologies and techniques created a growing demand for tools to manage and access massive amounts of data [2].

The Universal Decimal Classification (UDC) [8], developed in the early twentieth century by Belgian bibliographers Paul Otlet and Henri La Fontaine, is one of the earliest examples of an IR system. The UDC is a knowledge organization system that uses a hierarchical classification scheme to allow users to quickly find information on a specific topic [9].

The development of the first computer-based searching systems did not start until the late 1940s. During the Opening Plenary Session of The Royal Society Scientific Information Conference in 1948, J. E. Holmstrom mentioned a pre production version of Univac [10], a machine that is able to find text references associated with a subject code, both stored on magnetic steel tapes. It was expected to work at a rate of 120 words per minute and was a Boolean retrieval system [10].

Further development of computers and the emergence of electronic databases in the 1950s and 1960s resulted in the development of new, increasingly automated IR systems. In this period, the research proceeds within two topics, indexing and ranking, making it possible to create the first search engines accepting free-text queries [11]. The 1970s and 1980s saw the emergence of new IR techniques such as vector space models and probabilistic retrieval models. These models enabled IR systems to better understand the relationship between queries and documents, further improving the accuracy and relevance of search results [11]. Some of the ranking models that were developed in the 1980s are still commonly used today, such as BM25 [12].

In the late 1980s and early 1990s, academics worried that the size of document collections for testing were small compared to those of commercial search companies. This concern led to the formation of the Text REtrieval Conference (TREC), an annual event where several international research groups have been working to construct larger test collections than before [13].

The old weighting and ranking functions were not ideal for these new data sets. It became apparent that distinct collections required different ranking and weighting methods. Web search engines in the late 1990s proved this [10]. Additionally, the growth of the internet and the explosion of digital data in the 1990s and 2000s led to the development of new IR systems designed for web search and e-commerce [10]. Search engines such as Google became dominant players in the IR market, using advanced algorithms to deliver highly relevant and personalized search results.

Today, IR systems are ubiquitous, powering everything from search engines and digital assistants to recommendation systems and chatbots [14]. The field continues to evolve, with ongoing research into new techniques and approaches to improve the accuracy, relevance, and fairness of IR systems.

1.2 Background for research

Search engines determines the results for a given query. It is reasonable to expect search engines to be fair and impartial because they expose many individuals to information. However, search engine results do not always cover all perspectives on a search query topic, and they can be biased towards a particular viewpoint [15]. This happens because search engine results are returned based on relevance, which is calculated using many features and sophisticated algorithms where search neutrality is not always the focus.

In recent years, a lot of studies have proven that search engines are biased [4, 5, 16–18]). Pitoura et al. [7] mentions several studies proving systematic under representation of certain genders [19], ethnic [20] or other stereotypical groups.

There are far more examples. Angwin et al. [21] refers to a computer program that predicts the likelihood of committing a crime in the future based on the criminal's race rather than the criminal's previous crime record [21]. In this case, the program's bias may have consequences for the defendants' freedom because judges in several states in the United States use these scores during sentencing. Sadly, regardless of computer programs, judges tend to regard some stereotypical groups of people as having a high risk of repeat criminal behavior [22].

Therefore, it is essential to develop a fair information retrieval system; some of the reasons are listed below:

- Promoting equity: A fair IR system ensures that all users have equal access to information and that their search results are not influenced or skewed by characteristics such as their race, gender, or socioeconomic level [7]. This fosters equity and aids in the reduction of inequities in information availability.
- Preventing injury: An unfair IR system can affect people by maintaining stereotypes, reinforcing biases, and restricting access to critical knowledge. We may avoid these undesirable repercussions and ensure that our technology is not unintentionally contributing to societal harms by developing a fair IR system.
- Fulfilling legal and ethical obligations: Several nations have laws and regulations that make it illegal to discriminate on the basis of protected traits such as race, gender, and religion [22]. We can verify that we are achieving these legal requirements and respecting ethical standards by developing a fair IR system.

Summarizing the points above - developing a fair and relevant IR system is critical since it promotes equity, accuracy, and avoidance of harm while also meeting legal and ethical requirements.

1.3 Problem Statement

In the context of information retrieval systems, achieving fairness is a significant concern due to the potential for biased and unequal treatment of various user groups. For example, when users initiate search requests, the IR system's ranking algorithm may prioritize items from the largest or most prominent subtopic or aspect group, thereby neglecting minority aspects and resulting in a skewed discovery of information [23, 24]. Moreover, the positive feedback loop created by users' tendency to click on top results can perpetuate unfairness in rankings and recommendations.

To address these issues, a diversification framework can be employed to increase the coverage of topical aspects and enhance fairness. However, it is important to distinguish between the goals of diversity and fairness. While diversity focuses on serving users' interests by providing relevant information across different topics, fairness aims to moderate the exposure of information and resources to ensure equitable attention Singh and Joachims [25]. Consequently, fairness and diversity can sometimes be competing factors in IR systems Porcaro et al. [26].

Existing evaluation metrics primarily focus on individual factors such as relevance, diversity, and novelty, but lack a comprehensive assessment of overall fairness and utility. Evaluating an algorithm's effectiveness in achieving fairness while considering utility becomes crucial. However, the absence of a standard metric capturing such trade-offs poses challenges in comparing different algorithms.

The lack of a unified metric not only complicates evaluation but also hinders fairness optimization. Optimizing solely for utility may yield biased results, while prioritizing fairness can lead to decreased utility. Hence, algorithms that jointly optimize both factors are necessary. In this context, Gao et al. [1] proposed a novel fairness-aware evaluation metric, FAIR, to measure fairness and user utility simultaneously, allowing for a balanced assessment. This metric offers a new perspective for evaluating fairness-aware ranking results and can be generalized to various IR applications.

Gao et al. [1] suggested further research and improvement of FAIR; the suggestions will be presented, implemented and discussed in this thesis.

What impact can different utility metrics have on the performance of FAIR?

The purpose of this thesis is to study the important factors for designing a fair IR system. Section 2 presents the related work. Section 3 presents and describes the chosen baseline and an advanced method. The development process and the results are described and analyzed in Section 4. Section 5 summarizes the study and makes recommendations for future research.

Chapter 2

Related Work

The work presented in the thesis is strongly related to the topics of bias and fairness measurements in the context of information retrieval, standard evaluation metrics and post-processing fairness re-ranking algorithms. The work is building upon FAIR ε-greedy [1].

This chapter presents an overview of previous studies and research conducted in the field of fairness. By examining existing scholarship, valuable insights are gained, research gaps are identified, and the study is positioned within the broader context. Through a synthesis of literature, key themes, methodologies, and findings are explored, informing the research design and approach. The aim is to contribute to the field by addressing research gaps and advancing understanding of fairness.

2.1 Evaluation metrics in Information Retrieval

Traditional evaluation measures in information retrieval (IR) systems primarily focus on the quality and utility of search and recommendation results from the user's perspective. These metrics, such as Mean Average Precision (MAP), Discounted Cumulative Gain (DCG), and Rank-Biased Precision (RBP)[27], assess the relevance of the overall results to the user's information need, considering factors like precision, recall, and the position of relevant results.

Moffat and Zobel [27] introduced the evaluation metric called Rank-Biased Precision (RBP) to address limitations in existing measures for assessing information retrieval systems. The metric overcomes issues with recall that is not a reliable measure of user satisfaction because real system users do not have access to the complete set of relevant documents [28], and average precision that is derived from recall, hence, inherits the same limitations. It provides a more accurate evaluation of user satisfaction and improves stability properties. It also allows for quantifying experimental uncertainty when partial relevance judgments are available.

In addition to utility-based metrics, diversity and novelty metrics are used to address the ambiguity of multiple aspects of user queries. For example:

$\boldsymbol{\alpha}$-nDCG[29], prioritizes novelty and diversity in evaluation and rewards documents that are relevant to a wide range of novel aspects based on previously selected documents. As seen in Clarke et al. [29], the approach demonstrated effectiveness on a TREC question answering track-based test collection.

Intent-Aware (IA) metrics [30], such as Intent-Aware Expected Reciprocal Rank (ERR-IA) [31], assume a probability distribution over user intents for a query.

Agrawal et al. [30] proposed a systematic approach for diversifying search results - both web queries and documents can belong to multiple categories within a taxonomy-based information framework. Empirical evidence showed that algorithm outperformed commercial search engines based on the generalized metrics (traditional IR metrics such as Normalized Discounted Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR), and mean average precision (MAP) were extended to explicitly incorporate the value of diversification).

Chapelle et al. [31] introduced a novel editorial metric Intent-Aware Expected Reciprocal Rank (ERR-IA), which addresses the limitations of DCG. This metric accounts for the positioning of documents and implicitly discounts less relevant documents shown below highly relevant ones. ERR-IA was evaluated using query logs from a commercial search engine, and demonstrated its superior correlation with click metrics compared to other editorial metrics.

While these traditional metrics assess the utility of IR systems in fulfilling user information needs, they do not directly capture the notion of fair information exposure. They are unable to measure or evaluate the fairness of the system in terms of information distribution and exposure to different user groups.

2.2 Fairness notions and metrics

Previous studies have extensively examined fairness in information retrieval systems ([32–34]), particularly regarding the exposure of information related to sensitive attributes like gender and race. For example, Singh and Joachims [25] implements a framework that accommodates various fairness constraints, such as demographic parity and disparate impact. The objective is to ensure that different items receive equal exposure or exposure proportional to their utilities or impacts, based on the system designer's definition of fair distribution.

This work builds upon existing research ([35–37]) by considering topical aspects as sensitive attributes and focusing on fairness within the context of diversity, thereby connecting it with standard IR metrics.

Celis et al. [35] present a scalable algorithm with provable guarantees that maximizes utility while satisfying user-defined constraints. The framework addresses the problem of personalization in online platforms, which can lead to efficiency and higher revenue but also propagate biases and polarization; the framework allows users to constrain content selection and reduce polarization in personalized systems.

Diaz et al. [36] introduce the concept of expected exposure as a measure of attention received by ranked items. The work advocates for equal expected exposure, where all items of the same relevance grade should receive similar attention. This principle is beneficial for diverse retrieval objectives and fair ranking. Diaz et al. [36] propose an evaluation methodology based on expected exposure, relaxing traditional assumptions in information retrieval. This allows systems to produce distributions of rankings instead of fixed rankings.

Gao and Shah [37] address concerns regarding bias in search engines and the social responsibilities of information retrieval systems. The focus of the research is on achieving fairness in top-k diversity ranking, specifically statistical parity fairness and disparate impact fairness.

Numerous efforts have been made to propose fairness metrics in IR, which involve constraints such as distance and ratio between the proportion of a protected attribute and the overall attribute proportion (Yang and Stoyanovich [38]), pairwise comparisons considering utility and prediction errors (Beutel et al. [39]; Kuhlman et al. [40]; Yao and Huang [41]), and exposure distributions compared to the desired distribution (Geyik et al. [42]; Yang and Stoyanovich [38]).

Yang and Stoyanovich [38] introduce fairness measures for rankings and analyze them using controlled data. These measures were applied to real datasets, identifying instances of bias. Additionally, the work demonstrate how incorporating fairness measures into an optimization framework can improve fairness without compromising accuracy.

Beutel et al. [39] introduce novel metrics to assess algorithmic fairness in recommender systems. By measuring fairness through pairwise comparisons from randomized experiments, the metrics provide a practical approach to evaluate rankings. Additionally, it proposes a new regularisation method to enhance fairness during model training and improve rankings. Applying this regularization to a large-scale recommender system, Beutel et al. [39] demonstrate a significant improvement in pairwise fairness.

Kuhlman et al. [40] propose error-based fairness criteria for rankings, including Rank Equality, Rank Calibration, and Rank Parity. These criteria cover a range of fairness considerations and use rank-appropriate error metrics based on pairwise discordance.

Geyik et al. [42] propose measures to assess bias related to protected attributes and algorithms for fairness-aware re-ranking of results. The framework aims to achieve desired distributions of top-ranked results based on protected attributes, promoting fairness criteria such as equality of opportunity and demographic parity.

Yao and Huang [41] examine fairness in collaborative-filtering recommender systems, which can be affected by discriminatory patterns in historical data. Biased data may result in unfair predictions against minority user groups. To address this, authors introduce four new fairness metrics that tackle various forms of unfairness such as underrepresentation of women in science, technology, engineering, and mathematics. These metrics can be optimized by incorporating fairness terms into the learning objective.

Sapiezynski et al. [43] introduce a novel metric for auditing group fairness in ranked lists. It offers improved modeling of user attention and accommodates non-binary protected attributes. By considering the human factors and the entire sociotechnical system, we provide a more comprehensive evaluation of fairness.

Several studies have compared existing fairness metrics:

Chouldechova [44] examines fairness criteria applied to Recidivism Prediction Instruments (RPIs) and demonstrates that it is challenging to satisfy all criteria when recidivism rates vary among groups. It also highlights how disparate impact can occur when an RPI does not achieve error rate balance.

Garg et al. [45] present commonly used fairness metrics within a shared mathematical framework and provide new insights into their relationships. Metrics such as statistical parity, equalized odds, and predictive parity for comparing binary predictions and outcomes can only be simultaneously satisfied under limited conditions and require careful consideration of base rate ratios. Equalized odds and predictive parity are incompatible when base rates differ, and equalized odds and statistical parity can only be met when true positive and false positive rates are equal.

Hinnefeld et al. [46] approach bias detection as a causal inference problem using observational data. Their study identifies sampling bias and label bias as the main causes of bias and evaluate the effectiveness of six fairness metrics in detecting each type of bias. Based on the findings, Hinnefeld et al. [46] propose best practice guidelines for selecting the most suitable fairness metric to detect bias. When ground truth positive rates differ between classes, it is challenging to interpret fairness metric results. In imbalanced scenarios, practitioners should consider causal reasoning for metric selection, using Normalized Mutual Information for sample bias detection. Detecting label bias in imbalanced cases is difficult, and therefore Disparate Impact is not suitable for imbalanced scenarios.

Raj et al. [24] describe various fair ranking metrics in a unified notation, allowing for direct comparison of their assumptions, objectives, and design choices. The AWRF$\boldsymbol{\Delta}$ metric is recommended for single rankings as it supports multiple protected attributes and allows for adaptability in various aspects. The EED metric (introduced in Diaz et al. [36]) is suggested for achieving demographic parity in sequences, while EEL, EER from Diaz et al. [36], and IAA metrics from Biega et al. [23] are recommended for equal opportunity in sequences based on their compatibility with multinomial groups and soft associations.

In contrast to metrics that solely capture fairness, this work focuses on integrating fairness into a unified metric that also considers standard IR metrics. As a result, FAIR provides a better balance between utility and fairness. Chapter 3 offers a comprehensive and detailed description of FAIR.

Related work Liu et al. [47] proposes combining accuracy and fairness in the reward function for reinforcement learning, aiming to maintain the accuracy-fairness trade-off in interactive recommendation systems. However, their focus is on optimization approaches rather than an evaluation metric that can be directly applied to measure a given rank list. This contribution lies in developing a specific metric for evaluating fairness-aware ranking. Another closely related work is Diaz et al. [36], which introduces expected exposure to capture fairness and user models based on ERR and RBP. In contrast, this work is situated in static ranking evaluation and accommodates various use cases without assuming stochastic ranking models.

2.3 Fairness ranking algorithms

Fairness ranking has been extensively explored as an optimization problem in various studies [25, 48, 49]. Many fairness ranking algorithms aim to optimize system utility while adhering to a set of fairness constraints.

Celis et al. [48] study the problem of ranking with fairness or diversity constraints. Contribution include approximation algorithms that consider constraints on attribute distribution while maximizing the rank quality metric. Unlike previous work, this algorithm runs efficiently and produces solutions with small constraint violations. These results are based on insights from constrained matching problems and common ranking metrics.

Zehlike et al. [49] introduce the Fair Top-k Ranking problem that aims to select the best candidates while ensuring group fairness. The definition of ranked group fairness ensures that the proportion of protected candidates remains statistically above a given threshold in every prefix of the top-k ranking. The algorithm produces fair top-k rankings and proves its effectiveness on various datasets, including German Credit[1] and SAT[2]. This is the first algorithm that uses statistical tests to address biases in the representation of under-represented groups in ranked lists.

These algorithms re-rank items based on fairness constraints using an existing ranking or estimation of item utility (Gao and Shah [37]), or incorporate fairness as a regularization in the objective function [35, 39, 50–53].

Asudeh et al. [51] present a system that helps users select criterion weights to achieve fair rankings. By representing ranking functions as points in a multi-dimensional space, it efficiently identifies regions that satisfy fairness criteria. The system provides feedback to users, indicating whether their proposed ranking function meets the desired fairness criteria and suggesting minimal modifications if needed.

Mehrotra et al. [50] propose a framework to evaluate recommendation policies that balance relevance and fairness without extensive A/B testing in the context of music streaming services. By considering user disposition towards fair content, the study identifies recommendation policies that improve supplier fairness without significant impact on user satisfaction.

Wan et al. [53] analyze the correlation between user interactions and product marketing images in e-commerce datasets and find that marketing strategies can introduce bias in collaborative filtering algorithms. To mitigate this bias and improve recommendation fairness without sacrificing accuracy, Wan et al. [53] propose a framework that addresses marketing bias in recommender systems.

Yao and Huang [52] examine fairness in collaborative-filtering recommender systems, where biased data can result in unfair predictions for minority users. The study introduce four new fairness metrics to address different forms of unfairness.

Several studies show that optimizing for one metric often leads to a decrease in another [25, 36, 50].

[1] https://archive.ics.uci.edu/

[2] https://www.qsleap.com/sat/resources/sat-2014-percentiles

In Diaz et al. [36] it is mentioned that increased exposure resulting from nonuniform ranking has negative consequences, including fairness concerns for content producers, potential impact on the quality of service for different user groups, the risk of overlooking important content, limited diversity in users' information experiences, and the promotion of rich-get-richer effects. Additionally, nonuniform exposure can shape users' perception of relevant information.

Singh and Joachims [25] propose a flexible framework for exploring fairness constraints in rankings, considering concepts such as demographic parity, disparate treatment, and disparate impact. It highlights the trade-off between user utility and the rights of ranked items and discusses the challenges related to individual fairness, estimated utilities, the cost of fairness in terms of effectiveness, and the feasibility of fair solutions in extreme conditions.

The correlation between fairness and system utility depends on the data characteristics, making constrained maximization algorithms potentially suboptimal in achieving fairness and utility. For example, Gao and Shah [34] address the problem of presence of bias in IR systems, leading to the optimization of utility and fairness constraints. The performance of optimization algorithms, however, depends on the data, making it important to understand the solution space and its effects. The research proposes a framework that efficiently estimates the solution space, and demonstrates the application of the framework in facilitating analyses and decision-making for optimizing fairness and relevance.

Similar to this work, Mehrotra et al. [50] and Celis et al. [35] propose randomized algorithms for jointly optimizing fairness and relevance. FAIR directly optimizes the proposed integrated metric, benefiting from its ability to capture the balance between utility and fairness as inferred from the evaluation metric. The algorithm is specifically designed for fairness ranking and can be generalized to the recommendation setting.

2.4 Methods for Achieving Fairness

Pitoura et al. [7] divided methods for generating fair ranking output into the three following categories:

- Pre-processing methods: These techniques aim to transform the data to address underlying bias or discrimination. They are generally agnostic to specific applications and focus on mitigating bias in the training data. The bias in the data may stem from decisions made during the data collection process or deviations from the intended use of the data.

- In-processing methods: This category involves modifying existing algorithms or developing new ones to achieve fair rankings and recommendations. The objective is to eliminate bias and discrimination during the model training process. In-processing methods incorporate fairness considerations into the objective function of the algorithm, such as through the inclusion of fairness terms or the imposition of fairness constraints. However, they do not provide explicit guarantees of fairness in the resulting ranked outputs.

- Post-processing methods: These techniques operate on the output of the ranking or recommendation algorithm. These methods will be described in detail in the following section

2.4.1 Postprocessing approaches

Biases or discriminatory patterns in the initial ranking can lead to unequal treatment or outcomes for various groups [7]. To address this challenge, postprocessing approaches have been developed, they take an initial ranking and a specification of fairness requirements as input. These approaches aim to produce a new ranking that satisfies the fairness criteria while respecting the original ranking as closely as possible.

One of the techniques focuses on interchanging positions within a ranking to address biases and promote fairness [54]. It involves identifying groups or individuals who are consistently ranked lower due to biases and swapping their positions with those ranked higher, while preserving the relative ordering of other items. This approach aims to mitigate the impact of discriminatory biases and ensure a more equitable ranking outcome. A good example of such rank swapping is the algorithms proposed by Gupta et al. [54].

Rank rescaling is an approach that involves adjusting the positions or scores of items in a ranking to address biases and promote fairness [7]. By redistributing the positions while maintaining the relative ordering, this approach aims to correct biased rankings while preserving the overall structure of the initial ranking.

FAIR can be categorized under another approach, which involves incorporating fairness constraints into the optimization process [7]. By formulating fairness requirements as constraints, the system can generate a new ranking that satisfies the specified fairness conditions while adhering to the original ranking as closely as possible. This method ensures fairness by explicitly considering and addressing the biases present in the initial ranking.

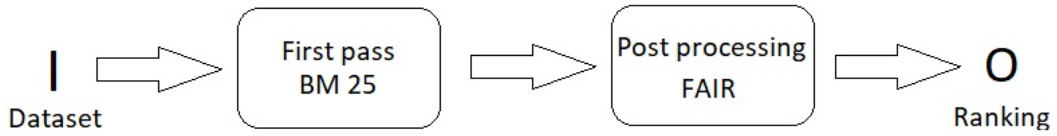
FAIR metric aims to unify fairness with standard information retrieval utility metrics, considering factors such as precision, diversity, novelty, and search intent. By incorporating fairness constraints into the ranking algorithm, FAIR ensures that the resulting ranking satisfies the specified fairness conditions while respecting the initial ranking to the extent possible[1].

Chapter 3

Approach

The forthcoming chapter presents a comprehensive and in-depth analysis of the approaches. Section 3.1 provides a detailed description of the chosen baseline, followed by a comprehensive depiction of the advanced method in section 3.2. Subsequently, this chapter concise portrayal of the datasets, followed by an elaborate description of the selected pre-procesing method.

The approach can be simplified into three main steps. First, the dataset undergoes preprocessing and indexing. Second, an initial ranking is obtained by applying the BM25 retrieval method. Finally, the FAIR post-processing algorithm is applied to generate a new ranking.



3.1 Baseline Method

Baseline method for this thesis is Fairness-Aware IR (FAIR) ε-greedy re-ranking algorithm introduced in Gao et al. [1]. The algorithm will be described in detail in section 3.3, while this section will mainly focus on the description of the FAIR metric.

FAIR metric addresses the limitations of fairness metrics described in Chapter 2 by unifying fairness with standard Information Retrieval (IR) utility metrics. Previously-mentioned metrics fail to consider important aspects such as precision, diversity, novelty, and search intent, resulting in an inability to capture the interplay between system utility and fairness. Consequently, they do not serve as comprehensive measures for evaluating system performance. Algorithms directly optimized for fairness metrics often sacrifice optimal utility, while those attempting to balance fairness and utility as separate and competing factors present their own challenges. In contrast, FAIR treats utility and fairness as integrated components within a unified metric. This design choice allows for a flexible approach that does not rely on assumptions about the relationship between utility and fairness. By optimizing FAIR, the trade-offs between utility and fairness can be effectively balanced in orthogonal scenarios, while simultaneously optimizing both aspects when they are interdependent.

The baseline FAIR score is calculated using the following formula[1]:

$$\text{FAIR} = \frac{1}{M} \sum_{i=1}^{k} \frac{\text{IRM}(i)}{d_{\text{KL}}(D_{\text{ri}}||D^*)} + 1$$

where M represents the total number of queries in the dataset, k denotes the rank position, i is the index variable ranging from 1 to M, $IRM(i)$ stands for the Information Retrieval Metric at position i, and $D_{\text{KL}}(D_{\text{ri}}||D^*)$ represents the Kullback-Leibler divergence between D_{ri} and D^*, capturing the difference or distance between the ranking distribution D_{ri} and the desired or target ranking distribution D^*.

Gao et al. [1] mention that FAIR aims to serve as a general fairness-aware metric, with the specific focus on intent or diversity captured by the IRM term. For example, IRM can be replaced with metrics like MAP and RBP in order to capture precision. To compute intent-aware FAIR, IRM can replaced with a metric like ERR-IA [31]. To emphasize diversity and novelty, IRM can be replaced with metrics like Novelty- and Rank-Biased Precision (NRBP) [55] and $\boldsymbol{\alpha}$-nDCG [29].

3.1.1 IRM

The computation of the FAIR metric involves the initial step of calculating the Information Retrieval Metric (IRM). For the baseline, fairness-aware metric will be designed by adopting the model from $\boldsymbol{\alpha}$-nDCG, as described by Gao et al. [1]. In this model, a query is regarded as ambiguous or comprising multiple aspects, where each aspect represents an information need. Diversity is quantified as the number of covered information needs, while novelty measures the amount of new information.

A detailed calculation formula for $\boldsymbol{\alpha}$-nDCG, which involves assessing the relevance of documents based on diversity aspects, computing cumulative gain, addressing position bias, and approximating the Ideal DCG to obtain the final score representing relevance and information gain considering the ranking position, is provided in Gao et al. [1] and Clarke et al. [29]. FAIR metric with $\boldsymbol{\alpha}$-nDCG may be written as follows:

$$\text{FAIR} = \frac{1}{\text{IDCG}} \sum_{i=1}^{k} \left(\frac{1}{\log_2(i+1)} \cdot \frac{G[i]}{\text{dKL}(D_{r_i}||D^*)} + 1 \right)$$

α-nDCG is a valuable metric for assessing and improving fairness in search and ranking systems and offers several advantages as a fairness-aware metric [29]. It considers multiple aspects within a ranking, allowing for fairness evaluation based on diverse dimensions of relevance. It provides flexibility in balancing relevance and diversity, accommodating specific requirements. α-nDCG captures the importance of ranking position, reflecting real-world impact. It is compatible with existing evaluation techniques and offers interpretability, facilitating discussions and decision-making processes.

3.1.2 KL divergence

Kullback-Leibler divergence is a versatile utility metric in information retrieval, providing a quantitative measure of similarity between user queries and document rankings Zhai and Massung [56]. Its ability to handle uncertainty and incorporate user feedback makes it a valuable tool for evaluating and improving the utility of IR systems.

Advantages of Kullback-Leibler (KL) divergence as a utility metric in information retrieval include its ability to measure the relevance between user queries and document rankings, its incorporation of user feedback for iterative improvement, and its handling of uncertainty through probabilistic models.

However, there are a few limitations to consider. KL divergence relies on accurate estimation of probability distributions that can be challenging in practice. It may not directly capture other aspects of utility, such as user satisfaction or task completion. Therefore, it is often used in combination with other metrics to provide a more comprehensive evaluation of information retrieval systems.

3.2 Advanced Method

The advanced method employed in this thesis involves substituting the Information Retrieval Metric (IRM) with alternative suggested metrics to evaluate their impact on the performance of FAIR. α-nDCG, that was used in the baseline, will be replaced with MAP, RBP and ERR-IA.

3.2.1 MAP

MAP (Mean Average Precision) is an evaluation metric commonly used in information retrieval and ranking tasks [56]. It provides a measure of precision at different recall levels, making it suitable for assessing the effectiveness of ranked retrieval systems.

The concept of MAP is based on the notion of precision, which represents the proportion of relevant documents among the retrieved ones. However, unlike traditional precision that considers only a single cutoff point, MAP considers precision at multiple recall levels.

$$\text{MAP} = \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{k=1}^{n_i} \text{Precision@}k \times \text{Relevance@}k}{\text{Number of Relevant Items}}$$

To calculate MAP, we first compute the average precision (AP) for each query or information need [57]. The average precision is obtained by computing the precision at each rank position where a relevant document is retrieved and then taking the average of these precision values. In other words, AP accounts for both the order and completeness of the retrieved documents.

Once we have the AP values for all queries, we calculate the mean of these average precision values, resulting in the MAP score. MAP ranges from 0 to 1, where a higher value indicates better performance.

MAP is advantageous because it captures the ranking quality of a system by considering the precision achieved at different recall levels. It rewards systems that retrieve relevant documents early in the ranking, as precision values are higher at the beginning of the list. This makes MAP particularly useful when the order of retrieval results is important, such as in search engine rankings or recommendation systems.

In the context of FAIR evaluation, incorporating MAP as one of the metrics in the FAIR framework would emphasize precision and the retrieval of relevant documents. The fairness aspect would be assessed based on how well the system achieves balanced precision across different subtopics, ensuring fairness in exposure to various information needs.

3.2.2 RBP

Rank Biased Precision (RBP), introduced in Moffat and Zobel [27], is an evaluation metric used to assess the quality of ranked retrieval systems. It takes into account the user's preference for examining documents at different ranks and incorporates a stopping rule that allows users to stop examining documents based on a given probability distribution.

RBP measures the precision of a ranking system by assigning higher weights to documents at the top of the ranking. The metric is calculated by summing the precision values at each rank, weighted by a probability distribution that favors early ranks. The probability distribution is typically modeled using a persistence parameter, which determines the extent to which the weights decrease as the rank position increases.

Mathematically, RBP can be expressed as follows [27]:

$$RBP(p) = (1 - p) \sum_{i=1}^{\infty} (p^{i-1} \times \mathrm{r}(i))$$

In this expression, p represents the persistence parameter, which controls the weight assigned to each rank. The term r(i) represents the relevance of the document at rank i, typically represented as a binary value (1 for relevant, 0 for non-relevant). The summation is performed over all ranks, starting from the top-ranked document.

RBP combines both precision and the user's preference for early examination of documents. It provides a measure of the system's ability to present relevant documents at higher ranks, reflecting the user's behavior and information-seeking preferences. By incorporating RBP into the FAIR metric, the evaluation of fairness considers how well the system balances the presentation of subtopics based on the user's stopping probability, ensuring fairness in exposure across different stopping points.

3.2.3 ERR-IA

ERR-IA (Expected Reciprocal Rank - Intent-Aware) is an evaluation metric introduced in Chapelle et al. [31] that takes into account both relevance and user intent. It is an extension of the ERR metric[30], which is designed to capture the satisfaction of user intents during the information retrieval process.

ERR-IA considers that users have different intents or information needs when conducting a search. It models the examination behavior of users and assigns a probability to each rank position indicating the likelihood of a user examining an item at that position. The metric assumes that users will stop examining the results once they find a relevant document that satisfies their intent.

ERR-IA calculates the expected reciprocal rank (ERR) by multiplying the relevance of each item by the examination probability at that rank position. It then sums up these values to compute the overall ERR-IA score. The examination probabilities are typically modeled using a decay function that decreases as the rank position increases.

By incorporating intent-awareness, ERR-IA provides a more nuanced evaluation of search systems[31]. It takes into consideration not only the relevance of the retrieved documents but also the satisfaction of user intents. This makes ERR-IA particularly useful in scenarios where understanding and fulfilling user intents is crucial, such as personalized search or recommendation systems.

The ERR-IA score is calculated as follows[31]:

$$\text{ERR-IA} = \sum_{i=1}^{N} \left(\frac{1}{i} \prod_{j=1}^{i-1} (1 - R(j) \cdot p(j)) \cdot R(i) \cdot p(i) \right)$$

In this equation, N represents the total number of items or documents in the ranking. The product term captures the probability that the user has not found a relevant document up to rank i-1, while the (1/i) factor represents the reciprocal rank discount. The relevance of each item and the examination probability at each rank position are multiplied together to compute the contribution of that item to the overall ERR-IA score.

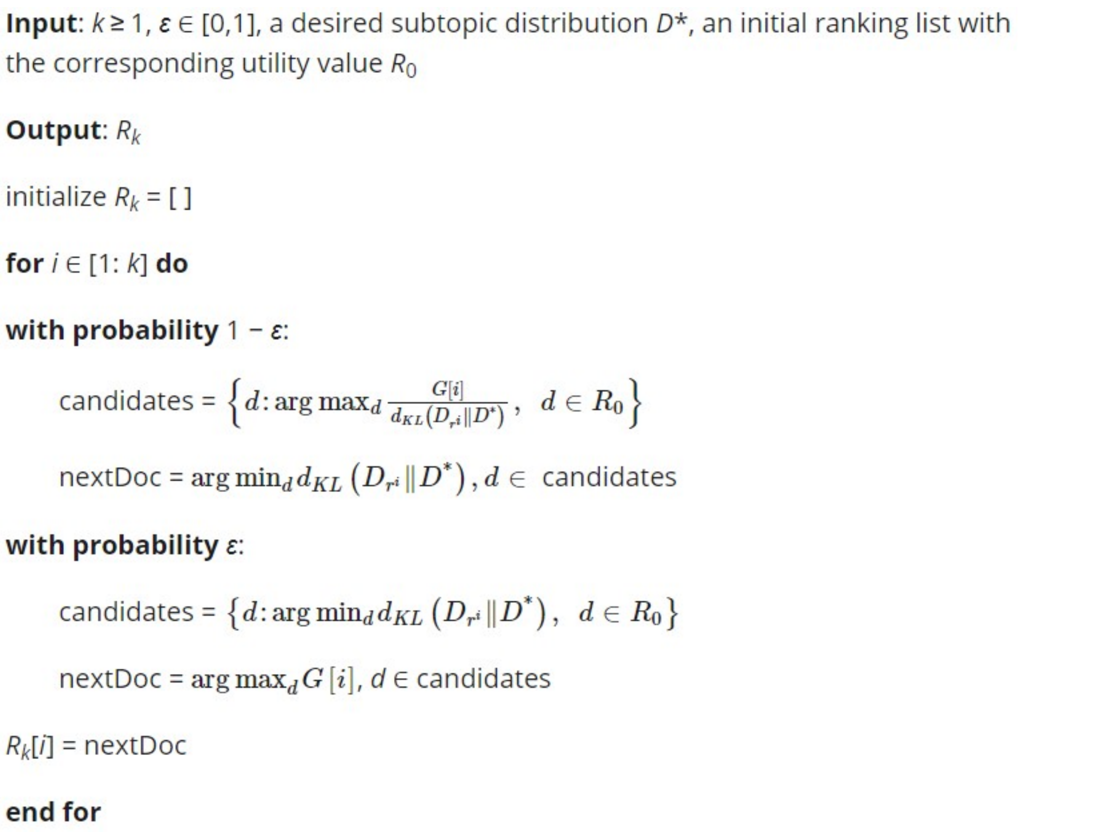
This formula quantifies the expected satisfaction of user intents by considering both relevance and examination probabilities. It accounts for the likelihood of users stopping at different ranks based on the probability distribution and captures the cumulative satisfaction of intents throughout the ranking.

The use of ERR-IA in the FAIR metric would contribute to the fairness evaluation by ensuring equitable representation of different intent categories in the ranking. It would assess how well the system satisfies user intents across various demographic or interest groups, promoting fairness in information retrieval and recommendation processes.

3.3 FAIR ε-greedy re-ranking algorithm

Gao et al. [1] propose FAIR ε-greedy algorithm that aims to achieve fairness-aware re-ranking in the top-k results. This algorithm builds upon the concept of the ε-greedy algorithm, which seeks to strike a balance between fairness and utility in search and recommendations. This balance between fairness and the system's utility is customized by utilizing proposed FAIR metric.

At each rank i, the algorithm optimizes for the diversity and novelty gain discounted by bias and rank position, denoted as $\frac{G[i]}{d_{\mathrm{KL}}}$. With probability 1 - ε, the algorithm selects candidates based on this optimization, while with probability ε, it explores documents that minimize the KL-divergence d_{KL}. The FAIR ε-greedy algorithm follows this process for k ranks, ultimately generating the re-ranked list Rk.



```
Input: $k \geq 1$, $\varepsilon \in [0,1]$, a desired subtopic distribution $D*$, an initial ranking list with
the corresponding utility value $R_0$

Output: $R_k$

initialize $R_k$ = [ ]

for $i \in [1: k]$ do

with probability $1 - \varepsilon$:

    candidates = $\left\{d: \arg\max_d \frac{G[i]}{d_{KL}(D_{r^i} \| D^*)},\ \ d \in R_0\right\}$

    nextDoc = $\arg\min_d d_{KL}\left(D_{r^i} \| D^*\right), d \in$ candidates

with probability $\varepsilon$:

    candidates = $\left\{d: \arg\min_d d_{KL}\left(D_{r^i} \| D^*\right),\ \ d \in R_0\right\}$

    nextDoc = $\arg\max_d G[i]$, $d \in$ candidates

$R_k[i]$ = nextDoc

end for
```

There are two special cases worth noting. First, when $\varepsilon = 0$, the algorithm becomes a simple greedy algorithm that always optimizes for the FAIR metric, effectively balancing α-nDCG and d_{KL}. Second, when $\varepsilon = 1$, the algorithm becomes a simple greedy algorithm that minimizes the KL-divergence at each rank, disregarding α-nDCG. Although increasing fairness may improve diversity for certain datasets, α-nDCG is not considered in this scenario [37].

Importantly, the FAIR ε-greedy algorithm operates as a ranking algorithm without relying on gold standard relevance judgment labels. In the absence of such labels, it is possible to reference the default ranking provided by a retrieval system. It is assumed that the top-ranked items in the default ranking are all relevant, with higher ranks indicating higher relevance. Similarly, in recommender systems, the top recommended items in the default set are assumed to have higher utility, such as a higher click-through-rate or a higher probability of being purchased.

3.4 Datasets

3.4.1 MovieLens

The MovieLens dataset is a widely used and well-known benchmark dataset in the field of recommender systems [58]. It consists of movie ratings provided by users, along with associated movie metadata. The dataset contains a diverse collection of movies from different genres, spanning various time periods.

The MovieLens dataset offers a rich source of information for evaluating and developing recommendation algorithms. It includes explicit user ratings that provide insights into individual preferences and opinions. Additionally, the dataset includes demographic information about the users, such as age and gender, enabling the exploration of personalized recommendation approaches [58].

The latest version of the MovieLens dataset, MovieLens 25M, contains approximately 25 million ratings from about 162,000 users on around 62,000 movies[1]. Along with the ratings, the dataset provides metadata about the movies, including genres, release year, and tags assigned by users. This metadata enables content-based recommendation techniques and genre-based analysis.

Due to its popularity and availability, the MovieLens dataset has been extensively used for benchmarking and comparing different recommendation techniques [1, 35, 59, 60]. Its widespread adoption has fostered the advancement of recommender system research and enabled the development of novel algorithms and evaluation methodologies.

3.4.2 TREC2019 Fair ranking track dataset

The dataset used in the TREC 2019 Fair ranking track is the Semantic Scholar (S2) Open Corpus from the Allen Institute for Artificial Intelligence [61].

Semantic Scholar Open Corpus (S2ORC) is an extensive collection of 81.1 million academic papers written in English, covering a wide range of academic disciplines [62]. This corpus offers comprehensive metadata, abstracts, resolved bibliographic references, and structured full text for 8.1 million open access papers. The full text is enriched with

[1] https://grouplens.org/datasets/movielens/25m/

automatically-detected inline mentions of citations, figures, and tables, all linked to their respective paper entities. S2ORC combines papers from various academic publishers and digital archives into a unified source, resulting in the largest publicly available machine-readable academic text collection to date [62].

3.4.3 TREC2009 Web track dataset

ClueWeb09[2] is a large-scale web document collection that was created for information retrieval research purposes. It consists of approximately one billion web pages and covers a wide range of topics and domains in ten languages. ClueWeb09 was developed as part of the TREC (Text Retrieval Conference) initiative and has been widely used in various information retrieval tasks[3] and evaluations. The dataset provides researchers with a valuable resource for studying web search algorithms, information retrieval techniques, and other related fields[63]. Its size and diversity make it a valuable benchmark for evaluating the effectiveness of retrieval systems and developing new approaches to web search and information retrieval[64].

3.5 Prepossessing and Indexing

As previously mentioned, the MovieLens dataset was utilized in the original research [1]. Notably, the results obtained from this dataset were considerably poorer in comparison to the other datasets primarily due to its high sparsity. Consequently, it becomes intriguing to investigate whether any of the proposed advanced methods will yield superior performance on this particular dataset. Thus, to ensure a fair and accurate comparison, the exact same preprocessing and indexing methodology employed by Gao et al. [1] is applied to the MovieLens dataset. This methodology is described in the following chapter 4.2.1.

Datasets were initially indexed and retrieved using Elasticsearch with the BM25 algorithm. Subsequent subsections provide a detailed description of the algorithms used.

3.5.1 BM25

When it comes to information retrieval and search engines, finding the most relevant documents or web pages based on a user's query is of utmost importance. To achieve this,

[2] `https://lemurproject.org/clueweb09.php/`
[3] `https://trec.nist.gov/data/web09.html`

various ranking algorithms have been developed, and one such widely used algorithm is Best Match 25 (BM25) [65].

BM25, also known as Okapi BM25, was first introduced in 1994 by Stephen Robertson and Stephen Walker as an improvement over the Okapi Term Weighting Scheme[66]. It is designed to address the limitations of traditional term frequency-inverse document frequency (TF-IDF) ranking algorithms by incorporating additional factors for better relevance scoring.

BM25, as described by Robertson and Zaragoza [65], incorporates several key components that contribute to its relevance scoring. Firstly, it takes into account the frequency of a term within a document, employing a logarithmic function to mitigate the impact of excessively high term frequencies. Secondly, BM25 normalizes the document length based on the average length of documents in the collection, ensuring that longer documents do not possess an unfair advantage. Lastly, weights are assigned to each query term based on its frequency and inverse document frequency (IDF), thereby influencing the relevance score of documents.

BM25 calculates the relevance score of a document by considering the query terms, their weights, and the document's term frequencies [67]. The final score is a combination of these factors, providing a ranking that reflects the relevance of each document to the user's query.

BM25 offers several advantages over traditional ranking algorithms [65]. It provides flexibility through customizable parameters, such as term saturation and field-length normalization. BM25 performs well across various document types and datasets, making it suitable for different domains. Moreover, the scoring formula of BM25 allows for better interpretability and analysis.

BM25 has found extensive applications in search engines, recommendation systems, document retrieval, and question-answering systems [65]. Its effectiveness and versatility have made it a popular choice in information retrieval research and industry implementations [56].

Researchers have proposed enhancements and variants of BM25 to address specific challenges and improve its performance. For example, Lv and Zhai [68] proposes an extension of BM25 - BM25L - which "shifts" the term frequency normalization formula to boost scores of very long documents. These include incorporating term proximity, utilizing query expansion techniques, and adapting BM25 for specialized domains.

3.5.2 Elasticsearch

Elasticsearch[4], a widely adopted distributed search and analytics engine, offers powerful indexing and scoring capabilities that enable efficient retrieval of relevant information from large datasets.

Indexing is the process of storing and organizing data in a way that facilitates fast and accurate search operations [69]. Elasticsearch uses a flexible data model, allowing users to define mappings that describe the structure of their documents and specify how fields should be indexed and analyzed [70]. During the indexing process, Elasticsearch tokenizes the text, applies analyzers to process and normalizes the tokens, and builds an inverted index to enable efficient search operations.

The BM25 algorithm takes into account both the user's query and the characteristics of the indexed documents to calculate the relevance score. By leveraging the BM25 scoring algorithm, Elasticsearch enhances search relevance and improves the quality of search results. It allows users to fine-tune relevance parameters by adjusting BM25-specific parameters such as the k_1 and b values to better align with their specific use cases and desired search behavior [70].

Enhancing search relevance in Elasticsearch with the BM25 scoring algorithm involves considering various factors [71]. First, term frequency rewards documents that contain query terms multiple times, indicating higher relevance. Second, inverse document frequency assigns higher weights to terms that appear in a limited number of documents, highlighting their significance. Additionally, field length normalization ensures that longer documents are not favored solely based on length. Lastly, document length normalization prevents shorter documents from dominating search results by considering their length relative to the average document length in the index[70].

[4] https://www.elastic.co/

Chapter 4

Experimental Evaluation

4.1 Evaluation Metrics

Several metrics were employed by Gao et al. [1] to evaluate the performance of the approach, including FAIR, nDCG, RBP, KL-divergence (KL), and nDRKL. The top-k scores for each metric were reported with k values of 10, and 50 to account for various rank positions. To ensure a comprehensive and rigorous analysis, this thesis employs the same set of evaluation metrics.

α-nDCG and RBP are utilized as utility functions to assess the system's ability to retrieve relevant results for users, considering the impact of rank position. For fairness evaluation, KL-divergence (KL) was employed as a direct indicator of the disparity between the group distributions in the ranking and the desired distributions.

A modified version of normalized Discounted KL-divergence (nDKL) [38, 42], referred to as normalized Discounted Reciprocal KL-divergence (nDRKL), was introduced by Gao et al. [1] to measure rank-biased unfairness. $nDRKL$ discounts the bias at each rank position, with a higher discount on bias for higher ranks. This metric highlights highly biased results at earlier ranks. $nDRKL$ is derived from nDKL by replacing the KL-divergence with its reciprocal at each rank position. The calculation for $nDRKL$ in the top-k ranking is defined as:

$$nDRKL = \frac{1}{Z}\sum_{i=1}^{k}\left(\frac{1}{\log_2(i+1)}\right)\left(\frac{1}{dKL(D_{ri}||D^{*})+1}\right)$$

Here, $dKL(D_1||D_2)$ represents the KL-divergence between distribution D_1 and D_2, and Z is the sum of $(1/\log_2(i+1))$ over the range of i. Adding 1 to dKL ensures the avoidance of divide-by-zero issues. The $nDRKL$ metric has a desirable property where its value ranges between 0 and 1, with an optimal score of 1 at rank 1 when the KL-divergence is zero. Higher scores indicate better fairness. Moreover, $nDRKL$ discourages bias at top rank positions by encouraging lower KL-divergence at higher ranks. The $nDRKL$ metric reports the average performance by averaging the scores over all queries.

4.2 Experimental Results

4.2.1 MovieLens - original approach

Initial preprocessing of MovieLens dataset was done following the approach of Morik et al. [72].

Fairness in the dataset is addressed in relation to production companies. The dataset contains movies, and the two production companies with the highest number of movies, MGM and Warner Bros., are selected for analysis. Movies with limited ratings were excluded, resulting in a partially filled ratings matrix. This matrix consists of 10,000 users, 100 movies, and 46,515 user-item interactions, with a sparsity level of approximately 95.35% [1].

To ensure fairness, the demographic parity fairness constraint is employed, where the desired distribution D* is set to be the group distribution in the entire movie collection for a query/user. It should be noted that different fairness notions can be incorporated into our metric by adjusting D* to different distributions [1].

In the experiments by Gao et al. [1], the Fairness of exposure in ranking (FOE) algorithm is utilized for MovieLens. FOE[25] is a versatile framework that utilizes probabilistic rankings and linear programming to compute the ranking that maximizes utility under various fairness constraints. It is considered a state-of-the-art reranking framework based on group fairness constraints. To adapt FOE to the recommendation task, the query-document relevance scores are replaced with user-item rating scores.

Results

The results for MovieLens with original approach are presented in Table 4.1. Due to low values, the results are provided with four significant figures.

Table 4.1: MovieLens

Metric	$k = 10$	$k = 50$
FAIR α-nDCG	0.0176	0.0577
FAIR MAP	0.0279	0.0653
FAIR RBP	0.0243	0.0598
FAIR ERR-IA	0.0602	0.0911
KL	0.0730	0.0308
nDRKL	0.8204	0.9121
nDCG	0.0230	0.0641
RBP	0.0157	0.0158

4.2.2 BM25

The following subsection shows the results for the approach described in Chapter 3 for all three datasets.

Results

The results for MovieLens, S2ORC and ClueWeb09 are presented in Table 4.2 Table 4.3 and Table 4.4. Also now, the results for MovieLens are provided with four significant figures.

Table 4.2: MovieLens

Metric	$k = 10$	$k = 50$
FAIR α-nDCG	0.0124	0.0337
FAIR MAP	0.0219	0.0593
FAIR RBP	0.0189	0.0512
FAIR ERR-IA	0.0547	0.0814
KL	0.0548	0.0185
nDRKL	0.8046	0.8671
nDCG	0.0230	0.0641
RBP	0.0146	0.0146

4.3 Analysis

4.3.1 General observations

After the research, it is clear that the choice of the utility metric in FAIR should be based on the specific dataset. The results indicate that ERR-IA appears to be the best

Table 4.3: S2ORC

Metric	$k = 10$	$k = 50$
FAIR α-nDCG	0.692	0.729
FAIR MAP	0.756	0.793
FAIR RBP	0.634	0.685
FAIR ERR-IA	0.564	0.655
KL	0.003	0.001
nDRKL	0.860	0.921
nDCG	0.884	0.927
RBP	0.971	0.974

Table 4.4: ClueWeb09

Metric	$k = 10$	$k = 50$
FAIR α-nDCG	0.598	0.634
FAIR MAP	0.638	0.687
FAIR RBP	0.612	0.646
FAIR ERR-IA	0.583	0.623
KL	0.165	0.129
nDRKL	0.835	0.877
nDCG	0.989	0.998
RBP	0.994	0.999

choice for the MovieLens dataset, regardless of the approach used. However, this metric exhibited the poorest performance for the other datasets. On the other hand, MAP showed better results than the baseline for both the ClueWeb09 and S2ORC datasets, while RBP performed better for ClueWeb09 but worse for S2ORC.

The impact of the rank threshold, denoted as k, on the utility and fairness metric scores was investigated for each dataset and ranking algorithm. As the value of k increased, there was a consistent trend of improvement observed in all utility and fairness metrics. This finding suggests that the inclusion of more items in the ranking list positively influences the performance of both utility and fairness measures.

Comparing different ranking algorithms applied to MovieLens, FAIR demonstrated its effectiveness for all IRMs in highlighting the dimensions that exhibited differences in performance. This distinction was particularly pronounced in the fairness metrics, as indicated by the results in Table 4.1 and Table 4.2. By emphasizing the dimension with more distinctive performances, FAIR provides valuable insights into the trade-offs between utility and fairness in ranking algorithms.

In conclusion, the findings presented in this thesis underscore the significance of FAIR as a fairness-aware metric in information retrieval tasks. It effectively captures the challenges posed by datasets with extreme sparsity and showcases the trade-offs between utility and fairness. While FAIR demonstrates promising capabilities, there are opportunities for further exploration and refinement; some suggestions are presented in Section 5.1.

4.3.2 MovieLens - original approach

Although similar experiments were performed on three different datasets, it is reasonable to discuss the results for MovieLens in more detail. MovieLens achieved the poorest results in the original research due to sparsity, and it was therefore interesting to recreate the original approach to investigate whether another IRM had any impact on the results.

Once again, it was observed that MovieLens exhibited notably poorer utility scores compared to other datasets, as indicated by significantly lower nDCG and RBP values. This highlights the greater challenges faced in generating accurate and relevant rankings for MovieLens.

FAIR, as a fairness-aware metric, effectively captured this limitation by assigning the lowest score to MovieLens. However, all of the suggested utility metrics improved the FAIR score for the dataset, although it still remained significantly smaller than the scores for the other datasets. This observation supports the argument for adapting IRM metrics in FAIR according to the specific dataset characteristics.

Chapter 5

Conclusions

FAIR metric requires further investigation and evaluation using diverse datasets and applications to assess its ability to capture user models in retrieval and recommendation tasks. The metric's current design exhibits certain limitations that warrant improvement in future research.

One of the limitations that was not addressed in this thesis pertains to the applicability of FAIR to effectiveness metrics; these are defined as a sum over the ranking. To address this, the outer sum can be removed, and the utility metric can be considered as the numerator while the fairness metric serves as the denominator.

Moreover, it is important to note that FAIR cannot be regarded as a universal metric applicable to all scenarios. This was clearly proved in the results of this research, as different metrics were performing better on different datasets.

When utility and fairness metrics exhibit interdependencies, careful consideration and elimination of these interdependencies are necessary to ensure meaningful integration. Instead of treating utility and fairness as independent factors with no knowledge of their relationship, it is preferable to directly model the trade-offs between utility and fairness as a metric.

Furthermore, there are instances where two algorithms yield identical FAIR scores, despite one algorithm demonstrating significantly higher utility but extremely low fairness, while the other exhibits the opposite performance in terms of utility and fairness. In such cases, the FAIR metric fails to differentiate between these algorithms. Addressing this limitation could involve fine-tuning the weights assigned to utility and fairness.

The current design of the FAIR metric imposes restrictions on the types of metrics that can be incorporated, particularly when considering variants of precision, diversity, and novelty-based metrics. Extending FAIR to handle unbounded relevance or fairness metrics necessitates further exploration.

It is vital to clarify that the use of KL-divergence in FAIR and evaluation metrics should not be misconstrued as evaluating a metric using another metric. Rather, the intention is to design an algorithm that optimizes a combination of two metrics and subsequently evaluate its performance with respect to each individual metric.

In order to simplify the discussion surrounding fairness and bias, the terms are used interchangeably. It is important to note that FAIR primarily emphasizes group fairness rather than individual fairness. The question of "fair to whom?" plays a central role in fairness considerations. By focusing on group fairness, it becomes possible to account for diversity and representation of information sources or topics.

5.1 Future research suggestions

The following is a curated list of potential directions for future research, encompassing the key points mentioned previously. This list provides potential research directions that can contribute to advancing the field of fairness-aware information retrieval and recommendation systems. Further investigations and collaborations in these areas will enhance understanding, address limitations, and promote the development of more robust and equitable systems.

- Conduct extensive evaluations across diverse datasets and applications to assess the effectiveness and generalizability of the proposed fairness-aware metric (FAIR) in capturing user models in retrieval and recommendation tasks.
- Metric refinement: Explore further modifications to the FAIR metric design to address its limitations, such as removing the outer sum in effectiveness metrics.
- Investigate approaches to handle interdependencies between utility and fairness metrics, ensuring that their integration is meaningful and reflective of the desired trade-offs.
- Develop algorithms that minimize deviations from the desired distribution and incorporate fairness scores into the normalizer, facilitating a more nuanced evaluation of rankings and recommendations.
- Investigate the applicability of FAIR to other types of metrics, expanding the range of metrics that can be integrated into FAIR.

- Refine the weights assigned to utility and fairness in the FAIR metric to address situations where algorithms with different utility and fairness profiles yield identical FAIR scores.

- Investigate the relationship between fairness and bias, elucidating their differences and connections within the context of information retrieval. Develop strategies to effectively address bias within the fairness framework.

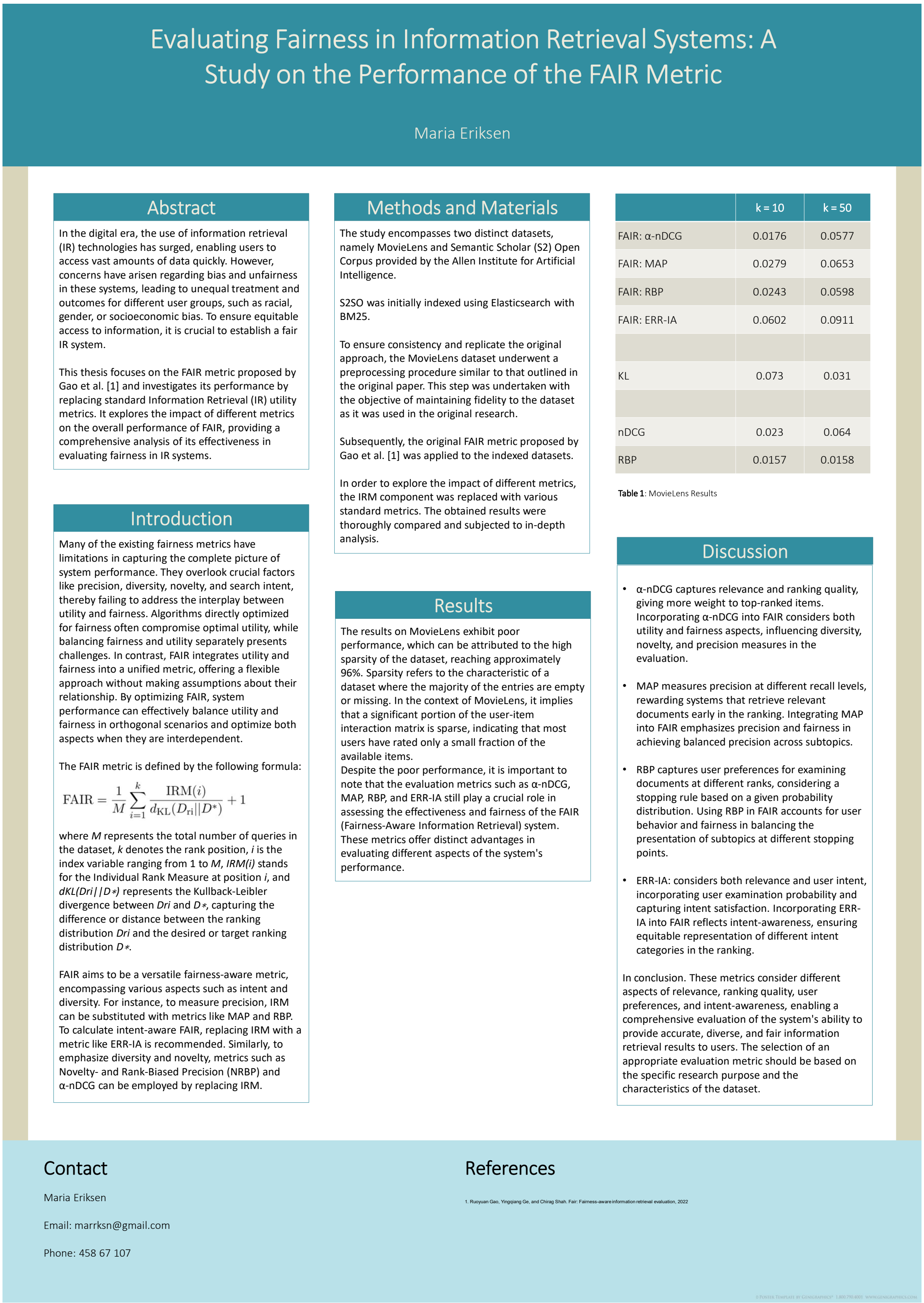
- Examine the implications of group fairness on diversity and representation of information sources or topics. Develop approaches that balance the considerations of both end-users and information sources.

- Investigate ways to extend FAIR to handle additional dimensions of fairness, such as temporal fairness (as discussed in Melton [73]), intersectional fairness (as shown in Foulds et al. [74]), or fairness across multiple user groups.

- Continuously evaluate and refine the FAIR metric based on empirical studies, user feedback, and emerging research findings, adapting it to evolving understandings of fairness and advancements in information retrieval techniques.

Appendix A

Poster



Evaluating Fairness in Information Retrieval Systems: A Study on the Performance of the FAIR Metric

Maria Eriksen

Abstract

In the digital era, the use of information retrieval (IR) technologies has surged, enabling users to access vast amounts of data quickly. However, concerns have arisen regarding bias and unfairness in these systems, leading to unequal treatment and outcomes for different user groups, such as racial, gender, or socioeconomic bias. To ensure equitable access to information, it is crucial to establish a fair IR system.

This thesis focuses on the FAIR metric proposed by Gao et al. [1] and investigates its performance by replacing standard Information Retrieval (IR) utility metrics. It explores the impact of different metrics on the overall performance of FAIR, providing a comprehensive analysis of its effectiveness in evaluating fairness in IR systems.

Introduction

Many of the existing fairness metrics have limitations in capturing the complete picture of system performance. They overlook crucial factors like precision, diversity, novelty, and search intent, thereby failing to address the interplay between utility and fairness. Algorithms directly optimized for fairness often compromise optimal utility, while balancing fairness and utility separately presents challenges. In contrast, FAIR integrates utility and fairness into a unified metric, offering a flexible approach without making assumptions about their relationship. By optimizing FAIR, system performance can effectively balance utility and fairness in orthogonal scenarios and optimize both aspects when they are interdependent.

The FAIR metric is defined by the following formula:

$$\mathrm{FAIR} = \frac{1}{M}\sum_{i=1}^{k}\frac{\mathrm{IRM}(i)}{d_{\mathrm{KL}}(D_{\mathrm{ri}}||D^{*})}+1$$

where *M* represents the total number of queries in the dataset, *k* denotes the rank position, *i* is the index variable ranging from 1 to *M*, *IRM(i)* stands for the Individual Rank Measure at position *i*, and *dKL(Dri||D∗)* represents the Kullback-Leibler divergence between *Dri* and *D∗*, capturing the difference or distance between the ranking distribution *Dri* and the desired or target ranking distribution *D∗*.

FAIR aims to be a versatile fairness-aware metric, encompassing various aspects such as intent and diversity. For instance, to measure precision, IRM can be substituted with metrics like MAP and RBP. To calculate intent-aware FAIR, replacing IRM with a metric like ERR-IA is recommended. Similarly, to emphasize diversity and novelty, metrics such as Novelty- and Rank-Biased Precision (NRBP) and α-nDCG can be employed by replacing IRM.

Methods and Materials

The study encompasses two distinct datasets, namely MovieLens and Semantic Scholar (S2) Open Corpus provided by the Allen Institute for Artificial Intelligence.

S2SO was initially indexed using Elasticsearch with BM25.

To ensure consistency and replicate the original approach, the MovieLens dataset underwent a preprocessing procedure similar to that outlined in the original paper. This step was undertaken with the objective of maintaining fidelity to the dataset as it was used in the original research.

Subsequently, the original FAIR metric proposed by Gao et al. [1] was applied to the indexed datasets.

In order to explore the impact of different metrics, the IRM component was replaced with various standard metrics. The obtained results were thoroughly compared and subjected to in-depth analysis.

Results

The results on MovieLens exhibit poor performance, which can be attributed to the high sparsity of the dataset, reaching approximately 96%. Sparsity refers to the characteristic of a dataset where the majority of the entries are empty or missing. In the context of MovieLens, it implies that a significant portion of the user-item interaction matrix is sparse, indicating that most users have rated only a small fraction of the available items.
Despite the poor performance, it is important to note that the evaluation metrics such as α-nDCG, MAP, RBP, and ERR-IA still play a crucial role in assessing the effectiveness and fairness of the FAIR (Fairness-Aware Information Retrieval) system. These metrics offer distinct advantages in evaluating different aspects of the system's performance.

	k = 10	k = 50
FAIR: α-nDCG	0.0176	0.0577
FAIR: MAP	0.0279	0.0653
FAIR: RBP	0.0243	0.0598
FAIR: ERR-IA	0.0602	0.0911
KL	0.073	0.031
nDCG	0.023	0.064
RBP	0.0157	0.0158

Table 1: MovieLens Results

Discussion

- α-nDCG captures relevance and ranking quality, giving more weight to top-ranked items. Incorporating α-nDCG into FAIR considers both utility and fairness aspects, influencing diversity, novelty, and precision measures in the evaluation.
- MAP measures precision at different recall levels, rewarding systems that retrieve relevant documents early in the ranking. Integrating MAP into FAIR emphasizes precision and fairness in achieving balanced precision across subtopics.
- RBP captures user preferences for examining documents at different ranks, considering a stopping rule based on a given probability distribution. Using RBP in FAIR accounts for user behavior and fairness in balancing the presentation of subtopics at different stopping points.
- ERR-IA: considers both relevance and user intent, incorporating user examination probability and capturing intent satisfaction. Incorporating ERR-IA into FAIR reflects intent-awareness, ensuring equitable representation of different intent categories in the ranking.

In conclusion. These metrics consider different aspects of relevance, ranking quality, user preferences, and intent-awareness, enabling a comprehensive evaluation of the system's ability to provide accurate, diverse, and fair information retrieval results to users. The selection of an appropriate evaluation metric should be based on the specific research purpose and the characteristics of the dataset.

Contact

Maria Eriksen

Email: marrksn@gmail.com

Phone: 458 67 107

References

1. Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. Fair: Fairness-aware information retrieval evaluation, 2022

Bibliography

[1] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. Fair: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 73(10):1461–1473, 2022.

[2] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

[3] R. Khaziev, B. Casavant, P. Washabaugh, A. A. Winecoff, and M. Graham. Recommendation or discrimination?: Quantifying distribution parity in information retrieval systems. *ArXiv*, 2019.

[4] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[5] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377, 2020.

[6] Swapnil Kangralkar. Types of biases in data. *Towards Data Science*, 57(6):102377, 2021.

[7] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, 31(1):431–458, 2022.

[8] Mohinder Partap Satija. Universal decimal classification: past and present. *DESIDOC Journal of Library Information Technology*, 28(6):3–10, 2008.

[9] Alex Wright. Paul otlet: A visionary to unify global knowledge. *Journal of Scientometric Research*, 6(3):212–214, 2017.

[10] Mark Sanderson and W. Bruce Croft. The history of information retrieval research. *IEEE Data Engineering Bulletin*, 100(1):1445–1451, 2011.

[11] Michael Lesk. The seven ages of information retrieval. *International Federation of Library Associations and Institutions*, 1996.

[12] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[13] Ellen M. Voorhees and Donna K. Harman. Trec: Experiment and evaluation in information retrieval. *Computational Linguistics*, 32(4):563–567, 2005.

[14] Meennapa Rukhiran and Paniti Netinant. Automated information retrieval and services of graduate school using chatbot system. *International Journal of Electrical Computer Engineering*, 15(5):5330–5338, 2022.

[15] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018.

[16] Jahna Otterbacher, Jo Bates, and Paul Clough. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 6620–6631, 2017.

[17] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828, 2015.

[18] Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. Measuring societal biases from text corpora with smoothed first-order co-occurrence. In *Proceedings of the international aaai conference on web and social media*, volume 15, pages 549–560, 2021.

[19] Andres Ferraro, Xavier Serra, and Christine Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 249–254, 2021.

[20] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

[21] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016. Accessed: 2023-03-01.

[22] Charles W. Ostrom, Brian J. Ostrom, and Matthew Kleiman. *Judges and discrimination: assessing the theory and practice of criminal sentencing.* Final Grant Report to the National Institute of Justice, 2004.

[23] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414, 2018.

[24] Amifa Raj, Connor Wood, Ananda Montoly, and Michael D Ekstrand. Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311*, 2020.

[25] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, page 2219–2228, 2018.

[26] Lorenzo Porcaro, Carlos Castillo, Emilia Gómez, and João Vinagre. Fairness and diversity in information access systems. *arXiv preprint arXiv:2305.09319*, 2023.

[27] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1), dec 2008.

[28] Carol L Barry. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159, 1994.

[29] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, 2008.

[30] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, page 5–14, 2009.

[31] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 621–630, 2009.

[32] Aldo Lipani. Fairness in information retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1171–1171, 2016.

[33] Amin Bigdeli, Negar Arabzadeh, Shirin SeyedSalehi, Morteza Zihayat, and Ebrahim Bagheri. Gender fairness in information retrieval systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3436–3439, 2022.

[34] Ruoyuan Gao and Chirag Shah. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 229–236, 2019.

[35] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 160–169, 2019.

[36] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 275–284, 2020.

[37] Ruoyuan Gao and Chirag Shah. Toward creating a fairer ranking in search engine results. *Information Processing Management*, 57(1):102138, 2020.

[38] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, 2017.

[39] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 2212–2220, 2019.

[40] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*, WWW '19, page 2936–2942, 2019.

[41] Sirui Yao and Bert Huang. New fairness metrics for recommendation that embrace differences. *arXiv preprint arXiv:1706.09838*, 2017.

[42] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search and recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 2221–2231, 2019.

[43] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 553–562, 2019.

[44] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[45] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666, 2020.

[46] J Henry Hinnefeld, Peter Cooman, Nat Mammo, and Rupert Deese. Evaluating fairness metrics in the presence of dataset bias. *arXiv preprint arXiv:1809.09245*, 2018.

[47] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. In *Advances in Knowledge Discovery and Data Mining*, pages 155–167, 2020.

[48] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

[49] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1569–1578, 2017.

[50] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness amp; satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2243–2251, 2018.

[51] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1259–1276, 2019.

[52] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30:2921–2930, 2017.

[53] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. Addressing marketing bias in product recommendations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 618–626, 2020.

[54] Ananya Gupta, Eric Johnson, Justin Payan, Aditya Kumar Roy, Ari Kobren, Swetasudha Panda, Jean-Baptiste Tristan, and Michael Wick. Online post-processing in

rankings for fair utility maximization. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 454–462, 2021.

[55] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Advances in Information Retrieval Theory*, pages 188–199. Springer Berlin Heidelberg, 2009.

[56] ChengXiang Zhai and Sean Massung. *Text data management and analysis: a practical introduction to information retrieval and text mining.* Morgan & Claypool, 2016.

[57] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278, 2007.

[58] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), dec 2015.

[59] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 445–453, 2021.

[60] Tao Yang and Qingyao Ai. Maximizing marginal fairness for dynamic learning to rank. In *Proceedings of the Web Conference 2021*, pages 137–145, 2021.

[61] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. Overview of the trec 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*, 2019.

[62] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, 2020.

[63] Mark D Smucker, Charles L Clarke, and Gordon V Cormack. Experiments with clueweb09: Relevance feedback and web tracks. Technical report, WATERLOO UNIV (ONTARIO) DEPT OF MANAGEMENT SCIENCES, 2009.

[64] Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. Chatnoir: a search engine for the clueweb09 corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1004–1004, 2012.

[65] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[66] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[67] Stephen E Robertson and Steve Walker. Okapi/keenbow at trec-8. In *TREC*, volume 8, pages 151–162, 1999.

[68] Yuanhua Lv and ChengXiang Zhai. When documents are very long, bm25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1103–1104, 2011.

[69] Tim Miller. How does indexing work. *Chartio*, 2021.

[70] Elasticsearch. http://https://www.elastic.co/, . Accessed: 2023-03-01.

[71] Elasticsearch. practical bm25, part 2, . URL https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables. Accessed: 2023-03-01.

[72] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 429–438, 2020.

[73] Hayden Melton. Understanding and improving temporal fairness on an electronic trading venue. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 1–6, 2017.

[74] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2020.