FAIR LAYOUTS IN INFORMATION ACCESS SYSTEMS

PROVIDER-SIDE GROUP FAIRNESS IN RANKING BEYOND RANKED LISTS

by

Amifa Raj

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Computing

Boise State University

August 2023

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Amifa Raj

Thesis Title: Fair Layouts in Information Access Systems
Provider-Side Group Fairness in Ranking Beyond Ranked Lists

Date of Final Oral Examination: 5th July 2023

The following individuals read and discussed the dissertation submitted by student Amifa Raj, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Michael D. Ekstrand, Ph.D.	Chair, Supervisory Committee
Sole Pera, Ph.D.	Member, Supervisory Committee
Edoardo Serra, Ph.D.	Member, Supervisory Committee
Casey Kennington, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Michael D. Ekstrand, Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

DEDICATION

To my loving grandmother

ACKNOWLEDGMENT

Working on my PhD was a long journey of growth and self-development. During this time, I was blessed to find many supportive people around me. I am immensely grateful to my advisor Dr. Michael Ekstrand who was always patient and kind to me. He ensured a supportive environment for me to grow and learn as a researcher and as a person as well. He was my inspiration to get involved into various academic activities which helped me enjoy the PhD journey and made my research life more fulfilling. I would like to show my gratitude towards Dr. Sole Pera for her motivation and constructive criticisms. Her valuable feedback and guidance always nudged me to the right direction.

I am grateful to my colleagues in the People and Information Research Team (PIReT) for all the support, valuable discussions, and feedback. It was a pleasure to learn from them and share my ideas with them. I have collaborated with amazing researchers during my PhD and it was an honor to work with such great minds.

The journey of my PhD came with lots of ups and downs and I have always found my friends and family as a strong support for me. My fiancée and best friend Devan Karsann has always been a source of my motivation and positive energy and his feedback on my writing helped me to improve my writing. I am deeply grateful to my friend Fariha Moomtaheen for all the emotional support.

My parents taught me to dream big and provided everything to pursue and attain

my dreams. Their love and encouragement was indispensable for me to keep going. I want to show my appreciation to my sister, Saifa for inspiring me to do a PhD. My grandmother for her unconditional love and prayers. And lastly, I am grateful to all the blessings from God.

This work is based upon work supported by the National Science Foundation under Grant No. IIS 17-51278.

ABSTRACT

Information access systems, such as search engines and recommender systems, often display results in ranked order based on their estimated relevance. The fairness of these rankings has received attention as an important evaluation criteria along with traditional metrics capturing constructs such as utility or accuracy. Fairness has many facets, including provider and consumer-side fairness at both group and individual levels. Research on provider-side group fairness involve concerns regarding measurement and optimization of fairness in ranking. Although there are several fair ranking metrics to measure provider-side group fairness based on various "sensitive attributes", multiple open challenges still exist in this area to consider. Moreover, the fair ranking research mostly focuses on linear layouts when items are displayed in single-column list, often overlooking fairness issues in other layouts such as grid view.

In my dissertation, I work on the area of provider-side group fairness in ranking in information access systems. I seek to understand the fairness concepts and practical applications of existing fair ranking metrics and find ways to improve the metrics. My work will aid researchers and practitioners in selecting fair ranking metrics by pointing out the strengths, limitations, applicability and reliability of the metrics. Moreover, I contribute to the advancement of fair ranking metrics by considering various ranking layout models and further contribute to provider-side group fairness optimization in ranking in widely-used but seldom-studied grid layout.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:

INTRODUCTION

Information access systems (IAS), such as search and recommender systems, are prevalent mechanisms of accessing relevant information from the large amount of information available online[1]. For example, Google is one of the most widely used search engines and Netflix is a popular streaming service which recommends movies and TV shows. Users interact with these systems to satisfy their *information needs* which the systems infer by analyzing user queries (explicit), user preferences (user history inferred from users' past interaction), and/or context of the requests. In response to a user request, IAS often present results in top-N ranked lists based on their relevance to user need and other measures of item utility and relationships (e.g. similarity, as in maximum marginal relevance [28]). The quality of the ranked results are often evaluated by traditional metrics (estimating accuracy or utility) that measure system's ability to find items that are relevant to user information need. However, other constructs such as diversity [204], novelty [40], and fairness [62] have proven important to develop a better and more complete understanding of the behavior of IAS and allow researchers and practitioners to evaluate IAS on ethical and social concerns by going beyond immediate user satisfaction.

[1]Portions of this dissertation reuses material from the author's published work: [142], [143], [141], [144] and [140], consistent with the ACM Author Agreement.

Users are not only stakeholders who benefit from the IAS, however. Systems *expose* items by displaying them in a ranking to consumers or users for selecting, purchasing, or consuming. Item providers or producers get benefits or profits from users noticing and interacting with their items, but user attention varies with the ranking position [165]; users tend to engage more with items at the top of the list (position bias) [198]. Systems do not always fairly distribute exposure among items, thus causing disadvantage to item providers through *disparate exposure* allocation [52]. Some providers or group of providers get more exposure than others in a way that is somewhat unfair. Item providers are often associated with sensitive attributes such as age, gender, or race and they are categorized in groups based on these demographic attributes. Disparate exposure can disadvantage providers at both individual and group level based on their sensitive attributes and can reflect historical discrimination such as racial or gender discrimination.

The broader goal of my dissertation research is to improve group fairness of ranking in IAS for providers of items retrieved or recommended across multiple layout paradigms. With that goal, I work on the advancement of provider-side group fairness measurement techniques in ranking of IAS and provide insights on measuring and optimizing group-fairness for providers in popularly-used but inadequately-studied display layout models.

My work provides comprehensive knowledge on theoretical and practical applications of the fair ranking metrics while identifying their strengths, limitations and applicability; further providing informed guidance on the fairness-task specific metric selection process. Moreover, I work to improve group fairness in ranking by considering one of the most emergent open questions regarding fairness measurement and

optimization in grid layout in IAS. I do not aim to propose one universal metric for every issue; rather, I want to produce knowledge needed to design fair ranking evaluations and intervention addressing some of the challenges in fairness in ranking in IAS. With that goal, I design my research in the following steps:

Analyzing Fair Ranking Metrics Various metrics have been proposed to measure (un)fairness in rankings, with respect to protected groups of item providers for different contexts; some in the form of fairness metrics, others as fairness constraints [103, 62]. However, there is no comparative and empirical analysis of these fair ranking metrics showing their conceptual differences and their applicability to specific scenario or real-world IAS dataset. Consequently, it is challenging for researchers and practitioners to find the suitable metric(s) for specific fairness tasks and implement them in real-world IAS ranking scenarios without knowing the challenges and applicability of the existing metrics.

To fill that gap, I conduct a comprehensive and comparative analysis of the fair ranking metrics that have been proposed so far in fair ranking research literature. This work shows conceptual similarities and differences among the metric by describing them in a common notation. The study further connects the gaps between theoretical and practical application of these metrics by implementing them in the same real-world IAS datasets under the same experimental setup. I also conduct a sensitivity analysis to evaluate the metrics stability regarding their dependency on external factors and identified the best suited metric(s) in different fairness contexts. My purpose for this analysis is to support researchers and practitioners in their metrics selection process and identify further improvements needed in the area of measuring provider-side fairness in ranking.

Beyond Linear Ranked List Fairness From studying existing fair ranking metrics, I observe that current fair ranking metrics to measure provider-side group fairness in ranking are designed for linear layout, usually vertical representation of results. However, many IAS use other layouts such as grid-view. There is no prior study on measuring group fairness in grid layout which can show what happens to fairness measurement in ranking with different layout models.

I address these issues by identifying various ranking layouts in IAS and further implement fair ranking metrics for those scenarios by considering layout-appropriate user attention models. Moreover, I identify factors that are important to consider while designing and implementing fair ranking metrics in grid layout in order to generate trustworthy and valid fairness measurements. This work elicits the applicability of the fair ranking metrics in various ranking layout scenarios and how the fairness scores change across layout models. The purpose of this component is to aid the development of fair ranking metric(s) that are able to address broader issues of real-world IAS applications and the applicability and reliability of the metrics on frequently-used but little-studied layout models.

Generalized User Browsing Models By conducting a sensitivity analysis of the fair ranking metrics and implementing them in grid layout, we recognize that the user browsing model is a crucial component in fair ranking metric design which is used to infer user provided attention to items in various positions in ranking [165, 52, 154, 16]. Various user browsing models have been proposed based on user browsing behavior while interacting with ranked results in both linear and grid layouts [122, 43, 192]. With a similar underlying concept, these models differ in their component dependency and parameter settings.

In my work, I identify multiple user browsing models for ranking in IAS and unify them in a generalized framework. The generalized framework of the user browsing model can be re-configured based on tasks, ranking layouts, and available components and it can be further extended to more advanced ranking scenarios and user browsing behaviors. My goal is to provide theoretical knowledge on the existing user browsing models and aid IAS evaluators in implementing user browsing models by designing a single generalized browsing model that can be re-configured based on their requirements.

Fairness-Aware Grid-Based Ranking The final stage of my dissertation works towards improving provider-side fairness in grid layout in IAS. There are multiple re-ranking techniques to optimize ranked results for utility, diversity, or fairness [95, 2, 111]. However, the existing re-ranking techniques are suitable for linear ranked list overlooking the widely-used grid layout. The concerns regarding fairness optimization of ranking in grid layout have not been well-studied.

My work contributes towards filling this gap by providing a preliminary analysis on provider-side fairness optimization in grid layout. I provide a grid-aware re-ranking technique by incorporating grid-layout suitable user browsing models in an existing re-ranking algorithm. Both fair ranking and effectiveness metrics to measure fairness and utility respectively are modified to take grid layout suitable user browsing behavior into account. The modified grid-aware re-ranking algorithm is then used to optimize provider-side group fairness in grid layout with minimum utility loss. Moreover, this work identifies the impact of grid-layout specific factors such as user browsing models and device sizes on fairness optimization in grid layout. The goal of this work is to generate knowledge about designing and implementing re-ranking techniques for grid

layout to improve provider-side group fairness in ranking.

My dissertation work contributes to provider-side group fairness in IAS ranking by providing the first (to our knowledge) empirical analysis of the fair ranking metrics regarding metric formulation and implications in real-world IAS. My work provides several future research directions that were identified from the conceptual analysis and implementations of the metrics. Moreover, by going beyond the linear ranked list and considering ranking in grid layout, my work has expanded the applicability of fair ranking measurements and techniques from rankings in linear layouts to grid layouts which is commonly used in modern search and recommendation applications. By providing a simple re-ranking technique to optimize ranking in grid layout for provider-side fairness, my work will draw attention to a seldom-explored research area on fairness in IAS ranking and provide further research directions in that area.

CHAPTER 2:

BACKGROUND

My research work contributes in measurement and optimization of fairness in ranking in information access systems, particularly concerning provider-side group fairness. In this chapter, I introduce some of the relevant concepts regarding IAS workflow, algorithmic fairness, and fairness and bias in ranking in IAS, and related research work in these areas.

2.1 Information Access Systems

This section provides an overview on how IAS generate ranked results and how these ranked results are evaluated.

There is a large volume of information available on the internet and more focused corpora such as e-commerce inventories, but users often want to access a limited amount of resources that are suitable for their *information need*. Information access systems (IAS) such as search engines and recommender systems help users to find relevant resources by retrieving items from a large pool of information [116, 150, 62] thus reducing information overload. Users interact with IAS with specific information requests to satisfy their information need. Users can express their requests in various ways such as explicitly providing their query by writing in an unstructured text

form [193] or using voice search [183]. Users' past interactions with systems is a rich source of user information that is commonly used to infer users' information preference to replace or supplement explicit queries [25]. Moreover, user context such as, location, time, and demographic information are also used to infer user information need [161, 187, 113, 180, 147, 175].

In search engines, users explicitly describe their need through a query and the system retrieves items or documents relevant to the query [116, 102]. In recommendation scenarios, users do not explicitly describe their information need, rather systems infer user information preference from their past interaction with systems to recommend items that will satisfy their need [4, 150]. While interacting with items in recommender systems, users can provide both explicit (ratings and reviews) and implicit (click and purchase) feedback and user preferences or requests are often inferred from their past interactions with systems [63, 200].

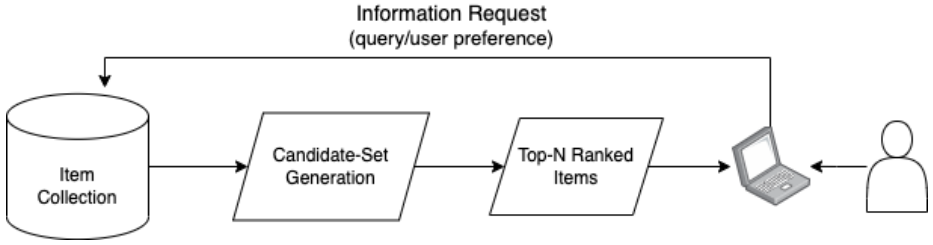


Figure 2.1: Information access systems framework

In a multi-stage IAS, given a query or user preference, IAS identifies a candidate-set of relevant items from a collection of items by measuring similarities or estimating relevance of the items [48, 9, 36, 38]. From the retrieved candidate-set of items, the top-most relevant items are presented to the user [45]. Figure 2.1 shows the basic workflow of an IAS.

Search engines estimate relevance or measure similarities between queries and

items or documents by using various information retrieval models such as Boolean, probabilistic, vector space, and inference network models [151]. In a recommendation scenario, user interacted items are used to determine similar users or items to a particular user by estimating relevance or computing similarity scores [150, 139]. In item-based recommendations, the candidate-set for an active user is generated by finding items that are similar to user-interacted items [156]. For example, two items can be considered similar if they are interacted with by the same user. In user-based algorithms, top-most similar users are identified to generate a candidate-set from their interacted items. For example, two users can be considered similar if they interact with the same items. In matrix-factorization algorithms, the latent features of user and items are used to predict user preferences [101]. Figure 2.2 shows examples of user-user and item-item similarities. Collaborative-filtering [60, 101],

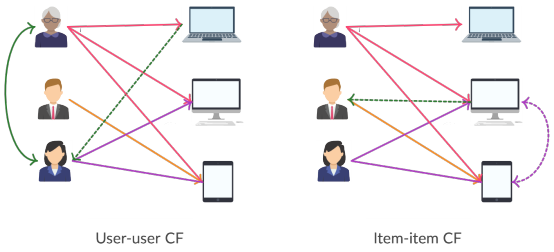


Figure 2.2: Two types of collaborative-filtering algorithms

content-based [131] (figure 2.3), or knowledge-based [24] models are some commonly used families of algorithms to identify similar items or users from user interactions. Most of these models use an *user-item interaction matrix* (figure 2.4) to create user and item profiles which are used to estimate relevance or measure similarities [101, 20]. In content-based recommendations, item metadata such as title, description, tags, or user reviews are also used to create and enrich user and item profiles or feature vectors

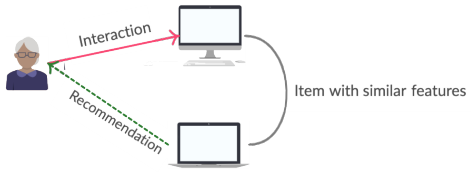


Figure 2.3: Content-based recommendations

[157, 6]. Several natural language processing (NLP) techniques such as tokenization, stemming, and stop-words removals are applied to process user queries and item information. Traditional *keyword-based* approaches such as *TF-IDF* [146] and *BM-25* [171] or embedding-based techniques such as *word2vec* [148], or advanced language models such as *BERT* [34] are popularly used techniques to analyze and represent text.

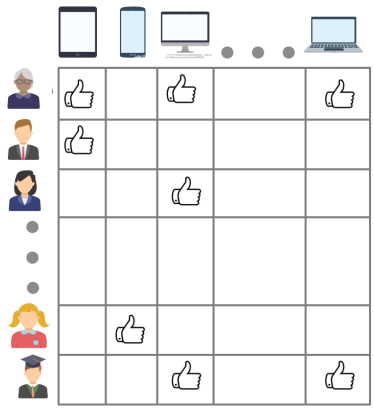


Figure 2.4: User-item interaction matrix

After generating a candidate-set by computing relevance of items from the similarity scores between information requests and items, relevant items are ranked based on their relevance scores. The top-N most relevant items are displayed to users in a ranked order [45, 44]. These ranked results can be displayed in various layouts such

as list or grid [192, 165]. Learning to rank (LTR) are popular ranking methods to optimize the ranking by training models for the target evaluation measures [110, 105].

Evaluation of IAS Evaluating ranked results is crucial in the process of designing and developing an effective and efficient IAS and this process can be done both in offline and online settings [164, 90, 138]. Online evaluation such as A/B test [100, 82] is widely used in commercial settings in a controlled experimental setup [90, 181] and this process analyzes user experience in real-time [133, 49]. On the other hand, in offline settings, systems are evaluated in a supervised manner where the ranked results are split into train and test sets [153, 30]. Offline evaluation is easier to conduct and less time consuming than online evaluation [133, 3]. Traditionally ranked results are evaluated based on user satisfaction or utility and several offline effectiveness metrics have been proposed to measure quality of ranking [123, 83, 177]. *Precision*, *recall*, *ndcg*, and *reciprocal-rank* are some of the commonly used offline effectiveness metrics [138, 35]. These metrics estimate how successfully systems satisfy user information needs with ranked results [196]. Some metrics measure the error in predicting relevance (rating) of documents for a given information request (for example, mean absolute error [89]) while other metrics evaluate the top-N ranked results and consider the position of relevant documents in the top-N ranking [182].

However, soon researchers and practitioners realized that in order to develop a whole picture of the performance of the ranked results in IAS, the systems need to be evaluated by going beyond accuracy or utility [119]. Diversity [204, 179], novelty [40], fairness [136], and serendipity [70] are some of the other objectives that are often used to evaluate IAS with respect to other important aspects of user experience and social impact. Diversity, serendipity, and novelty ensure that users are having

diverse, surprising, and new experiences respectively through the ranked results [95, 77]. Fairness, on the other hand, focuses on the social and ethical issues regarding ranked results in IAS because IAS may introduce or reflect systematic bias through their results [62, 52]. Fairness is an important and challenging concern regarding performance of IAS and has received significant attention in recent years with several metrics proposed to identify and measure fairness in ranking [142, 202, 66].

Top-N Ranking Evaluation Ranking position plays an important role in the evaluation process of ranking in IAS [124]. The top-N effectiveness metrics are used to investigate systems based on their ability of identifying relevant items and displaying those relevant items in top positions in ranking [178]. Moreover, fair ranking metrics often evaluate rankings based on their ability of fairly allocating relevant items in ranking [63, 165, 52, 195]. Hence, both effectiveness and fair ranking metrics take ranking position into account while evaluating a ranking. Most of these metrics rely on the approximation of user browsing behavior to infer user attention which is used to determine position weight in ranking [124, 16, 165, 52, 154].

For a given ranking, ranking positions have different weight depending on how users browse a ranking because user attention varies across ranking positions [112]. For example, users provide higher attention to the top ranked items than items at lower ranked positions [125]. There are several research works on analyzing and understanding user browsing behavior in ranking and there exists multiple user browsing models to compute position weight in ranking [122, 33, 192, 43]. These studies also showed that user browsing behavior depends on various factors such as ranking layouts, task, or item metadata [124, 192, 170, 12, 53]. Most of the research works on evaluating ranked results considering user browsing models are suitable for linear

(single-column) ranking. However, systems often display items in other layouts such as grid [192, 197], hence evaluating grid-based ranking by considering grid layout suitable user browsing models is still an important area to explore.

2.2 Algorithmic Fairness

This section provides a brief review on algorithmic fairness to better understand the basic concepts regarding fairness definitions, impact, mitigation, and measurement issues.

The definition of algorithmic fairness is difficult to construct; it depends on the fairness context and task [74, 160, 71]. Friedman and Nissenbaum [74] defined bias in computer systems as the systematic and unfair discrimination against certain individual or group entities by denying opportunity and assigning unfair outcomes. Mitchell et al. [121] defined algorithmic biases in form of both statistical bias (systematic mismatch between the model output and real-world) and societal bias (systematic discrimination towards groups reflecting social bias). Systems can discriminate at both group and individual levels. Items can be categorized into groups based on sensitive attributes, these sensitive attributes are also called *protected* attributes and include race, gender, religion, age and other demographic attributes. Deldjoo et al. [50] provide several categorizations of fairness definitions with examples; they identified the following categorizations that are often used to define fairness concepts in literature: *group vs individual* [73], *process vs outcome* [199], *direct vs indirect* [129], *statistical vs predictive parity* [154], *static vs dynamic* [109], and *associative vs causal* [108].

Research on algorithmic fairness often involves bias detection [5, 97], quantification [15, 202], source identification [11, 120], and mitigation [57, 136, 185]. This research area generally includes the following high-level fairness concepts:

- Fairness Level: What notion of fairness or dimension is considered? For example, systematic unfairness can happen at both group and individual level.
- Fairness Side: Who is impacted by the unfairness of systems? For example, both users and items in systems can receive systematic discrimination.
- Potential Harms: What kind of harm or negative impact is caused by systematic unfairness? For example, systems can discriminate in resource allocation or can misrepresent certain demographic groups through results.
- Fairness Application: What type of algorithms are studied regarding fairness issues? For example, the algorithmic fairness concept includes fairness issues in classification algorithms and ranking algorithms.
- Fairness Target: What is the fairness goal? For example, systems ensuring equal opportunity for users with similar qualities or mitigating social stereotypes in results.

One of the categorizations that is particularly relevant to my work is group vs individual fairness. Individual fairness addresses the goal that similar individuals should (statistically) receive similar decisions, but crucially depends on a robust construct of similarity with respect to the task for which decisions are made, and there is currently no consensus in assignment of task-relevant similarity among individuals [19, 57]. Group fairness aims to provide similar service for members of different

groups; this is often framed as ensuring a *protected group* is not treated unfairly with respect to a *dominant group* [56]. Group membership is often defined by *sensitive attributes* such as race, gender, or ethnicity. In group fairness, considering one sensitive attribute at a time can ignore fairness towards members of the intersection of two groups [21].

In group fairness scenarios, systematic bias by algorithms can occur through both indirect and direct discrimination; *disparate impact*, *disparate treatment*, and *disparate mistreatment* are common notions of unfairness to differentiate between these concepts. Disparate treatment happens when different groups receive different treatment based on their protected or sensitive attributes, whereas disparate impact happens when a system produces different outcomes for different groups [67, 199].

This systematic bias or discrimination by algorithms can cause both distributional and representational harm [96]. Distributional harm refers to the discrimination in resource allocation or distribution [39], whereas representation harm refers to the misrepresentation of individuals or groups [129]. In 2015, Amazon stopped using AI-based recruiting systems after they found out that the system was unfairly scoring candidate resumes reflecting gender stereotypes associated with occupation [41] which is an example of systems causing both distributional and representational harm.

Research work on algorithmic fairness includes fairness issues in classification and ranking algorithms. In classification algorithms, entities are assigned to a certain class based on their observed features. One common example of bias in classification algorithm is in the task of identifying the risk of recidivism. It is not possible to accurately identify the phenomena since the outcome is unobservable. Hence, classification algorithms are used to assess a defendant's likelihood of committing a

crime based on their other observable aspects and previous studies showed that the classification model can cause discrimination in outcome based on demographic attributes [115, 54] especially disadvantaging African-Americans over white defendants [39]. Similar problems appear in other classification algorithms such as loan approval, college admission decisions, and hiring applications [134, 46, 159].

In ranking algorithms, fairness concerns involve both ranked items and users who are interacting with the ranked results. Item position in ranking affects the advantage that items receive from the system because items at the top position receive more exposure and user engagement than items at lower positions [165]. However, items with similar qualities do not always get the same position in the ranking, thus causing *position bias* [43, 166]. Moreover, systematic discrimination in positioning items in ranking can happen based on the sensitive attributes of items such as race and gender[165, 186, 202]. On the other hand, user experience with ranking can vary across users [1] and users can receive biased ranked results based on their sensitive attributes [184, 126, 107]. Moreover, systems may misrepresent items through ranked results [66]. For example, women are significantly under-represented in Google image search results for "CEO" [68].

As I explained in section 2.1, IAS display results in ranking to users. Since algorithms may reflect bias in ranking, fairness is a critical and important issue to consider in IAS research.

2.3 Fairness in Information Access Systems

This section provides a literature review on fairness in IAS showing the importance and the challenges of evaluating ranking considering fairness and highlights the research gaps in this area.

IAS may unfairly discriminate across items while displaying them to users which can happen at both individual [176] and group levels [76, 162]. Bias in IAS can appear in following ways:

- IAS can recommend or retrieve items with high user interaction thus causing *popularity bias* [176] and reflecting *rich gets richer* effect.
- Systems can cause *disparate exposure* where items with similar quality do not receive similar exposure because of their sensitive attributes, for example, in micro-lending loan recommendation systems, borrowers from minority groups can receive lower exposure [111] causing disparate exposure [52].
- Additionally, users may receive unfair ranked results based on their group membership such as users receiving biased job recommendations based on their gender identity [152, 47].
- Systems can manifest and propagate pre-existing social stereotypes through search results or recommendations [141, 65, 114]. For example Noble [129] showed that search engines often negatively represent "black women" in retrieved results.

By denying fair distribution of opportunities and misrepresenting entities reflecting social bias, ranking in IAS can cause both distributional and representational harm.

Ekstrand et al. [63] provide a thorough survey on fairness and bias issues in IAS; Pitoura et al. [136] provide a taxonomy of fairness definitions particularly focusing on ranking and recommendations and Wang et al. [186] provide a taxonomy of fairness definitions in the context of recommendations. Deldjoo et al. [50] provide overviews of the research on fairness in recommender systems and highlight the potential future directions in this area.

Fairness in IAS involves the concerns of multiple stakeholders since systems involve users who interact with the systems to consume items and producers who create or provide the items that users are interacting with. For example, in a book recommendation system, users are interacting with books in the ranked list in order to read (consume) them and each book is associated with book providers, such as authors or publishers. Burke [26] and Sonboli et al. [167] introduced the terms *provider fairness* and *consumer fairness*. Consumer fairness refers to the user-centric fairness who interact or consume items in the ranking [184, 126] and provider fairness considers fairness for item producers, creators, or providers in IAS [136]. Item providers or producers are often associated with sensitive group attributes and items can be treated unfairly based on their providers' groups membership.

Fairness in Ranking in IAS As previously discussed in section 2.1, search engines and recommender systems often present the retrieved or recommended results in a ranked list form based on relevance to user information need or preference and bias can appear in ranking algorithms [202]. Ranking may introduce or reflect systematic bias; fairness in ranking in IAS focuses on social and ethical concerns in these systems for both providers and consumers [62, 103, 130]. Provider-side fairness in ranking involves systematic discrimination against providers while displaying items in ranking.

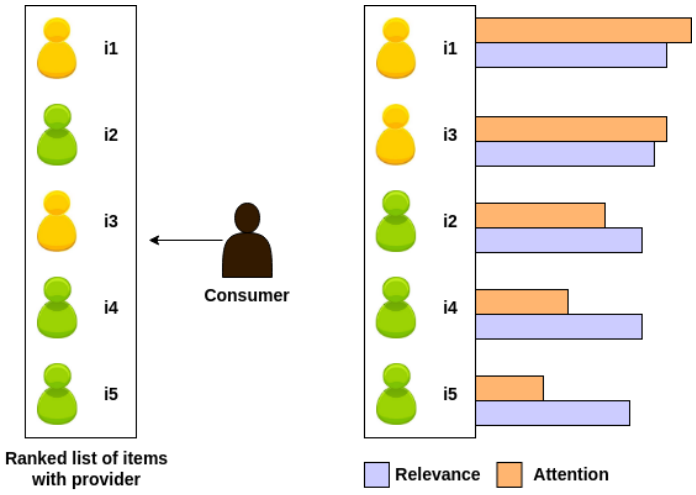


Figure 2.5: Item exposure and relevance distribution ranked results in IAS

In ranking, items receive *exposure* through their position in ranked lists and users provide *attention* to items which varies across positions of items in the ranked list [154, 122]. Users tend to engage with items in the top-ranked position resulting in *position bias* [198]. Figure 2.5 shows how user attention varies with the position of items in ranking and how relevance and attention for items varies with positions. A small change in ranking position may lead to significant changes in the *attention* paid to a result [16, 136]. Since the items are usually associated with item producers or providers, the discrepancy in attention affect the the economic return to item providers. Thus systems can have *disparate impact* on item provider through ranking [16]. For example, in the Spotify[1] music recommender system, music artists make profit from user interactions (play, share, or like) with their songs. Consequently, if a song does not get enough exposure or user attention in recommended ranked lists, it can economically disadvantage song artists. Moreover, systems can show systematic discrimination in allocating exposure across items based on their sensitive attributes

[1]https://open.spotify.com/

associated with items providers.

One way to measure this unfairness is *demographic parity* or *statistical parity*, which argues that systems should derive balanced outcomes irrespective of sensitive group attributes like race or gender ensuring similar decisions for both *protected* and *dominant* groups [73]. Another goal is to achieve *equality of opportunity* where qualified subjects should receive equal probability of favorable outcome regardless of group membership [87]. Disparate mistreatment or equalized odds defines the difference in error rates [199].

These systematic biases in ranking can appear due to existing bias in data and during the relevance estimation stage. Bias mitigation from ranked results can include removing bias from underlying data (pre-processing), auditing algorithms for bias in training phase (in-processing), and improve fairness in ranked results (post-processing) [136]. Improving fairness in ranked results or the post-processing methods often involve optimization of rankings for fairness [76, 166, 16, 195]. Various fairness aware re-ranking techniques have been introduced to optimize the ranked results considering fairness [111, 59, 78, 127].

2.4 Organization and Position

The focus of this work is on **provider-side group fairness** in ranking in IAS where the fairness concern is the *distributional harm* regarding unfair exposure distribution that item providers can receive based on their group membership. Fairness in ranking can be defined with respect to the discrepancy in incorrectly-distributed exposure for different groups of item providers in ranked results based on their sensitive attributes.

The existing fair ranking metrics significantly contribute to the area of measuring fairness in ranked output in IAS. However, there are open questions regarding which metrics to use, how to implement them, and how reliable they are. Furthermore, there exist open problems that still need to be considered in designing and applying fair ranking metrics in a real-world IAS setup. For example, measuring fairness in ranking in grid layout and optimizing the grid layout for fairness by incorporating grid layout suitable metrics are still open research questions to explore.

My work contributes to the fair ranking metric research by informing the research community about the strengths and limitations of fair ranking metrics and providing guidance to researchers and practitioners in selecting task-specific metrics and implementing them in the real-world IAS setup. Several others have also contributed fair ranking syntheses from various perspectives. Chapter 4 provides a comprehensive analysis of fair ranking metrics and more thorough comparison of my synthesis with those of my peers. Moreover, my work contributes to the advancement of measurement of group-fairness in ranked results in IAS by considering user browsing models that are suitable for grid layouts. Chapter 5 provides the modified fair ranking metrics with grid-based browsing models to measure provider-side fairness in grid layout. The existing user browsing models are conceptually similar, hence I generalize the existent user browsing models for both linear and grid layouts. Chapter 6 provides the generalized framework of user browsing models which can be configured based on ranking layouts, parameter settings, and the availability of several components. Lastly, I contribute to the fairness in ranking research by working on the optimization of ranking in grid layout for provider-side fairness. Previous research on fairness aware re-ranking techniques mostly focus on linear layouts ignoring fairness optimization in

ranking in other layouts. I work towards filling up this gap by providing fairness aware re-ranking techniques for ranking in widely-used grid layout in Chapter 7. Through my work on fairness in grid layouts, I provide potential research directions in this important but seldom-studied research area.

CHAPTER 3:

DATASET AND NOTATIONS USED IN THIS PAPER

In my dissertation research, I use real-world IAS datasets for all of the experiments. For some experiments, I use datasets for both search and recommendation scenarios. Moreover, since all the works described in this paper are on provider-side group fairness in ranking in IAS, I use consistent notations throughout the paper to provide a cohesive story of my PhD research. This chapter provides the description on the datasets I used in my work and the notations that are used throughout this paper.

3.1 Dataset

For most of the experiments, we use multiple datasets from real-world IAS setup considering both search and recommendations scenarios.

Search Systems For search systems, we use the dataset from the *TREC 2020 Fair Ranking Track* [18], which includes submitted runs (document rankings in response to a query) from participants in the TREC conference for both retrieval and re-ranking tasks. The participants' systems had to retrieve relevant documents for queries from

Table 3.1: Summary of experiment data.

Dataset	Systems	#Users	#Items	#Test Users	$\|\mathcal{G}^+\|$	$\|\mathcal{G}^-\|$
Amazon	4	8,026,324	2,268,142	5000	217032	490953
GoodReads	4	870,011	1,096,636	5000	177359	282857
Fair TREC20 Re-rank	23	195	2112	195	294	1632
Fair TREC20 Retrieval	5	189	2112	195	294	1632

Table 3.2: nDCG scores of the recommendation algorithms

Systems	Amazon	GoodReads
Item-Item	0.08	0.23
User-User	0.13	0.24
WRLS	0.10	0.26
BPR	0.03	0.13

the Semantic Scholar[1] corpus; documents are associated (soft group association) with author demographic information (socio-econnomic status of author's country) which was manually annotated by NIST assessors. The participants submitted multiples sequences of ranking for each query. We consider each submission to be an individual system. The participants provided details about their submitted systems in notebook papers published in the TREC proceedings [68, 118, 155, 99, 7].

Recommender Systems For a recommendation scenario, we use two user-book interaction datasets from *GoodReads* [183] and *Amazon* [117], integrated with the PIReT Book Data Tools[2] [59] to obtain author metadata. Table 3.1 shows the summary of datasets used in this paper. For both datasets, we generate 1000 personalized book recommendations for 5000 users using four collaborative filtering (CF) algorithms: user-based CF (UU [88]), item-based CF (II [51]), matrix factorization (WRLS [172]), and Bayesian Personalized Ranking (BPR [149]), as configured by

[1]https://www.semanticscholar.org/

[2]`https://bookdata.piret.info`

Ekstrand and Kluver [59]. We used *Lenskit for Python* [58] to generate recommendations using user's implicit feedback for items (binary feedback that was a 1 if the user added a book to one of their shelves). Table 3.2 shows the Normalized Discounted Cumulative Gain or $nDCG$ scores [92] for the recommendation systems where a higher score represents better recommendations.

Author gender identity is the sensitive attribute for our experiments. Due to limitations of the underlying data set [61], this is a discrete but possibly unknown binary gender attribute; we acknowledge the importance of more faithful representations of gender in research [135], and the metrics that we study can all be used with a larger set of gender identities as well as mixed or partial membership when such data can be obtained.

3.2 Notations

These notations help to describe an information access systems that retrieves or recommends a ranked list L of n documents or items $d_1, d_2, \ldots, d_n \in D$ in response to requests (e.g. queries in a search system or users and/or contexts in a recommender system) $q_1, q_2, \ldots, q_n \in Q$ (notation summarized in table 3.3). Items may have an associated request-specific relevance score $y(d|q)$, and the system may estimate this by a predictor $\hat{y}(d|q)$.

Providers are associated with one (or more) of g groups. We represent this by giving each item an alignment vector $\mathcal{G}(d) \in [0,1]^g$ (s.t. $\|\mathcal{G}(d)\|_1 = 1$) indicating its group association; generalizing from a categorical variable to a vector allows soft association (mixed or partial membership) or uncertainty about membership [154].

Table 3.3: Summary of notation.

$d \in D$	document or item
$q \in Q$	request (query or user)
L	ranked results of N documents from D
$L(i)$	the document in position i of linear (1-column) layout
$L^{-1}(d)$	rank of document d in linear layout
$\text{row}(d)$	row number of document d in grid layout
$L(\text{k}, \cdot)$	items in kth row in grid layout
$L(\text{k}, \text{c})$	items in row k and column c in grid layout
$y(d\|q)$	relevance of d to q
g	number of groups
$\mathcal{G}(d)$	group alignment vector
$\mathcal{G}(L)$	group alignment matrix for documents in L
$\mathcal{G}^+(L)$	set of documents in protected group in L
$\mathcal{G}^-(L)$	set of documents non-protected group in L
$\hat{\mathbf{p}}$	target group distribution
$\mathbf{a}_L$	attention vector for documents in L
$\mathbf{a}_L(d)$	position weight of d in L
$\boldsymbol{\epsilon}_L$	the exposure of groups in L ($\mathcal{G}(L)^T\mathbf{a}_L$)
E_i	event: user examines the item at position i
S_i	event: user selects the item at position i
A_i	event: user abandons the process after examining the item at position i
K_k	event: user skipping the kth row.

Table 3.4: Parameters of browsing Models and the range of parameter values

Parameter	Name	Description	Values
ψ	Selection Probability	Probability of selecting an item at position i	{0.1, 0.2, ..., 0.9}
α	Abandon Probability	Probability of abandoning the process.	{0.1, 0.2, ..., 0.9}
λ	Continuation Probability	Probability of continuing to the position i	{0.1, 0.2, ..., 0.9}
γ	Skipping Probability	Probability of skipping an entire row.	{0.1, 0.2, ..., 0.9}
β	Decay	Incorporate slow browsing tendency for grid layout	{1.1, 1.2, ..., 2.0}

We generalize $\mathcal{G}(d)$ to a list function, with $\mathcal{G}(L)$ denoting an $n \times g$ alignment matrix whose rows correspond to the items in L and columns are groups. In the case of definitively-known membership in a binomial pair of groups, $\mathcal{G}^{+}(L)$ denotes the set of items in L in the "protected" group and $\mathcal{G}^{-}(L)$ the remaining items (dominant group). The rest of the notations are explained later in the paper in the proper context.

CHAPTER 4:

FAIR RANKING METRICS ANALYSIS

As previously mentioned in chapter 2, *fairness* is a relatively new but important aspect of rankings to measure, joining a rich set of metrics that go beyond traditional accuracy or utility constructs to provide a more holistic understanding of IAS system behavior[1]. To measure the (un)fairness of rankings, particularly with respect to the protected group(s) of producers or providers, several metrics have been proposed in the last several years. However, an empirical analyses of these metrics concerning the applicability, usability, and reliability of these metrics is still lacking. We aim to bridge the gap between theoretical and practical application of these metrics. In this work, we describe several fair ranking metrics from the existing literature in a common notation, enabling direct comparison of their approaches and assumptions, and empirically compare them on the same experimental setup and data sets in the context of information access tasks. We also provide a sensitivity analysis to assess the impact of the design choices and parameter settings that go in to these metrics and point to additional work needed to improve fairness measurement.

Kuhlman et al. [103] compare selected fair ranking metrics for measuring the *statistical parity* of rankings (whether they provide equal exposure to different groups),

[1]This chapter is adopted and modified from author's previously published work [143] and [142] where I collaborated with Dr. Michael Ekstrand

and Zehlike et al. [202] provides a thorough conceptual survey of fair ranking constructs, but there has not yet been a systematic comparison of group fairness metrics for ranked IAS outputs (where the system provides different rankings in response to different information needs — both prior comparisons focus on rankings for a single need), or direct comparisons within the same data set and experiment.

Moreover, many of the metrics have been tested primarily on small and/or synthetic data sets that are not reflective of real-world information access applications and experiments. Realistic experimental settings present challenges for applying many metrics, including incomplete data (for both relevance and group membership) and the occurrence of edge cases such as a group with no retrieved (or relevant) items. Metrics need to be robust and usable in such situations in order to be practically useful in experiments and for auditing deployed applications. Metric results may also be heavily influenced by parameter choices and experimental designs. This is an important factor to consider when choosing framework-applicable metrics because metrics which are significantly sensitive towards external factors or design choices are more complex to apply, as those decisions must be calibrated appropriately. Therefore, despite the progress in metrics for measuring fairness, both practitioners and researchers may have difficulty finding the most applicable metric for their problem setting and its requirements.

In this chapter, we seek to fill this gap: to provide a comparative analysis of fairness metrics in the context of information access, to better inform the community of their relative strengths and weaknesses, and facilitate both better application of existing metrics and further research to advance the state of the art in measuring ranking fairness. Our goal is not to identify universally best metrics; the essentially

contested nature of fairness [160] implies such a quest is futile. Rather, we want to connect fair ranking metrics with applications by providing insight into how to measure the provider-side group fairness of the ranked outputs in actual search and recommender systems experiments using these metrics. Further, there is not a fixed ground truth to use to assess external support for a metric — there are many potential fairness objectives with varying degrees of compelling arguments. We therefore seek to inform the discussion through internal support: documenting and comparing the structure of the metrics, and their varying behaviors over real data, to assess and suggest what strengths and weaknesses we can for each metric.

Our aim is to do for provider-side group fairness what Friedler et al. [72] did for fair classification metrics; this complements the thorough conceptual survey of fair ranking constructs and interventions in a general ranking setting by Zehlike et al. [202] and Kuhlman et al. [103]. We provide a concise treatment of fair ranking metrics specifically focused on measuring fairness in information access settings, and implement these metrics in a common experimental setting to show their results on the same data, systems, and tasks. This enables us to investigate the following:

- What is needed to apply these metrics to real IAS outputs, which often have missing data (including relevance judgments and group annotations), may have highly imbalanced outputs or relevant sets, or exhibit other edge-case behavior?

- What are the actual substantive differences between these metrics, once superficial differences in framing and notation are resolved?

- What are the design decisions and parameters involved, and how sensitive are the resulting metrics to those decisions?

- What are the empirical differences in how these metrics assess the relative fairness of different recommendation algorithms or retrieval runs?

In this work, we make four contributions:

- We describe rank fairness metrics in a unified notation for information access, identifying similarities and differences.
- We identify gaps between the conceptual form and the practicalities of applying the metrics to both search and recommender system evaluation experiments.
- We directly compare the outcomes of these metrics with the same data and experimental settings[2].
- We conduct sensitivity analysis to assess the impact of design choices and external factors on these metrics.

From our results we highlight the strengths and limitations of the metrics, finding that some of them are particularly sensitive to edge cases and/or parameter settings. We conclude with recommendations for choosing metrics from the current state of the art for different experimental settings, and pointers to further research that is needed to fill out our understanding of fair ranking measurement.

4.1 Related Work

In this section we introduce brief summary of previous research concerning the group-side provider fairness measurement issue in ranking in IAS.

[2] `https://github.com/BoiseState/rank-fairness-metrics`

Provider-side Fairness in Ranking Research on provider-side fairness often focuses on the phenomenon of disparate exposure in ranking where systems shows discrimination in distributing exposure across items. To identify if the uneven distribution of exposure is unfair, several research works have been done on fairness in ranking, considering both individual and group fairness [16, 52, 165]. The fairness target of these works consider *equality of opportunity* where the system should ensure equal outcome for items with similar qualities [87]. Some of these works focused on measuring fairness in ranking and proposed metrics for that purpose which I discuss in this chapter.

Provider-side Group Fairness in Ranking Item producers or providers are often associated with sensitive attributes like age, gender, and race etc and item exposure can vary based on their association with particular protected attributes [80, 188, 111, 78]. In figure 2.5, items are categorized into two groups- yellow and green and in that ranking, we can see that items from group yellow are ranked over group green, thus causing uneven distribution of exposure for these two groups.

Therefore, the fair ranking research area concerning provider-side group fairness mostly target *demographic parity* and *equality of opportunity* [127] and several theoretical and practical methods have been proposed to measure fairness in this regard [16, 154, 52, 14, 128, 165, 195, 202, 63, 31, 103]. In demographic or statistical parity, the fairness metrics compares group distribution in ranking with target population [195, 154]. In equality of opportunity, the fairness metrics measure if items with similar relevance are receiving unequal exposure or attention based on their group membership, thus this family of metrics compare the received attention or exposure across group of items or item providers [165, 52, 16]. Furthermore, in any single rank-

ing only one item may be placed at the top of the list, and will be the only item to accrue the benefits of first position regardless of the merit of the second item. Hence, it is impossible to ensure equal exposure for items in a single ranking which led to the introduction of *stochastic ranking* or distribution over rankings [52].

Measuring fairness in ranking is a developing research area [158] with challenges like multinomial protected attributes, non-binary group membership or soft-group association, missing group labels [79], or uncertainty in demographic inference [80]. Intersectional-group fairness is gaining attention to ensure fair ranking for multiple sensitive attributes simultaneously [81, 64]

The existing metrics significantly contribute to the area of measuring provider-side fairness in ranked output in IAS. However, there are questions that remain regarding which metric to use, how to implement them, and how reliable they are. Furthermore, there exist open problems that still need to be considered in designing and applying fair ranking metrics in real-world IAS setup. This work contributes to the fair ranking metric research by informing the research community about the strengths and limitations of fair ranking metrics and providing guidance to researchers and practitioners in selecting task-specific metrics and implementing them in real-world IAS setup.

4.2 Fair Ranking Metrics

We begin by describing several fair ranking metrics, summarized in table 4.1 and table 4.2, in a common framework and notation. This enables direct comparison of their designs and theoretical behavior, and facilitates easier implementation in IAS experiments. In some cases, we assign new name for metrics based on their

functionality, purpose, and comparability within our synthesis.

4.2.1 Problem Formulation

We consider an information access system that retrieves or recommends a ranked list L of n items $d_1, d_2, \ldots, d_n \in D$ in response to requests (e.g. queries in a search system or users and/or contexts in a recommender system) $q_1, q_2, \ldots, q_n \in Q$ (notation summarized in table 3.3). $y(d|q)$ denotes the request-specific relevance score for an item d, and the system may estimate this relevance by a predictor $\hat{y}(d|q)$.

Item are associated with item provider or producers and item providers are associated with one (or more) of demographic g groups. The group membership of items are represented by giving each item an alignment vector $\mathcal{G}(d) \in [0,1]^g$ (s.t. $\|\mathcal{G}(d)\|_1 = 1$); generalizing from a categorical variable to a vector allows soft association (mixed or partial membership) or uncertainty about membership [154]. We generalize $\mathcal{G}(d)$ to a list function, with $\mathcal{G}(L)$ denoting an $n \times g$ alignment matrix whose rows correspond to the documents in L and columns are groups. In the case of definitively-known membership in a binomial pair of groups, $\mathcal{G}^+(L)$ denotes the set of items in L in the "protected" group and $\mathcal{G}^-(L)$ the remaining items (dominant group).

Our goal is to measure *exposure* (sometimes called *attention*) for each item, content provider, or group receives, and assess the fairness of this distribution to ensure *demographic* or *statistical parity* (ensures comparable outcomes across groups) or *equality of opportunity* (ensures equal treatment based on merit or utility irrespective of the group membership). Accounting for the decreasing attention users are likely to pay to items at deeper rank positions (*position bias*) requires a browsing model; some metrics build this implicitly into their structure, while others explicitly model it as a *position weight vector* $\mathbf{a}_L$ for L. Table 4.4 describes the various weighting schemes

Table 4.1: Summary of fair ranking metrics.

Metric(s)	Goal	Weighting	Target
PreF_Δ [195]	Each prefix representative of whole ranking	—	$\hat{p}$ from full ranking
AWRF_Δ [154]	Weighted representation matches population	Geometric	configured $\hat{p}$
FAIR [201]	Each prefix matches target distribution	—	binomial $\hat{p}$
logDP [165]	Exposure equal across groups	Logarithmic	equality
logEUR [165]	Exposure proportional to relevance	Logarithmic	$\propto$ utility
logRUR [165]	Discounted gain proportional to relevance	Logarithmic	$\propto$ disc. utility
IAA [16]	Exposure proportional to predicted relevance	Geometric	$\propto$ est. utility
EEL, EER [52]	Exposure matches ideal (from relevance)	Cascade, RBP.	f(utility)
EED [52]	Exposure well-distributed	Cascade[a], RBP.	equality
PAIR [14, 128]	Pairwise preference accurately modeled across groups	—	equal accuracy

[a]Cascade weighting also incorporates relevance into exposure, even if exposure is not compared to relevance.

Table 4.2: Summary of fair ranking metrics.

Metric(s)	Binomial?	Range	More Fair
PreF_Δ [195]	Dep. on Δ[a]	$[0, 1]$	0
AWRF_Δ [154]	Dep. on Δ	$[0, 1]$	0
logDP [165]	Yes	$(-\infty, \infty)$	0
logEUR [165]	Yes	$(-\infty, \infty)$	0
logRUR [165]	Yes	$(-\infty, \infty)$	0
IAA [16]	No	$[0, \infty)$	0
EEL, EER [52]	No	$[0, \infty)$	EEL 0, EER >
EED [52]	No	$[0, \infty)$	0
PAIR [14, 128]	No	$[0, 1]$	0

[a] Δ_{RD} and Δ_{RD} both require binomial protected group attributes, but Δ_{KL} generalizes.

Table 4.3: Parameters of weighting Models for computing $\mathbf{a}_L(d)$ and the range of parameter values with default values

Parameters	**Values**	**Browsing Models**	**Default Values**
Abandon Probability α	{0.1, 0.2, ..., 0.9}	Cascade	0.5
Stopping Probability ψ	{0.1, 0.2, ..., 0.9}	Cascade Geometric	0.5
Patience Probability λ	{0.1, 0.2, ..., 0.9}	RBP	0.5

used by the metrics we survey. The resulting exposure is then sometimes compared with a *target distribution* $\hat{\mathbf{p}}$ that represents across groups. There are several ways of computing $\hat{\mathbf{p}}$, including strict group equality, an estimate of the population of actual or potential content providers, or the distribution among providers of relevant items.

4.2.2 Statistical Parity in Single Rankings

We begin with metrics that assess the fairness of a single ranking and only measure exposure equity without considering relevance (that is, they target *statistical parity*). These metrics can be aggregated over the rankings produced by a system, e.g. by taking the mean, to produce an overall system fairness score.

Table 4.4: Default weighting models for computing $\mathbf{a}_L(d)$

Metric	Model	Formula
AWRF, IAA	Geometric	$\psi(1-\psi)^{L^{-1}(d)-1}$
logDP, logEUR, logRUR	Logarithmic	$1/\log_2 \max\{L^{-1}(d), 2\}$
EER, EED, EEL	RBP	$\lambda^{L^{-1}(d)}$
EEL, EED, EER	Cascade	$\alpha^{L^{-1}(d)-1} \prod_{j\in[0,L^{-1}(d))} [1-\psi\,(y(L(j)\|y))]$

Table 4.5: Distance functions for comparing distributions.

Distance Function	$\hat{\mathbf{p}}^a$	Formula
$\Delta_{\mathrm{ND}}(L, \hat{\mathbf{p}})$	Binomial	$\frac{\lvert\mathcal{G}^+(L)\rvert}{N} - \hat{\mathbf{p}}$
$\Delta_{\mathrm{RD}}(L, \hat{\mathbf{p}})$	Binomial	$\frac{\lvert\mathcal{G}^+(L)\rvert}{\lvert\mathcal{G}^-(L)\rvert} - \frac{\hat{\mathbf{p}}}{1-\hat{\mathbf{p}}}$
$\Delta_{\mathrm{KL}}(L, \hat{\mathbf{p}})$	Multinomial	$D_{\mathrm{KL}}(\hat{\mathbf{p}}(L)\|\hat{\mathbf{p}})^b$
$\Delta_{\mathrm{AD}}(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}})$	Binomial	$\lvert\frac{\lvert\mathcal{G}^+(L)\rvert}{N} - \hat{\mathbf{p}}\rvert$

[a] Binomial $\hat{\mathbf{p}}$ is a scalar probability of the protected group.
[b] K-L divergence; $\hat{\mathbf{p}}(L)$ is the probability distribution of groups in L.

The simplest way to measure the fairness of a single ranking is to measure the proportion of items in each group [59], but this does not account for position bias. Yang and Stoyanovich [195] proposed a family of statistical parity measures that incorporate position bias by averaging parity over successive prefixes of the ranking; we call this the *prefix fairness* family (PreF_Δ). These metrics are optimized when the representation in each prefix matches the target $\hat{\mathbf{p}}$ as closely as possible, as measured by a distance function Δ; Yang and Stoyanovich used the full ranking's composition as $\hat{\mathbf{p}}$, and instantiate PreF_Δ with distance functions Δ_{ND}, Δ_{RD}, and Δ_{KD} (from Table 4.5) to yield different members of the family. The metric is defined as

$$\mathrm{PreF}_\Delta(L) = \frac{1}{Z} \sum_{i=10,20,30,\ldots}^{N} \frac{\Delta(L_{\leq i}, \hat{p})}{\log_2 i} \tag{4.1}$$

where normalizing scalar $Z = \max_{L'} \mathrm{PreF}'_\Delta(L', \hat{p})$ (taken over all L' with the same

length and group composition as L, where PreF'_Δ is the prefix fairness function without the normalizer), scaling PreF_Δ to the range $[0, 1]$ where 1 is the maximum unfairness. Δ_{KL} has the advantage of allowing multinomial protected attributes and soft group association. PreF_Δ does not work when $\mathcal{G}^-(L) = \emptyset$, and Δ_{RD} does not work when $\mathcal{G}^-(L)$ is small. Z is also troublesome to compute with incomplete group membership data.

Zehlike et al. [201] propose a similarly-motivated group fairness constraint for a single list and fixed membership in binomial groups: L satisfies the FAIR constraint if for every prefix $L_{\leq k}$ with $1 \leq k \leq N$, the protected group is not statistically significantly under-represented. Unlike PreF_Δ, FAIR does not penalize over-representing the protected group. We convert this constraint into a metric by taking the average of the binomial probabilities:

$$\begin{aligned} \mathrm{FAIR}(L) &= \frac{1}{N} \sum_{k=1}^{N} P_{\mathrm{Binomial}} \left(m \leq |\mathcal{G}^+(L_{\leq k})| \hat{p}, k \right) \\ &= \frac{1}{N} \sum_{k=1}^{N} \sum_{j=1}^{|\mathcal{G}^+(L_{\leq k})|} \binom{k}{j} (\hat{p})^j (1-\hat{p})^{k-j} \end{aligned} \tag{4.2}$$

Sapiezynski et al. [154] provide a more general metric for single-list fairness by using an explicit (and configurable) position weight model instead of embedding the browsing model in the metric structure. Given an alignment matrix $\mathcal{G}(L)$ and suitably normalized position weight vector $\mathbf{a}_L$, $\boldsymbol{\epsilon}_L = \mathcal{G}(L)^{\mathrm{T}} \mathbf{a}_L$ is a distribution that represents the cumulative exposure of the various groups in L. The resulting unfairness metric, which we call *Attention-Weighted Rank Fairness* (AWRF_Δ), is the difference between

this exposure distribution and the population estimator:

$$\text{AWRF}_{\Delta}(L) = \Delta(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}}) \tag{4.3}$$

AWRF_{Δ} allows soft association and multinomial protected attributes. The distance function in Table 4.5 depends on application context; for assessing a particular protected class representation, difference in probability is suitable distance.

4.2.3 Statistical Parity in Multiple Rankings

In many cases, fair exposure cannot be achieved in a single ranking, because the attention paid to rank positions often decreases more steeply than the utility (relevance) of items [16, 52]. One solution is to measure fairness over sequences or distributions of rankings so providers have comparable opportunity to be exposed in at least some sessions or responses. This approach can be modeled as a request-dependent distribution (or *policy*) $\pi(L|q)$ over rankings [165, 52]. We extend this to include a distribution over requests $\rho(q)$, so a sequence of rankings $L_1, L_2, \ldots, L_{\tilde{n}}$ [16] is a series of draws from the distribution $\rho(q)\pi(L|q)$. The group exposure within a single ranking from Eq. 7.3, $\boldsymbol{\epsilon}_L = \mathcal{G}(L)^T \mathbf{a}_L$, is the fundamental building block of these metrics, along with their expected value:

$$\boldsymbol{\epsilon}(q) = \text{E}_{\pi}[\boldsymbol{\epsilon}_L] = \sum_{L} \pi(L \mid q)\boldsymbol{\epsilon}_L$$

$$\boldsymbol{\epsilon}_{\pi} = \text{E}_{\pi\rho}[\boldsymbol{\epsilon}_L] = \sum_{q} \rho(q)\boldsymbol{\epsilon}(q)$$

Singh and Joachims [165] and Diaz et al. [52] each propose metrics for measuring statistical parity over ranking policies. Neither metric incorporates a target distribu-

tion; they are optimal when all groups are equally exposed. Demographic parity [DP, 165] measures the difference in exposure between two groups[3]:

$$\text{DP} = \boldsymbol{\epsilon}_\pi(\mathcal{G}^+)/\boldsymbol{\epsilon}_\pi(\mathcal{G}^-) \tag{4.4}$$

Expected exposure disparity [EED, 52] ensures well-distributed exposure by measuring the inequality in exposure distribution across groups with the L_2 norm:

$$\text{EED} = \|\boldsymbol{\epsilon}_\pi\|_2^2 \tag{4.5}$$

4.2.4 Equal Opportunity in Multiple Rankings

So far, none of the metrics we have discussed account for the utility of the ranked results — rankings do well by exposing providers regardless of the utility of their items. The intuition behind incorporating utility, articulated independently by Singh and Joachims [165] and Biega et al. [16], is that exposure should be proportional to relevance: if an item or a group contributes 10% of the relevance to a request (user and/or query), it should receive approximately 10% of the exposure. This is a ranked analog of the *equality of opportunity* construct from fair classification [87]: outcome is conditionally independent of group given utility.

To measure deviation from this goal, Singh and Joachims [165] propose two metrics. The *exposed utility ratio* (EUR)[4], measures deviation from the goal that each

[3]The original paper presented a constraint, not a metric, for demographic parity; we have implemented it as a ratio to be consistent with the other metrics.

[4]Singh and Joachims [165] used the terms "disparate treatment ratio" and "disparate impact ratio" for EUR and RUR, respectively, but this terminology is not consistent with the use of these terms in the broader algorithmic fairness literature as we understand it. Exposure the system gives to providers is an impact, not a treatment. We have changed the names to hopefully reduce confusion going forward.

group's exposure is proportional to its contributed utility (measured by $\Upsilon(\mathcal{G}) = \mathrm{E}_\rho[\frac{1}{g}\sum_{i\in g} y(d|q)]$):

$$\mathrm{EUR} = \frac{\boldsymbol{\epsilon}_\pi(\mathcal{G}^+)/\Upsilon(\mathcal{G}^+)}{\boldsymbol{\epsilon}_\pi(\mathcal{G}^-)/\Upsilon(\mathcal{G}^-)} \tag{4.6}$$

The *realized utility ratio* (RUR) incorporates utility into both numerators and denominators by measuring whether the *discounted* utility contributed by each group ($\Gamma(\mathcal{G}) = \sum_{d\in\mathcal{G}} \mathrm{E}_{\pi\rho}[\mathbf{a}_L(d)y(d|q)]$) is proportional to its total utility:

$$\mathrm{RUR} = \frac{\Gamma(\mathcal{G}^+)/\Upsilon(\mathcal{G}^+)}{\Gamma(\mathcal{G}^-)/\Upsilon(\mathcal{G}^-)} \tag{4.7}$$

As they are based on ratios between group metrics, EUR and RUR do not support multinomial protected groups or soft association.

Biega et al. [16] present the *amortized attention* construct to measure exposure over the sequence of rankings. This compares rank exposure with expected utility $\hat{\Upsilon}$ (computed with system-predicted utility $\hat{y}(d|q)$) instead of ground truth relevance assessments $y(d|q)$), measuring whether the system allocates exposure proportional to the utility it estimates items to have. Deviations from this goal are measured by taking the L_1 norm of the group exposure-utility differences, yielding the *Inequity of Amortized Attention* (IAA) metric:

$$\mathrm{IAA} = \|\boldsymbol{\epsilon} - \hat{\Upsilon}\|_1 \tag{4.8}$$

Diaz et al. [52] build on this to integrate relevance in a different way. Rather than relate exposure directly to relevance, they use relevance to derive *target exposure* based on an ideal policy τ that assigns equal probability to all rankings that place

items in non-decreasing order of relevance and 0 (or miniscule) probability to all other rankings. The target exposure $\boldsymbol{\epsilon}^*$ is the expected exposure under the ideal policy ($\boldsymbol{\epsilon}^* = \mathrm{E}_{\tau\rho}[\boldsymbol{\epsilon}_L]$). They take the squared Euclidean distance between system expected exposure and target exposure, yielding the *Expected Exposure Loss*:

$$\mathrm{EEL} = \|\boldsymbol{\epsilon}_\pi - \boldsymbol{\epsilon}^*\|_2^2 \tag{4.9}$$

$$= \|\epsilon_\pi\|_2^2 - 2\boldsymbol{\epsilon}_\pi\boldsymbol{\epsilon}^* + \|\boldsymbol{\epsilon}^*\|_2^2 \tag{4.10}$$

The decomposition in Eq. 4.9 yields *expected exposure relevance* $\mathrm{EER} = 2\boldsymbol{\epsilon}_\pi^{\mathrm{T}}\boldsymbol{\epsilon}^*$ (measuring the alignment of exposure and relevance, higher values represent better alignment) along with EED. Neither IAA nor the EE metrics distinguish between group over- or under-exposure; for both, 0 is perfectly fair and larger values are unfair, with no preferential treatment given to a protected group.

The common thread between these metrics, articulated by Diaz et al. [52], is that for a fixed information need, differences in exposure between items with the same relevance grade results in unjustifiably unfair outcomes. Relating exposure to relevance sets the goal that items of comparable relevance should have comparable opportunity to be exposed, as measured by expected or amortized exposure over repeated rankings.

4.2.5 Pairwise Metrics

Beutel et al. [14] and Narasimhan et al. [128] define fairness objectives over pairwise orderings instead of entire rankings; like Bayesian Personalized Ranking [149], this treats the ranking problem as a binary classifier to predict relative ordering of pairs. Pairwise fairness is then defined in terms of the pairwise accuracy for ranking relevant

items in different groups:

$$A_{\mathcal{G}^1>\mathcal{G}^2} = P\left(d_1 \succ_L d_2 \mid d_1 \succ_{y|q} d_2, d_1 \in \mathcal{G}^1, d_2 \in \mathcal{G}^2\right) \tag{4.11}$$

where

$$d_1 \succ_L d_2 = L^{-1}(d_1) < L^{-1}(d_2) \qquad d_1 \text{ ranks above } d_2$$

$$d_1 \succ_{y|q} d_2 = y(d_1 \mid u) > y(d_2 \mid q) \qquad d_1 \text{ more relevant than } d_2$$

A ranking satisfies *pairwise equal opportunity* if pairs of items are equally likely to be ranked consistently with their relevance regardless of the group membership of the items in the pair. This can be measured by the group's pairwise accuracy with respect to all items ($A_{G_i>:}$), its *inter-group* accuracy ($A_{G_1>G_2}$), or its *intra-group* accuracy ($A_{G_1>G_1}$). Given protected and unprotected groups, we can define a fairness metric as the difference in pairwise accuracy:

$$\text{PairAcc} = A_{G^->:} - A_{G^+>:}$$

$$\text{IntraAcc} = A_{G^->G^-} - A_{G^+>G^+}$$

$$\text{InterAcc} = A_{G^->G^+} - A_{G^+>G^-}$$

4.2.6 Assessing Metric Design

Rendering metrics in a common notation shows that the metrics are quite similar in their basic concepts. The fundamental construct — weighted exposure — is the same across most metrics (pairwise fairness being an exception), and they differ primarily

in how they relate exposure to relevance and how they aggregate and compare exposure distributions. The following questions help identify more precisely what their salient differences are and how those may relate to particular IAS applications and experimental settings.

Does the metric incorporate relevance? EEL, EER, EUR, RUR, IAA, and PAIR directly incorporate relevance into metrics; others strictly measure statistical parity. It is desired depending on the precise task and evaluation goal. Statistical parity metrics are useful for measuring relative fairness of rankings already optimized for utility, particularly when there is no relevance information available or the relevant sets for a query are large. They can also be used to detect discrepancies that may indicate unfairness in relevance data (if relevance data is unfair, such as by systematically under-estimating the relevance of a group's documents, a metric that relates exposure to relevance will use the unfair relevance to justify unfair disparities in exposure). However, using such metrics in isolation for evaluation or optimization may reduce ranking quality.

How does it handle missing data? Real-world data sets are often incomplete, missing relevance and/or group labels for many documents. Metrics that are less sensitive to that problem will be easier to apply in such cases. Missing relevance data affects EUR, RUR, EER, and EEL like it does classical IAS evaluation metrics such as nDCG; the straightforward but biased approach is to treat items with unknown relevance as irrelevant ($y = 0$). IAA's use of system-estimated relevance allows it to sidestep this problem.

Missing group labels require different handling. For many metrics, we can include unlabeled items when computing attention weights but exclude them from further

analysis, or treat "unknown" as an additional group identity. Unknown data is a more significant problem for PreF_Δ family because it treats a list with fewer than 10 known-group items as maximally fair, and the straightforward way of computing Z — make the ranking maximally unfair by putting all protected items last — does not work in the face of missing data.

How does it respond to edge cases? Realistic IAS experiments bring a number of important edge cases, such as groups with no items relevant to or retrieved for a request. Ratio-based metrics and distance functions are particularly vulnerable to these problems; the EUR metric and the Δ_{RD} distance function, for example, approach infinity as the number of non-protected-group items retrieved goes to zero. RUR is even more brittle, as it requires nonzero relevance from retrieved non-protected-group items to avoid infinity, and both it and EUR can be infinite or undefined if the set of relevant items from either group is zero.

Reformulation of DP*,* EUR *and* RUR*:* Since these three metrics are ratios, their maximally fair point is 1, with a nonlinear relationship between values favoring and disfavoring the protected group, hindering interpretability; further, they approach ∞ if the dominant group exposure is close to 0. To improve interpretability, we take the logs of these ratios, so 0 is fair and distance is symmetric in either direction; and we address the empty-group problem by adding a small damping constant to both sides of the ratio. This yields the following reformulation for DP:

$$\text{logDP} = \log\left(\epsilon(G^+) + 10^{-6}\right) - \log\left(\epsilon(G^-) + 10^{-6}\right)$$

logEUR and logRUR are defined equivalently. As log ratios, values greater than 0 indicate a bias in favor of the protected group.

What is the target? PreF_Δ, FAIR, AWRF_Δ, EEL, and EER provide flexibility in determining how the (un)fairness of exposure is ultimately assessed through selection of the target distribution, while targets are implicitly baked in to the structure of others. This configurability is useful because it allows the metric to be adapted to the fairness requirements of a particular task, although it can impair comparability between experiments.

How does the metric compare the system with the target? Some metrics (AWRF_Δ and PreF_Δ) use an explicit distance function to compare distributions, while others use ratios of specific proportions or norms of differences in distributions. Norms and selected distance functions (such as Δ_{KL}) can accommodate soft association, while ratios and distance functions based on binomial probabilities require definitive membership in binomial groups. They can be adapted to some multi-group situations if only one group's exposure needs to be considered.

What user model does it use? Most metrics allow different position weighting strategies to be selected, both in its structure and its parameters. This configurability allows the metric to be adapted to specific application but introduces potential sensitivity towards the choices of weight functions and parameter values. PreF_Δ and FAIR are not configurable, as position weighting is built-in (as in PreF_Δ and $FAIR$) or unavailable (in PAIR).

The conceptual analysis of fair ranking metrics in a common framework and unified notations shows that the metrics within same fairness goal are not compatible for every experimental setting and fairness task. Figure 4.1 shows the common design decomposition of fair ranking metrics and various ways of incorporating those factors into measurement.

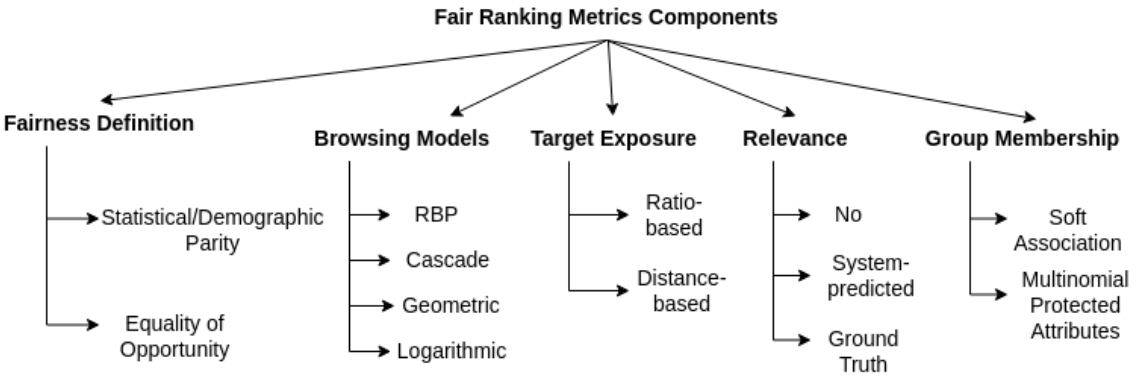


Figure 4.1: Fair ranking metrics design decomposition

4.3 Experimental Setup

We now turn from our analytical treatment of the metrics to an empirical comparison, using each of them (except for PreF_{Δ}, due to its difficulties with missing labels and soft membership) in real-world IAS experiments for three tasks across two problem settings:

1. Personalized book recommendations, measuring fairness with regards to author gender.

2. Scholarly article retrieval (both retrieval and re-ranking of short candidate sets) based on queries, measuring fairness with regard to the economic development of the author's country (as a proxy for the research resources available).

We further carry out a sensitivity analysis to understand how experimental outcomes change in response to design decisions and parameter values in the metrics.

This section describes the experimental setup itself, and the considerations we had to make when adapting the metrics in this setting. For our recommendation experiments, we used the *GoodReads* data [183] and for search experiments, we used

submitted runs and evaluations from the *TREC Fair Ranking Track* 2020 [18]. The description of the dataset used in in this chapter are provided in chapter 3. Table 3.1 shows summary statistics for each dataset.

4.3.1 Recommendation (GoodReads)

We measure the fairness of each recommendation list with respect to the gender of the book's author, extracted from Virtual Internet Authority File (VIAF)[5] (as described by Ekstrand and Kluver [59]). Group membership in this data is binary but incomplete, so we considered female authors to be the protected group G^+ and male authors G^- for all two-group metrics (unknown-author books are therefore ignored).

Metric Implementation For AWRF$_\Delta$, we used Δ_{AD} (following the original presentation [154]), and the distribution of male and female authors among the set of books in the data set as the population estimator. For IAA and the EE metrics, we treat unknown gender as a third author group.

Pairwise accuracy does not just depend on the top-N list — it is a function of the system's overall ranking between items. Therefore, we did not compute it from ranked output, but rather directly computed it from the recommendation model's scores for a sample of items. For each test item, we sampled 10,000 items not rated by the target user as negative examples, and used these to estimate the probability of correct orderings. This proved relatively efficient for our experiment size.

4.3.2 Search (FairTREC)

The runs from FairTREC data covered two tasks (re-ranking and full retrieval). We considered each submitted run as an individual system and used the given sequences

[5]http://viaf.org/viaf/data/

of rankings for each system. For the re-ranking task, we only used one run from each participating team. The details about the systems can be found in participants' notebook papers [68, 118, 155, 99]

Metric Implementation Unlike GoodReads, in the FairTREC data, each document has a soft association with the economic development level of its author(s), and thus we could not implement logDP, logEUR, and logRUR. Additionally, IAA uses system predicted relevance as ground truth, which makes it inapplicable in TREC setup (because systems do not provide scores for all items). We could not test these metrics on FairTREC because we do not have access to full rankings or the systems' relevance scores.

4.4 Empirical Results

We now present the results of our experiment, using both the metrics in their default configurations and conducting a sensitivity analysis with respect to weighting methods and parameters.

4.4.1 Direct Comparison

We begin by directly comparing the metrics with default parameter settings from their original papers to see how they assess each system in our experiments. Table 4.4 shows the default configuration of the metrics and table 4.3 shows the default values for the parameters. This comparison allows us to get a first view of the differences in results using each metric as originally presented, with minimal adjustments for practical implementation (see Section 4.2). We applied AWRF_{Δ} with two target distributions; AWRF_{Δ} computes $\hat{\mathbf{p}}$ from the distribution of providers in the full data set, while

AWRF_equal targets equal representation of protected and unprotected groups.

Figure 4.2 shows whether the metrics agree or disagree, and if this agreement is consistent across experiments along with their Kendall *tau* correlations. This figure does not show results from metrics that only worked on one experiment, but the metrics do not show clear consensus across datasets; there are substantial differences in their system orderings, and metrics that agree in one experiment don't often don't agree on others. The most consistently agreeing pair is EEL and AWRF_{Δ} with an equal-exposure target (positive correlation in all experiments, and comparatively high correlation in three of them). Figures 4.3 shows all the metric results from our experiments.

From this analysis we observe two things:

- Metrics frequently disagree on system orderings.
- Metrics that agree in one experiment don't necessarily agree on the others. The most consistently-agreeing pair is FAIR and AWRF_{Δ} metric, the two single-list metrics we study.

4.4.2 Sensitivity Analysis

Section 4.2 demonstrates that the fair ranking metrics often incorporate several design choices such as weighting strategies and parameter settings. However, this does not tell us how much difference these choices make in practice; if a metric is highly sensitivity towards design choices, it is more difficult to make correct configuration decisions (particularly in the absence of external guidance for those choices), increasing the complexity of applying it and the likelihood of error. To further analyze the applicability and sensitivity of these metrics, we need to know to what extent these

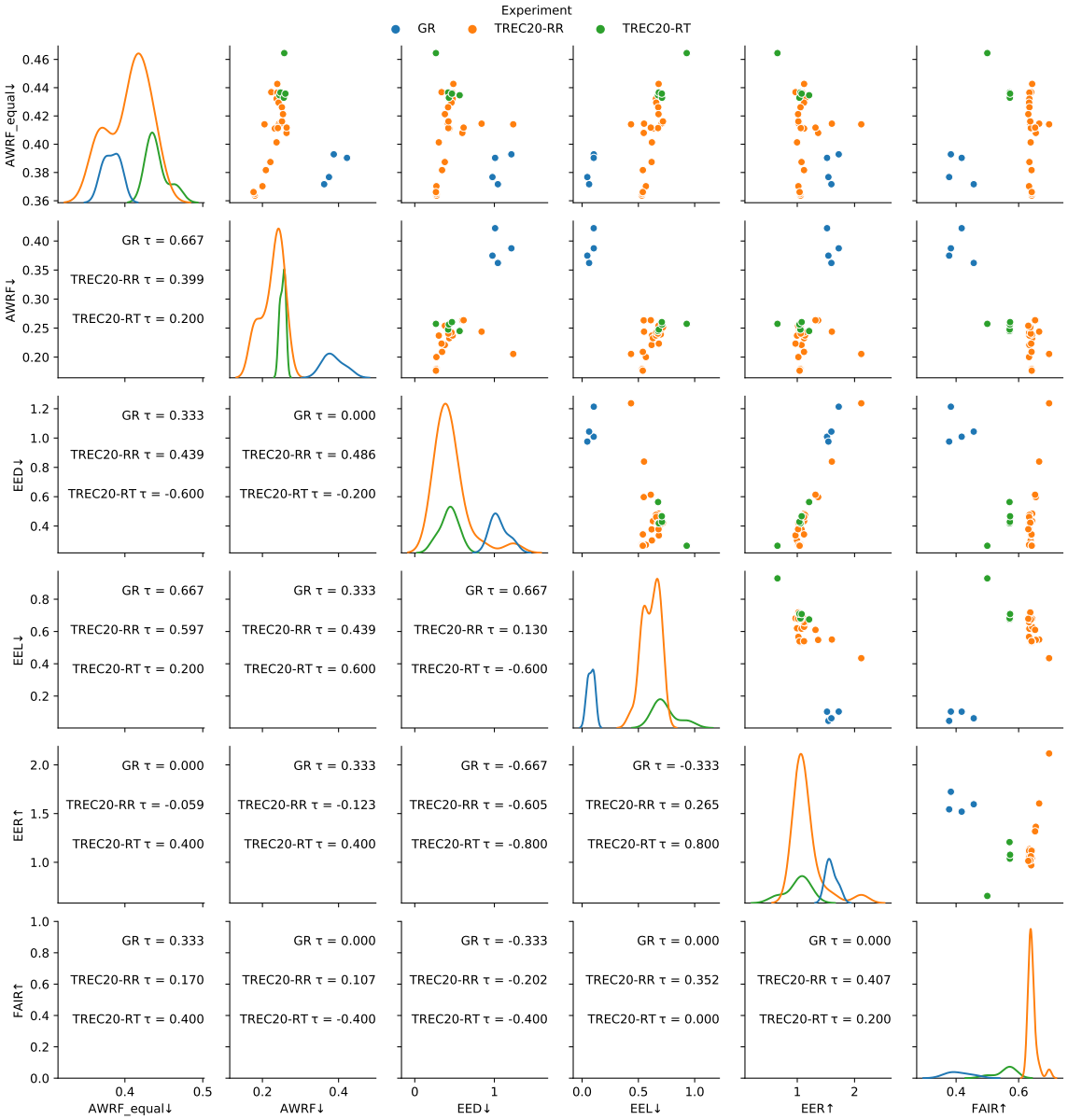
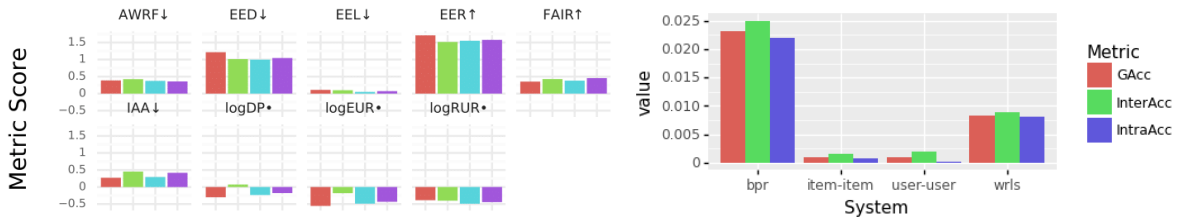
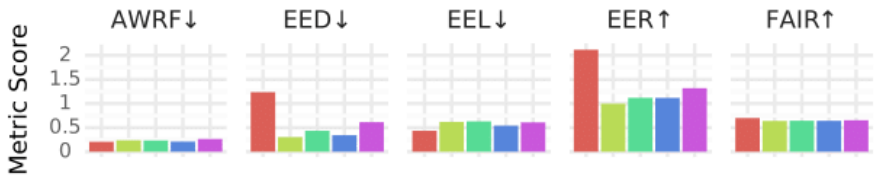


Figure 4.2: Metric results and correlations. Arrows indicate the direction of increasing fairness. Correlations computed with Kendall's τ-c within each experiment, ordering systems according to each metric's directionality.

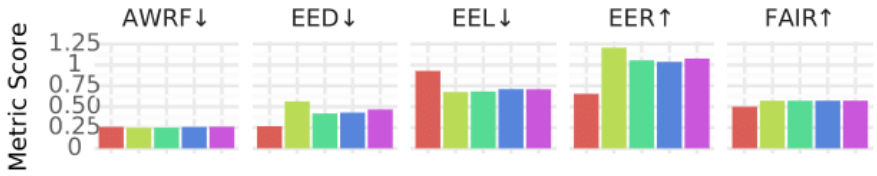
metrics are dependent on their decision choices. We now turn to understanding the impact of design decisions and parameter settings *within* each metric.



(a) GoodReads recommendation task



(b) FairTREC reranking task



(c) FairTREC full retrieval task

Figure 4.3: Fairness metrics *GoodReads* and *FairTREC* datasets using their original configurations. Arrow indicates direction of maximal fairness; · means 0 is fair. System identities are not relevant to our results.

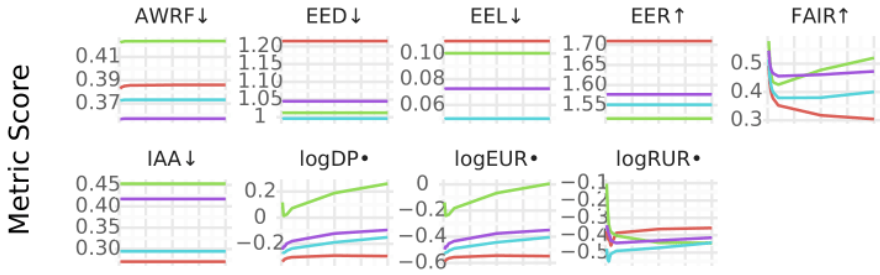
As we noted in Section 4.2, the exposure-based metrics and AWRF_Δ combine position weights and relevance in various ways; each was presented with particular position weighting strategy, but could be applied to any other. Further, most weighting strategies have parameters that affect the strength of the discounting. We test the sensitivity of metrics and conclusions drawn from them to the choice of ranked-list size, position weight formula, patience parameter, and stopping probability.

Size of Ranked List To observe the sensitivity towards ranked list size we apply the metrics lists of varying sizes (10–1000 for GoodReads recommendation, and 10–100 for FairTREC full retrieval). Fig. 4.4 shows the outcome of fairness metrics with the change of ranking length. We observe that:

- Changing ranked list size had no effect on any metric applied to FairTREC.
- AWRF_Δ, *IAA*, *EEL*, *EED*, and *EER* are mostly stable as the list length changes in the GoodReads recommendation experiment; they show slight changes through length 50, but without affecting system ordering, and then stabilize.
- *logDP, logEUR* and *FAIR* (on GoodReads) change notably, including reordering algorithms, as the list size changes.

Sensitivity towards ranked-list size of ratio-based metrics and FAIR in recommendation task indicates the need of studying metric dependency on relevance and group information availability.

Weighting Strategy For position-weighted metrics, we applied each metric to all four position weight models: *rbp*, *cascade*, *geometric*, and *logarithmic* (summarized



(a) GoodReads recommendation task

(b) FairTREC full retrieval task

Figure 4.4: Metric results with the change of ranked-list size.

in Table 4.4). Figure 4.5 shows the outcome of fairness metrics with the change of position weighting strategy. We use a continuation probability of 0.5 for the patience parameter and a stopping probability of 0.5. From these results, we observe:

- For the GoodReads recommendation task, logDP, logEUR, and AWRF_{Δ}, systems show differences with the change of weighting strategy, whereas for IAA, EER and EED, algorithms remain stable and did not show much disagreement across different weighting strategies. logRUR and EEL show extreme sensitivity towards the change of weighting model.

- In FairTREC reranking (Fig. 4.5(b)), systems show small differences but generally maintain system orderings across weighting models. We observe a few changes in order (e.g.AWRF_{Δ} from cascade to logarithmic) but these are be-

tween systems already very close.

- In FairTREC full retrieval (Fig. 4.5(c)), systems are generally stable across position models.

From the analysis, we observe that browsing models can have effects over some metrics' behavior to some extent, specially on EEL and logRUR. However, this analysis does not let us conclude that these metrics which showed stability over various weighting strategies will act uninfluenced with the change of parameters in weighting strategies. For further investigation, we measure the metrics by changing the parameter values.

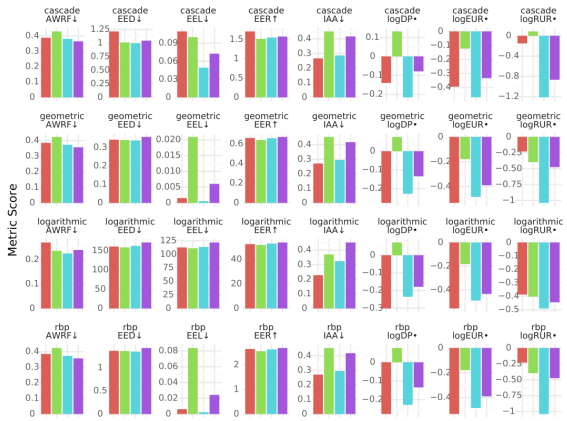
Patience Parameter Figure 4.6 presents the response of the metrics across patience parameter changes for the *rbp* and *cascade* weightings where we can see the following patterns:

- AWRF_{Δ}, EEL, EER, and EED show sensitivity towards the patience parameter following the same pattern in all three tasks (full retrieval, reranking, and recommendations)

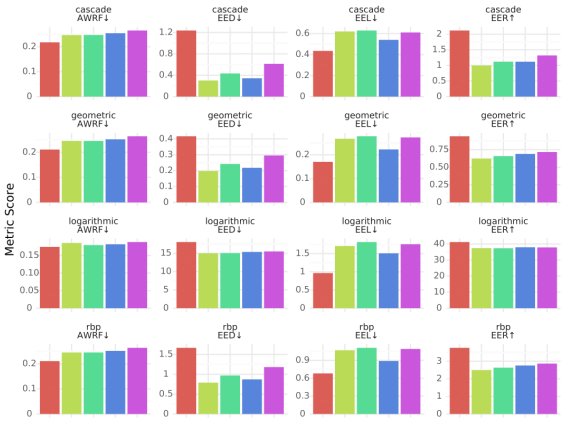
- In EEL, EED, and EER, systems show mild separation with each other following the same pattern across weighting strategies.

- On GoodReads recommendation tasks, logDP, logEUR, and IAA show substantial separation between systems; they also preserve system order as the parameter changed but the differences between systems shifted. The systems follow a similar pattern across weighting strategies.

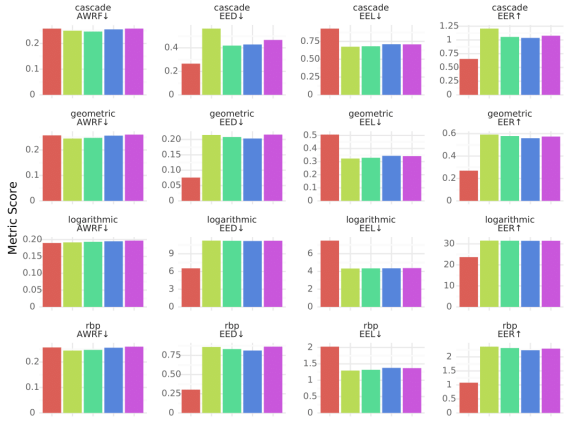
- logRUR is extremely sensitive to patience parameter changes.



(a) GoodReads recommendations task

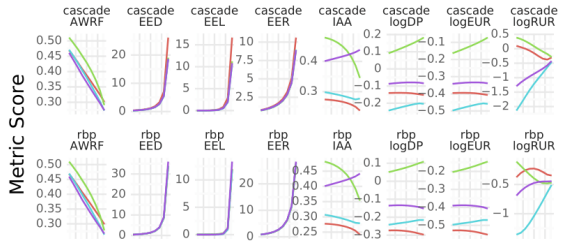


(b) FairTREC reranking task

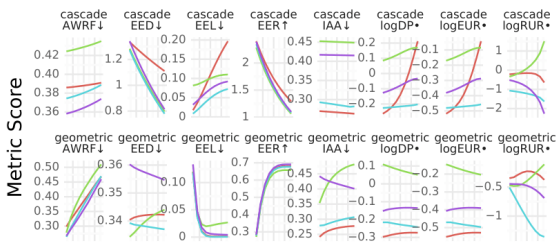


(c) FairTREC full retrieval task

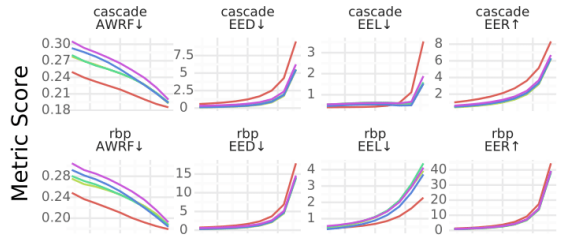
Figure 4.5: Metric results with the change of weighting strategy.



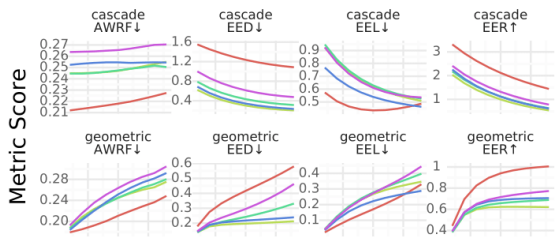
(a) Patience parameter on recommendation tasks



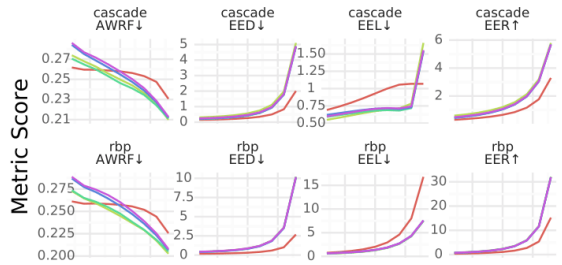
(b) Stopping probability on recommendation tasks



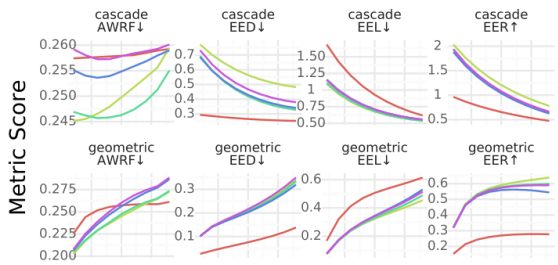
(c) Patience parameter on reranking task



(d) Stopping probability reranking task



(e) Patience parameter on full retrieval task



(f) Stopping probability on full retrieval task

Figure 4.6: Metric results with the change of external parameters.

Stopping Probability

Figure 4.6 shows the outcome of fairness metrics on the generated recommendations for *geometric* and *cascade* position weight models and the sensitivity towards the change of stopping probability. We have made the following observations from the charts:

- In FairTREC full retrieval and reranking tasks, metric results change with stopping probability. However, the systems did not vary significantly in the changing pattern.

- On the GoodReads recommendations task (figure 4.6(b)), IAA, logDP and logEUR shows sensitivity with the change of stopping probability and the sensitivity is notable in the cascade weighting strategy.

- In all three tasks, systems show complete inversion across weighting strategies for EEL, EED, and EER. In EEL, the pattern of sensitivity towards the stopping parameter is different between recommendation and ad-hoc tasks.

- logRUR is extremely sensitive to patience parameter changes.

Almost all metrics show sensitivity towards parameter value changes, which imply the necessity of identifying optimal parameter settings while implementing these metrics.

Overall, we observe that metrics do vary in their responses with the change of design choices, however, IAA, EER, EED and AWRF_{Δ} showed the most stability.

4.5 Discussion and Recommendations

We started this project with three goals:

1. Identify requirements to implement the fair ranking metrics in actual search and recommendation frameworks.

2. Identify similarities and both analytical and empirical differences among metrics to inform the metric selection process.

3. Identify the observable effects of different changes in the metric design or configuration.

Our analysis provides significantly more in-depth knowledge about the fairness goals, requirements, implementations, and effect of design decisions. In summary, our key findings are the following:

- Many metrics are remarkably similar in their underlying concept of fairness.

- Metric implementation highly relies on crucial factors such as group size, ranked list size, item relevance information, and group membership.

- Certain design choices can make metrics vulnerable to edge cases. For example, ratio-based metrics have difficulties with empty groups and zero values, such as a ranking that has no retrieved items from one of the groups.

- Despite having similar fairness goals, these metrics can differ in their sensitivity towards external factors.

This still leaves the question, however, of what we should do in the present to measure (un)fairness in ranking from real IAS datasets using a fair ranking metric. We propose to use metrics compatible with the following criteria:

- Allow multinomial protected attributes. Such metrics are applicable to a wider range of fairness settings, and choosing one means that the metric is not a reason to use a binary simplification of a multinomial attribute, such as gender.
- Allow soft group association (mixed or partial membership).
- Be stable with respect to design choices.

This last point is to support ease of use; if a metric is highly sensitive to design choices such as the attention weighting model, then its validity depends more strongly on the correctness of those choices. While a metric's validity with respect to the fairness objective in a particular application setting is the most important factor, given two comparably-appropriate metrics we would prefer one that is more robust to misspecification in its configuration. Based on these requirements, combined with our observations in sections 4.2 and 4.4, we make recommendations for different measurement goals and context based on the current state of the art and knowledge about fair ranking metrics:

Single Rankings All single-ranking metrics we considered are statistical parity metrics — they do not incorporate relevance. From our analysis, AWRF_{Δ} seems the most generally useful, because it supports multinomial protected attributes with soft assignment, and is adaptable to multiple attention models, target distributions, and difference functions. We are not yet able to make concrete recommendations for the choice of a difference function.

Demographic Parity in Sequences logDP and EED measure statistical parity on sequences of rankings. **EED** seems more generally useful because of its support for multinomial groups with soft membership, and was relatively robust with respect to design choices.

Equal Opportunity in Sequences The logEUR, logRUR, IAA, EER, and EEL metrics use relevance to measure (un)fairness in sequence of ranking, aiming at some version of equality of opportunity. We currently recommend using **EER** and **EEL** because of their support for multinomial groups with soft assignment, and comparative robustness. IAA shows comparable stability and can be adapted to multinomial groups and soft assignment; exploring that possibility is future work. In each context, position weighting model should be chosen based on user behavior in the expected context of use.

4.6 Conclusion and Future Direction

This chapter presents a comparative analysis among several fairness metrics recently introduced to measure fair ranking. We discuss the metric formulations and implications in an integrated notation and present the first (to our knowledge) empirical comparison of fair ranking metrics for recommendation and search systems in common data sets and fairness goals. We believe this comprehensive presentation and comparison among metrics will help future researchers and practitioners to make more informed decisions about metric choice and configuration.

Our findings from this empirical analysis point to several directions for future research. Further work is needed on the limitations we observe from implementing

the metrics in real data: implications and corrective methods for missing or sparse relevance information of items and missing [98], ambiguous, or multiple group associations [80] are not yet well-understood. Moreover, the instabilities we observe in our sensitivity analysis point to the need to work towards designing robust and efficient fair ranking metrics and developing a body of research that can lend external support for choosing where in the design space best meets a particular fairness goal. We also expect simulation studies will yield a much deeper insight into the differences we observed while applying metrics across different tasks and datasets, understanding more thoroughly the impact of factors like relevant set size, soft association, and missing relevance information, among others.

Significant progress has been made in the last few years on measuring the fairness of rankings, but more work is needed in order to understand how best to design and apply these metrics in wider settings. For example, measuring fairness in grid-based ranking layout has received limited attention even thought grid design is a widely used ranking layout in IAS. The following chapter addresses this gap where I describe my work on the advancement of fairness measurement in ranking by considering the applicability and reliability of fair ranking metrics in grid layout.

CHAPTER 5:

MEASURING PROVIDER-SIDE GROUP FAIRNESS IN GRID LAYOUT

The thorough analysis on the various components and design factors of existing fair metrics in chapter 4, shows that user attention (and probability of interaction) is an important factor in measuring fairness in ranking. User attention varies with items' positions in a ranking and governs the items' *exposure* to the users. Moreover, ranking design, task, or item metadata can affect user browsing behavior in ranking and the resulting attention [124, 192, 170, 12, 53]. However, the browsing models I discussed in previous chapter are suitable for linear or vertical ranked list, thus limiting the applicability of fair ranking metrics in wider range of ranking design such as, grid layout. It is unknown whether and how existing fair ranking metrics for linear layouts can be applied to grid-based displays. In this chapter[1], I explain our work on measuring *provide-side group* fairness in *grid* ranking layout. We extend existing fair ranking concepts and metrics to study provider-side group fairness in grid layouts, presenting an analysis of the behavior of these grid adaptations of fair ranking metrics, and study how their behavior changes across different grid ranking

[1]This work was done in collaboration with Dr. Michael Ekstrand. This work was submitted to RecSys'23 but got rejected and currently under-revision.

layout designs and geometries.

In chapter 4, I discussed provider-side group fairness in the context of ranking in IAS where provider-side exposure fairness is concerned with whether or not exposure and its benefits are distributed fairly [16, 52, 165, 142]. An "equality of opportunity" goal [87] would be to ensure that two providers whose items are equally useful to a user's information need have the same opportunity to be exposed to and consumed by that user, but systems do not always meet this criteria, instead providing *disparate exposure* [52]. Moreover, from analyzing existing fair ranking metrics, we realize that how users browse ranked results and provide attention to the exposed items in ranked results is important to know while measuring provider-side group fairness in ranking. For example, in rankings in linear layout, users tend to pay more attention to top-ranked items [198], so those items are more effectively visible to users and accrue greater benefit to their providers. Most of the existing metrics discussed in chapter 4 to measure fairness of exposure (or related constructs) in ranked lists [142] are designed for linear — usually vertical — layout models (figure 5.1(a)). However, many systems use other ranking layouts such visual grids or voice responses.

Grid layouts (figure 5.2) are particularly popular for streaming media platforms and image search, but also appear elsewhere; unfortunately, there has been little work to determine how to measure group fairness in such layouts, or how to measure fairness under different layouts (more than one of which can be employed in the same system). Measuring fairness of a ranking in grid layout using existing metrics by simply mapping the position of items in a grid layout to a linear layout can be problematic because user attention to items as a function of position varies between layout models [37]. For the same set of recommended or retrieved items, user attention

varies depending on how the items are being displayed, affecting item exposure and therefore the fairness of that exposure. Using fair ranking metrics without taking layout-specific user browsing behaviour into consideration may provide unreliable and erroneous results.

Furthermore, based on the device (phone, tablet, TV, laptop, etc.) used to interact with an information access system, the geometry of grid layouts varies, often re-ranking the list as the number of available columns changes. There are also multiple methods for adjusting the grid layout; for example, when moving from a wider to a narrower screen, some systems *truncate* the list at the right-side while others *re-wrap* the entire list where the right-most items appear in the left-most position. The impact of these layout adjustments on fairness scores is unknown. In summary, researchers and developers of IAS using grid layouts have little to work with when trying to reason about how the system layouts affect equity of exposure or how to apply the various metrics that have been developed to this setting.

In this chapter, we seek to fill this gap and broaden the applicability of fair ranking metric research by extending fair ranking metrics to grid layouts, providing the first (to our knowledge) study of metrics for this widely-used but under-studied paradigm. We adapt existing metrics to grid layouts by incorporating user attention models that are appropriate for grid displays in order to provide researchers and practitioners insights into what to expect when translating existing concepts from linear to grid layouts and lay groundwork for future research on measuring and providing fair exposure in grid display and interaction formats. Our goal is not to provide a metric recommendation that is suitable for grid layout rather, our study will guide researchers and practitioners in identifying required components to implement existing

fair ranking metrics in grid layout and understanding the applicability and reliability of the metrics in grid layout by providing insights on how the fairness scores change across layout models. Moreover, our findings will advance future research directions regarding fairness measurement considering various ranking layouts. The purpose of this work is to aid the development of fair ranking metric(s) that are able to address broader issues of real-world IAS applications by incorporating a frequently used layout model.

In this work, we observe what happens to group fairness for a list of recommended items with the change of layout model by answering the following research questions:

- **RQ1.** Do fairness measurements remain consistent across layout models?
- **RQ2.** Do rankings optimized for fairness in linear layouts remain fair in grids?
- **RQ3.** How do fairness scores change as grid size changes?

  - RQ3.a. Does the fair ranking metric score change when the grid layout is truncated?

  - RQ3.b. Does the fair ranking metric score change when the grid layout is re-wrapped?

  - RQ3.c Does the change in group-fairness score with column size reduction remain consistent across truncation and re-wrap approach?

The main contributions of this work are to:

- Describe various types of layouts that are often used to display retrieved or recommended items in IAS.

- Provide modified fair ranking metrics which incorporate suitable browsing models to measure fairness in a given layout model.
- Provide insights on fairness score consistency and applicability across layouts.
- Describe the impact of column reduction approaches on fairness scores within a ranking in grid layout.

5.1 Problem Formulation

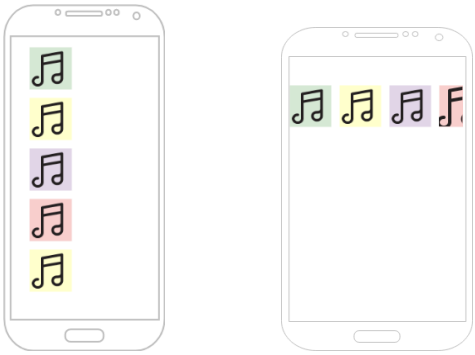
We consider an information access system that recommends or retrieves n items $d_1, d_2, \ldots, d_n \in D$ in response to information requests from users $q_1, q_2, \ldots, q_m \in Q$ based on their relevance to the request $y(d|q)$ and presents the results in a layout L (either 1-column, as in a classical linear layout, or a multi-column layout). Items are associated with producers or providers who in turn can be associated with demographic attributes identifying them with one or more of g groups. We model group membership of items with group alignment vector $\mathcal{G}(d) \in [0,1]^g$ (s.t. $\|\mathcal{G}(d)\|_1 = 1$) forming a distribution over groups; this allows for mixed, partial, or uncertain membership in an arbitrary number of groups. Table 3.3 summarizes the notation used in this paper.

5.1.1 Ranking Layouts

The items ranked and displayed to a user in response to their information preference come from a multi-stage process. IAS first analyzes the user's information request (contextual data, and historical information like preferences inferred from a user's past interaction), selects a set of candidate items, ranks those items based on their estimated utility with respect to user's need, and finally presents them to users in

ranking layout (the simplest of which is a list of the ranked results). To fairly allocate exposure between different items, even though only one can be placed in the coveted first position in a single ranking, the system may employ a *stochastic ranking policy* (distribution over rankings) and present a draw from this distribution to the user [52].

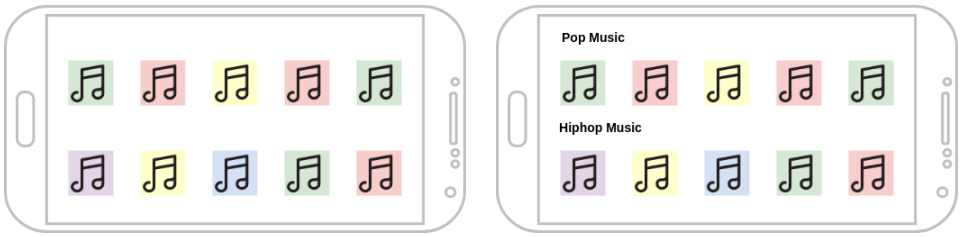
Once items are scored and ranked for a particular user information request, there are various *layouts* in which these results can be displayed. In this work, we consider layouts in $r \times c$ grids, where r is the number of rows and c the number of columns; this encapsulates at least four distinct models.



(a) Linear Vertical Layout

(b) Linear Horizontal Layout

Figure 5.1: Various types of linear layout models



(a) Wrapped Grid Layout

(b) Multi-list Grid Layout

Figure 5.2: Various types of grid layout models

Linear Layouts

Items are displayed in a single linear list. These come in two varieties:

Vertical Ranking Model Items are displayed in a multi-row single-column ranked list ($r \times 1$, see Figure 5.1(a)). Users generally see items from top to bottom. The layout may be split into multiple pages.

Horizontal Ranking Model Items are displayed in single-rows with multiple column lists following $1 \times c$ pattern. Users see items from left to right. In Figure 5.1(b) recommender results are ranked in a single row ranked list

Grid Layout

Items are displayed in multiple rows and columns ($r \times c$). These also come in multiple varieties:

Wrapped Grid Items are displayed as a single ranking in an $r \times c$ grid, without being categorized into groups. The grid is formed by displaying the items in order horizontally and starting a new row when the display runs out of space. Figure 5.2(a) shows a grid list of recommended books.

Multi-ranking Grid Items are displayed in multiple rows, often based on categories or recommendation sources, and each row consists of a ranked list of items. In figure 5.2(b), recommended items are categorized by genre which may facilitate users to find them from their preferred categories.

We focus on **wrapped grid** layout in this work due to the better availability of browsing and attention models for this problem setting. Further work is needed to provide usable models of user browsing behavior with multi-ranking grids before we can attempt to measure their fairness.

5.1.2 Fair Ranking Metrics

We followed the metric recommendations from our comparative analysis of fair ranking metrics in chapter 4 [142], and study two metrics: *Attention-Weighted Rank Fairness* [AWRF$_\Delta$, 154] to measure *statistical parity* in single ranking (averaging over multiple rankings to measure overall system fairness), and *Expected Exposure Loss* [EEL, 52] to measure equal opportunity in sequences or distributions over rankings. These metrics measure the distribution of exposure (based on estimated user attention) across provider groups to measure the fairness of rankings. They represent user attention with a *position weight* assigned to each item in a ranking.

Both metrics rely on a model of user attention (estimating the attention a user is likely to give to items at different positions in ranking) in order to measure fairness; it is important to know how users browse and interact with different positions in the ranked layout. Several studies have used user eye gaze tracker to study user browsing behavior [163, 53, 191, 203]. Some studies used user click behavior to infer browsing behavior of users [191, 55] with respect to ranking positions. Simple models of user browsing behavior, commonly used in information retrieval metrics and described in the next sections, determine these weights based on items' position in the ranking (along with other information, such as the relevance of preceding items).

AWRF$_\Delta$ is suitable to measure provider-side fairness in single ranking and it measures the difference between group exposure and configurable *population estimator*

(representing the ideal distribution of exposure over groups) using a distance function Δ. The exposure for each groups $\boldsymbol{\epsilon}_L$ is derived from the attention vector and the group alignment matrix ($\boldsymbol{\epsilon}_{\mathrm{L}} = \mathcal{G}(\mathrm{L})^{\mathrm{T}}\mathbf{a}_{\mathrm{L}}$) which aggregates the attention given to items of each group in proportion to their group membership as represented by the alignment vector:

$$\mathrm{AWRF}_{\Delta}(L) = \Delta(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}}) \tag{5.1}$$

EEL is suitable for *stochastic* ranking policy where fairness is measured over user-dependant distribution over rankings $\rho(L|q)$ since it is not possible to achieve equal exposure in single ranking [52]. It can be drawn as distribution over rankings $L_1, L_2, \ldots, L_{\tilde{n}}$ from the distribution over requests $\rho(q)\pi(L|q)$ [142]. EEL uses available relevance information to derive a *target exposure* $\boldsymbol{\epsilon}_\tau$, based on an ideal policy τ where relevant items are sorted in non-decreasing order in ranking and exposure is fairly distributed across the relevant items. Using the $\boldsymbol{\epsilon}_L$ of each ranking, the system exposure is derived as $\boldsymbol{\epsilon}_\pi = \sum_L \pi(L|q)\boldsymbol{\epsilon}_L$. EEL is computed as the squared Euclidean distance between system exposure $\boldsymbol{\epsilon}_\pi$ and target exposure $\boldsymbol{\epsilon}_\tau$:

$$\mathrm{EEL} = \|\boldsymbol{\epsilon}_\pi - \boldsymbol{\epsilon}_\tau\|_2^2 \tag{5.2}$$

EEL further decomposes into two constituent metrics, EER and EED, where *Expected Exposure Disparity* (EED) measures demographic parity and *Expected Exposure Relevance* (EER) shows the the extent to which exposure is bound to relevant items [52]. We focus only on the EEL metric in this work.

For both AWRF_{Δ} and EEL, the fairness goal is to provide fair exposure across

groups, thus they measure fairness by comparing exposure distribution with target distribution (EEL) or population estimator (AWRF_Δ) which is not dependant on layout. Hence, we can apply these metrics in both linear and grid layout by incorporating layout-suitable browsing models and measuring fairness with the same target distribution regardless of layout.

5.1.3 Linear Browsing Models

In linear layout, users typically browse the list from top to bottom [43], with the probability that they will continue (and thus view more items) decreasing as they move down the list. This assumption underlies many common metrics for recommendation effectiveness, including $nDCG$ [92] and MRR [28]. There is a variety of models of this scanning behavior with decaying attention; *cascade* and *geometric* are commonly-used click models to estimate user interaction probability with ranking positions. These models have been employed to construct evaluation metrics to measure utility [122, 29, 33, 10] or item exposure [52, 16, 154] in rankings.

Moffat and Zobel [122] proposed the *rank-biased precision* (RBP) evaluation metric to weight precision based on user attention to different ranking positions. This metric used a geometric browsing model with a *continuation probability* λ to estimate the probability of users moving to the next item (position) or stopping (click) at that position; the visit or interaction probability exponentially decreases with ranking positions. Biega et al. [16] proposed a modified version where the position weight decays geometrically with each position having the same probability of being stopped (clicked). In this model, the visiting probability of item d in position $L^{-1}(d)$ is determined by:

$$P_{\text{geometric}}[V_d] = \lambda^{L^{-1}(d)} \tag{5.3}$$

Table 5.1: Parameters of Weighting Models for computing $\mathbf{a}_L(d)$ and the range of parameter values

Parameters	Values	Browsing Models	Default Values
Skipping Probability γ	{0.1, 0.2, ..., 0.9}	Row Skipping	0.5
Continuation Probability λ	{0.1, 0.2, ..., 0.9}	Cascade Geometric	0.5
Slow parameter β	{1.1, 1.2, ..., 2.0}	Slower Decay	1.9

Craswell et al. [43] proposed the *cascade* click model where users will view position i if they have skipped items before that position, and whether users will click or skip a position depends on the relevance of the item in that position and the relevance of items in previous positions. Chapelle et al. [33] proposed a cascade-based metric *expected reciprocal rank* (ERR) by extending the cascade model to include the probability of users terminating the entire process as an *abandonment* probability that decays geometrically. In the cascade model, users will visit item d if they did not stop at any position before that item in the ranked list which is determined by item relevance. The continuation probability λ is now a function of relevance, and the probability of visiting d is given by:

$$P_{\text{cascade}}[V_d] = \prod_{j \in [0, L^{-1}(d))} \lambda\left(y\left(L\left(j\right)|q\right)\right) \tag{5.4}$$

5.1.4 Grid-based Browsing Models

Users do not interact with grid-based displays in the same way they interact with linear displays — treating the display as a linear ranking read from left to right and top to bottom is not an accurate model of user browsing behavior and attention. Several studies have been performed to understand how users allocate attention to

different items in grid layout.

Existing Literature on User Browsing Behavior in Grid Layouts

Tatler [173] observed that users have a tendency of *central fixation* where they tend to put more attention in middle of the screen than on the edges. Djamasbi et al. [53] found that users usually focus on results located at the top left-hand side and proceed in an *F-shaped* reading pattern, but the viewing pattern varies based on task, content, and complexity of web pages. Shrestha and Lenz [163] showed that in image-based web pages users do not always show the *F-shaped* viewing pattern and identified the need to consider page content while understanding user viewing patterns. Zhao et al. [203] observed user gaze pattern in recommender system with grid-based interfaces to predict user preference; their eye-tracking study infers that users show an *F-pattern* while interacting with grid-based interfaces rather showing a *center effect* but they showed that pattern can vary depending on task.

Xie et al. [192, 191] performed eye-tracking studies to understand user attention in grid-based image search results and observed that users tend to put more attention in middle position rather than results in left or rightmost positions (*middle-bias*). Moreover, in grid-view, user attention decreases at a slower speed than in linear search results (*slower-decay*) and users may not go through every row; they often directly interact with results after skipping previous rows (*row-skipping*).

The studies mentioned above mostly focus on understanding user viewing patterns in grid-based interfaces with the goal of providing and measuring user satisfaction. There is limited research work concerning fairness issues when IAS results are displayed in grid layout. Guo et al. [85] proposed de-biasing techniques for grid-based

product search result pages in e-commerce systems; consistent with the studies above, they observed that user attention follows *row-skipping* and *slower-decay* while interacting results in grid layout and modified cascade browsing model to account for these in estimating user attention. Balyan et al. [12] argued that in e-commerce grid-based product search results, item meta information has an impact on user viewing behavior, and incorporated item metadata into their un-biased learning to rank technique.

Adapting Browsing Models to Grid-Based User Behavior

Since the previous studies showed that user attention varies between applications depending on task, domain, device, and details of the layout, considering multiple viable models from existing literature will provide insights useful to researchers and practitioners in various contexts, as they can apply an appropriate model for their systems. For this analysis, we consider two such behaviors: *row-skipping* (RS) and *slower-decay* (SD) in the context of wrapped grid layout; we leave central fixation, multi-list rankings, and incorporating multiple browsing model adjustments simultaneously to future work.

Since both AWRF_Δ and EEL use position weights to capture user browsing behavior, we can adapt them to ranking in grid layout by adapting the browsing models. We adapt both the *cascade* and *geometric* browsing models to account for *row-skipping* (RS) and *slower-decay* (SD).

For *row-skipping* behavior, the visiting probability of item d at $\text{row}(d)$ and ranking position $L^{-1}(d)$ depends on the skipping probability of a row γ; for each of the k rows before $\text{row}(d)$, the user either continued through that row, or skipped it with probability γ. If users visited items in a row, that implies that a particular row was

not skipped. With that assumption, visiting probability of item d in cascade-based row-skipping model considering relevance:

$$P_{RS(\text{cascade})}[V_d] = \left[\prod_{k=0}^{\text{row}(d)}(1-\gamma)\prod_{i\in L(\text{k},\cdot)}\lambda(y(L(i)|q)) + \prod_{k=0}^{\text{row}(d)}\gamma\right]\prod_{i\in\text{row}(d)}\lambda(y(L(i)|q)) \tag{5.5}$$

The visiting probability of item d in geometric-based row-skipping model is given by[2]:

$$P_{RS(\text{geometric})}[V_d] = \left[\prod_{k=0}^{\text{row}(d)}(1-\gamma)\prod_{i\in L(\text{k},\cdot)}\lambda + \prod_{k=0}^{\text{row}(d)}\gamma\right]\prod_{i\in\text{row}(d)}\lambda \tag{5.6}$$

With the *slower-decay* browsing behavior, visiting probability of items across a row in a grid layout decays more slowly than in a vertical linear list, but jumps when the user moves to the next row. This is modeled by a decay parameter β to modify the continuation probability for items in ranked results based on the row in which they appear. The visiting probability of item d in cascade-based slow-decay model is:

$$P_{SD(\text{cascade})}[V_d] = \min(\beta^{\text{row}(d)}\prod_{i=[0,L^{-1}(d)]}\lambda(y(L(i)|q)), 1) \tag{5.7}$$

The geometric visiting probability of item d with slower decay is (derived by [85]):

$$P_{SD(\text{geometric})}[V_d] = \min(\beta^{\text{row}(d)}\prod_{i=[0,L^{-1}(d)]}\lambda, 1) \tag{5.8}$$

Table 5.1 shows the parameters and range of values we consider to measure at-

[2]This model is derived by [85]) where they referred the model as *cascade click model*. However, in our paper, we referred the model as *geometric* to keep the conceptual consistency.

tention weight of items in ranking.

5.1.5 Changing Grid Layouts

Based on the device the user is using to interact with the system, grid layout can be converted into a size suitable for a particular device using two different approaches: *truncation*, where each row is truncated and item off-screen are no longer displayed, and *re-wrapping*, where the rows are re-wrapped so the items that would be off-screen are moved to the next row. These approaches may differ in their influence on the fairness of the resulting display. For example, if ranked results are presented in a 4×5 wrapped grid layout, the attention distribution over groups may vary when the same results are presented in 5×4 wrapped grid layout or in a 4×4 layout. Users need to scroll more to see other items, affecting item visibility and exposure. To observe the impact of column size and column reduction approaches on group fairness score, we change the column size for a given ranking in grid layout using both *truncation* and *re-wrap* approaches.

5.2 Experimental Setup

Our central goal is to understand how measurements and optimizations for classical linear layouts apply to grid layouts, both to apply existing methods and to identify where further research is needed to support fairness in these widely-used layouts. To answer our research questions, we conduct several experiments by implementing the metrics with adaptations for user behavior in grid layouts and using them to measure outputs in real-world IAS datasets covering both search and recommendation scenarios.

5.2.1 Dataset

In this work, we use datasets from both search and recommendations scenario. For search systems, we use the Fair Ranking Track 2020 dataset and for recommendation scenario, we use two datasets from GoodReads and Amazon. The details of the datasets are provided in chapter 3. Both recommendation datasets have incomplete relevance judgements and incomplete group labels. We follow common practice and consider documents without relevance data as non-relevant, and treat missing group labels as a separate unknown category in our experiments.

5.2.2 Methodology

Across several different scenarios, we measure the fairness of recommendation and retrieved ranked results.

RQ1. Consistency of Fair Ranking Metric Scores Across Layouts To observe the consistency of fair ranking measurements across layouts, we implement AWRF_{Δ} and EEL with user attention models modified to account for wrapped grid layout.

- For a given set of recommended or retrieved items, we represent the items in linear-vertical layout and 5-column wrapped grid layout.

- We measure fairness using each of the metrics in their default parameter settings for both layouts.

- We compare the metric scores to observe if and how fairness scores change with the choice of layout models and to what extent.

RQ2. Consistency of Fairness Optimized Ranking Across Layouts To better understand the fairness score differences across layouts, we want to identify if fairness remains consistent across layouts — whether a ranked list optimized to be fair for a certain layout model remains fair for other layout models.

- We apply group-fairness aware *GreedyEQ* re-ranking technique from Ekstrand et al. [61] on ranked results to generate optimized linear vertical ranked lists.
- Then we represented the optimized linear ranked lists into 5-column grid layout and measure group fairness in both linear and grid layouts.
- This experiment shows the persistence of fairness scores of a ranked list across layouts.

RQ3. Consistency of Fair Ranking Metric Scores Across Devices Depending on devices, column size of grid layout is changed (reduced) to fit the screen size. Users may access a system from different devices with different screen size (such as phones, laptops. and TV set-top boxes), and thus the grid layout may be adjusted to fit the user's current screen, usually by reducing the number of columns displayed. As noted in Section 5.1.5, column reduction can be done by either *truncating* or *re-wrapping* the rows. These methods may have different impacts on the fairness scores of system outputs. Further, fairness scores may change as column size changes regardless of approach. To see the impact of column size on group fairness score and the fairness score consistency across column-reduction approaches,

- We represent the set of retrieved and recommended items in grid layout changing the column size in 10, 8, 6, 5, 4, 3 using both truncation and re-wrap approaches.

- We compare fairness scores across column sizes and also across the reduction approaches.
- Moreover, to identify the impact of initial column size on metric score for reduced column grid layout, we measure fairness with two different initial column sizes.

5.3 Results and Discussion

We now present the results of our investigation into the behavior of fair ranking measurements applied to ranked results in grid layouts.

RQ1: Do fairness measurements remain consistent across layout models? Figure 5.3 shows the fair ranking metric scores change with the change of layout models n both search and recommendation scenarios. Metric scores change within grid layouts with the change of browsing behaviors (*row-skipping*, *slower-decay*). AWRF_{Δ} is not consistent across grid adjustments to browsing models, keeping the same order of systems, but the *cascade* and *geometric* browsing models rank systems in a different order. In both *cascade* and *geometric* browsing models, EEL scores with *row-skipping* model show notable inconsistency; this shift is significantly greater than the shift seen in AWRF_{Δ}. Metric scores are consistent between grid layout with SD(geometric) browsing model and list layout under the *geometric* browsing model.

Implications From *RQ1*, we have following observations:

- Fair raking metric scores are highly dependent on layout and user browsing model.

- Within a layout, metric score further varies across user browsing behavior.

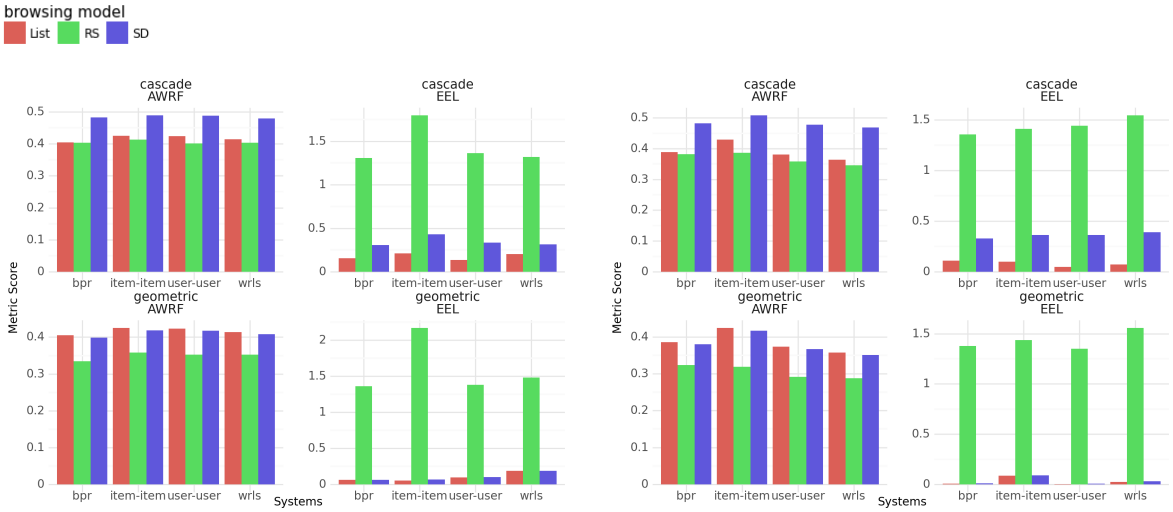
- Since user attention is one of the required components of AWRF_{Δ} and EEL implementation and user attention for ranking positions is determined by user browsing behavior, it is important to consider accurate browsing model while applying these metrics.

RQ2: Do rankings optimized for fairness in linear layouts remain fair in grids? In *RQ2*, we examine how models that are optimized for fairness under linear layout score when measured under a group layout. From Figure 5.4, we see that AWRF_{Δ} scores are consistent across layouts specifically with *geometric* browsing model. EEL score for a fairness optimized ranking can vary across layouts depending on user browsing models. Within a grid layout, EEL with the *row-skipping* browsing model provides different fairness scores and rankings than *slower-decay*.

Implications From *RQ2*, we made following observations:

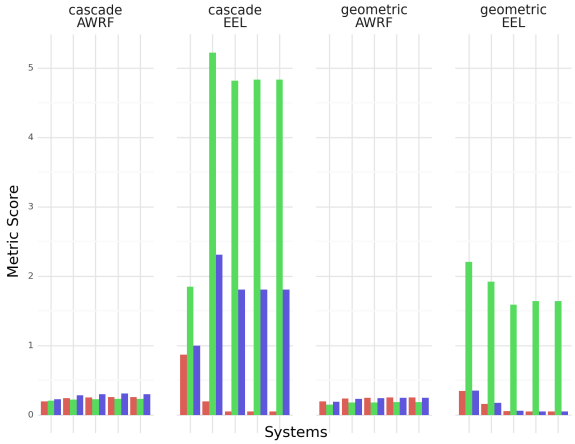
- A ranking that is fair in linear layout can be represented as unfair depending on the user browsing behavior we are assuming while measuring fairness. This reinforces the need to incorporate accurate user browsing models in fairness measurement techniques.

- Without considering layout-suitable browsing models, fair ranking metrics will provide unreliable fairness scores.



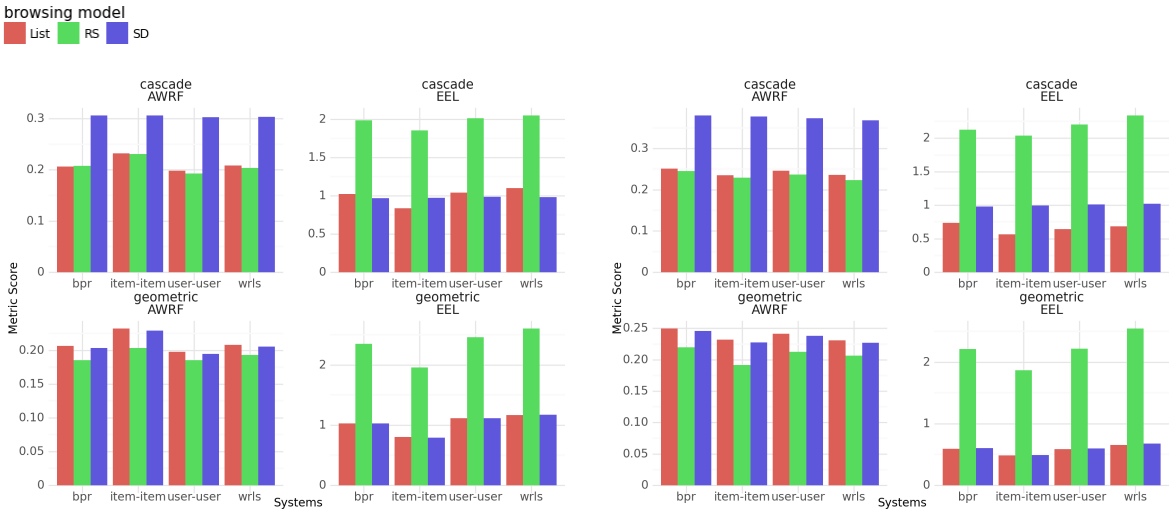
(b) Amazon Recommendations

(c) GoodReads Recommendations



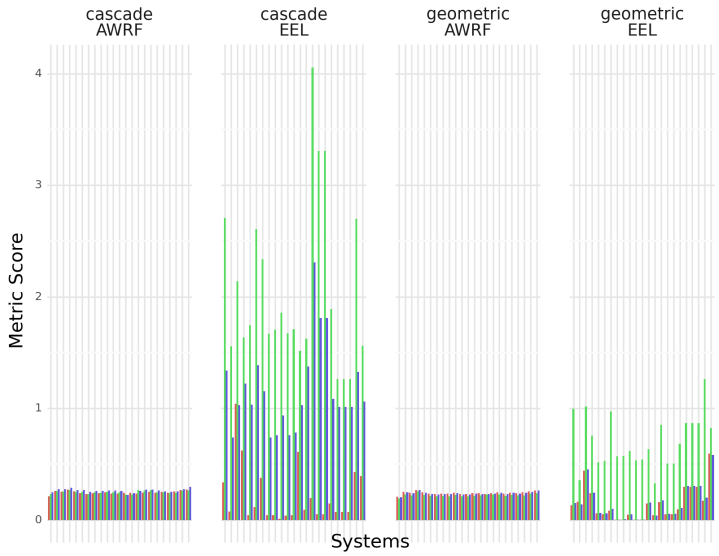
(d) FairTREC Retrieval Task

Figure 5.3: Metrics results with the change of weighting strategy



(b) Amazon Fairness-aware Re-Ranked Recommendations

(c) GoodReads Fairness-aware Re-Ranked Recommendations



(d) FairTREC Re-ranking task

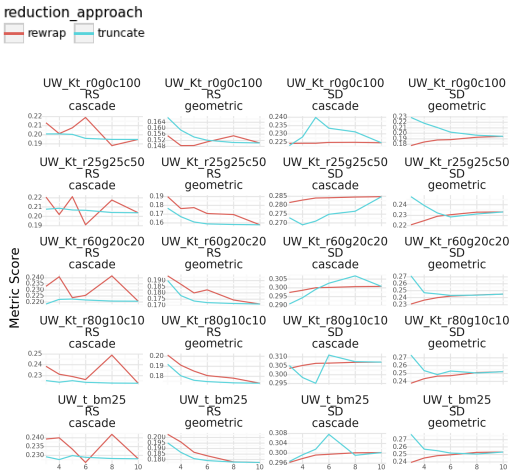
Figure 5.4: Metrics results with the change of weighting strategy in optimized ranking

RQ3: How do fairness scores change as grid size changes? *RQ3* shows how the metric score changes with the column size and the approach for reducing columns within a grid layout. Figure 5.5 shows that metric score changes with column sizes and they differ in their changing pattern with column reduction approaches. In all three datasets, the impact of column sizes on metric score varies across systems.

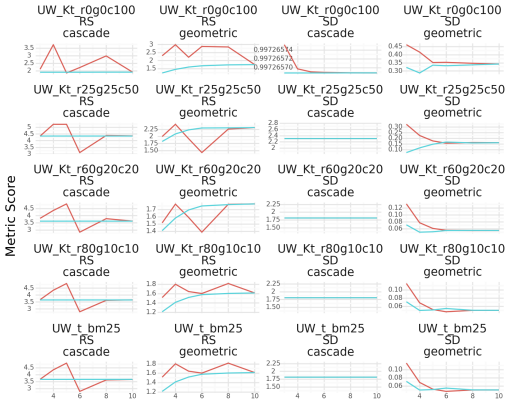
RQ3.a. Does the fair ranking metric score change when the grid-based list is truncated? When columns are reduced using the *truncate* approach, metrics show some stability towards column size in both search and recommendations for most of the systems. However, column size has more impact on AWRF_{Δ} scores than EEL with the *truncate* approach.

RQ3.b. Does the fair ranking metric score change when the grid-based list is re-wrapped? When columns are reduced using the *re-wrap* approach, AWRF_{Δ} shows high sensitivity towards column sizes in both search and recommendations for most of the systems.

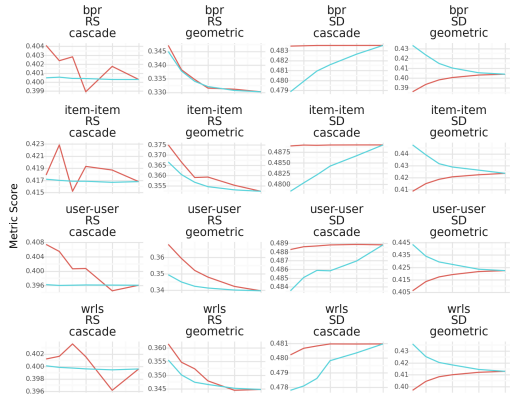
RQ3.c Does the change in group-fairness score with column size reduction remain consistent across truncation and re-wrap approach? Metric scores vary with the change of column sizes and the direction of this change is different between column reduction approaches. However, for some systems, metric scores with both column reduction approaches converges at some column sizes. The metrics are consistent across systems. In the *truncate* approach, the starting column size has an impact on metric score changing pattern. Figure 5.6 shows the EEL scores in GoodReads dataset (showing same pattern in Amazon dataset) with the change of



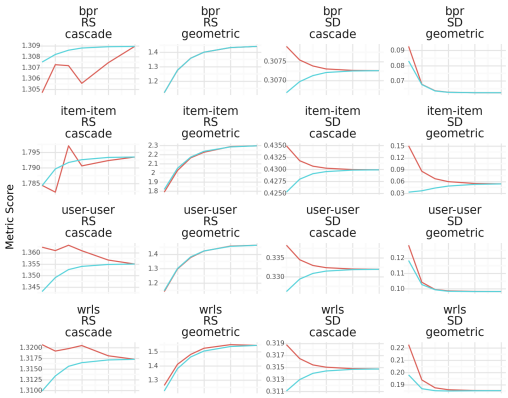
(b) AWRF$_\Delta$ in FairTREC retrieval task



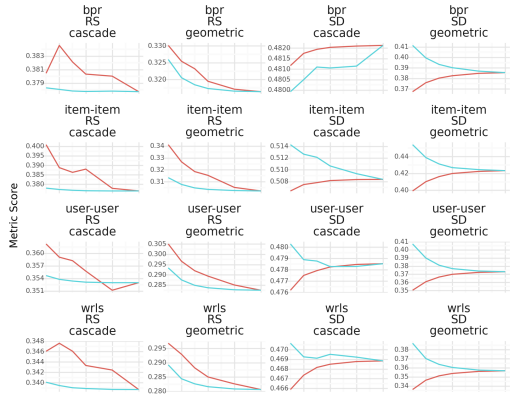
(c) EEL in FairTREC retrieval task



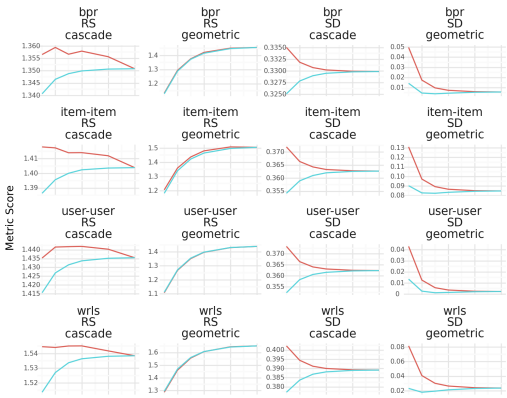
(d) AWRF$_\Delta$ in Amazon Recommendations



(e) EEL in Amazon Recommendations

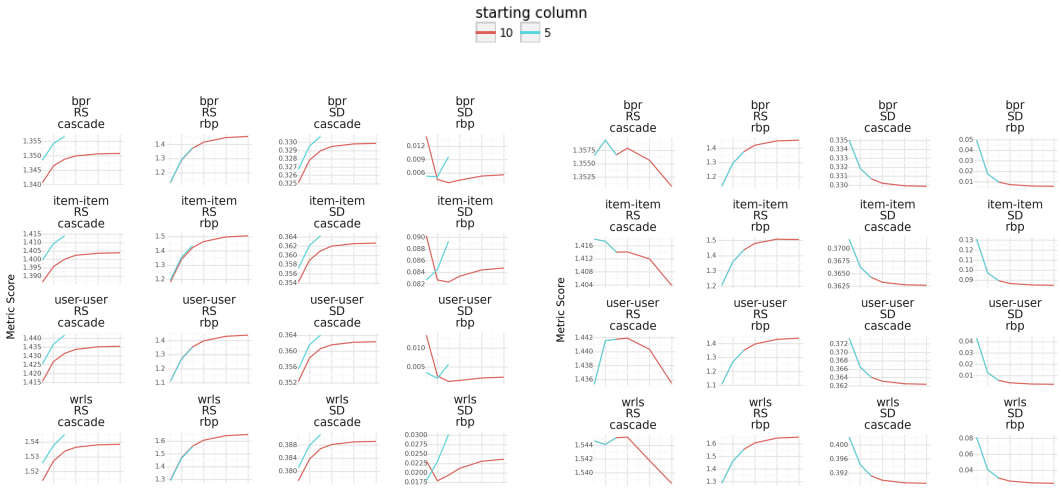


(f) AWRF$_\Delta$ in GoodReads Recommendations



(g) EEL in GoodReads Recommendations

Figure 5.5: Metrics results with the change of column sizes across column reduction approaches



(b) Impact of starting column on EEL in truncate column reduction

(c) Impact of starting column on EEL in rewrap column reduction

Figure 5.6: Impact of starting column on EEL **across column reduction approaches**

column size and the pattern of this change varies with the starting size. In *re-wrap* column truncation approach, the starting column size has no affect on metric score consistency. We do note that the *truncate* approach is primarily used with multi-list layouts in practice, while our results here are for wrapped layouts; however, finding that the use of truncation has significant effects on fairness has implications for fair layouts regardless of the initial grid layout method.

Implications From *RQ3* we have following observations:

- Device is an important factor in measuring fairness.
- With the change of device (column size) fairness scores show high sensitivity which indicates the importance of carefully selecting column-reduction approaches while re-ranking the grid layout.

5.3.1 Discussion

Our findings provide insights on implementation and reliability of fair ranking metrics in grid layout. Our analysis provides knowledge on how metric behavior changes across ranking layouts and across column-reduction approaches within grid layouts as well. Our results suggest that metrics can vary in their consistency across ranking layouts (AWRF_{Δ} was more consistent across layouts than EEL). However, a metric that is consistent across layouts may not be stable across device sizes within a particular grid layout (EEL was more consistent across column sizes than AWRF_{Δ}, while for AWRF_{Δ}, the consistency of metric results notably varies depending on the column-reduction approach.) Therefore, our results advise researchers and practitioners to pay close attention to ranking layout, device sizes, and column-reduction approaches while using a metric to measure fairness in ranking. Even though AWRF_{Δ} metric score is consistent across layouts to some extent, while using AWRF_{Δ} in grid layouts, practitioners should pay attention to column sizes and column-reduction approaches. While using EEL to measure fairness in ranking, ranking layout must be taken into account but the reduction approach has less impact on the measurements.

Furthermore, our results indicate that metrics can be highly affected by user browsing behavior. Since the concept of provider-side fairness in ranking often relies on the attention users pay to items in different positions, it is important to use accurate models of user attention behavior when measuring provider-side fairness in ranking. It is therefore necessary to develop a clear and detailed understanding of user browsing behavior in order to generate valid and trustworthy fairness score using fair ranking metrics. Our work is not able to directly provide those measurements, but provides a first analysis of what to expect when applying existing measurements

with the current public state of knowledge in user behavior modeling.

5.4 Conclusion

In this chapter, we consider a gap in the state of the art in measuring the provider-side fairness of rankings by considering grid layouts. We apply existing fair ranking metrics in linear and grid layouts to identify their consistency across layout models. Our results show that metrics scores are dependant on user browsing models and ranking layouts. Moreover, within grid layout, metric scores are inconsistent across column sizes and column reduction approaches.

Researchers and practitioners can not just apply an arbitrary user browsing behavior in order to measure the fairness of a system's rankings with existing fair ranking metrics; they need to account for users' likely browsing behavior for a particular layout, as the differences in assumed or observed behavior affect the fairness measurement results. This highlights the need for additional research, in particular eye-tracking and similar studies to develop and validate models of user browsing behavior for different devices, domains, and other display parameters (such as the type and quantity of item metadata provided in the display). Further, many recommendation applications use multi-list grid layouts, but there is almost no public research on browsing behavior that enables realistic models for such layouts.

Developing reliable fair ranking metrics for a range of layout configurations, and the browsing models needed to implement them, will help ensure that this widely-used display format for information access outputs provide equitable exposure and opportunity to content creators. The results we presented here will help researchers

and practitioners to identify potential risks and considerations in applying the existing metrics, and lay the groundwork for further research to make information access systems, in all their varied displays and interaction formats, fair.

CHAPTER 6:

UNIFIED BROWSING MODELS FOR LINEAR AND GRID LAYOUTS

In the previous chapters (chapter 5), I described various types of ranking layouts that are used to display the results along with several existing user browsing models that are used to implement fair ranking metrics. The analysis of various browsing models in chapter 5 for both linear and grid layouts shows that the underlying concepts of these browsing models are often similar, including varying components and parameter settings. In this chapter[1], I seek to leverage that similarity to represent multiple browsing models in a generalized, configurable framework which can be further extended to more complex ranking scenarios. We describe a probabilistic user browsing model for ranking linear layout, show how this can be configured to yield models commonly used in current evaluation practice, and generalize this model to also account for browsing behaviors in grid layouts. This model provides configurable framework for estimating the attention that results from user browsing activity for a range of IAS evaluation and measurement applications in multiple formats, and also identifies parameters that need to be estimated through user studies to provide

[1]This work was done in collaboration with Dr. Michael Ekstrand. This was submitted to ICITR'23 but was rejected and currently under-revision.

realistic evaluation beyond ranked lists.

As previously described, IAS can display results in a linear ranked list (figure 5.1) or ranked items can be displayed in a grid view with multiple rows and columns (figure 7.1). In a linear ranking layout, items are displayed in a single-column list whereas in grid layout, items are presented in multiple rows and columns. Depending on the ranking layout, item position changes in the displayed page which affects user attention and interaction with items. Users do not provide equal attention to every item that are exposed in the ranking and user attention varies based on item position [198]. Moreover, user attention varies across ranking layouts as well [191]. Thus user browsing behavior describing how users interact with ranking helps to estimate approximate user attention provided to items at various ranking positions.

Ranked results are often evaluated based on one or both criteria: user satisfaction such as maximum marginal relevance [28, 32, 122] and social and ethical issues such as fairness [121, 63, 142, 96]. User browsing behavior is a significant component in evaluation metrics construction. Evaluation metrics regarding effectiveness [123] (e.g. *reciprocal rank* [42] or *nDCG* [93]) and fairness (e.g. *equal exposure* [52, 165] or *statistical parity* [154]) of rankings take user browsing behavior into consideration since user attention changes with ranking position and items with similar relevance do not necessarily get the same attention in ranking. Hence, user browsing behavior is used to estimate the probability of user engagements with ranking positions which helps to measure item utility [122] and exposure [165] in ranking.

There are research works on understanding user browsing behaviors and these studies often involve eye-tracking [203, 53] and click models [85] and to date, we have multiple user browsing models. However, these browsing models work with the same

underlying concept that user attention changes with ranking positions but they differ on their use of components such as, relevance information, and external parameters settings. Moffat et al. [123, 124] identified three interchangeable functions that can be used to describe user behaviors in ranking and showed that effectiveness metrics can be represented by those functions. Their study focused particularly on linear ranking layout and generalization of effectiveness metrics for ranking.

In this work, we unify many of the extant user browsing models into a single model that accounts for both linear and grid layouts showing how particular models from the literature can be derived from specific parameterization of our model and extending them to a wider range of SERP designs. We identify the conceptual similarities among these models and disintegrate them into components and parameters. We provide a generalized framework of user browsing models that allows researchers and practitioners to re-configure the core structure based on their required components. This structure can be further extended to more complex ranking layout scenarios.

6.1 User Browsing Behaviors in Linear Layouts

Several research works observed user browsing behaviors when reviewing linear layouts, particularly vertically-oriented linear lists, and proposed user *browsing models* to approximate the attention users are likely to pay to different items in a linear ranking. Two commonly-used models are the *geometric* [122] and *cascade* browsing model [43], each of which estimates the probability that a user will attend to the item in a particular ranking position. These models are developed on three fundamental assumptions:

- Users browse a linear ranking from top to bottom.
- User attention decays with ranking positions.
- Users stop the process once they select an item.

Given a ranking, both geometric and cascade models approximate the probability of the user continuing to the next item. With similar underlying concepts, these models can be described using state transition model (figure 6.1). Moffat and Zobel [122] presented a similar model of user browsing behavior in linear ranked list; our model distinguishes between the *selection* and *abandon* events, a distinction we use in Section 6.1.2. Table 3.3 presents the list of notations used in this paper. Given a linear ranked list L, for a given position i, user u can take following actions:

1. Examine (E): The user *examines* (or "visits", "'views", or "inspects" the item at the current position (E_i is the event of examining the item at position i).
2. Select (S): The user *selects* (usually by clicking or tapping; other work has also used the terms "stop" or "click") the item at the current position.
3. Abandon (A): The user terminates their browsing process without selecting an item.
4. Continue (C): If the user has not select the item at the current position or abandoned the process, they move to the next position.

The probability of continuing to the item at position $i + 1$ depends on two fundamental probabilities: (1) the probability of selecting the item at position i and (2) the probability of abandoning the page after examining item at position i. Some

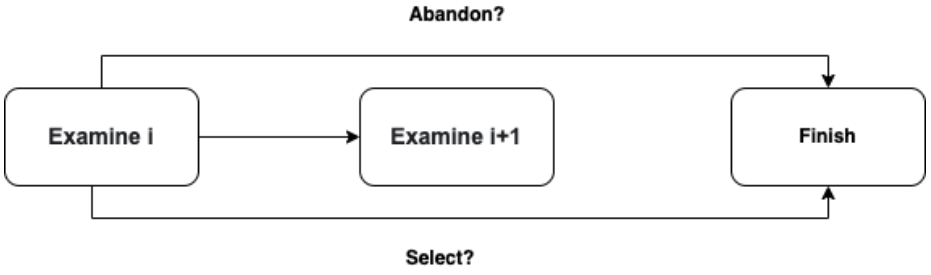


Figure 6.1: State transition model of user browsing a linear layout

treatments use other probabilities as the fundamental parameters; we discuss this more later in section 6.2.

Selection Probability In figure 6.1, the event "select" has the probability of selecting the item at position i; at a particular position, this probability is conditional on user *examining* the item, and it can also depend on the *relevance* of that item to the query. The conditional probability of user selecting item at position i is:

$$P\left[S_i|E_i, y(L(i)|q)\right] = \psi(i) \tag{6.1}$$

Hence, the *selection* probability can be defined as a function of relevance:

$$\psi(i) = \left\{ \begin{array}{ll} \psi_{\text{rel}} when & y = 1 \\ \psi_{\neg \text{rel}} & y = 0 \end{array} \right\} \tag{6.2}$$

When relevance is not considered, the probability of selecting the item at position i is a constant: $\psi_{\neg \text{rel}}(i) = \psi$. where S is the event of selecting or clicking on an item and E refers to the event of examining or viewing that item. This selection probability can be modified based on the availability and the type of relevance information.

Abandon Probability A user can abandon the page with *abandon* probability α which can be fixed at each position that has the effect of being cumulative over the ranking positions. The *abandon* probability can be derived as position-based exponential decay and it can also be extended as a conditional probability function which can be dependant on relevance of items.

Therefore, the continuation probability or the probability of moving to the next position can be derived from selection probability and abandon probability of the current position. Users will continue to the item at position $i+1$ if they have not selected the item at position i and did not abandon the process after examining the item and position i. For a given linear ranking L, the generalized model for predicting the probability of user examining the item d at a particular position i when user attention decays exponentially is:

$$P[E_i] = (1-\alpha)^{i-1} \prod_{j \in [1,i-1)} (1 - \psi(j)) \tag{6.3}$$

6.1.1 Static User Browsing Models

In static user browsing models, only item position is taken into account to estimate user attention to items in ranking [122, 33]. Moffat and Zobel [122] considered a *geometric* browsing model to propose *rank-biased precision* metric (*RBP)* which is an effectiveness metric for linear ranking. In their assumed browsing model, user attention decays exponentially with ranking positions and the probability of the user continuing to the next positions depends on the probability of a user selecting item at the current positions. They proposed a *persistence* or *continuation* probability λ to derive the possibility of user continuing to the next position. In this model, the

continuation probability is not dependant on relevance of items and it is considered as a constant, hence, their proposed model is not distinguishing between *selection* probability and *abandon* probability. In this model, users will always examine the item at the first position and hence the probability of examining item at ith position is λ^{i-1}. For a given ranking L of size N, *rank-biased precision* can be derived as

$$\mathrm{RBP}(L) = (1-\lambda) \sum_{i=1,2,\ldots N} y\left(L(\mathrm{i})|q\right) \lambda^{i-1}$$

where $y\left(L(\mathrm{i})|q\right)$ denotes the relevance of item in position i to the user request q. Since the *persistence* probability λ was not dependant on relevance in a geometric user browsing model, the probability of examining item d at ranking position i is:

$$P_{\mathrm{geometric}}[E_i] = \prod_{i \in [1, i-1)} \lambda$$

which can be derived from equation 6.1 without considering relevance information and *abandon* probability.

$$P[E_i] = \prod_{j \in [1, i-1)} (1 - \psi_{\neg \mathrm{rel}}) \tag{6.4}$$

Biega et al. [16] proposed another version of a *geometric* model where the attention decays geometrically and each position has the equal probability of being selected. Whether we want to consider examine probability of prior positions or not, we can represent both versions of *geometric* model through equation 6.1.

6.1.2 Cascade Models

Craswell et al. [43] proposed another user browsing model that incorporates item relevance when estimating users' item selection behavior. Their proposed *cascade* click model (called *adaptive* by [123]) incorporates item relevance into the selection process, so that users are much more likely to select a relevant document than an irrelevant one; the probability of examining an item at a particular ranking position therefore depends on the relevance of items in previous positions. Specifically, the probability of the user clicking or selecting an item d is a function of $y(d|q)$ (the relevance of d for a given query q, which may be binary or graded). The event of user selecting an item at position i depends on the probability of user selecting an item at position i and users skipped (did not select) all the items prior to that position which are dependant on the relevance of items. Hence, the the probability of user clicking or selecting an item at position i can be derived from the fundamental relevance-dependant selection probability ψ_{rel}:

$$P[S_i] = \psi_{\text{rel}}(i) \prod_{j \in [1,i-1)} (1 - \psi_{\text{rel}}(j))$$

In both the cascade and geometric models, user attention decays exponentially with ranking position, but in the cascade model the user is much more likely to stop at a relevant item, so the examination and selection probabilities at a particular position differ from ranking to ranking, and jump at the positions of relevant items.

$$P_{\text{cascade}}[E_i] = \prod_{j \in [1,i-1)} (1 - \psi_{\text{rel}}(j)) \tag{6.5}$$

where $\psi_{\text{rel}}(j)$ depends on the relevance of the item in position j to the user request q and the relevance can be binary or graded.

Chapelle et al. [33] use a cascade model to derive an effectiveness metric, *expected reciprocal rank* (ERR). Unlike *RBP* metric, *ERR* depends on the relevance of items to infer probability of a user selecting an item at particular ranking position and directly derives effectiveness from those probabilities instead of using them to weight a precision metric. For a given ranking of size N,

$$ERR(L) = \sum_{i=1,2..N} \frac{1}{i} P[S_i]$$

where $P[S_i]$ is the probability of clicking or selecting an item at position i which is same as the *cascade* click model.

Chapelle et al. extended the cascade model through an additional parameter, an *abandonment* probability modeling the probability of user terminating their browsing regardless of whether they have selected an item (either to abandon the search entirely, or to reformulate their query). In this extended *cascade* model, the probability of an user examining an item at position i is:

$$P[E_i] = (1-\alpha)^{i-1} \prod_{j \in [1,i-1)} (1 - \psi_{\text{rel}}(j)) \tag{6.6}$$

Therefore, the probability of examining the item at position i can be derived using equation 6.3 incorporating abandon probability and relevance information.

Table 6.1: Parameters of browsing Models and the range of parameter values

Parameter	Name	Description	Values
ψ	Selection Probability	Probability of selecting an item at position i	{0.1, 0.2, ..., 0.9}
α	Abandon Probability	Probability of abandoning the process.	{0.1, 0.2, ..., 0.9}
λ	Continuation Probability	Probability of continuing to the position i	{0.1, 0.2, ..., 0.9}
γ	Skipping Probability	Probability of skipping an entire row.	{0.1, 0.2, ..., 0.9}
β	Decay	Incorporate slow browsing tendency for grid layout	{1.1, 1.2, ..., 2.0}

6.1.3 Unifying Ranking Browsing Models

We can see that the geometric and cascade browsing models are capturing the same fundamental ideas, with the difference that the cascade model incorporates item relevance into *selection* probabilities. We can therefore derive these browsing models from our state model with two main probability parameters: the *selection* probability and the *abandon* probability, where the *selection* probability may or may not depend on item relevance. *Abandon* probability can also be extended as a conditional probability of relevance. Our generalized browsing model can be configured to implement various browsing models in the following ways:

- To use the geometric browsing model, the relevance component of *selection* probability function and the *abandon* probability will be ignored (setting the value as $\alpha = 0$).
- To use the cascade browsing model, the relevance component of *selection* probability can be binary or graded.
- Both models can further incorporate the *abandon* probability by setting appropriate parameter values.

Table 6.1 shows the list of discussed parameters that can be incorporated in browsing models and their range of acceptable values. With suitable parameter choices, our

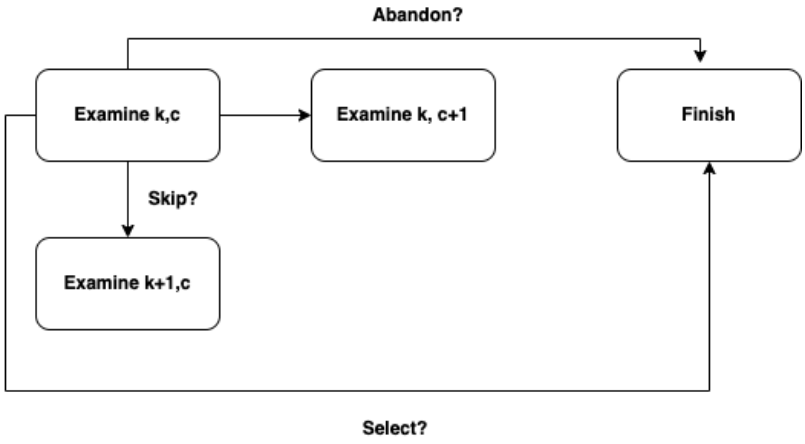


Figure 6.2: State transition model of user browsing a grid layout

generalized model can therefore realize a wide range of probabilistic attention models both from the literature and yet to be devised.

6.2 Extending Generalized Framework to Grid Layout

In this section, we further extend the generalized user browsing model for linear ranking layout to grid layout.

6.2.1 Linear Layout is Single-Column Grid Layout

In grid layout, users have similar actions as linear ranking layout with an additional possibility of *skipping row* action. Unlike linear layout, in grid layouts, users can skip an entire row and move to the next row. We denote this *row-skipping* event as K. User actions in grid layout can be demonstrated by figure 6.2, the state-transition model. Therefore, for a given ranking in grid layout L the probability of examining item at position $L(\mathrm{k}, \mathrm{c}+1)$ (row is k and the column is $c+1$) depends on:

- the probability of selecting the item at position $L(\mathrm{k}, \mathrm{c})$,
- the probability of abandoning the process after examining item at position $L(\mathrm{k}, \mathrm{c})$, and
- the probability of skipping row k after examining item at position $L(\mathrm{k}, \mathrm{c})$.

For a 1-column grid layout or linear vertical layout, the row skipping probability can be ignored. Hence, the continuation probability of or the probability of moving to the next position in grid layout can be derived from *selection* probability, *row-skipping* probability, and *abandon* probability at the current position. The generalized conditional probability of user selecting the item position i in grid ranking L is:

$$P\left[S_{L(\mathrm{k,c})}|E_{L(\mathrm{k,c})}, y\left(L(\mathrm{k}, \mathrm{c})|q\right)\right] = \psi(L(\mathrm{k}, \mathrm{c})) \tag{6.7}$$

6.2.2 User Browsing Models for Grid Layouts

Several studies have sought to understand user browsing behavior in grid layouts, identifying that user browsing behavior in a grid layout is different than it is in linear layouts. Users do not necessarily view top-to-bottom while interacting with a grid-based interface; rather, they show various distinct tendencies such as *central fixation* [173] and *F-shaped browsing* [53]. Tatler [173] did an user eye-movement study and showed that users have the tendency of *central-fixation* while viewing images showing on screen and that tendency can persist for grid-based image search results. Eye-tracking studies on grid-based web-pages [53] and on grid-based recommendations [203] show that users often follow *F-shaped* viewing pattern but both studies indicate that user browsing behavior depends on task and content. In grid-based image

search results, Xie et al. [191, 192] observed several user browsing behavior tendencies through eye-tracking study; (1) *slower-decay* where user attention decays at a slower rate in grid layout than in linear layout, (2) *row-skipping* where users skip a row and move to the next row, and (3) *middle-bias* where users put more attention to items at the middle positions. However, Shrestha and Lenz [163] and Balyan et al. [12] highlighted the need of considering page content while studying user behavior for web-pages and item meta information for e-commerce product search results displayed in a grid layout.

6.2.3 Generalized Browsing Model

Linear vertical layout can be treated as 1-column grid layout, and the existent browsing models for ranking in linear vertical will further be modified for grid layouts (and still capture the linear behavior when the number of columns is set to 1). Guo et al. [85] observed a similar user browsing pattern as Xie et al. [192] in grid-based e-commerce products search results and they proposed modified *geometric* (equation 6.4)[2] browsing models incorporating *slower-decay* and *row-skipping* grid-based layout specific user browsing behaviors to generate a user attention model suitable for grid-based e-commerce search results.

For *slower-decay* (SD), a *decay* parameter β is used to incorporate users slow browsing patterns for ranking in grid layout. These parameters can be plugged into any linear browsing model to make the browsing model suitable for the grid layout. The visiting probability of item d at ranking position i where the row number is $r(i)$

[2]The original study referred the model as cascade click model. However, in our paper, we referred the model as geometric to keep the conceptual consistency.

and column is $c(i)$ which is:

$$P_{SD}[E_i] = \min(\beta^{r(i)} \prod_{j=[1,i-1]} (1 - \psi(j), 1)) \tag{6.8}$$

The selection probability may or may not incorporate relevance judgements or estimates. The parameter β can be modeled by the *abandon* probability with the assumption of the probability of abandoning items increasing with rows; as users move vertically in grid layout, the probability of abandoning the process increases. Hence, the *slower-decay* user browsing behavior can be modeled by adjusting *abandon* probability where the *abandon* probability will increase vertically with rows but the horizontal (columns) *abandon* probability will remain the same across the columns in each row.

$$P_{SD}[E_i] = \alpha^{r(i)-1} \prod_{j=[1,i-1]} (1 - \psi(j), 1)) \tag{6.9}$$

Users' *row-skipping* (RS) behavior is incorporated in browsing models with an assumption that if users *examine* any item in a row, that particular row is not *skipped*. The parameter γ determines the probability of *skipping* each of the rows before $r(i)$. If users *examined* or *selected* any item in a particular row, that means that row was not skipped. Hence the examining probability of item d at ranking position i (row $r(i)$, column $c(i)$) in the generalized browsing model is:

$$P_{RS}[E_i] = \left[\prod_{k=1}^{r(i)-1} (1 - \gamma) \prod_{j \in L(\mathrm{k},\cdot)} (1 - \psi(j)) + \prod_{k=1}^{r(i)-1} \gamma \right] \prod_{i \in L(\mathrm{k},\cdot)} (1 - \psi(i)) \tag{6.10}$$

Therefore, if we want to incorporate all the components and browsing behaviors, the generalized browsing model of estimating the probability of a user examining an item at position i in a given ranking is:

$$P[E_i] = P[\neg A_{r(i)}] \left[\prod_{k=1}^{r(i)-1} P[\neg K_k] \prod_{j \in L(\mathrm{k},\cdot)} P[\neg A_j | \neg S_j, E_j] + \prod_{k=1}^{r(i)-1} P[K_k] \right] \prod_{i \in L(\mathrm{k},\cdot)} P[\neg A_i | \neg S_i, E_i] \tag{6.11}$$

Our generalized browsing model can be configured to implement various browsing models and ranking layouts in the following ways:

- To consider *slower-decay* user browsing behavior without *row-skipping* tendency, the *row-skipping* probability γ can be ignored with the value of 0.

- To consider a linear vertical layout or single-column ranking, the generalized model can be used by ignoring *row-skipping* and *slower-decay* behavior. In that case, the probability of skipping a row will be fixed ($\gamma = 0$) and the column abandon probability will be 0 with a constant row abandon probability.

- The configuration of the discussed grid-layout suitable browsing models depend on the parameterization of *selection* probability, *row-skipping* probability, and *abandon* probability. Table 6.1 show the potential values of the mentioned parameters.

- Based on the availability of relevance, the *selection* probability $\psi(i)$ can be relevance dependant (*cascade*) or constant (*geometric*).

- The browsing models can incorporate *abandon* probability α at each position of the ranking.

- The abandon probability can be derived as a function of relevance of items.

This generalized browsing model can be further extended to incorporate various browsing patterns and complex ranking layouts.

- The generalized browsing model can be extended to incorporate *middle-bias* which is a grid-layout suitable user browsing behavior by increasing *selection* probability for the middle positions in rankings. Xie et al. [192] modified *selection* probability by considering that as a normal distribution.

- This generalized browsing model can also incorporate an *F-shaped* user browsing tendency by adjusting *skipping* probability γ and *selection* probability ψ. However, we need accurate user browsing pattern to derive the appropriate value of the parameters.

6.3 Conclusion and Future Work

In this work, we identify user browsing models for ranking in linear layout in information access systems and show that the existent user browsing models are conceptually similar and they can be generalized and configured based on ranking layouts, availability of component like relevance information, and parameter settings. We provide generalized configurable framework of the browsing models that can be extended to grid layout by considering grid-layout suitable browsing behaviors. The proposed unified framework relies on configurable parameters such as *selection* probability,

abandon probability, *row-skipping* probability, and *decay* parameters and thus various browsing behavior in various ranking layouts can be represented by calibrating these parameters.

Our analysis indicates the importance of knowing accurate user browsing behaviors for various ranking layouts to estimate optimal parameter values and user-eye-tracking studies in various ranking scenarios can help in this area. This work relies on multiple common and existing assumptions of user browsing behaviors (mentioned in section 6.1) excluding other possible user browsing behaviors that are seldom studied. For example, we assumed that the process will end once the user selects an item. However, users may select an item and return to the result page. Future user studies can consider this browsing behavior so that browsing models can incorporate this *multi-select* user behavior; our theoretical treatment can be extended to account for it by adding additional probability modeling whether the user continues or stops their browsing after selecting an item.

Furthermore, in grid layout, users can skip a row even after examining some items in that particular row and examining items in a row may not always follow a left-to-right pattern. Future research work on understanding user browsing behaviors in grid layout can focus on identifying users' row-skipping behavior and their patterns in examining items in rows. Then the generalized browsing models can be further modified and configured depending on users' browsing patterns in grid-layout. Therefore, studies on user browsing behavior can emphasize on inferring optimal parameter setting to generate reliable user attention scores for any given ranking layouts. Users' browsing behavior in a *multi-list* grid layout is still under-studied and the categories or genres can have an affect on users browsing behavior. Hence, *wrapped* grid-layout suitable

browsing models may not be applicable in *multi-list* grid layouts which indicate the need to understand user browsing behaviors in *multi-list* ranking scenarios.

CHAPTER 7:

OPTIMIZING GRID LAYOUT FOR PROVIDER-SIDE FAIRNESS

In chapter 5, we modify fair ranking metrics to measure provider-side group fairness in ranking in grid layouts and show the consistency of metric scores across ranking layouts and browsing models as well. We observe whether a ranking that is optimized for fairness in a particular layout still remains fair with the change of layout or not and from figure 5.4, we find a linear ranked list that has been optimized for fairness may not remain fair when the results are displayed in a grid layout. This observation emphasizes the need of considering layout suitable browsing models while optimizing ranked results for fairness. In this chapter[1], I undertake that work and propose and test methods for producing rankings that are optimized for fairness in grid layouts by using browsing models suitable to grid layout.

In IAS, ranked results are often optimized for fairness and relevance to preserve a balance between user satisfaction and fairness and there exist several fairness aware re-ranking techniques [61, 201, 52, 111, 78]. *TREC Fair Ranking Track 2019* [17] and *2022* [64] provided multiple fairness-aware re-ranking tasks for which participants

[1]I collaborated with Dr. Michael Ekstrand for this work. We will be preparing this chapter for submission to WSDM 2023 or a similar venue after the defense

optimized their information retrieval systems by optimizing for both fairness and relevance. However, these efforts are limited to linear ranked lists and the problem of optimizing ranking in grid layouts for fairness has received limited attention so far. Chen et al. [37] proposed a re-ranking technique when recommended items are displayed in a 2-D grid, but fairness was not in their scope. Re-ranking technique suitable for grid layout to optimize provider-side group fairness is still unknown.

Moreover, as seen in chapter 4 and chapter 5, fair ranking metric scores vary depending on user browsing behavior and user browsing behavior varies across ranking layouts. For example, user attention decays more slowly in grid layouts than in linear ranked lists [192]. Hence, for the same set of ranked retrieved or recommended results, items can receive different user attention depending on how they are displayed to users. There is limited research on understanding user browsing behavior in grid layout in IAS but the browsing models that are available have not yet been incorporated into fairness ranking strategies.

As previously stated in chapter 5, the geometry of grid layout changes depending on user devices and the column size (number of columns) changes with screen sizes. For example, the streaming service *Disney+* displays movie recommendations in 3 columns on a phone, in 5 columns on a laptop, and in 4 columns on TV. In production, systems often use whole-page optimization strategies where the ranking is optimized for the device form factor when it is known in advance [8]. However, re-sizable or re-orientable devices like laptop browsers, tablets, and phones etc. still need resizing strategies. To adapt the layout to a particular device size, items are re-ranked by either re-wrapping or cutting-off from the right side. The same set of recommended or retrieved items are re-ranked based on device size to fit the screen. As a consequence,

the exposure or attention items receive from ranked results vary with the change of user device. Hence, with the change of column sizes the fairness score will be different for provider groups. Moreover, optimizing a grid ranking for fairness with a particular column size may not remain fair when the column size or user device changes. Systems need to re-rank items in ranked results for device-suitable column sizes to preserve fairness across devices.

Items can be presented in two ways in a grid layout: wrapped (figure 7.1(a)) and multi-ranking lists (figure 7.1(b)). In a multi-ranking list, items are ranked in a list of lists based on categories. Displaying items in groups facilitates users to find them from their preferred categories. In a wrapped grid layout, items are ranked in multiple rows and columns without being categorized into groups. A fairness-aware re-ranking technique designed for wrapped grid layouts may not be applicable for multi-list grid layout. There is no user eye-tracking studies to show how user allocate attention in multi-list grid layouts specifically for search results or recommendations scenarios. However, having a re-ranking method for general grid-based browsing models will yield an optimization approach that can be fine-tuned for more precise attention models.

Since fairness score varies across layouts and an optimized linear layout for fairness does not remain fair in a grid layout, we need to optimize fairness for grid layouts considering grid layout-suitable user browsing models. Our work will contribute towards filling the gap in optimization of fairness in grid layouts by providing the first re-ranking technique to optimize provider-side group fairness in grid layouts.

In this work, we work on optimizing ranking in grid layout for provider-side group fairness in wrapped grid layout. We adapt a commonly used re-ranking techniques

which is suitable for linear layout and modify that for grid layout by incorporating grid-layout suitable browsing models. Since designing fairness-aware re-ranking techniques for ranking in grid layouts depends on ranking design, user browsing behavior, and column size or user device, we observe the impact of column sizes and browsing models on grid-aware re-ranking techniques.

We answer the following research questions in these regards:

- **RQ1.** Does incorporating grid-aware browsing models to existing re-ranking technique improve fairness for ranked results in grid layout?

- **RQ2.** Does a ranking in a grid layout optimized for fairness in a device remain fair for other devices?

- **RQ3.** How can we optimize ranking in grid layouts for various screen sizes?

- **RQ4.** Do browsing models have an affect on the optimization of ranking in grid layouts for fairness?

By providing a simple and initial re-ranking approach for general grid layouts, we contribute to the improvement for provider-side group fairness in grid layouts. If more specific user attention models are developed in the future for ranking in grid layouts, they can be plugged into the proposed re-ranking models. Our analyses help practitioners to design a more fine-tuned re-ranking approach for grid layout in IAS considering item metadata, tasks and domain and elicit several future research directions towards fairness concerns in multi-ranking grid layouts.

7.1 Related Work

In this section, I review background work on re-ranking and fairness-aware re-ranking techniques in information access systems; the fairness-specific background is provided in chapter 2.

7.1.1 Re-Ranking Techniques

System retrieved or recommended results are often optimized for quality measures derived from these evaluation criteria; for example, re-ranking the initial ranking using various optimization approaches [132, 86]. Moreover, in multi-stage ranking architectures, systems retrieve an initial set of candidate items based on relevance and then various re-ranking techniques are applied to generate the final ranking [84, 69, 137]. Several re-ranking and *learning to rank* (LTR) approaches have been proposed to optimize ranking for utility [23, 22, 27, 61, 75, 111, 106, 132, 190, 174]. LTR methods learn to rank based on scoring functions which is used to determine an optimized ranking; individual items, list, or pair of ranked items are considered to measure loss function against ideal ranking. Depending on the design of the loss function, the LTR approaches are categorized into pairwise, point-wise, and list-wise approaches [132, 189].

Pairwise approaches are often based on the change in ranking quality with the swap of each pair of items in ranking [149]. In RankNet [23] and LambdaMART [22], ranking quality is optimized by predicting an optimal ordering for each pair of items in ranked list before generating the final ranking. In point-wise optimization approaches, the ranking model is trained to minimize loss function determined from each individual item score [106]. In this approach, each of the item in candidate set

is scored independently based on the target quality. Unlike previous two approaches, list-wise approaches consider the entire ranked list and the ranking function is trained on the entire list based on the minimization of the loss function [27, 104, 194].

7.1.2 Optimizing Ranking for Fairness

Fairness optimization in ranking often involves trade-off between utility and fairness score [165, 52, 14, 57] where various LTR and re-ranking approaches focus on improving fairness score with minimum utility loss. There are several approaches to optimize ranking for fairness in IAS that are often categorized into *pre-processing*, *in-processing*, and *post-processing* (typically re-ranking) techniques [136]. In pre-processing approaches, the potential bias in dataset or training labels are investigated in order to identify and mitigate bias in ranking [168, 94]. In in-processing approaches, the IAS algorithms or models are optimizing for fairness or a combination of fairness and utility in the training phase [128, 14] and LTR methods can be used in this regard. In post-processing approaches, already ranked results are optimized for fairness by applying re-ranking techniques to improve or optimize a fairness objective. Items from the initial ranked results are used to generate a new ranking following fairness objective such as target distribution [195]. Several constraint optimization approaches have been used to re-rank the initial ranked results and generate an optimized ranking [111, 165, 52]; the optimization constraints often includes both user satisfaction metrics and fair ranking metrics to preserve a balance between fairness and utility.

Various fair ranking metrics are used to measure fairness in ranking and to determine the target fairness score. There are several *greedy* optimization techniques that are proposed to generate fairness-aware rankings with minimum impact on ranking

quality [61, 78]. Liu et al. [111] proposed a personalized fairness-aware re-ranking algorithm for micro-lending recommendations where each item from the initial ranking will be assigned to a position in the ranking based on the optimization or maximization of personalization and group fairness. Singh and Joachims [165] and Diaz et al. [52] considered exposure of provider-side in ranking in their fairness-aware ranking optimization techniques. However, all the approaches I discussed above are proposed and implemented in linear ranked results when items are displayed in single-column list.

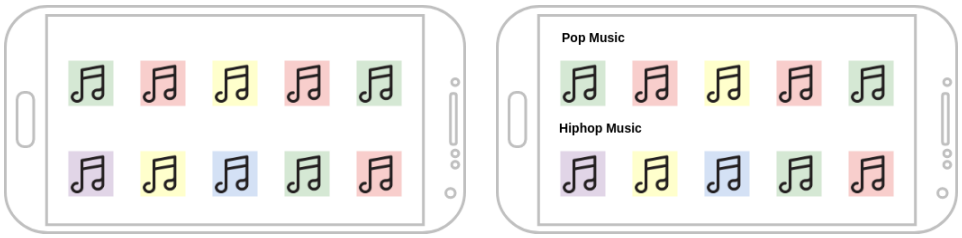
In this work, we modify a widely used pairwise swap re-ranking technique to optimize ranking in grid layout for provider-side group fairness.

7.2 Problem Formulation

In this section we introduce the experimental settings for optimizing grid-based ranking for fairness.

In this work, we consider a recommender system that recommends n items $d_1, d_2, \ldots, d_n \in D$ in response to information requests from users $q_1, q_2, \ldots, q_m \in Q$ based on their relevance to the request $y(d|q)$ and presents the results in a wrapped grid layout L. Items are associated with producers or providers who in turn can be associated with demographic attributes identifying them with one or more of g groups. We model group membership of documents with group alignment vector $\mathcal{G}(d) \in [0,1]^g$ (s.t. $\|\mathcal{G}(d)\|_1 = 1$) forming a distribution over groups; this allows for mixed, partial, or uncertain membership in an arbitrary number of groups. Table 3.3 presented in chapter 3 summarizes the notation used in this paper.

The ranking will be optimized for provider-side group fairness while preserving a balance between utility and fairness. Hence, we use grid layout-aware browsing models in both fair ranking metrics to measure fairness in ranking and in an effectiveness metric to measure utility in ranking.



(a) Wrapped Grid Layout (b) Multi-list Grid Layout

Figure 7.1: Various types of grid layout models

Ranking Layout IAS display ranked results in both linear or grid layout and in grid layout (figure 7.1) items are displayed in multiple rows and multiple columns. Grid layouts can further be classified into *wrapped* and *multi-list* layout. In multi-list grid layout, items are displayed in a grid but each row represents different genre or category (figure 7.1(b)). In wrapped grid layout, items are not categorized into any categories (figure 7.1(a)), rather the recommended or retrieved items are displayed horizontally in multiple rows.

There is no research yet (to our knowledge) on understanding user browsing behavior in multi-list grid ranking in IAS settings. Hence, in this work, we are providing fairness aware re-ranking method considering wrapped grid layout which can be further extended for multi-list layout with suitable browsing models.

User Browsing Model Users do not provide equal attention to every position in ranked results [16], item at the lower position of a ranked list will not receive similar attention as the item at the top position. Since user attention varies across positions in ranked results, the position weight for each position in ranking depends on how users browse the displayed ranked results.

There are several user browsing models to demonstrate user browsing behavior in linear ranked lists. *Cascade* [43] and *geometric* [122] are two popularly used user browsing models to infer the the probability of user visiting an item in a particular position in ranking. These models are explained in chapter 6 showing that these models differ in their underlying components and parameter settings but can be cast as different configurations of the same model. In both geometric and cascade models, user attention or position weight decays exponentially with ranking positions but in cascade browsing model, user selection probability is a function of item relevance.

To implement grid layout-aware evaluation metrics, it is important to understand how users provide attention to items in grid layout or how user attention changes across items when they are displayed in grid layout. Xie et al. [191] showed various user browsing behaviors in grid layout in e-commerce search results and they showed users show *row-skipping*, *slower-decay*, *middle-bias* while browsing items in grid layout. Users often skips rows while browsing ranked results in a grid layout and they tend to show higher attention to the middle position of columns in ranking. Moreover, user attention decays slowly across items in grid layout than linear list.

In chapter 5, I described our proposed modified version of geometric and cascade browsing models incorporating grid layout-suitable *row-skipping* and *slower-decay* user browsing behaviors and I implemented fair ranking metrics in grid layouts by

incorporating grid-aware browsing models. In chapter 6, a generalized configurable framework of user browsing model is provided to estimate user attention in both linear and grid layout based on required components and selected browsing behavior.

In this work, we use the generalized framework of user browsing models and reconfigure that to adapt *row-skipping* and *slower-decay* user behavior in *geometric* browsing model to measure position weight in grid layout. Table 6.1 describes the parameters for browsing models with their range of values.

As noted in chapter 6, for a given ranking in grid layout, the visiting probability of item d in geometric-based row-skipping model is:

$$P_{RS(\text{geometric})}[V_d] = \left[\prod_{k=0}^{\text{row}(d)} (1-\gamma) \prod_{i \in L(\text{k},\cdot)} (1-\psi) + \prod_{k=0}^{\text{row}(d)} \gamma \right] \prod_{i \in \text{row}(d)} (1-\psi) \tag{7.1}$$

and the geometric visiting probability of item d with slower decay is:

$$P_{SD(\text{geometric})}[V_d] = \min(\beta^{\text{row}(d)} \prod_{i=[0,L^{-1}(d)]} (1-\psi), 1) \tag{7.2}$$

Target Fairness The purpose of this work is to provide a preliminary approach to develop fairness-aware re-ranking techniques for ranking in grid layout, so I focus on fairness optimization in a single-ranking setting, leaving fair grid layouts in stochastic settings for future work. To measure provide-side group fairness in single ranking layout, we follow recommendations from the comprehensive analysis of fair ranking metrics in chapter 4 [142] and use AWRF. Sapiezynski et al. [154] proposed attention-weighted rank fairness or AWRF which measures the difference between group exposure and configurable target distribution $\hat{\mathbf{p}}$ which represents the ideal ex-

posure distribution over groups. Attention vector and the group alignment matrix is used to derive group exposure $\boldsymbol{\epsilon}_L$ ($\boldsymbol{\epsilon}_{\mathrm{L}} = \mathcal{G}(\mathrm{L})^{\mathrm{T}}\mathbf{a}_{\mathrm{L}}$) by aggregating the attention given to items of each group in proportion to their group membership as represented by the alignment vector. Since our distribution difference function in bounded by $[0, 1]$, we invert it so that AWRF $= 1$ at maximal fairness to be more directly comparable to the effectiveness metrics:

$$\mathrm{AWRF}(L) = 1 - \Delta(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}}) \tag{7.3}$$

Target Utility To measure utility in ranking, we consider an effectiveness metric that consider items position weights in measurement. Moffat and Zobel [122] proposed *rank-biased precision* (RBP) which combined a geometric browsing model with binary relevance to measure the overall effectiveness of a ranking in a manner similar to nDCG, but with a re-configurable browsing model. The source of relevance can be the actual relevance judgement which generates RBP or system estimated relevance which generates $\hat{\mathrm{RBP}}$. For a given ranking L, the rank-biased precision metric score is

$$\hat{\mathrm{RBP}} = \psi \sum_{i=[0,L^{-1}(d)]} y(L(i)|q)\dot{(}1-\psi)^{i-1} \tag{7.4}$$

where $y(L(i)|q)$ is the systems estimated relevance score for the item in position i and the stopping probability ψ is decaying exponentially with ranking position. This metric can be adapted to measure $\hat{\mathrm{RBP}}$ in grid layout by incorporating grid layout suitable browsing behavior. Thus, we modify the attention model used in this metric by considering geometric-based row-skipping model (equation 7.1) and geometric-

based slower-decay (equation 7.2).

7.2.1 Re-Ranking Algorithm

Pairwise swapping re-ranking is a commonly used post-processing approach that we adapt to optimize ranking in grid layout for provider-side group fairness. For a given initial ranking L, we optimize the ranking by considering alternative ranking position for each pair of ranked items and finally generate a fairness-aware ranked result L'. Starting from the top of the list, for each position i, we consider each potential swap with positions $j > i$, items swap their position and temporarily generate a new ranking $L_{i \leftrightarrow j}$ keeping all the other items at the same place. Then we measure the lift in fairness as $\Delta\text{AWRF}(L, L_{i \leftrightarrow j})$ and the loss in utility as $\Delta\text{RBP}(L, L_{i \leftrightarrow j})$.

$$\Delta\text{RBP}(L, L_{i \leftrightarrow j}) = \text{RBP}(L_{i \leftrightarrow j}) - \text{RBP}(L) \tag{7.5}$$

$$\Delta\text{AWRF}(L, L_{i \leftrightarrow j}) = \text{AWRF}(L_{i \leftrightarrow j}) - \text{AWRF}(L) \tag{7.6}$$

Thus for each of the position i, we select the best swap by solving the maximization of lift function, $F(i \leftrightarrow j | i, j \in \{1, ..., N\}, i < j)$:

$$F(i \leftrightarrow j) = \arg \max_{j \in i, ..., N} \{\Lambda \Delta\text{AWRF}(L, L_{i \leftrightarrow j}) \cdot (1 - \Lambda)\,(1 - \Delta\text{RBP}(L, L_{i \leftrightarrow j}))\} \tag{7.7}$$

Algorithm 1 shows the formal algorithm for optimizing grid-ranking for provider-side group fairness. In each iteration, item in position i is temporarily swapped with items that are in higher position than i and for each swap, it measures the AWRF improvement and inverse RBP loss. The swap that gives the maximum lift in fairness score with minimum utility loss is selected to generate a new ranking. Λ is used as a

configurable balancing factor between fairness and utility.

Algorithm 1 Fairness-Aware Re-ranking for Grid Ranking

Require: initial ranking L, user q, estimated relevance score $y(L|u)$, balancing factor Λ

Ensure: Re-ranked L'

procedure RE-RANK(L)
  $L' \leftarrow L$
  measure AWRF(L)
  measure RBP(L)
  for $i \in 1, ..., N$ **do**
    for $j \in i, ..., N$ **do**
      swap items in position i and j to generate $L_{i \leftrightarrow j}$
      measure ΔRBP$(L, L_{i \leftrightarrow j})$
      measure ΔAWRF$(L, L_{i \leftrightarrow j})$
    end for
    $i' = \arg\max_{j \in i,...,N} \{\Lambda \Delta \text{AWRF}(L, L_{i \leftrightarrow j}) \cdot (1 - \Lambda)\,(1 - \Delta RBP(L, L_{i \leftrightarrow j}))\}$
    if $i' \neq i$ **then**
      $L^{'} \leftarrow L_{i \leftrightarrow i'}$
      AWRF$(L) \leftarrow$ AWRF$(L_{i \leftrightarrow i'})$
      RBP$(L) \leftarrow$ RBP$(L_{i \leftrightarrow i'})$
    end if
  end for
  return L'
end procedure

7.3 Experimental Setup

In this work, our goal is to observe whether and how the provider-side group fairness improves in ranking when we modify existing re-ranking technique to be grid layout-aware and apply that to optimize ranking in grid layout for fairness and utility. To answer our research questions, we perform several experiments. We modify the pairwise swap re-ranking technique to be grid-aware by incorporating grid-layout

suitable browsing models. We further implement the modified grid-aware re-ranking algorithm on real-world IAS dataset to observe how the algorithm performs in real-world IAS scenario.

7.3.1 Dataset

In this work, we use *GoodReads* book dataset described in chapter 3.

7.3.2 Methodology

We optimize provider-side group fairness in grid layout using the modified re-ranking technique considering two types of user browsing models. We also observe the affect of device sizes or column sizes on fairness optimization in grid layout.

RQ1. Improvement of Fairness in Grid Layout To observe the group fairness score improvement for provide-side fairness in grid layout,

- We implement the fair ranking metric AWRF to measure fairness in single ranking. We use distribution of male and female authors in book dataset to compute target distribution $\hat{\mathbf{p}}$. We compare the improvement of AWRF score in the re-ranked grid ranking where 1 is the highest score of fairness.
- To measure utility, we implement effectiveness metric RBP.
- Both AWRF and RBP are implemented with grid-layout suitable browsing models, *row-skipping* and *slower-decay* with column size 5.
- We use 0.5 as the default value of the fairness-utility balancing parameter Λ.

RQ2. Consistency of Optimized Fairness Across Devices As previously discussed, based on user devices, column size of grid layout changes. For example,

Goodreads shows book recommendations in grid layout and the column size changes across devices; books are displayed in 5 columns on laptop, 2 columns on phone, and 9 columns on iPad. Hence, the system can display the same set of items in various column sizes depending on user device. Re-ranking the items by taking device size into consideration can help to preserve fairness across devices because optimizing the ranked results in grid layout for a particular device may not remain fair for other devices.

- We observe if and how the optimized fairness score from a re-ranked grid layout of column size n changes in other columns sizes.
- We optimize the grid-based ranked results with column size of 5 and use that fairness-aware re-ranked results to measure provider-side group fairness by changing column size to 2, 3, 4, 7, and 9.

RQ3. Preserve Fairness Across Devices Since item exposure varies across column sizes in grid layout which affect the fairness score for provider groups, we want to preserve provider-side fairness across devices. With that goal,

- We implement the grid-aware re-ranking technique for multiple column sizes to maintain group fairness across user devices and observe the change in fairness optimization with the change of column sizes.
- We implement the grid-aware re-ranking algorithm for grid ranked results with common columns sizes of 2, 3, 4, 5, 7, and 9.

RQ4. Impact of Browsing Models on Fairness Optimization To observe the impact of browsing models on fairness optimization in grid layout, we implement

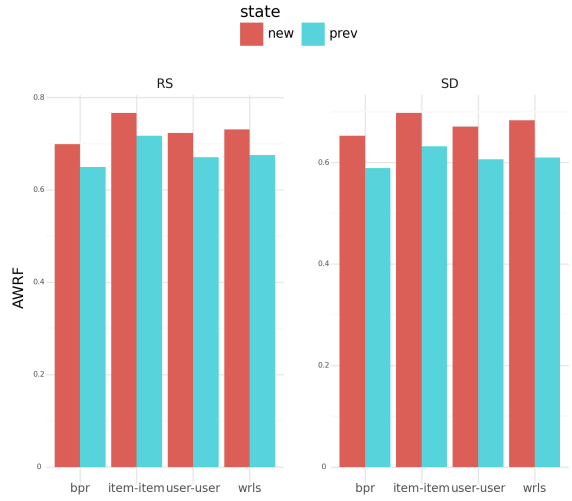
both group fairness metric and effectiveness metric incorporating grid-layout suitable *row-skipping* and *slower-decay* browsing models with their default parameter settings. Table 4.3 shows the default parameters values for these browsing models.

7.4 Results and Discussion

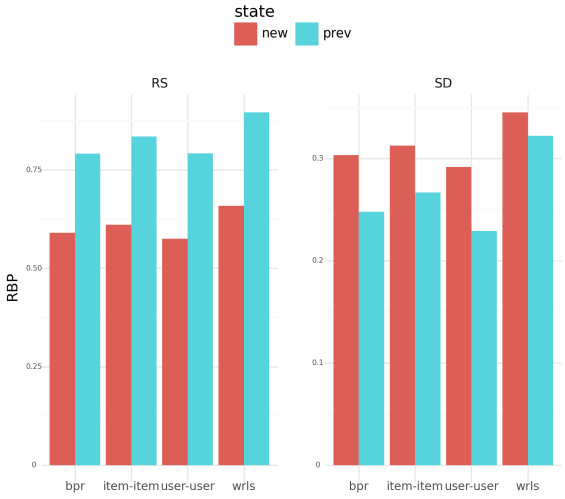
This section provide the results from our experiments.

RQ1 *Does incorporating grid-aware browsing models to existing re-ranking technique improve fairness for ranked results in grid layout?*

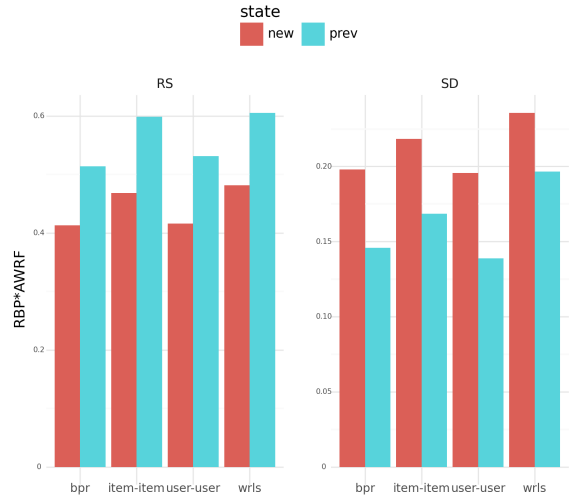
Figure 7.2(a) shows that the AWRF score increases in all the recommendation algorithms for both *row-skipping* and *slower-decay* browsing models. We do *paired t-test* [91] to observe the significance of this fairness improvement and find that for the algorithms in both browsing models, the AWRF score improvement is statistically significant with $p_{val} < 10^{-20}$. We round up the p-values at $\alpha = 0.05$ with Benjamini-Hochberg correction [13]. In both browsing models, the fairness score varies across recommendation algorithms during both pre and post-optimization showing the same patterns. For all the recommendation algorithms, the fairness scores improves significantly for ranking in grid layout when we consider grid-layout suitable browsing models. Figure 7.2(b) shows the RBP score and Figure 7.2(c) shows the RBP$*$AWRF score differences in between pre and post-optimization. For the *slower-decay* browsing model, the combined score improves in all the algorithms and the utility score improves after re-ranking. This observation emphasizes the importance of using grid-aware re-ranking technique while optimizing ranked results displayed in grid layout.



(a) AWRF score difference



(b) RBP score difference



(c) Fairness-Utility trade-off

Figure 7.2: Pre and post-optimization fairness and utility scores in grid layout with column size 5

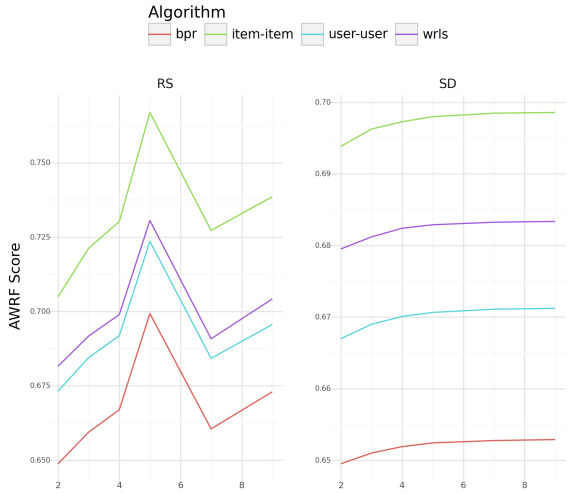


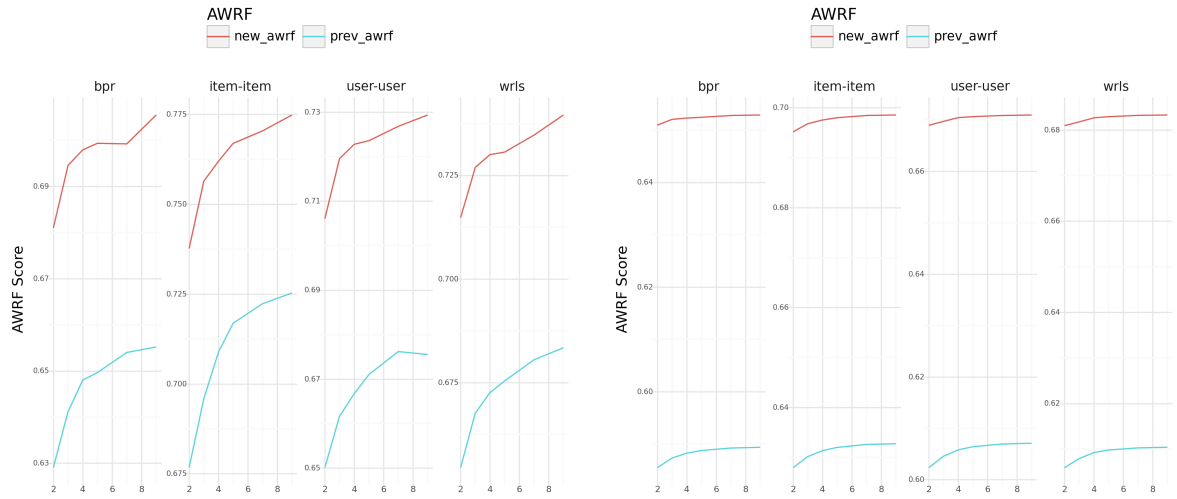
Figure 7.3: An optimized grid layout with column size 5 is not fair for other column sizes.

RQ2. *Does a ranking in grid layout optimized for fairness in a device remain fair for other devices?*

RQ2 shows the impact of column sizes on fairness optimization in grid layout. Figure 7.3 shows how fairness score for an optimized ranking changes with column sizes. A fairness-aware re-ranked 5-column grid layout does not remain fair when the column size is different and this pattern is true for all the algorithms. The pattern is more notable in *row-skipping* browsing model for all the algorithms. This result implies the need of considering appropriate column size to preserve fairness for the same set of ranked items across devices.

RQ.3 *How can we optimize ranking in grid layout for various screen sizes?*

Figure 7.4 shows the improvement in fairness scores after optimizing ranking in the grid layout for various column sizes and the result shows a consistency in fairness improvement across various column sizes. For all the considered column sizes, AWRF



(a) *Row-skipping* browsing model (b) *Slower-decay* browsing model

Figure 7.4: Improvement in fairness across column sizes in grid-aware browsing models

score improve significantly in all the recommendation algorithms ($p_{val} < 0.0001$ rounded at $\alpha = 0.05$ with Benjamini-Hochberg correction) after optimizing ranking in grid layout using grid-aware browsing models. By looking at figure 7.4, we can see that fairness score varies with the change of column sizes and this pattern remains consistent even after optimization in all the algorithms for both browsing models. This result shows that, fairness optimization of a given grid layout of column size n should consider the same column size while measuring position weight using browsing models to improve fairness in that ranking.

RQ.4. *Do browsing models have an affect on the optimization of ranking in grid layout for fairness?*

Figure 7.5 shows fairness improvement pattern across grid-layout suitable browsing models: *row-skipping* and *slower-decay*. With both *row-skipping* and *slower-decay*

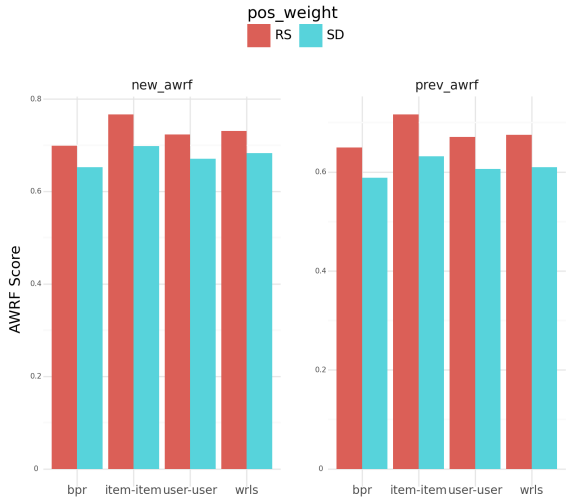


Figure 7.5: AWRF **score varies across browsing models.**

user browsing models, fairness scores are improved in all the algorithms and the algorithms show the similar pattern is fairness score across the browsing models before and after optimization. Even though fairness score improves with both browsing models through grid-aware fairness optimization, figure 7.5 shows that the fairness score differences across column sizes is more notable for *row-skipping* browsing model than for *slower-decay*. In figure 7.2(b) and in figure 7.2(c), we observe that the utility score improves for all the algorithms when we consider *slower-decay* browsing model but that improvement does not hold for *row-skipping* browsing model. This result shows that fairness score varies with the change of browsing model, thus fairness optimization for a grid layouts requires to consider suitable grid-aware browsing models.

7.4.1 Discussion

This work provides a preliminary design of grid-aware re-ranking techniques to optimize provider-side group fairness in grid layout. Through our experiments we provide

insights on the impact of device sizes and browsing models on fairness optimization in grid layout. Grid-aware re-ranking for fairness has not received enough attention in fair ranking research especially concerning provider-side groups fairness. Hence, an analysis on the implementation of fairness optimization technique in grid layout will help researcher and practitioners in designing more advanced grid-aware re-ranking strategies.

We have made several observations from our analysis. Our work shows that it is possible to improve fairness in grid layout if we can make re-ranking techniques grid-aware by incorporating grid-layout suitable browsing models. However, our results also show that, the improvement in fairness score can vary depending on user browsing models. This observation highlight the importance of considering suitable browsing models while measuring and optimizing group fairness in grid layout. Understanding how users browse grid layout and identifying various browsing tendencies can help developing more accurate fairness optimization technique for grid layout.

Moreover, results from our analysis show that, device size is an important factor in improving fairness in grid layout. Optimizing provider-side group fairness in ranking in grid layout by considering a particular column size will not remain fair when the column size is different. Hence, a ranked result which is optimized for fairness while displaying in phone will not remain fair while displaying in a laptop. Optimizing a grid layout considering a default column size will not provide fair outcome for provider groups across devices such as phone, TV, and laptop. Therefore, to preserve fairness across devices, a retrieved or recommended results displayed in a particular device needs to be re-ranked considering the appropriate column size while displaying in another device.

Implementing fairness-aware re-ranking technique with different column sizes show that the fairness improved in all the algorithms for both browsing models. However, our results show that the consistency of fairness score across column sizes varies based on browsing models. For *row-skipping* browsing model, the fairness score varies notably across column sizes but for *slower-decay*, the fairness scores are more consistent across column sizes. This observation emphasizes the need of selecting suitable browsing model and column size while optimizing ranking in grid layout for provider-side group fairness.

7.5 Conclusion

In this chapter, we work towards filling a gap in the research area of provider-side group fairness in ranking in IAS by studying fairness improvement in grid layout. We modify a widely used fairness-aware re-ranking technique to make it grid-aware by incorporating grid-layout suitable user browsing models. We implement the modified grid-aware re-ranking technique in real-world IAS dataset to observe the fairness improvement in ranking in grid layout. Our analysis shows that device size and user browsing models are crucial factors in designing fairness-aware re-ranking technique to optimize provider-side group fairness in grid layout in IAS.

This work opens up several potential research directions in improving provider-side fairness in grid layout. Our work shows the importance of using accurate user browsing models in fairness optimization for grid layout. User browsing behavior in ranking in grid layout has not received much attention yet, hence, further research work on understanding user browsing behavior in grid layout will help ensuring fair-

ness in grid layout with minimum utility loss.

Moreover, in this work, we do not consider multi-list grid layout where items are displayed in multiple categories. Re-ranking technique designed for wrapped grid-layout may not work for multi-list grid layout because in multi-list grid, each rows represents different genre or categories. Moreover, same item can appear in multiple rows. Hence, future work is needed to optimize multi-list grid ranking for fairness by considering unique features and suitable user browsing models for multi-list ranking.

I believe this work will provide researcher and practitioners an guideline on what to expect while designing an optimization technique for fairness in grid layout and what factors to consider carefully.

CHAPTER 8:

CONCLUSION

In my dissertation, I work on the improvement of fairness in ranking in information access systems and the provider-side group fairness is the scope of my work in fairness. My research work broadly focuses on the problem of measuring fairness in ranking and optimizing ranked results for fairness. Prior to this dissertation, there were several metrics for assessing fairness in ranking, but (1) there was little guidance to select or apply them to real-world IAS setup; and (2) they were limited to linear layout, while ignoring widely-used ranking in grid layouts. My work addresses both of these gaps making it significantly easier to select and implement fair ranking metrics to measure provider-side fairness in ranking in real-world IAS and identifying important work needed in the future to ensure IAS are fair to item providers. Furthermore, my work seeks to bridge a research gap in the area of provider-side group fairness in ranking in IAS by modifying widely used fairness-aware re-ranking technique to make it grid-aware to optimize provider-side group fairness in grid layout. This work provides insights on designing grid-aware re-ranking techniques.

8.1 Contributions

My research work contributes to the area of provider-side group fairness in ranking in following ways:

- We provide a comprehensive and comparative analysis of existing fair ranking metrics that are proposed to measure provider-side group fairness in ranking. By describing the metrics using unified notations and framework we show conceptual similarities and differences among these metrics which help researchers and practitioners to develop deeper insights on the fairness goal, assumptions, and applicability of these metrics.

- By implementing the metrics using real-world IAS dataset considering both search and recommendation scenarios we identify the required components and challenges in implementations of these metric in the real-world IAS data. Moreover, implementing these metrics under same experimental setup using same dataset helps to directly compare their applicability.

- By conducting sensitivity analysis considering external factors such as parameter settings and position weight, we show the impact of design choices on metric results. This comparative analysis on the sensitivity of the metrics helps to identify the vulnerability and reliability of the existing metrics.

- The empirical analysis on metric design, applicability, and reliability provides informed guidelines on fairness-task specific metric selection process. Our recommendations help researchers and practitioners in their decision making process while selecting a metric that matches with their requirements.

- From the comprehensive analysis of fair ranking metrics, we identify several potential research directions in this area that can help to improve the usability and reliability of the fair ranking metrics.
- We expand the applicability of fair ranking metrics by implementing them in widely-used but seldom-studied ranking structure which is grid layout. We provide modified versions of the fair ranking metrics to measure provider-side group fairness when items are displayed in grid layout.
- Considering various grid-layout suitable user browsing models into fair ranking metrics provides insights on the applicability of these metrics in grid layouts. Moreover, showing how the fairness score changes across browsing models, we provide an idea on the consistency and reliability of the fair ranking metrics while applying to grid layouts.
- Showing the impact of column sizes and column reduction approaches on the fair ranking metric scores in grid layouts help researcher and practitioners to better understand the important factors to implement fair ranking metrics in grid layout.
- Modifying the linear layout suitable fair ranking metrics to measure fairness in grid layout highlights several potential research gaps. Findings from this work emphasize the need of investigating and identifying suitable user browsing behavior in grid layout to develop trustworthy and valid fair ranking metrics for various layouts.
- We recognize the existing user browsing models that are suitable for linear and grid layouts and identify the conceptual similarities among the browsing models.

We provide a generalized framework that can be configured to account for both linear and grid layouts. This unified framework of browsing models can further be extended to more advanced browsing models.

- Our analysis on the existing user browsing models provide insights on their required components and parameter settings. Depending on the task and ranking layout, researchers and practitioners can use the generalized framework of user browsing models by re-configuring the proposed unified structure.

- Our research contributes towards the improvement of provider-side group fairness in ranking by developing a primary and general provider-side group fairness-aware re-ranking technique for grid layout in IAS. This work is a starting point of provider-side fairness improvement in ranking in grid layout which is a commonly used layout in many IAS.

- We modify a widely-used re-ranking technique which is suitable for ranking in linear layout by incorporating grid layout suitable user browsing models. By making the existing re-ranking technique grid-aware, we provide a simple way to optimize ranking in grid layout for provider-side fairness.

- By implementing grid-aware re-ranking technique in real-world IAS dataset, we identify crucial factors to consider while designing fairness-aware re-ranking algorithms for ranking in grid layout.

- Our analysis on the impact of device sizes and user browsing models on the grid-aware re-ranking technique provides insights on the viability and reliability of a ranking in grid layout that is optimized for fairness.

8.2 Future Work

Our comparative analysis on measuring provider-side group fairness in ranking elicits several future research directions towards the advancement of fair ranking metrics. Moreover, our preliminary work on the fairness measurement and optimization issues in a limited-studied but widely-used ranking layout paves the way for new avenues of research to enhance fairness in ranking within the grid layout. Further research as immediate next steps from our work can include:

- Developing fair ranking metrics to be stable towards external factors such as browsing models and parameter settings.
- Most of the metrics are sensitive towards the data sparsity or missing data problem such as missing relevance information and missing group label. Further work is needed to develop fair ranking metrics that will be stable towards these issues.
- A simulation study on the fair ranking metrics is needed to better understand the sensitivity of the metrics towards design choices such as ranking size, group information, relevance availability, and position weights.
- Since user browsing model is one of the most important factors in measuring fairness in ranking, further work is needed to better understand user browsing behavior in ranking for various layouts. User eye-tracking studies can help to show how user browsing behavior changes with ranking layouts, tasks, and item meta information in IAS.
- How users browse multi-list grid layout is not well-explored yet, hence, it is

important to identify user browsing behavior in multi-list grid layout to generate reliable and accurate fairness score for ranking in that setting.

- Future research work on designing fairness-aware re-ranking technique for grid layout should include multi-list grid layout and investigate how the categories or various rows play role in fairness optimization.

Towards Fair Ranking in IAS The fair ranking research has a long way to go with several issues to focus on. The common goal of the research in this area is to improve fairness in ranking in IAS by mitigating bias. With that goal, fair ranking research includes the concerns regarding identification, measurement, and mitigation of bias in ranking. Our research on measuring and optimizing fairness in ranking especially concerning grid layout is an initial step towards the advancement of fairness in ranking in various layouts. Moreover the comprehensive analysis of fair ranking metrics highlights multiple limitations that need more extensive research work in future. Some of the potential long-term research ideas include:

- Having the ground truth or item relevance information is a crucial part of measuring fairness in ranked lists because several metrics incorporate relevance information to measure fairness in ranking. However, relevance information can carry societal bias reflecting social stereotypes and prioritizing one group over another, thus influencing the result of fair ranking metrics. Future work can address that issue to help design fair ranking metric(s) that will take the potential existence of underlying societal historical bias in relevance into consideration while measuring fairness.

- There are fair ranking metrics to measure fairness concerning unfair exposure

distribution or distributional harm. However, IAS can reflect and reinforce representational bias such as social stereotypes through ranking. Since the existing fair ranking metrics will not be suitable to measure representational bias in ranking, future research work should focus on the issue of measuring representational bias associated with items in ranking . Issues concerning how stereotypes are propagated in ranking, the harm from this phenomenon, and the mitigation of social stereotypes from ranked results in IAS are limitedly explored. Working on these concerns will help improve overall fairness of ranking limiting both distributional and representational harm.

- With the advancement of various manners of user interaction with IAS, the dimension of fairness in ranking issue is also changing. Users nowadays use voice search to interact with systems and the retrieved results can also carry social biases. However, the concern regarding the existence of bias in voice search results have not been explored yet. Identifying and measuring bias in more advanced modalities or formats such as voice search results is important to study to develop an comprehensive understanding of fairness in ranking in IAS.

8.3 Concluding Remarks

Fairness is a complicated concept and fairness-aware ranking in IAS is still a developing research area with multiple open problems. We believe my dissertation work will lay a valuable foundation for this vital and ongoing work. Our empirical analysis on fair ranking metrics will assist researchers and practitioners in understanding the

concepts of fair ranking metrics, the fairness assumptions while measuring provide-side group fairness in ranking, and the challenges in their implementation; this will provide an informed guidance in fair ranking metric selection process. Our research on measuring and optimizing provider-side group fairness in ranking in grid layout has opened a significant and impactful realm of exploration. By gaining insights into the complexities and possibilities within this space, our work has opened up new avenues for addressing crucial challenges and discovering solutions. We look forward to research that explore the open concerns in fair ranking research area while identifying new challenges in order to ameliorate the fairness issues in ranking in IAS.

CHAPTER 9:

PUBLICATION TARGETS

Table 9.1 shows my PhD dissertation status and target. I presented my research progress and status at the Doctoral Symposium at RecSys 2022 [140].

Table 9.1: Publication status and target

Chapter	Status	Publication/Target
Analyzing Fair Ranking Metrics	Published	1. FAccTRec 2020 [143] 2. SIGIR 2022 [142] (full paper)
Beyond Linear Layout Fairness	Working on the rejection reviews	RecSys 2023
Generalized Framework of User Browsing Models	Working on the rejection reviews	ICTIR 2023
Fairness-aware Grid-based Ranking	Targeting	WWW 2023

REFERENCES

[1] H. Abdollahpouri, R. Burke, and B. Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555*, 2019.

[2] G. Adomavicius and Y. Kwon. Toward more diverse recommendations: Item re-ranking methods for recommender systems. In *Workshop on Information Technologies and Systems*, pages 79–84. Citeseer, 2009.

[3] Y. Afoudi, M. Lazaar, and M. Al Achhab. Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, 113:102375, 2021.

[4] C. C. Aggarwal and C. C. Aggarwal. An introduction to recommender systems. *Recommender systems: the textbook*, pages 1–28, 2016.

[5] S. Alelyani. Detection and evaluation of machine learning bias. *Applied Sciences*, 11(14):6271, 2021.

[6] H. Alharthi, D. Inkpen, and S. Szpakowicz. A survey of book recommender systems. *Journal of Intelligent Information Systems*, 51:139–160, 2018.

[7] B. Almquist. Macewan university at the trec 2020 fair ranking track.

[8] C. Alvino and J. Basilico. Learning a personalized homepage. 2015. URL `https://netflixtechblog.com/learning-a-personalized-homepage-aa8ec670359a`.

[9] N. Asadi and J. Lin. Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 997–1000, 2013.

[10] A. Ashkan and C. L. Clarke. On the informativeness of cascade and intent-aware effectiveness measures. In *Proceedings of the 20th International Conference on World wide web*, pages 407–416, 2011.

[11] R. Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018.

[12] A. Balyan, A. Singh, P. Suram, D. Arora, and V. Srivastava. Using product meta information for bias removal in e-commerce grid search. *IEEE Data Eng. Bull.*, 44(2):81–91, 2021.

[13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[14] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International*

Conference on Knowledge Discovery & Data Mining, pages 2212–2220, 2019. doi: 10.1145/3292500.3330745.

[15] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019.

[16] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 405–414, 2018. doi: 10.1145/3209978.3210063.

[17] A. J. Biega, F. Diaz, M. D. Ekstrand, and S. Kohlmeier. Overview of the trec 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*, 2019.

[18] A. J. Biega, F. Diaz, M. D. Ekstrand, and S. Kohlmeier. Overview of the trec 2019 fair ranking track. *arXiv preprint arXiv:2003.11650*, 2020.

[19] R. Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 514–524, 2020. doi: 10.1145/3351095.3372864.

[20] J. Bobadilla, A. Hernando, F. Ortega, and J. Bernal. A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 38 (12):14609–14623, 2011.

[21] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[22] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

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

[24] R. Burke. Knowledge-based recommender systems. *Encyclopedia of library and information systems*, 69(Supplement 32):175–186, 2000.

[25] R. Burke. Hybrid web recommender systems. *The adaptive web: methods and strategies of web personalization*, pages 377–408, 2007.

[26] R. Burke. Multisided fairness for recommendation. July 2017. URL `http://arxiv.org/abs/1707.00093`.

[27] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[28] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998. doi: 10.1145/290941.291025.

[29] B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*, pages 903–912, 2011.

[30] P. Castells and A. Moffat. Offline recommender system evaluation: Challenges and new directions. *AI Magazine*, 43(2):225–238, 2022.

[31] C. Castillo. Fairness and transparency in ranking. In *ACM SIGIR Forum*, volume 52, pages 64–71. ACM New York, NY, USA, 2019.

[32] P. Chandar, F. Diaz, and B. St. Thomas. Beyond accuracy: Grounding evaluation metrics for human-machine learning systems. In *Advances in Neural Information Processing Systems*, 2020.

[33] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 621–630, 2009.

[34] S. Chatterjee and L. Dietz. Bert-er: Query-specific bert entity representations for entity ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1466–1477, 2022.

[35] M. Chen and P. Liu. Performance evaluation of recommender systems. *International Journal of Performability Engineering*, 13(8):1246, 2017.

[36] R.-C. Chen, L. Gallagher, R. Blanco, and J. S. Culpepper. Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proceedings of the 40th Interna-*

tional ACM SIGIR Conference on Research and Development in Information Retrieval, pages 445–454, 2017.

[37] S. Chen, X. Zhang, X. Chen, Z. Li, Y. Wang, Q. Lin, and J. Xu. Reinforcement re-ranking with 2d grid-based recommendation panels. *arXiv preprint arXiv:2204.04954*, 2022.

[38] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

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

[40] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. 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.

[41] Y. Cooper. Amazon ditched ai recruiting tool that favored men for technical jobs. *The Guardian*, 2018.

[42] N. Craswell. Mean reciprocal rank. *Encyclopedia of database systems*, 1703, 2009.

[43] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

[44] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46, 2010.

[45] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010.

[46] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications, 2018. URL `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0`

[47] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[48] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[49] F. H. Del Olmo and E. Gaudioso. Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, 35(3):790–804, 2008.

[50] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, and D. Zanzonelli. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, pages 1–50, 2023.

[51] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004. doi: 10.1145/963770.963776.

[52] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 275–284, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411962. URL https://doi.org/10.1145/3340531.3411962.

[53] S. Djamasbi, M. Siegel, and T. Tullis. Visual hierarchy and viewing behavior: An eye tracking study. In *International conference on human-computer interaction*, pages 331–340. Springer, 2011.

[54] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[55] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 331–338, 2008.

[56] C. Dwork and C. Ilvento. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*, volume 3, 2018.

[57] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through

awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[58] M. D. Ekstrand. Lenskit for python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2999–3006, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412778. URL https://doi.org/10.1145/3340531.3412778.

[59] M. D. Ekstrand and D. Kluver. Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction*, feb 2020. doi: 10.1007/s11257-020-09284-2. URL https://md.ekstrandom.net/pubs/bag-extended.

[60] M. D. Ekstrand, J. T. Riedl, J. A. Konstan, et al. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173, 2011.

[61] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, and D. Kluver. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 242–250, 2018.

[62] M. D. Ekstrand, A. Das, R. Burke, and F. Diaz. Fairness in information access systems. *arXiv preprint arXiv:2105.05779*, 2021.

[63] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, et al. Fairness in information

access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2): 1–177, 2022.

[64] M. D. Ekstrand, G. McDonald, A. Raj, and I. Johnson. Overview of the trec 2022 fair ranking track. *arXiv preprint arXiv:2302.05558*, 2023.

[65] A. Epps-Darling, H. Cramer, and R. T. Bouyer. Artist gender representation in music streaming. In *ISMIR*, pages 248–254, 2020.

[66] A. Fabris, A. Purpura, G. Silvello, and G. A. Susto. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377, 2020.

[67] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[68] Y. Feng, D. Saelid, K. Li, R. Gao, and C. Shah. University of washington at trec 2020 fairness ranking track. *arXiv preprint arXiv:2011.02066*, 2020.

[69] M. A. Fligner and J. S. Verducci. Multistage ranking models. *Journal of the American Statistical association*, 83(403):892–901, 1988.

[70] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of documentation*, 2003.

[71] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[72] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287589. URL https://doi.org/10.1145/3287560.3287589.

[73] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, mar 2021. ISSN 0001-0782. doi: 10.1145/3433949. URL https://doi.org/10.1145/3433949.

[74] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

[75] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[76] R. Gao and C. 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*, pages 229–236, 2019.

[77] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260, 2010.

[78] S. C. Geyik, S. Ambler, and K. Kenthapadi. Fairness-aware ranking in search &

recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2221–2231, 2019.

[79] A. Ghazimatin, M. Kleindessner, C. Russell, Z. Abedjan, and J. Golebiowski. Measuring fairness of rankings under noisy sensitive information. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2263–2279, 2022.

[80] A. Ghosh, R. Dutt, and C. Wilson. *When Fair Ranking Meets Uncertain Inference*, page 1033–1043. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. URL https://doi.org/10.1145/3404835.3462850.

[81] A. Ghosh, L. Genuit, and M. Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR, 2021.

[82] C. A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 2015.

[83] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(12), 2009.

[84] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42, 2022.

[85] R. Guo, X. Zhao, A. Henderson, L. Hong, and H. Liu. Debiasing grid-based product search in e-commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2852–2860, 2020.

[86] A. Gupta, E. Johnson, J. Payan, A. K. Roy, A. Kobren, S. Panda, J.-B. Tristan, and M. Wick. Online post-processing in rankings for fair utility maximization. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 454–462, 2021.

[87] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[88] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 230–237, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130961.

[89] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[90] K. Hofmann, L. Li, F. Radlinski, et al. Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval*, 10(1):1–117, 2016.

[91] H. Hsu and P. A. Lachenbruch. Paired t test. *Wiley StatsRef: statistics reference online*, 2014.

[92] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[93] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA, 2017.

[94] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[95] M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1): 1–42, 2016.

[96] C. Kate. The trouble with bias. 2017. URL `https://www.youtube.com/watch?v=fMym_BKWQzk`.

[97] M. Kay, C. Matuszek, and S. 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.

[98] O. Kırnap, F. Diaz, A. Biega, M. Ekstrand, B. Carterette, and E. Yilmaz. Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the Web Conference 2021*, WWW '21, page 1065–1075, New York, NY, USA,

2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450080. URL https://doi.org/10.1145/3442381.3450080.

[99] T. Kletti and J.-M. Renders. Naver labs europe at trec 2020 fair ranking track. 2020.

[100] R. Kohavi and R. Longbotham. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

[101] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[102] G. J. Kowalski. *Information retrieval systems: theory and implementation*, volume 1. springer, 2007.

[103] C. Kuhlman, W. Gerych, and E. Rundensteiner. Measuring group advantage: A comparative study of fair ranking metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*, 2021.

[104] Y. Lan, Y. Zhu, J. Guo, S. Niu, and X. Cheng. Position-aware listmle: A sequential learning process for ranking. In *UAI*, pages 449–458, 2014.

[105] H. Li. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

[106] P. Li, Q. Wu, and C. Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. *Advances in neural information processing systems*, 20, 2007.

[107] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*, pages 624–632, 2021.

[108] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang. Fairness in recommendation: A survey. *arXiv preprint arXiv:2205.13619*, 2022.

[109] D. Liang, L. Charlin, and D. M. Blei. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*, 2016.

[110] T.-Y. Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[111] W. Liu, J. Guo, N. Sonboli, R. Burke, and S. Zhang. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 467–471, 2019.

[112] Y. Liu, C. Wang, M. Zhang, and S. Ma. User behavior modeling for better web search ranking. *Frontiers of Computer Science*, 11:923–936, 2017.

[113] A. Livne, E. S. Tov, A. Solomon, A. Elyasaf, B. Shapira, and L. Rokach. Evolving context-aware recommender systems with users in mind. *Expert Systems with Applications*, 189:116042, 2022.

[114] M. Makhortykh, A. Urman, and R. Ulloa. Detecting race and gender bias in visual representation of ai on web search engines. In *Advances in Bias and Fairness in Information Retrieval: Second International Workshop on Algorithmic Bias in Search and Recommendation, BIAS 2021, Lucca, Italy, April 1, 2021, Proceedings*, pages 36–50. Springer, 2021.

[115] K. Mallari, K. Inkpen, P. Johns, S. Tan, D. Ramesh, and E. Kamar. Do i look like a criminal? examining how race presentation impacts human judgement of recidivism. In *Proceedings of the 2020 Chi conference on human factors in computing systems*, pages 1–13, 2020.

[116] C. D. Manning. *Introduction to information retrieval*. Syngress Publishing,, 2008.

[117] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[118] G. McDonald and I. Ounis. University of glasgow terrier team at the trec 2020 fair ranking track. In *The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings*, volume 1266, 2020.

[119] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[120] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[121] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Prediction-Based decisions and fairness: A catalogue of choices, assumptions, and definitions. Nov. 2018. URL `http://arxiv.org/abs/1811.07867`.

[122] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27, 2008.

[123] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 659–668, 2013.

[124] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS)*, 35(3):1–38, 2017.

[125] B. B. Murdock Jr. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482, 1962.

[126] M. Naghiaei, H. A. Rahmani, and Y. Deldjoo. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 770–779, 2022.

[127] P. Nandy, C. Diciccio, D. Venugopalan, H. Logan, K. Basu, and N. El Karoui. Achieving fairness via post-processing in web-scale recommender systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 715–725, 2022.

[128] H. Narasimhan, A. Cotter, M. R. Gupta, and S. Wang. Pairwise fairness for ranking and regression. In *AAAI*, pages 5248–5255, 2020.

[129] S. U. Noble. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press, 2018.

[130] G. K. Patro, L. Porcaro, L. Mitchell, Q. Zhang, M. Zehlike, and N. Garg. Fair ranking: a critical review, challenges, and future directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1929–1942, 2022.

[131] M. J. Pazzani and D. Billsus. Content-based recommendation systems. *The adaptive web: methods and strategies of web personalization*, pages 325–341, 2007.

[132] C. Pei, Y. Zhang, Y. Zhang, F. Sun, X. Lin, H. Sun, J. Wu, P. Jiang, J. Ge, W. Ou, et al. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM conference on recommender systems*, pages 3–11, 2019.

[133] L. Peska and P. Vojtas. Off-line vs. on-line evaluation of recommender systems in small e-commerce. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 291–300, 2020.

[134] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[135] C. Pinney, A. Raj, A. Hanna, and M. D. Ekstrand. Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access. *arXiv preprint arXiv:2301.04780*, 2023.

[136] E. Pitoura, K. Stefanidis, and G. Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28, 2021.

[137] J. Qin, J. Zhu, B. Chen, Z. Liu, W. Liu, R. Tang, R. Zhang, Y. Yu, and W. Zhang. Rankflow: Joint optimization of multi-stage cascade ranking systems as flows. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 814–824, 2022.

[138] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674, 2010.

[139] S. K. Raghuwanshi and R. K. Pateriya. Recommendation systems: techniques, challenges, application, and evaluation. In *Soft Computing for Problem Solving: SocProS 2017, Volume 2*, pages 151–164. Springer, 2019.

[140] A. Raj. Fair ranking metrics. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 742–743, 2022.

[141] A. Raj and M. D. Ekstrand. Fire dragon and unicorn princess; gender stereotypes and children's products in search engine responses. *arXiv preprint arXiv:2206.13747*, 2022.

[142] A. Raj and M. D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736, 2022.

[143] A. Raj, C. Wood, A. Montoly, and M. D. Ekstrand. Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311*, 2020.

[144] A. Raj, A. Milton, and M. D. Ekstrand. Pink for princesses, blue for superheroes: The need to examine gender stereotypes in kid's products in search and recommendations. *arXiv preprint arXiv:2105.09296*, 2021.

[145] A. Raj, B. Mitra, N. Craswell, and M. D. Ekstrand. Patterns of gender-specializing query reformulation. *arXiv preprint arXiv:2304.13129*, 2023.

[146] J. Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[147] F. Rehman, O. Khalid, and S. A. Madani. A comparative study of location-based recommendation systems. *The Knowledge Engineering Review*, 32:e7, 2017.

[148] N. Rekabsaz, B. Mitra, M. Lupu, and A. Hanbury. Toward incorporation of relevant documents in word2vec. *arXiv preprint arXiv:1707.06598*, 2017.

[149] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 452–461, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

[150] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[151] A. Roshdi and A. Roohparvar. Information retrieval techniques and applications. *International Journal of Computer Networks and Communications Security*, 3(9):373–377, 2015.

[152] C. Rus, J. Luppes, H. Oosterhuis, and G. H. Schoenmacker. Closing the gender wage gap: Adversarial fairness in job recommendation. *arXiv preprint arXiv:2209.09592*, 2022.

[153] M. Sanderson et al. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375, 2010.

[154] P. Sapiezynski, W. Zeng, R. E Robertson, A. Mislove, and C. Wilson. Quantifying the impact of user attention on fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 553–562, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317595. URL `https://doi.org/10.1145/3308560.3317595`.

[155] M. F. Sayed and D. W. Oard. The university of maryland at the trec 2020 fair ranking track. 2020.

[156] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. *The adaptive web: methods and strategies of web personalization*, pages 291–324, 2007.

[157] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas. Music recommender systems. *Recommender systems handbook*, pages 453–492, 2015.

[158] T. Schumacher, M. Lutz, S. Sikdar, and M. Strohmaier. Properties of group fairness metrics for rankings. *arXiv preprint arXiv:2212.14351*, 2022.

[159] A. Selbst and S. Barocas. Big data's disparate impact. *California Law Review*, 104:671–732, September 2016.

[160] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL `https://doi.org/10.1145/3287560.3287598`.

[161] C. Shah, R. White, P. Thomas, B. Mitra, S. Sarkar, and N. Belkin. Taking search to task. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 1–13, 2023.

[162] D. Shakespeare, L. Porcaro, E. Gómez, and C. Castillo. Exploring artist gender bias in music recommendation. *arXiv preprint arXiv:2009.01715*, 2020.

[163] S. Shrestha and K. Lenz. Eye gaze patterns while searching vs. browsing a website. *Usability News*, 9(1):1–9, 2007.

[164] T. Silveira, M. Zhang, X. Lin, Y. Liu, and S. Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813–831, 2019.

[165] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2219–2228, New York, NY, USA, 2018. Associa-

tion for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220088. URL https://doi.org/10.1145/3219819.3220088.

[166] A. Singh and T. Joachims. Policy learning for fairness in ranking. *Advances in neural information processing systems*, 32, 2019.

[167] N. Sonboli, R. Burke, M. Ekstrand, and R. Mehrotra. The multisided complexity of fairness in recommender systems. *AI magazine*, 43(2):164–176, 2022.

[168] R. Sonoda. A pre-processing method for fairness in ranking. *arXiv preprint arXiv:2110.15503*, 2021.

[169] L. Spear, A. Milton, G. Allen, A. Raj, M. Green, M. D. Ekstrand, and M. S. Pera. Baby shark to barracuda: Analyzing children's music listening behavior. In *Fifteenth ACM Conference on Recommender Systems*, pages 639–644, 2021.

[170] P. Sulikowski and T. Zdziebko. Horizontal vs. vertical recommendation zones evaluation using behavior tracking. *Applied Sciences*, 11(1):56, 2020.

[171] K. M. Svore and C. J. Burges. A machine learning approach for improved bm25 retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1811–1814, 2009.

[172] G. Takács, I. Pilászy, and D. Tikk. Applications of the conjugate gradient method for implicit feedback collaborative filtering. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 297–300, New York, NY, USA, Oct. 2011. Association for Computing Machinery. ISBN 9781450306836. doi: 10.1145/2043932.2043987.

[173] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14):4–4, 2007.

[174] M. Taylor, J. Guiver, S. Robertson, and T. Minka. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 77–86, 2008.

[175] J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In *Proceedings of the second acm international conference on web search and data mining*, pages 15–24, 2009.

[176] D. R. Turnbull, S. McQuillan, V. Crabtree, J. Hunter, and S. Zhang. Exploring popularity bias in music recommendation models and commercial steaming services. *arXiv preprint arXiv:2208.09517*, 2022.

[177] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells. On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 260–268, 2018.

[178] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells. Assessing ranking metrics in top-n recommendation. *Information Retrieval Journal*, 23:411–448, 2020.

[179] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, 2011.

[180] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval. Context-aware recommender systems for learning: a survey and future challenges. *IEEE transactions on learning technologies*, 5(4):318–335, 2012.

[181] J. Vinagre, A. M. Jorge, and J. Gama. Evaluation of recommender systems in streaming environments. *arXiv preprint arXiv:1504.08175*, 2015.

[182] E. M. Voorhees, D. M. Tice, et al. The trec-8 question answering track evaluation. In *TREC*, volume 1999, page 82, 1999.

[183] M. Wan and J. McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 86–94, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

[184] L. Wang and T. Joachims. User fairness, item fairness, and diversity for rankings in two-sided markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 23–41, 2021.

[185] X. Wang, Y. Zhang, and R. Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022.

[186] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

[187] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1009–1018, 2010.

[188] L. A. Wicht, J. Waldfogel, S. Waldfogel, et al. Playlisting favorites: Is spotify gender-biased? Technical report, Joint Research Centre (Seville site), 2018.

[189] L. Wu, C.-J. Hsieh, and J. Sharpnack. Sql-rank: A listwise approach to collaborative ranking. In *International Conference on Machine Learning*, pages 5315–5324. PMLR, 2018.

[190] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, 2008.

[191] X. Xie, Y. Liu, X. Wang, M. Wang, Z. Wu, Y. Wu, M. Zhang, and S. Ma. Investigating examination behavior of image search users. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 275–284, 2017.

[192] X. Xie, J. Mao, Y. Liu, M. de Rijke, Y. Shao, Z. Ye, M. Zhang, and S. Ma. Grid-based evaluation metrics for web image search. In *The world wide web conference*, pages 2103–2114, 2019.

[193] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18 (1):79–112, 2000.

[194] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398, 2007.

[195] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–6, 2017.

[196] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1561–1564, 2010.

[197] J. Yu, D. Tao, M. Wang, and Y. Rui. Learning to rank using user clicks and visual features for image retrieval. *IEEE transactions on cybernetics*, 45(4): 767–779, 2014.

[198] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 1011–1018, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772793. URL `https://doi.org/10.1145/1772690.1772793`.

[199] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva,

CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052660. URL https://doi.org/10.1145/3038912.3052660.

[200] E. Zangerle and C. Bauer. Evaluating recommender systems: Survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022.

[201] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. 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, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3132938. URL https://doi.org/10.1145/3132847.3132938.

[202] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021.

[203] Q. Zhao, S. Chang, F. M. Harper, and J. A. Konstan. Gaze prediction for recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 131–138, 2016.

[204] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

APPENDIX A:

NON-THESIS PUBLICATIONS

Besides working on the area of measuring fairness in ranking, I have conducted some research work on the existence of stereotypes in information access systems. Through my research work, I explore whether and how search engines and recommender systems replicate and reinforce gender stereotypes associated with children's products.

Published

- Raj et al. [145] is published at the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (*SIGIR 2023*). During my internship at Microsoft, I worked on understanding demographic group-based query-reformulation and the impact of search engine result page (SERP) on such reformulation. Users often reformulate their queries to precisely express their information need. In this work, we looked at demographic group-based query reformulation where user explicitly mention demographic group attributes such gender while reformulating their queries. We identified potential reasons for reformulating queries for particular groups and observed the connection between group representation in SERP and user need to reformulate their queries for that group. Our findings further shed light on the need of understanding bias in group representation in SERP and the impact on user demand mismatch.

- Raj et al. [144] is published at the 5th International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (*KidRec 2021*) Co-located with ACM IDC 2021. In this work, I argued that it is important to investigate the existence of pre-existing gender stereotypes in search engines and recommender systems. We provide real-world examples to strengthen our argument while specifically focusing on learning materials for

children. We provide potential research directions concerning this phenomena.

- Raj and Ekstrand [141] is published at SIGIR ecom: ACM SIGIR Workshop on e-Commerce (*SIGIR e-com 2022*). I explored e-commerce search systems to identify the tendency of reflecting and manifesting gender stereotypes associated with children's products. We generated an aggregated list of pre-documented gender stereotypes children's products to identify and measure gender stereotypes in search results and query suggestions. We provide preliminary methods for quantifying gender stereotypes in e-commerce search system and conducted our experiments across multiple e-commerce sites. Our findings provide initial evidence to the existence of gender stereotypes associated with kid's products in search results and query suggestions in e-commerce settings. This work is an initial step towards identifying and measuring gender stereotypes in IAS, particularly for children's products.

- Spear et al. [169] is published at the 15th ACM Conference on Recommender Systems (*RecSys 2021*). This is a collaboration with my research group members at *PIRet* where we looked for pattern in online music listening behavior of children. The purpose of this work is to improving music recommendations for children by providing insights on their music preference.

- Pinney et al. [135] is published at the ACM SIGIR Conference on Human Information Interaction and Retrieval (*CHIIR 2023*). I contributed in a research work where we looked at how gender has been used in information retrieval and user profiling research area. We collected paper published in renowned conferences on information retrieval and user profiling and identified whether, why,

and how gender has been used as a variable. We also analyzed how gender information has been inferred in the existing research. We categorized the papers based on This work indicated several problematic use of gender in research. For example, gender is often considered as binary where in real-world it is not binary. This work further provide guidance ethical and proper use of gender information in research work to avoid harmful outcome from misuse of gender.