Bias in book recommendation: A survey

SAVVINA DANIIL, Human-Centered Data Analytics, Centrum Wiskunde en Informatica, Amsterdam, Netherlands
GAURI BHAGWAT, DHLab, KNAW Humanities Cluster, Amsterdam, Netherlands
LAURA HOLLINK, Human-Centered Data Analytics, Centrum Wiskunde en Informatica, Amsterdam, Netherlands

'Automated book recommendation is a compelling application given the rich and long-standing human tradition of writing and reading. The documented discrimination in the publishing world raises questions on the types of bias that can arise when algorithms are trained on user-book interactions. Various papers have studied recommender systems bias based on datasets from platforms where users share their opinions on books. In this paper, we conduct the first survey of computer science research on bias in book recommender systems. We survey 40 papers from top conferences and journals using the structured PRISMA methodology and provide an overview of the field. We map the existing literature into bias types, subjects and actions, and present key metrics and bias mitigation methods. We find that there is limited targeted research, fragmented evaluation practices, and a lack of engagement with the unique characteristics of the literary domain. Finally, we propose avenues for future work and make concrete suggestions for interdisciplinary research.

CCS Concepts: • **Information systems → Recommender systems**; • **Applied computing → Arts and humanities**.

Additional Key Words and Phrases: Recommender Systems, Bias, Book Recommendation

1 Introduction

Online book recommendation has existed for a long time and influences worldwide consumption patterns. For example, Amazon-owned Goodreads is a book-centric social media platform very popular among readers internationally. The datafication of reading patterns which drives their recommender system also has a strong commercial interest for authors, publishers and booksellers [2]. Recently, virality seems to be an important factor in the book market, with the rise of online communities hosted in social media such as TikTok and YouTube that often decide which books will be chosen for publication [11]. In the public library sector, there are attempts to utilize AI systems that recommend books to users while ensuring that central library values are being adhered to [73]. In this context, there have been multiple calls to promote diverse books from librarian associations and Library & Information Science (LIS) researchers [12]. It is, therefore, pertinent to investigate what is the effect of automated suggestion systems on book consumption. Do they propagate the status quo in the book market? Do they expose users to options they would otherwise not be aware of?

Bias in book recommender systems has received some attention in computer science research, but to our knowledge no comprehensive literature review on this topic currently exists. There are general surveys of bias in either machine learning (e.g., Mehrabi et al. [41]) or recommender systems in particular (e.g., Chen et al. [7]). These surveys provide valuable frameworks that establish how bias emerges as a result of the interaction between various system components. However, book recommendation warrants targeted analysis distinct from

Authors' Contact Information: Savvina Daniil, Human-Centered Data Analytics, Centrum Wiskunde en Informatica, Amsterdam, Netherlands; e-mail: s.daniil@cwi.nl; Gauri Bhagwat, DHLab, KNAW Humanities Cluster, Amsterdam, Netherlands; e-mail: gauri.bhagwat@dh.huc.knaw.nl; Laura Hollink, Human-Centered Data Analytics, Centrum Wiskunde en Informatica, Amsterdam, Netherlands; e-mail: l.hollink@cwi.nl.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2770-6699/2026/11-ART
https://doi.org/10.1145/3858614

Just Accepted

these broader surveys, especially in order to support researchers who study book recommendation bias by providing a structured map of the research landscape. Recommender systems are an explicitly applied discipline, and conclusions are dependent on the unique characteristics of the domain at hand. Previous surveys cover a wide ground, and thus do not facilitate readers in recognizing what has been studied, what remains fragmented and where gaps exist when it comes to book recommendation specifically It is very relevant, for example, to examine whether some types of biases are more prevalent in book-specific research while others have never been systematically studied, and this information is not immediately discernible in existing surveys due to their generality. Bias needs to be considered on a domain basis, especially since it is a complex term that is known to be inconsistently used. This is regardless of whether the method proposed to measure or mitigate bias explicitly takes into consideration characteristics of the dataset that are beyond statistical properties. Even statistical properties like popularity manifest differently between domains. A movie being popular among general audiences is not equivalent to a book being popular given the difference in time dedication between the two and the downward trends of reading [21]. It follows that to ensure scientific advancement, it is beneficial to monitor the research output of measuring and mitigating recommendation bias as it pertains to the unique characteristics of the book domain.

In this survey, we aim to contribute to the field of bias in book recommender systems by presenting and summarizing the latest research, as well as highlighting what is missing and what avenues future research could follow.

Our research questions are as follows:

(1) What is the current state of research on bias perpetuated by algorithmic systems that recommend books?
(2) What are the challenges, research gaps, and future paths related to this research?

To answer these questions, we survey 40 papers from top-tier computer science venues (journals and conference proceedings) that study bias in book recommender systems. We present the process followed to collect these papers in a structured way with the help of the PRISMA methodology [53]. We provide general statistics across a set of dimensions such as publication details, prediction task, and type of feedback. We then categorize the papers based on the type of bias they investigate, interaction-based or attribute-based; the bias subject, item or user; and the action they perform, bias measurement or mitigation. Finally, we discuss insights, gaps, and future work that is relevant for the domain. Published research on bias in book recommender systems is limited compared to other media applications. In recent years the field has expanded, with significant attention being given to popularity bias. However, we notice that studies often do not sufficiently engage with the fact that they are studying bias in a *book* dataset and rather usually treat it as one more benchmark dataset frequently used by the recommender systems community. Additionally, computer science research does not seem to build on LIS work around what the risks of automated systems that recommend books are. This observation leaves a lot of room for fruitful between-fields collaborations in the future.

The paper is structured as follows. In section 2, we provide background on recommender systems, bias, and recommendation in the library. In section 3, we discuss other surveys in recommender systems. In section 4, we describe our methodology for collecting and analyzing papers. In section 5, we present the results of our survey. In section 6, we discuss our insights into gaps and challenges of the domain. In section 7, we provide a short summary and conclusion.

2 Background

2.1 Recommender systems research

In the vast and ever-expanding information space, users struggle with finding items relevant or useful to them. Recommender systems' main role is to automate the process of suggesting items to users in a given context. Various features of a system are considered when such a process takes place, from characteristics of the items

Just Accepted

to user demographics and consumption history to requirements of the platform that hosts them. Research into recommender systems often operates by simulating and studying one or more systems, due to the often limited access to real systems. Adopting a simulation process assumes the existence of features that an algorithm can exploit to make recommendations. Researchers employ various strategies to construct datasets with features: they scrape web applications, access APIs set up by these applications, do user studies, and more [93]. These datasets in practice often take the form of *user - item - interaction*, with *interaction* being *click*, *score*, *purchase*, or other. Depending on the domain and source, further knowledge is sometimes available, such as information on the user (e.g., gender, age) and item (e.g., publication year, genre, creator). Algorithms are trained to construct high-dimensional relationships between users and items in the dataset via computational techniques such as encoding user history (*collaborative filtering*), analyzing item features (*content-based filtering*), or hybrid combinations [62]. Given access to a dataset, researchers evaluate a set of algorithms usually by training them on one part of the dataset and testing them on another, as is commonly done in machine learning. Finally, researchers report on the algorithm's 'success' in modeling the dataset based on calculating a set of metrics in the recommended lists produced by the algorithm [25].

2.2 Bias in recommendation

Algorithmic success in recommendation is relative. Traditionally, algorithms have been trained to accurately predict the structure of the test set by, for example, producing ranked lists of recommendations that closely resemble what users have actually interacted with in the test set. As the field progressed, various beyond-accuracy metrics have been proposed with the goal of encapsulating more than what accuracy metrics do [40]. Criteria such as diversity and novelty are considered equally critical components of success, both with user satisfaction and business interests in mind. At the same time, the concept of bias has received attention recently, both in recommender systems and AI applications in general. According to Ferrara [19], bias is defined as a "systematic error in decision-making processes that results in unfair outcomes" that stems from AI systems learning biased patterns in the training data and accordingly replicating them in their outcomes.

In recommender systems, bias can appear in various stages of the process. Chen et al. [7] distinguish three points where bias occurs: in the data, in the model, and in the results. *Data* bias occurs due to the fact that data is usually collected observationally, and it therefore does not necessarily match the underlying preference distribution of the users (e.g., selection bias, position bias). *Model* bias is a result of the assumptions made by a model to generalize beyond the training data. Finally, *result* bias refers to the discrimination towards a group of users or items in the output of the recommendation. Measuring result bias involves the estimation of whether a system malfunctions for a set of users or items that have common properties. There are multiple ways to view said bias, such as underrepresentation, inability to provide satisfactory recommendations, and stereotyping. Chen et al. [7] propose that two important types of result bias have been studied, namely *popularity bias* and *unfairness*. They define them as follows:

- **Popularity Bias**. Popular items are recommended even more frequently than their popularity would warrant.
- **Unfairness**. The system systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others.

There are various ways to categorize bias in recommender systems research. In this survey, we are interested in bias in the results and we build on Chen et al.'s categorization by generalizing it into *interaction-based* bias and *attribute-based* bias based on measurement strategy (see Section 5.1). Interaction-based bias relates to patterns within the user-item interaction data. It is based on properties like item popularity, user activity level, or rating distributions that can be observed directly from the interaction data without adding external information. For example, a system suffering from popularity bias recommends popular books disproportionately often. In contrast,

Just Accepted

attribute-based bias involves characteristics external to the interaction data, such as author gender or user age, which must be added separately to measure. For instance, a system might produce worse recommendations for younger users. Bias impacts multiple stakeholders; for example, a system that suffers from popularity bias might recommend popular items even to users who prefer niche items [1]. In this survey, we distinguish between *item* and *user* bias subjects.

2.3 Recommendation in the library

Libraries have always served as institutions dedicated to knowledge dissemination and cultural preservation, ensuring equal access to information through carefully structured classification systems, catalogs, and personalized librarian recommendations [58]. As digital technologies increasingly mediate the way users access information and knowledge, libraries have adapted by integrating information technology while upholding their core values of transparency, fairness, and inclusivity [67, 94]. The expansion of digital library collections and user-driven discovery tools has prompted technological interventions that balance traditional library functions with modern user expectations and needs [60].

One significant shift has been the transformation of book discovery processes. Users accustomed to commercial platforms like Goodreads now anticipate similar recommendation services from libraries [46]. Traditionally, recommendations in libraries relied on expert curation, classification frameworks, and reader advisory services, which helped users find materials suited to their interests, academic level, or thematic preferences [64]. However, these methods were inherently constrained by collection size and human effort. The introduction of Online Public Access Catalogs (OPACs) improved book retrieval through structured metadata but lacked the ability to offer personalized suggestions. AI-driven recommender systems showed the possibility to transform the process into a more dynamic and user-centric experience [74]. Libraries contain extensive user data, maintaining detailed records of patrons, including demographic information and reading habits which can collectively amount to millions of records [52, 59]. Because of such extensive and rich data they are in a unique position to develop personalized recommendation systems. The implementation of such systems in libraries has several advantages, top most among them being improved accessibility. Personalized book discovery can help users navigate extensive collections more efficiently and engage with materials that align with their interests [47]. Librarians also benefit from these systems, as they get assistance in curating personalized reading lists for patrons, streamlining the advisory process [68]. Moreover, automated recommendation models can facilitate the discovery of lesser-known works, ensuring that underrepresented authors and forgotten titles gain visibility within library collections.

Despite these advantages, such systems also raise concerns regarding bias, inclusivity, and fairness. Libraries, as institutions are fundamentally committed to equal access and being fair to all [4]. Hence they need to be especially mindful of the biases that can emerge when automating the recommendation process. One key issue is bias in training data. Many recommendation systems rely on a subset of a big dataset containing metadata, indexing, lending data, user ratings, and reviews, which may already reflect existing disparities [27]. A study on the Library of Congress subject headings found that, while online catalogs improved discoverability, subject headings still failed to adequately represent African American Studies resources [9]. If similar metadata and subject headings are used in training recommender systems, they may perpetuate the same biases, leading to underrepresentation or misrepresentation of certain works. Biases of this nature are not new to libraries. Collection bias arises when certain voices and perspectives are underrepresented in library holdings, often due to historical publishing trends and selection practices [27]. Cataloging bias manifests in the way books are classified and indexed, with traditional systems like the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) reflecting Eurocentric structures that can marginalize non-Western and diverse knowledge systems [38]. Bias can be implicit when it is unconsciously introduced by library staff during collection development and cataloging.

Just Accepted

This bias stems from their own cultural and social influences, which can shape the materials they select and how they categorize them, even without intentional prejudice [70].

A real-world example of the risks associated with biased AI-driven book recommendation systems can be seen in the 2024 Fable AI controversy [20, 24]. Fable, a reading, tracking, and social media app similar to Goodreads, deployed an OpenAI software to generate personalized end-of-year reading summaries for users. However, the AI exhibited significant racial and gender biases, producing offensive and exclusionary remarks such as:
"Your journey dives deep into the heart of Black narratives and transformative tales, leaving mainstream stories gasping for air. Don't forget to surface for the occasional white author, okay?"

Given that Fable has been integrated into multiple public library memberships in the United States[1], this incident underscores the potential risks of unchecked bias in library book recommendation systems. It highlights how biases in training data and algorithmic decision-making can lead to stereotyping and exclusion, reinforcing dominant literary trends rather than supporting diverse narratives. To avoid similar pitfalls it is important to consider not only technical adjustments but also the documented biases and non-inclusive practices. Fortunately, libraries have studied these biases and challenges extensively: for example, Warburton et al. [79] discussed the structural biases in the traditional hierarchical library classification systems. Cirasella [8] identified four structural barriers that inhibit equitable representation in collections, while Olson [50] has examined how naming in indexing systems can marginalise certain identities and worldviews. To address these issues, libraries have developed structured frameworks and tools to evaluate and reduce bias. Resources such as the Anti-racist Library Collection Building Workbook by the University of Colorado Boulder Libraries [72] provide guidelines to help professionals assess their collections for inclusivity. Such a framework can be directly applied to enhance the fairness of recommendation algorithms. For instance, questions such as *"Are selection filters potentially eliminating materials that don't conform to a white scholarly worldview?"* can be adapted to identify and address biases in AI systems. Various frameworks of enacting inclusivity and justice within library services have been proposed by LIS scholars [28, 37, 39, 86].

Even though the LIS research provides valuable insights into the challenges of delivering inclusive book recommendations [64] and can highlight the potential risks [10], the intersection between LIS and computer science research appears narrow. We hypothesize that this is due to difference in terminologies used to describe similar concepts. For example, LIS researchers may use terms like "injustice, "inequality", "disparity" to talk about "bias", or "library discovery," "information seeking," to describe "recommender systems". We used LIS-specific terminology in our searches such as library discovery, discovery system, cataloging bias, and AI in libraries, but the results focused on broad or conceptual treatments of AI rather than studies isolating recommendation bias as a research topic. Our searches showed that very little of this scholarship focuses specifically on bias in book recommender systems. Even when AI or algorithmic tools are mentioned, recommender systems typically appear only as one illustrative case of technological change, rather than as a primary subject of study. This reflects a disciplinary difference in scope and vocabulary and for this reason we incorporated LIS scholarship through a targeted, unstructured synthesis, that highlights its conceptual foundations and demonstrate how these insights can inform interdisciplinary work. In Section 9.3, we provide concrete recommendations towards interdisciplinarity in bias in book recommendation research.

3 Surveys in recommender systems

In this section, we discuss papers that either survey work on book recommender systems, or on bias in recommender systems, either generally or in a media-related domain. We focus on how they conduct paper collection and analysis, when explicitly described, as well as their main findings and conclusions.

[1]https://smcl.org/blogs/post/join-us-on-fable-a-book-club-app-for-social-reading/ (accessed 26 March 2025)

Just Accepted

Two papers conduct surveys on book recommender systems more broadly. Alharthi et al. [3] survey over 30 published papers that address book recommendation. They classify book recommender systems into 6 categories based on which book features they take into account. They list the following challenges and future research for the field in general: recommending textual items, the effect of mood, recommending books to non-readers, the long-tail problem, recommending audiobooks, assuming a reading model when recommending books, the confusion that translated books might cause, and others. Anwar et al. [5] specifically focus on machine learning-based book recommendation. They survey several databases such as Google Scholar and Scopus by querying terms like 'book recommender systems'. Out of the resulting 100 papers, they go through abstracts and conclusions to choose those where machine learning is used. They classify these papers into the following 6 categories based on the type of machine learning that is used: classification-based, KNN-based, SVM-based, neural network-based, association rule mining-based, and clustering-based. They trace the following future research perspectives: user privacy, incorporation of contextual information such as mood, and exploration of additional machine learning techniques. These surveys provide valuable insight into automated book recommendation, but do not examine issues of bias that arise from the process.

Topics such as bias and fairness have been the focus of recent surveys in recommender systems. Chen et al. [7] extensively review the topic of bias and debias in recommender systems, motivated by the field's fragmentation and the inconsistent ways in which researchers use the term 'bias' and its subcategories. They survey over 180 papers between 2010 and 2021 by querying top-tier recommender systems venues with keywords related to bias and recommendation, and traversing the citation graphs afterwards. They classify the papers into the following categories: data bias, which is present in the data, model bias, which is induced by the model itself, and result bias, which is present in the recommendation. They further break down the category of data bias into selection, exposure, conformity, and position bias, and result bias into popularity bias and unfairness, and present common debiasing methods for each subcategory. Finally, they discuss open issues relevant for future work, such as issues with evaluation, the fact that bias is not always a negative phenomenon, and trade-offs between accuracy and fairness in recommendation. Wang et al. [77] are prompted by the scatteredness of research on fairness in recommendation. They survey more than 60 papers on fairness in recommender systems published between 2017 to 2022 at top-tier information retrieval venues and group fairness definitions into process and outcome fairness, which they further group based on the target and concept. They also classify fairness issues based on the following concepts: subjects (user, item, joint), granularity (single, amortized), and optimization object (treatment, impact). They proceed to list various ways that researchers measure and implement fairness, as well as datasets commonly used in literature (they do not mention any book datasets). Finally, they pinpoint future directions for fairness research which they group into 4 pillars: definition, evaluation, algorithm design, and explanation. We are inspired by the frameworks constructed by these surveys to track the different kinds of bias and bias mitigation. Our added contribution is the emphasis on book datasets, as well as issues that specifically concern the book domain.

As we mentioned already, to our knowledge no survey on bias in book recommender systems has been published. However, there are papers that review bias in a different subdomain of media recommendation, namely music. Ospitia-Medina et al. [51] review literature around music recommendation to summarize how and why bias occurs in such systems, as well as put forward relevant guidelines. They perform two separate queries: one over the whole literature for mostly recent papers that are about (bias in) music recommender systems, and one for papers after 2022 that relate to musical datasets. They discuss recommendation techniques that have been used in music recommendation, and then zoom in on bias for which they collected 9 papers and which they categorize as preexisting, technical, and emergent in accordance with Friedman and Nissenbaum [22]. For each category of bias, they provide a list of guidelines on how it should be handled. In a mini review, Dinnissen and Bauer [16] adopt a stakeholder-centered perspective when surveying literature that addresses fairness in music recommender systems by classifying it based on the relevant stakeholder: user, provider, or multi-stakeholder.

Just Accepted

They also make a note of whether each paper was improvement-focused, what kind of methodology they used, what was the topic, which fairness attributes they considered, and which dataset they employed. As main findings, they highlight the abundance of work that only analyzes fairness instead of proposing improvements, the focus on gender bias despite a limiting binary view, the relevance of popularity bias, as well as the overuse of datasets that stem from Last.fm. Similarly to these surveys, we focus on a domain and distill the unique gaps and challenges.

4 Paper selection and analysis methodology

In this section, we describe the process followed for paper selection, as well as the difficulties we faced and the strategies that we eventually abandoned. Additionally, we present the dimensions that we defined in order to efficiently analyze the selected papers.

The field of recommender systems has a large body of work. A general search on Google Scholar for the term 'bias in book recommendation' yields more than one million results, the majority of which are irrelevant to our topic of interest; the terms 'book', 'bias' and 'recommend' are common enough due to polysemy and can therefore appear in many different scientific contexts. Our first attempt at addressing this issue was by restricting the terms 'book', 'bias' and 'recommend' (and their variants) to appear in the title of the paper when querying Google Scholar and Scopus, which resulted in 252 documents. However, after careful consideration of the results, we only came up with 5 papers; the rest were either duplicates, not retrievable, not related to the topic of bias in book recommendation, not a full scientific paper but rather a poster or slide deck, etc.

In order to appropriately (but not overly) restrict the literature space, as well as obtain high-quality research, we looked into a set of reputable registries that often publish research on recommender systems. Between October 9th and November 7th of 2024, we queried the following sources: the journals *Information Processing & Management* (Elsevier), *Transactions on Knowledge and Data Engineering* (IEEE), *User Modeling and User-Adapted Interaction* (Springer), and *Data Mining and Knowledge Discovery* (Springer), the proceedings of the *International Workshop on Algorithmic Bias in Search and Recommendation* (Springer) and the *European Conference on Information Retrieval* (Springer), and the entire *ACM digital library*. To target research into bias in book recommendation systems, we looked for papers that contained the terms 'bias' and 'recommender' as well as mentioned at least one of the following datasets with user-book interactions: Goodreads, Book-Crossing, Amazon-books, and LibraryThing. To our knowledge, most datasets with user-book interactions come from these web sources. We were inspired by PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [53] to follow a structured approach for the paper retrieval documentation. See figure 1 for how we reached the final number of 40 studies included in this survey.

The search on the selected sources resulted in 357 papers, which we combined with the 5 papers from Google Scholar and Scopus, and removed 2 duplicates. We then screened the papers by reading the title and abstract, as well as noting in what context they used each of the queried terms (i.e., 'bias' and 'recommender'). Accordingly, we discarded papers for being entirely out of scope (e.g., not being at all about bias). We could retrieve all 54 remaining papers. After carefully reading through them, we discarded 14 for the following reasons:

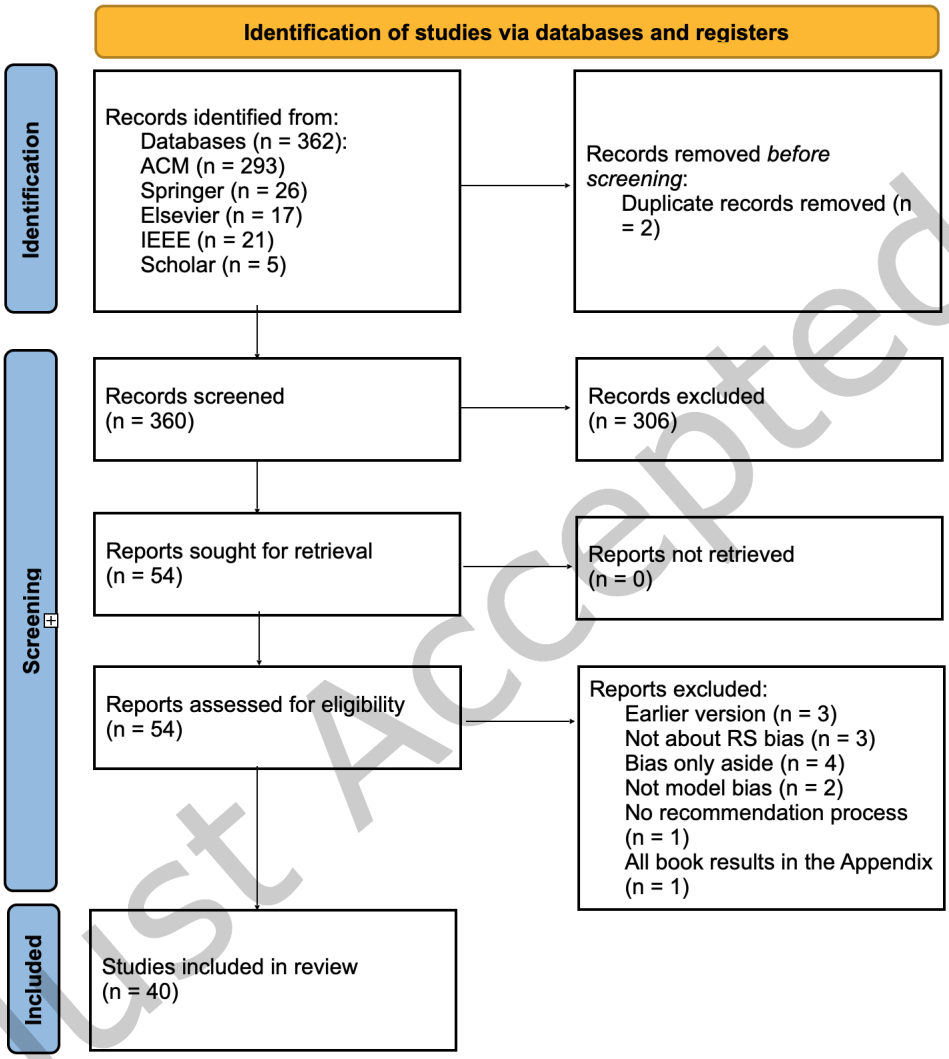
- 3 are earlier versions of papers also in the set (i.e., conference papers prior to becoming journal papers)
- 3 are not about bias in recommender systems
- 4 mention bias as a topic only tangentially related to their research
- 2 are not about model bias
- 1 is only tangentially related to recommender systems
- 1 includes all book dataset-related results only in the Appendix

Our final set includes 40 papers. To study these papers, we annotated them along the following set of dimensions:

- **Basic**: Title, Author(s), Link, Year of publication, Venue, Publisher

Just Accepted

Fig. 1. PRISMA 2020 flow diagram for systematic reviews showing how an initial collection of 362 studies were reduced to the final 40 studies included in this survey.



Just Accepted

- **Bias**: Problem statement, Bias definition, Bias subject(s) (User and/or Item), Bias type(s) (Interaction-based or Attribute-based), Acknowledgment of book domain
- **Data**: Dataset(s) used, Dataset preprocessing, Type(s) of feedback (Implicit or Explicit)
- **Algorithmic**: Algorithm(s) used, Prediction task(s) (Rating or Ranking), Framework (e.g., Python library or Original implementation), Hyperparameter tuning, Code availability
- **Methodology**: Action (Measuring or Mitigating bias), Method, Metric(s)
- **Results**: Result(s) on the book dataset

Since the *bias* dimension is the main topic of this survey, we will further explain its subcategories. *Problem statement* refers to the main issue that each paper attempts to address. *Bias definition* corresponds to whether the authors explicitly define (with their own words or through a citation) what bias means in the given system and what the definition is. *Bias subject* is equivalent to the stakeholder that is subject to the bias in question, which could be the user, the book, or more specifically the author, publisher, topic etc of the book. *Bias type* is the kind of bias that the paper addresses, either *Interaction-based bias* or *Attribute-based bias*. Additionally, for this dimension we keep track of which interaction or external aspect is of relevance, for example, book popularity, user activity (interaction-based bias), author gender, book topic (attribute-based bias) etc. Finally, *Acknowledgment of book domain* is used to note whether a given paper engages with the fact that the items at hand are books, either passively or actively.

The repository where we kept track of the collection and analysis process, including the final table with all papers and dimensions filled, has been made publicly available[2].

5 A framework for bias in book recommendation

This section presents a structured overview of the 40 papers included in this survey. Table 1 lists all the papers along with their main features. The majority of computer science research on bias in book recommender systems is within the last few years, and the papers we included are scattered across conferences and journals. Most papers study interaction-based bias, especially with item as a subject, and present a mitigation strategy. Figures 2, 3 and 4 show basic statistics about the papers included in the survey and can be found in the Appendix.

To systematize our analysis and identify research gaps, we organize bias in book recommender systems research across three of the dimensions listed in the previous section:

(1) **Bias type**. What kind of bias does the paper address?
(2) **Bias subject**. Which stakeholder suffers from said bias?
(3) **Action**. How does the paper engage with said bias?

In table 2, we divide the papers into categories across the dimensions of this framework.

5.1 Bias type

We categorize bias in book recommendation into **interaction-based** and **attribute-based** bias based on the type of feature it targets.

5.1.1 ***Interaction-based bias*** occurs when a recommendation algorithm is found to systematically produce unfair outcomes based solely on measurable properties of the user-item interaction matrix such as item popularity, user activity level, or ratings distribution. The key characteristic is that these properties are inherent to or directly derived from the interaction data and hence the presence and degree of bias can be assessed without adding any external labels or attributes.

This category is an extension of the *popularity bias* category recognized by Chen et al. [7]. While popularity is a crucial characteristic of the interactions, we found that in literature other interaction-related biases have also been studied. Accordingly, we propose that four types of interaction-based bias have so far been addressed in book recommendation: **popularity bias**, **blockbuster bias**, **activity level bias** and **popularity affinity bias**.[3]

Popularity bias is often defined as the phenomenon where items are recommended more frequently than their popularity would warrant [7]. Klimashevskaia et al. [30] propose that a system suffers from popularity bias when "the recommendations provided focus on popular items to the extent that they limit the value of the system

[2]https://github.com/SavvinaDaniil/BiasInBookRecommendationSurvey

[3]Note that these are the types of bias that have been studied so far in book recommender systems and not an exhaustive list of all potential biases.

Just Accepted

Paper	Bias Type		Bias Subject			Action		Venue	Year
	Interaction-based	Attribute-based	User	Item	Joint	Measurement	Mitigation		
[63]	☒			☒		☒		BIAS@ECIR	2025
[49]	☒				☒		☒	IPM	2024
[57]		☒		☒			☒	ECIR	2024
[82]	☒				☒		☒	RecSys	2024
[14]	☒				☒	☒		TORS	2024
[78]	☒			☒			☒	TOIS	2024
[55]	☒				☒		☒	TORS	2024
[6]	☒			☒			☒	RecSys	2024
[85]	☒		☒				☒	WWW	2024
[84]	☒		☒				☒	TOIS	2024
[18]		☒		☒			☒	UMAP	2024
[35]		☒	☒				☒	TKDD	2024
[91]		☒	☒				☒	WWW	2024
[66]	☒			☒			☒	Computer Science	2024
[65]		☒		☒			☒	JIIS	2024
[43]	☒				☒		☒	UMUAI	2023
[88]	☒		☒			☒		BIAS@ECIR	2023
[36]	☒				☒	☒		RecSys	2023
[80]	☒			☒			☒	SIGKDD	2023
[90]	☒			☒			☒	SIGKDD	2023
[26]		☒		☒			☒	IPM	2022
[34]	☒			☒			☒	TKDE	2022
[83]	☒			☒		☒		DMDK	2022
[45]	☒				☒	☒		BIAS@ECIR	2022
[31]	☒				☒	☒		BIAS@ECIR	2022
[61]	☒			☒			☒	RecSys	2022
[81]		☒	☒				☒	SIGKDD	2022
[54]	☒		☒				☒	SIGIR	2022
[33]		☒	☒				☒	TIST	2022
[69]	☒			☒			☒	WI-IAT	2022
[56]		☒		☒		☒		SIGIR	2022
[13]		☒		☒		☒		arXiv	2022
[42]	☒	☒	☒			☒		IPM	2021
[15]		☒	☒			☒		IPM	2021
[17]		☒		☒			☒	UMUAI	2021
[23]	☒			☒			☒	SAC	2021
[71]	☒				☒		☒	WSDM	2021
[89]	☒			☒			☒	WWW	2021
[29]		☒		☒		☒		WWW	2021
[44]	☒		☒			☒		TIST	2016

Table 1. Paper lookup table.

Just Accepted

Bias Dimensions		Related Work
Divided by Type	Interaction-based	[6, 14, 23, 31, 34, 36, 43–45, 49, 54, 55, 61, 63, 66, 69, 71, 78, 80, 82–85, 88–90]
	Attribute-based	[13, 15, 17, 18, 26, 29, 33, 35, 42, 56, 57, 65, 81, 91]
Divided by Subject	Item	[6, 13, 17, 18, 23, 26, 29, 34, 56, 57, 61, 63, 65, 66, 69, 78, 80, 83, 89, 90]
	User	[15, 33, 35, 42, 44, 54, 81, 84, 85, 88, 91]
	Joint	[14, 31, 36, 43, 45, 49, 55, 71, 82]
Divided by Action	Measurement	[13–15, 29, 31, 36, 42, 44, 45, 56, 63, 83, 88]
	Mitigation	[6, 17, 18, 23, 26, 33–35, 43, 49, 54, 55, 57, 61, 65, 66, 69, 71, 78, 80–82, 84, 85, 89–91]

Table 2. Lookup table of bias dimensions.

Just Accepted

or create harm for some of the involved stakeholders". These definitions are intentionally broad and encompass multiple views.

In book recommender systems research, a system is considered biased in favour of popular items in three different cases:

(1) *If an item's recommendation frequency depends on its popularity* [14, 31, 45, 61, 69]. This is the most direct manifestation: popular books appear in recommendations disproportionately often. When training on historical data, an algorithm may learn that popular books are "safe" choices that many users have interacted with. It therefore optimizes for these high-frequency items and neglects low-frequency ones.
(2) *If an item's recommendation accuracy depends on its popularity* [6, 23, 36, 43, 49, 80, 90]. This occurs when an algorithm performs better (in accuracy, recall, etc.) for popular items than for unpopular items. Popular books have more training examples, so a model might learn them better than unpopular books that have sparse data.
(3) *Based on other performance metrics that are considered sensitive to popularity bias* [14, 49, 69, 71, 78, 82]. The interpretation is that popularity bias reduces recommendation diversity, novelty, and coverage thus harming the overall system functionality.

Blockbuster bias is a specialized form of popularity bias proposed by Yalcin [83] that favours both popular *and* highly rated items. While characterizing an item as popular only considers how many interactions it has received, blockbuster status extends to how universally liked those interactions are. A book with 100 five-star ratings is a "blockbuster", while one with 100 mixed ratings is "popular" but not blockbuster.

In book recommender systems research, a system is considered biased in favour of blockbuster items in one primary case:

(1) *If an item's recommendation frequency depends on its blockbuster status* [83]. In practice, a model might learn that frequently and highly rated items are "safe" and optimizes to disproportionately recommended them to users.

Popularity affinity bias is the phenomenon where users are treated unfairly based on their affinity for popular items. It captures a discrepancy between users that is common in many domains including books: some users prefer mainstream items, while others tend to consume more niche material. Bias occurs when a model performs differently for users with different propensities for popular items.

In book recommender systems research, a system is considered biased against users with affinity for niche items in two different cases:

(1) *If a user's recommendations consist of items more popular than their preferences* [14, 45, 54]. Because of popularity bias, popular books might be globally recommended even to users who tend to prefer niche material. This misalignment harms their system experience.
(2) *If a user's recommendation accuracy depends on their popularity affinity* [14, 31, 49, 54, 88]. When a system more accurately recommends popular books to users than niche books, it follows that users with mainstream tastes receive better recommendations.

Activity level bias occurs when users are treated unfairly based on their activity level, i.e., the number of interactions they have in the training data. It is a known problem that recommender systems tend to underperform for *cold start* users due to lack of sufficient information on their preferences.

In book recommender systems research, a system is considered biased against users with affinity for niche items in three different cases:

(1) *If a user's recommendation accuracy depends on their activity level* [36, 43, 44, 54, 71, 84, 85]. Active users have rich histories which allows the model to better personalize their recommendations, while inactive users might get default generic recommendations.
(2) *If a user's recommendations consist of items more popular than their preferences* [54]. In practice, this occurs when a model falls back to recommending "safe" popular items to all sparse users, even those who tend to prefer niche items.
(3) *If system fairness is indirectly impacted* [55, 82].

5.1.2 ***Attribute-based bias*** occurs when a recommendation algorithm systematically produces unfair outcomes based on attributes of items or users such as author gender, author nationality, or user age, that must be enriched from external sources. The key characteristic is that these attributes exist outside the interaction matrix and require additional data to measure.

Both interaction-based and attribute-based biases emerge from the same mechanism: algorithms learn patterns from interaction data. The distinction is what we measure, not how the bias arises. In the case of protective or sensitive attributes, this alludes to what Mehrabi et al. [41] refer to as *indirect discrimination* which occurs when "protected groups, or individuals, still get to be treated unjustly as a result of implicit effects from their protected attributes". When a collaborative filtering system learns "popular items are good", it doesn't explicitly know author nationality labels. Yet if popular items have largely been authored by a given nationality, nationality bias emerges when we measure it. The bias mechanism is the same, but the measurement lens differs.

This category is an extension of the *unfairness* category recognized by Chen et al. [7]. While in practice most relevant bias is targeted towards human attributes, we keep it more general to account for other types of attributes. We propose that five types of attribute-based bias have so far been addressed in book recommendation: **author nationality bias**, **author gender bias**, **publisher location bias**, **user age bias** and **user gender bias**.[3]

Author nationality bias occurs when authors of certain nationalities are discriminated against in the recommendation process. This bias typically arises when popular items in datasets are predominantly written by authors from the country where these datasets stem from. When algorithms learn popularity patterns, they incidentally learn to over-recommend these authors.

A system is considered biased against certain author nationalities in one primary case:

(1) *If authors of a particular nationality are under-represented in recommendations compared to their presence in the training data* [13]. For example, if 30% of books in the dataset are by non-US authors but only 15% of recommendations feature non-US authors, the system exhibits author nationality bias.

Author gender bias can similarly be defined as discrimination against an author's gender in the recommendation process.

A system is considered biased against certain author genders in two different cases:

Just Accepted

(1) *If authors of a particular gender are under-represented in recommendations relative to the training data* [17, 29, 56, 57]. For example, if a user has rated books by both male and female authors equally, but the recommendations skew heavily toward male authors, this indicates bias against female authors.
(2) *If authors of a particular gender are under-represented in recommendations relative to other gender groups* [17, 18, 56]. This captures discrimination at the system level: the algorithm systematically favors one gender's books over another's.

Publisher location bias occurs when the recommendation system discriminates based on where a book was published or the publisher's geographic location. Publishers from certain regions or continents receive disproportionate visibility or lack thereof.

A system is considered biased against certain publisher locations in one primary case:

(1) *If books from publishers in certain geographic groups are under-represented relative to other groups* [26]. This expresses the extent to which publishers from a given region dominate historic interactions which propagates in the recommendations.

User age bias occurs when the recommendation system provides systematically worse service to users of certain ages.

In book recommender systems research, a system is considered biased against certain user age groups in three different cases:

(1) *If recommendation performance varies across user age groups* [15, 35, 81, 91]. Some age groups might receive better recommendations than others, suggesting the system is optimized for certain demographics.
(2) *If the recommendation outcomes differ depending on user age* [35]. For example, an algorithm might recommend more literary fiction to older users and more young adult fiction to younger users, regardless of what they actually want which is akin to *stereotyping*.
(3) *If certain age groups' preferences dominate the overall recommendation output* [33]. In practice, this might stem from the system implicitly optimizing for the most common user demographic while neglecting others.

User gender bias occurs when the recommendation system underserves users based on their gender.

In book recommender systems research, a system is considered biased against certain user genders in two different cases:

(1) *If recommendation accuracy varies across user gender groups* [15]. Some genders receive higher quality recommendations than others, indicating optimization for a particular demographic.
(2) *If certain user genders' preferences dominate overall recommendations* [33].

The categorization between interaction-based and attribute-based bias is very relevant for book recommendation. On the one hand, as discussed in Section 2.3, practices that are discriminatory towards users or authors based on their demographic characteristics are unwanted in book recommendation. On the other hand, such information is not generally publicly available, which often limits research to studying interaction properties as they are inherent to the datasets. As figure 4b shows, most papers study interaction-based bias, which can be attributed to the aforementioned limitation.

5.2 Bias subject

In line with other research on recommender systems values [77], we additionally categorize bias in book recommendation into **user**, **item**, and **joint** based on the subject it targets.

- **User bias** targets the users of a book recommender system, which corresponds to the book consumers, loaners, or cataloguers depending on the source of the dataset.

Just Accepted

- **Item bias** targets the books themselves, which corresponds to the authors, topics, characters, publishers, or other book properties.
- **Joint bias** targets the users and books simultaneously.

As seen in table 2, the majority of papers deal with item bias. This result is in line with Wang et al. [77] who found that most papers on fairness in recommendation approach the topic from an item perspective.

5.3 Action

We finally categorize bias research based on the goal into **measurement** and **mitigation**.

- **Measurement** includes studies that mainly measure bias, often by proposing a new metric or view of bias.
- **Mitigation** includes studies that, besides measuring a type of bias, additionally experiment with mitigation methods against it.

6 Findings on interaction-based bias

In this section, we discuss the findings of papers that examine interaction-based bias in book recommendation, either from a user or item perspective. Table 3 includes a summary of the different types of interaction-based bias along with their different views and metrics.

Bias Type	View / Description	Metrics
	Item Biases	
Popularity bias	Recommendation frequency is influenced by item popularity [14, 31, 45, 61, 69].	correlation [14, 31, 45], PopQ@1 [61], LT frequency [69], LT coverage [69]
	Recommendation performance varies across items with different popularity levels [6, 23, 36, 43, 49, 80, 89, 90].	recall per group [49], estimation error per group [43], LT HR [6, 80, 89, 90], LT NDCG [6, 80, 89, 90], LT precision [23], LT recall [23]
	Indirect impact of popularity bias on recommendation outcomes [14, 49, 69, 71, 78, 82]	coverage [14, 69, 71], novelty [49], NDCG [78], MRR [78], recall [78], P-fairness [82]
Blockbuster bias	Recommendation frequency is influenced by an item's blockbuster status [83].	frequency per group [83]
	User Biases	
Activity level bias	Recommendation performance varies across user groups with different activity levels [36, 43, 44, 54, 71, 84, 85].	estimation error per group [43], prediction difficulty per group [36], NDCG difference per group [54, 84], recall difference per group [44, 71, 84], coverage per group [44], IA MRR [85], IA HR [85], IA NDCG [85]
	Popularity of recommended items varies with user activity level [54].	$\%\Delta GAP$ [54]
	Indirect impact of activity level bias on recommendation outcomes [82].	C-fairness [82]
Popularity affinity bias	Recommendation performance varies across users with different popularity affinity [14, 31, 49, 54, 88].	recall per group [49], MAE per group [31], NDCG per group [14, 54, 88]
	Popularity of recommended items varies with users' popularity affinity [14, 45, 54]	$\%\Delta GAP$ [14, 45, 54]

Table 3. Interaction-based bias views and metrics grouped into item/user categories. 'LT' means long-tail (items) and 'IA' inactive (users).

Just Accepted

6.1 Item

6.1.1 Popularity bias.

Measurement. A common metric of popularity bias based on the frequency of recommendation view is the *correlation between popularity in profile and in recommendation*. Naghiaei et al. [45] measure popularity bias as a result of a set of collaborative filtering algorithms trained on a subset of Book-Crossing. They visually evaluate said correlation and find that certain algorithms do propagate popularity bias and others do not. On the contrary, Kowald and Lacic [31] on the same dataset and using some of the same algorithms find that all algorithms propagate popularity bias using the same visual evaluation. Daniil et al. [14] notice that different studies report different results when measuring popularity bias in a set of datasets from different domains, including Book-Crossing. They reproduce the studies and find that the divergence in findings is a result of many factors, such as evaluation strategy, that lead to different conclusions even based on the same dataset. Regardless, the characteristics of the domain play a role: compared to datasets from the other domains, Book-Crossing is more prone to popularity bias, according to correlation. Another relevant metric is the *coverage and recommendation frequency of long-tail items* [69], as well as *PopQ@1* which expresses the average popularity quantile of the first recommended item of each user [61]. Other metrics attempt to capture the impact of item popularity on recommendation performance. Liu et al. [36] borrow item response theory (*IRT*) from psychometric assessment practices to estimate whether traditional rank-based metrics reflect an algorithm's ability to model user preferences and assess the relationship between *item difficulty* and popularity. Unlike the other datasets, for the book dataset, they do not find any relationship between item difficulty and popularity, which they attribute to sparsity. Some studies divide items into groups based on their popularity and calculate *accuracy metrics for each group*, such as recall [49] and estimation error [43], while others especially highlight the system's *accuracy on the long-tail items*, with various accuracy metrics having been used [6, 23, 80, 89, 90]. Finally, many studies (additionally) indirectly measure popularity bias by assuming it has a negative impact on *novelty* [49], *coverage* [14, 69, 71], or *accuracy* in general [78].

Mitigation. A relatively large number of papers propose methods for mitigating popularity bias in book recommendation. Some studies focus on *increasing the recommendation frequency of long-tail items*. Rhee et al. [61] introduce a method called *Zerosum* that reduces popularity bias by regularizing the predicted scores to be equal across items a user prefers via an extension of BPR loss with a regularizer. For the book dataset, their method performs better (higher accuracy and lower PopQ@1) for most algorithms than other debiasing methods. Tang and Zhang [69] emphasize that it's not sufficient to recommend long-tail items, but also that the users don't receive the same recommendations again and again; *the long-tail items recommended should be sufficiently diverse*. They present *CADDP*, a long-tail item recommendation approach that extracts important feature embeddings to capture the common characteristics of multiple items clicked by users and ensures that they are relevant and diverse. *CADDP* achieves the best performance among the baselines in terms of presence and diversity of long-tail books in the recommendation.

Other studies attempt to *improve the system's performance on long-tail books*. Okamura et al. [49] note that dealing with popularity bias is a complicated task because popularity might imply quality, and hence it might not be desirable to mitigate it for all users. Based on some observations around the properties of the inner product model, they develop a method called *DirectMag* that facilitates both platforms and users to manipulate popularity bias post-model training with the help of a parameter γ. They find that their method generally exhibits high average novelty and accuracy. Additionally, it creates a trade-off between novelty and recall per popularity-level item group on the one side and accuracy on the other that the γ parameter helps deal with, though the trade-off is not consistent in the book dataset. Gabbolini et al. [23] offer the following observation: matrix factorization embeddings are very unstable, but especially so for long-tail items. They develop a variation of matrix factorization

Just Accepted

that mitigates instability issues and achieves better accuracy in the long-tail compared to the default version of MF and some other algorithms. Balasubramanian et al. [6] perform 'biased' user history synthesis: they augment the user profiles with a sampling strategy that is biased in favor of long-tail items and three different synthesis models. With exceptions, they find overall improvement of accuracy in the long-tail and in general. Zhang et al. [90] claim that despite the large amount of research around improving long-tail item recommendation, transferring these methods to industry practice is challenging due to the dive in performance that most of them introduce, as well as their complexity and high cost. They design Cross Decoupling Network (*CDN*) that reduces the difference between training and service data for both item distribution and users' preference given an item. *CDN* achieves significant performance on long-tail items compared to the baselines, while also maintaining good performance for all items. Zhang et al. [89] present a dual transfer learning framework that jointly learns knowledge transfer from both model-level and item-level (i.e. from head to long-tail items), called *MIRec*. Based on performance metrics, *MIRec* achieves the best performance on long-tail books. Wei et al. [80] focus on graph-based recommendation and present a Meta Graph Learning (*MGL*) framework for long-tail recommendation that successfully employs an auxiliary graph of item relations. *MGL* outperforms other methods in all performance metrics for long-tail books (including previously mentioned *MIRec*). Misztal-Radecka and Indurkhya [43] develop Bias Detection Tree (*BDT*), a model-agnostic technique that automatically detects combination of user-item attributes correlated with unequal treatment. They also propose a model selection technique that improves performance for biased attributes. For the book dataset, while several external features are considered, the most significant item bias is associated with low item popularity.

Finally, some studies *deal with popularity bias indirectly*. Wang et al. [78] develop *GBD*, a generative framework for bias disentanglement that constantly generates calibration perturbations during training to keep intermediate representations from being biased. GBD can encompass various types of bias. Among those considered, on the book dataset they only test popularity bias by employing a semi-synthetic version of the dataset that includes an unbiased subset. They find improvements over performance metrics with *GBD* on the unbiased subset, therefore reducing popularity bias. All other results are not on the book dataset. Togashi et al. [71] note that because the cold start problem in recommender systems is often solved by selecting unobserved interaction samples and assuming them to give a negative signal, this can lead to popularity bias. They propose *KGPL*, a method that leverages unobserved samples by using a knowledge graph (KG) with side information about the books and accordingly generates positive instances from the unobserved samples. They find that *KGPL* exhibits low coverage when trained on the book dataset, presumably due to the popularity bias present in the dataset. Rahmani et al. [55] notice that fairness-aware methods in recommender systems often address user and provider fairness separately, even though it's a two-sided marketplace. They propose *CP-FairRank*, a re-ranking algorithm that seamlessly integrates fairness constraints from both sides, with a focus on popularity from the item side. Even though not many results on the book dataset are included, their method performs well based on a two-sided fairness metric. On the topic of two-sided fairness, Xu et al. [82] propose *Ada2Fair*, a framework that extends the accuracy objective into a controllable loss over the interaction data, which boosts the exposure of niche book publishers. They find that their method achieves Pareto optimality over the baselines considering both accuracy and P-fairness.

6.1.2 Blockbuster bias. Yalcin [83] measure whether a set of ranking-based algorithms exhibit blockbuster bias and produce extensive and detailed results. They find that there is blockbuster bias in the book dataset itself, which is also distinct from popularity bias (i.e., the well-likedness is impactful). When it comes to the recommendations, they find that the algorithms tend to exhibit blockbuster bias when trained on the book dataset, which is not only caused by item popularity.

6.1.3 Other. We found three papers that differ from the rest and are insightful. Shen and Jiang [66] concern themselves with the fact that users tend to give extreme ratings when evaluating books that don't reflect book

Just Accepted

quality (a property that the authors consider intrinsic to a book), which according to research is especially a problem in children's books where the ratings are often done by the parents. They propose a hybrid model that combines a graph convolutional network and neural matrix factorization to accurately capture that bias. They consider ratings from top reviewers as unbiased standard and calculate rating and ranking metrics between that standard and their method's ability to calculate unbiased ratings. Their method yields the lowest MSE and rank error, as well as the best NDCG when a recommendation process is applied, compared to other methods. Rystrøm [63] attempt to empirically test the theory that companies have incentives to influence users so as to make them more predictable and use their recommender systems accordingly. They pose that studying a real-world system instead of performing a simulation is key to uncovering biases, as it allows validating theoretical claims about how recommender systems cause bias. They use the Barrier-to-Exit metric and a linear mixed-effects model to study the evolution of Amazon book ratings between 1998 and 2023. They find that it has become more difficult for users to change preferences, which they attribute to the system's effectiveness in enacting homogenization by potentially limiting exposure to diverse content. Lin et al. [34] discuss *Spiral of Silence*: the theory that ratings are missing in a biased way because users are more likely to submit ratings if they agree with the opinion climate. This leads to the majority opinion receiving growing popularity, while other opinions are pushed back. They verify the existence of the issue in Amazon-books, study its characteristics, and model the missing ratings in 4 different ways. They find monotonic trends for books with majority opinion, as well as that hardcore users feel more obligated to underrate a highly appreciated book than to save a criticized item. There was no recommendation process on the book dataset.

Synthesis. Most papers about interaction-based bias from an item perspective study popularity bias [6, 23, 31, 36, 43, 45, 49, 55, 61, 69, 71, 78, 80, 82, 89, 90]. The sparsity of the book datasets is noted to impact the result, especially in comparison with datasets from other domains [14, 36, 71, 83]. On a related note, some papers don't include all of the results on the book datasets in their main body, prioritizing datasets from other domains [34, 55, 61, 71, 78, 80, 89]. Most papers that mitigate popularity bias are explicitly concerned with maintaining good performance [49, 55, 61, 69, 78, 82], often with an additional emphasis on head items [6, 80, 89, 90]. Some papers specifically attempt to boost tail items by synthesizing user history that is biased in favor of tail items [6], transferring side information from knowledge graphs [71], or improving tail items' representations [23, 80, 89]. Another type of bias studied is blockbuster bias [83].

6.2 User

6.2.1 Popularity affinity bias.

Measurement. The metric used for assessing if the users' recommendations exceed the popularity of their profiles is $\%\Delta GAP$, which measures relative difference between average popularity in profile and in recommendation for each group of users [14, 45, 54]. The already discussed papers by Naghiaei et al. [45] and Daniil et al. [14], other than measuring general popularity bias, also measure the unfairness of it. Unfairness in this case refers to the fact that users with interest in niche books might get over-recommended popular books. They divide users into three groups, those with niche interests, those who prefer popular items, and everyone in between. Naghiaei et al. [45] find most algorithms to be unfair towards users with either niche or diverse tastes. In their reproduction work, Daniil et al. [14] also measure $\%\Delta GAP$ but don't present the results on the book dataset in the main body of the paper. To evaluate the extent to which the system performs better for mainstream users, studies also divide the users into group and measure the *accuracy per group*. Naghiaei et al. [45] find that all algorithms perform the best for the users who prefer popular items based on various performance metrics. Kowald and Lacic [31] measure MAE per user group and find the lowest recommendation accuracy for the niche group. Zhang et al. [88] extend a previous paper to also include user activity level in the consideration and then evaluate bias

Just Accepted

based on recommendation utility per user group. When it comes to the book dataset, popularity affinity seems to play a more critical role than user activity level since mainstream users tend to be less active but still receive better recommendations (based on NDCG), which they attribute to extreme sparsity.

Mitigation. In terms of mitigation, one paper evaluates *the effect of re-ranking* on $\%\Delta GAP$. Rahmani et al. [54] reproduce a user-oriented fairness study and do extensive experimentation to analyze its dependence on various aspects: domain, model, and user grouping. They find that the re-ranking method has different effect depending on whether the users were divided based on activity level or affinity for popular books. By comparing the datasets, they find that the higher number of average interactions per user can lead to better performance of the fairness model. The same paper also measures the difference of $NDCG$ between user groups to evaluate the divergence of performance for different popularity affinity levels. Okamura et al. [49], who proposed *DirectMag*, show that the γ parameter can be tweaked to address the needs of users that are diverse when it comes to their affinity for popular items according to *recall per user group*.

6.2.2 *Activity level bias.*

Measurement. Similarly to popularity affinity bias, $\%\Delta GAP$ is used to measure whether recommendation popularity exceeds profile popularity for a group of users, with the grouping in this case depending on activity level [54]. Other metrics are used to capture the impact of user activity level on recommendation performance. Most studies divide the users into groups based on their activity level and measure *performance per group* [36, 43, 44, 54, 71, 84], while others especially measure the system's *accuracy on the less active users* [85]. Moody and Glass [44] provide a framework for user-centered evaluation that divides users into quadrants based on 2 dimensions: their near-neighbor-based similarity and their activity level. They find the algorithm to perform the worst (in terms of recall) for users who have rated many books but are dissimilar to other users and best for users who have rated few books but have high similarity with other users. The other criteria (coverage and diversity) are not evaluated on the book dataset. The previously mentioned item response theory-inspired method developed by Liu et al. [36] is also used to assess the relationship between *user difficulty* and user activity level, but such a relationship is not found for the book dataset, which is attributed to sparsity. Yang et al. [84] claim that measuring fairness as similarity of average performance between groups is not sufficient because the variance might be high. Instead, they define a new type of fairness: *similarity between the performance distributions across different user groups*. Finally, some studies only note the indirect impact of inactive users on *C-fairness* [55, 82].

Mitigation. The aforementioned fairness reproduction study by Rahmani et al. [54] also estimates the impact of grouping the users based on activity level on a previously published re-ranking method's success in improving $\%\Delta GAP$, and the results differ from the popularity affinity grouping. They also measure the difference of $NDCG$ between user groups to evaluate the divergence of performance for different popularity affinity levels and find that the higher number of average interactions per user can lead to better performance of the fairness model. Yang et al. [84] pursue the goal of enacting distributional fairness via a generative adversarial training framework called *DistFair*, and theoretically and experimentally show its effectiveness. Their framework achieves the best fairness performance for active versus non-active users in most cases, while maintaining a small performance drop. Yang et al. [85] concern themselves with cross-domain recommendation (CDR) and pose that inconsistency in user activity leads to user embeddings that prioritize users with more frequent interactions. They propose *UCLR*, a user-aware contrastive learning framework for robust CDR, and report significant performance improvement for users with only one or two interactions in Amazon-books compared to the baselines. *KGPL*, the method developed by Togashi et al. [71] that uses a knowledge graph with side information about the books to generate unobserved positive instances, is found to perform better for cold start users than other KG-based methods. However, all KG methods struggle with the book dataset, which they attribute to high sparsity. The previously mentioned Bias Detection Tree developed by Misztal-Radecka and Indurkhya [43] detects activity level as the

Just Accepted

user feature that correlates the most with bias. Their model selection techniques also improve performance for less active users. Notably, these discrimination rules are different between the training and test sets. For instance, the least active users have the smallest training error and highest test error, which is probably an indication of overfitting in the training set. Rahmani et al. [55] evaluate *CP-FairRank* with user activity as the user-fairness concern, but there are limited results in the paper. Finally, *Ada2Fair* by Xu et al. [82] also boosts inactive users, which results in Pareto optimality over the baselines considering both accuracy and fairness.

Synthesis. All papers about interaction-based bias from a user perspective group the users either on (a) level of activity [44, 84, 85], (b) affinity for popular items [14, 31, 45, 49, 88], or (c) both [54]. Moody and Glass [44] additionally group the users based on near-neighbor-based similarity. Bias towards users with niche tastes is sometimes characterized as unfairness [14, 31, 45]. To measure user bias, some papers use the $\%\Delta GAP$ metric that expresses an increase in average popularity [14, 45, 54], while others compare the performance across user groups [31, 44, 49, 54, 85, 88]. Yang et al. [84] extend the notion of across-group fairness to account for differences in distribution.

6.3 Joint

Some of the papers discussed measure or mitigate interaction-based bias from a user and item perspective simultaneously. They either tackle popularity bias and activity level bias [36, 43, 71, 82], popularity bias and popularity affinity bias [14, 31, 45, 49], or both [55].

7 Findings on attribute-based bias

In this section, we discuss papers that concern themselves with attribute-based bias in book recommendation, either from a user or item perspective. Table 4 includes a summary of the different types of attribute-based bias along with their different views and metrics.

7.1 Item

7.1.1 Author nationality bias. This type of bias has only been studied by Daniil et al. [13], who investigate how popularity bias translates to attribute-based bias. They observe that since popularity is not a random phenomenon, it might be that popularity bias leads to other biases that are not obvious when additional user or item information is not available. They enrich a book dataset with author information using publicly available external resources, apply a recommendation process using various collaborative filtering algorithms, and compare popularity bias with bias towards author nationality. They find that popular books in the dataset are mainly written by US-based authors, and hence US-authored books tend to be over-recommended, especially so by algorithms vulnerable to popularity bias.

7.1.2 Author gender bias.

Measurement. Raj and Ekstrand [57] measures attention-weighted rank fairness (*AWRF*) which corresponds to the difference between a group's (in this case, gender of authors) exposure in the recommended lists compared to a target distribution, which they set to the gender distribution in the training ratings. Kırnap et al. [29] point to the problem of missing information when measuring bias: they observe that fair ranking metrics normally require knowledge of membership of items in groups that relate to a protected attribute, even though this information is rarely available or complete. To account for that issue, they propose a sampling strategy and estimation technique for fair ranking metrics by formulating an estimator that works well even with a limited number of labeled items. They apply their method to estimate a set of fair ranking metrics that aim for gender representation in the recommendation equal to the entire corpus ($\Delta_{\text{diff}}, \Delta_{\text{abs}}, \Delta_{\text{sq}}, \Delta_{\text{KL}}$) and the exposure of female authors ($\widetilde{E}_g$), and their estimation outperforms naive baselines. Other metrics compare the presence of male and female authors

Just Accepted

Bias Type	View / Description	Metrics
	Item Biases	
Author nationality bias	Under- or over-representation of author nationality groups relative to the training set [13].	frequency per group [13]
Author gender bias	Under- or over-representation of author gender groups relative to the training set [17, 29, 56, 57, 65].	gender bias [17], log-bias [65], AWRF [56, 57], $\Delta_{\text{diff}}, \Delta_{\text{abs}}, \Delta_{\text{sq}}, \Delta_{\text{KL}}$ [29], EUR, RUR, IAA, EEL [56]
	Under- or over-representation of author gender groups relative to other gender groups [17, 18, 56, 65].	gender bias [17], DP [56], BDV [18], log-bias [65], FAIR, EED, DP [56]
Publisher location bias	Under- or over-representation of publisher location groups relative to other groups [26].	frequency per group [26], exposure per group [26]
	User Biases	
User age bias	Recommendation performance varies across user age groups [15, 35, 81, 91].	MAD NDCG [15], MAD Precision [15], GF MAE [35, 81, 91], GF MSE [91]
	Recommendation outcomes differ depending on user age [35].	saACC [35]
	Preferences of certain user age groups dominate overall recommendations [33].	DIF [33]
User gender bias	Recommendation performance varies across user gender groups [15].	MAD NDCG [15], MAD precision [15]
	Preferences of certain user gender groups dominate overall recommendations [33].	DIF [33]

Table 4. Attribute-based bias views and metrics grouped into item/user categories.

Just Accepted

in the recommendation. Eleftherakis et al. [18] propose Bilateral Disparate Visibility (*BDV*), which represents the difference between the observed probability of recommending books written by one gender and that of recommending books written by another gender.

Raj and Ekstrand [56] perform the first empirical comparative analysis between fair ranking metrics using various datasets, including author-gender fairness assessment on Goodreads. Some of the metrics proposed only measure whether the genders are equally represented in the ranking(s) (*FAIR*, *DP*, *EED*), while others account for utility (*EUR*, *RUR*, *IAA*, *EEL*) or gender distribution in the book dataset (*AWRF*). They also perform sensitivity analysis to assess the impact of certain design choices. They generally find that metrics disagree on orderings of different algorithms' fairness, while they might agree for some experiments and not others. For Goodreads, they find that parameters such as the ranked list size and weighting strategy impact whether the system is considered fair towards authors of different genders. In an influential paper, Ekstrand and Kluver [17] apply a Bayesian model on user profiles and recommended lists (to account for size variances) and calculate the gender tendency. They find that, in Amazon-books, Book-Crossing, and Goodreads, female authors are less underrepresented in user-book interactions than in the unique books, particularly among the popular books. They compare the profile gender tendency with that in the recommended lists of a set of algorithms trained on implicit and explicit versions of the datasets and find that average balances are comparable to user profile average balances, but the algorithms differ when it comes to the distribution. Additionally, they find that explicit-feedback algorithms are much less responsive to their users' input profiles in terms of author gender distribution, likely due to the fact that they account for rating additionally to consumption patterns. Saxena and Jain [65] measure log-bias, which expresses the extent to which users tend to rate books written by one gender higher compared to another gender.

They report the presence of such bias in Book-Crossing and Amazon-books in favor of male authors and compare dataset log-bias with recommendation log-bias.

Mitigation. Some papers propose bias mitigation strategies to reduce differences in author gender distribution in the recommendations. Ekstrand and Kluver [17] apply simple re-ranking strategies on the recommended lists to meet gender balance goals. They find that their strategies are successful in bringing the gender distribution to about 50-50 with minimal accuracy loss. Saxena and Jain [65] address bias towards author gender given explicit feedback. They propose a model that quantifies gender bias in book ratings and subsequent recommendations and, accordingly, a model-agnostic approach to provide unbiased recommendation. Additionally, they incorporate user preference correction to maintain good model accuracy in rating and ranking prediction. They use estimated log-bias to measure whether the predicted ratings of male authors are higher than those of female authors. Their method achieves significant bias reduction with insignificant accuracy loss. Eleftherakis et al. [18] note that a source of demographic bias in neighbor-based collaborative filtering is that the neighbors may foster discriminatory patterns, and that some work that addresses this issue is of low efficiency and impacts accuracy. They propose two new algorithms that tackle efficiency and accuracy. They measure *BDV* on the recommendations based on the *young adult* subset of Goodreads (in this case male-written books are viewed as protected). Their method scales better compared to other fairness-aware methods, especially for Goodreads. Their method's mitigation impact is not as high for Goodreads than it is for the non-book dataset, but the imbalance (in favor of female authors) is still reduced. Raj and Ekstrand [57] attempt to reduce the difference in gender distribution between recommendation and training set, in the context of a grid-view recommendation. They develop a grid-aware re-ranking algorithm for provider-side fairness, which adopts existing techniques by taking into account grid features, and test it on a book dataset. They find that *AWRF* is improved when the re-ranking method is grid-aware, and the results depend on the browsing model, device size, and column size.

7.1.3 Publisher location bias. Gómez et al. [26] tackle **publisher location bias** based on the continent where the book was published, prompted by the geographic imbalance in the data. They measure disparate *visibility* and *exposure* for each continent-based group given a set of algorithms and propose a re-ranking approach that addresses both. They find that 90% of the books were produced in North America, with these books accounting for 93% of the ratings (similar to Daniil et al. [13] who noted the over-representation of American authors), and that this bias is propagated in the recommendation, but their mitigation method performs well.

Synthesis. Most papers about attribute-based bias from an item perspective study bias towards sensitive author characteristics, namely author gender [17, 18, 29, 56, 57, 65] and author nationality [13], while Gómez et al. [26] are concerned with publisher location. Some papers note the challenge of obtaining such information on the items [13, 29], while Saxena and Jain [65] use Genderize.io to estimate author gender. To assess bias towards a group of items, most papers measure how the algorithms treat disadvantaged versus advantaged groups, either by means of representation in recommendation [13, 17, 18, 26, 29, 56, 57] or predicted rating [65]. Unlike the other papers on author gender bias, Eleftherakis et al. [18] consider male authors as the disadvantaged group.

7.2 User

7.2.1 User gender.

Measurement. User gender bias based on the divergent performance view has been measured by mean average deviation of performance for different gender groups by Deldjoo et al. [15]. They systematically track data characteristics that may impact fairness (and accuracy) of collaborative filtering algorithms via regression-based explanatory modeling. User gender has small explanatory power for fairness on the book dataset. To measure whether the recommended items are fairly interacted with by users of different genders in their profiles, Li et al.

Just Accepted

[33] propose a new metric called Interaction Fairness (*IF*), for which they measure the difference between user profile and recommendation.

Mitigation. Li et al. [33] specifically look into fairness in sequential recommendation. They define a fairness-aware sequential recommendation task and propose a multi-task deep model, *FairSR*, which consistently produces the highest improvement in *IF* for different user gender groups while maintaining or even improving on performance metrics.

7.2.2 User age.

Measurement. Deldjoo et al. [15] measure the explanatory power of user age on fairness defined as mean average deviation of performance for different age groups, but similarly to user gender, they find small explanatory power on the book dataset. Other studies also measure deviation of performance metrics between user age groups, namely MAE [35, 81, 91] and MSE [91]. To estimate the impact of user age on the recommendation, Liu et al. [35] propose a new metric named *saAAC* (sensitive attribute accuracy), based on the items in the recommended list and the latent relationship between the content attributes and the users' sensitive attributes. Finally, Li et al. [33] measure the difference in *IF* for different user age groups.

Mitigation. Papers that mitigate bias towards user age in book recommendation by improving performance fairness are as follows. Wei and He [81] design a framework (*CLOVER*) that ensures fairness of meta-learned recommendation models (individual, counterfactual, and group fairness) and try to satisfy all three via a multi-task adversarial learning scheme. They find that meta-learning for cold start is more unfair than other cold start alleviation methods, and with *CLOVER* it becomes fairer towards users of different age while not sacrificing performance. Zhang et al. [91] are also concerned with unfairness in meta learning, however they're specifically interested in what is the critical factor that leads to unfairness. Through theoretical analysis, they show that the critical factor is group proportion imbalance. Then they propose a fairness-aware adaptive sampling framework of meta learning (*FAST*) which adjusts the sampling distribution for different user age groups during training, and theoretically prove convergence. Based on group fairness metrics among users of different age, *FAST* generally achieves more significant fairness improvements than other fairness methods (including *CLOVER*) and comparable utility across various meta-learning paradigms. In a different approach, Liu et al. [35] attempt to filter sensitive information so that it is not learned when constructing representations by using *DALFRec*, a fairness-aware algorithm that uses adversarial learning to mitigate the effect of sensitive information on both user and item side. They assess the algorithm's ability to reduce discrimination with *saAAC* (sensitive attribute accuracy) and additional performance fairness metrics. They measure relation between user age and item publication year and author. When it comes to user age, they achieve better fairness over traditional recommendations and other fairness methods. They also note that on the book dataset, the *saACC* of all algorithms is high, which indicates that the item distribution in the recommendations is highly correlated with users' age. Finally, the aforementioned multi-task deep model *FairSR* developed by Li et al. [33] also improves *IF* for different user age groups in a sequential recommendation task.

7.2.3 Other. To detect bias towards users, especially because minority preferences are discriminated against, Misztal-Radecka and Indurkhya [42] design a bias-aware hierarchical clustering algorithm that clusters users to find ones whose needs are not met by a black box recommender system. On top, they apply a post-hoc explainer that describes the most important features of these users. Among other results, they note that (NDCG) performance is lower for discriminated groups of smaller size than for bigger groups. However, the discriminated user features are not presented in this paper for the book dataset.

Synthesis. Most papers about attribute-based bias from a user perspective are concerned with either user age [35, 81, 91], or both user age and gender [15, 33]. Misztal-Radecka and Indurkhya [42] test user age and location,

Just Accepted

Dataset	Bias-related user features	Bias-related item features	Num. users	Num. items	Num. interactions	Source
Amazon-books	activity [36, 85], popularity affinity [49]	popularity [36, 49, 69], author gender* [17, 65]	8,026,324	2,330,066	22,507,155	Collected in 2018 from Amazon [48]
Book-Crossing	activity [43, 44, 54, 55, 71, 82, 84], popularity affinity [14, 31, 45, 54, 88], user age [15, 33, 35, 81, 91], user gender [15, 33]	popularity [6, 14, 23, 31, 43, 45, 55, 71, 78, 80, 82, 89, 90], author nationality* [13], author gender* [17, 65], publisher location* [26]	278,858	271,379	1,149,780	Collected in 2004 from the Book Crossing community [92]
GoodReads		popularity [61], author gender* [17, 18, 29, 56, 57]	876,145	2,360,655	228,648,342	Collected in 2017 from Goodreads [75, 76]
Goodbooks		blockbusterness [83]	53,424	10,000	5,976,479	Collected in 2017 from Goodreads [87]

Table 5. Bias-related features and properties of 4 book recommendation datasets. * means that the researchers enriched the datasets to acquire this information.

Just Accepted

but do not disclose which user features were found to be prone to discrimination in the book dataset. To assess bias, some studies measure difference in performance for different user groups (i.e., group fairness) [15, 42, 91]. Relevant and case-specific metrics are also proposed [33, 35].

7.3 Joint

None of the papers surveyed deal with attribute-based bias towards items and users simultaneously.

8 Datasets and reproducibility in book recommendation

In this section, we take a closer look into the datasets that researchers use to evaluate bias. Table 5 includes all datasets along with the features that researchers have evaluated bias towards, as well as basic information on them. These datasets stem from three sources: Amazon, Book-Crossing and Goodreads. Book-Crossing is widely used despite having been created more than 20 years ago. All the datasets are very sparse, which often prompts researchers to preprocess them before use, for example, by performing k-core filtering (i.e., removing

Dataset	Bias type	Measurement/Metric	Mitigation/Metric
Amazon-books	Popularity	popularity-difficulty correlation [36]	recall per group [49], novelty [49], coverage [69], LT frequency [69], LT coverage [69]
	Activity level	activity-difficulty correlation [36]	IA MRR [85], IA HR [85], IA NDCG [85]
	Popularity affinity		recall per group [49]
	Author gender		gender tendency [17], log-bias [65]
Book-Crossing	Popularity	correlation [14, 31, 45]	estimation error per group [43], P-fairness [82], coverage [71], LT precision [23], LT recall [23], NDCG [78], MRR [78], recall [78], LT HR [6, 80, 89, 90], LT NDCG [6, 80, 89, 90]
	Activity level	recall per group [44]	NDCG per group [54, 84], recall per group [71, 84], $\%\Delta GAP$ [54], estimation error per group [43], C-fairness [82]
	Popularity affinity	$\%\Delta GAP$ [14, 45], MAE per group [31], NDCG per group [88]	NDCG per group [54], $\%\Delta GAP$ [54]
	Author nationality	frequency per group [13]	
	Author gender		gender tendency [17], log bias [65]
	Publisher location		frequency per group [26], exposure per group [26]
	User age	MAD NDCG [15], MAD precision [15]	DIF [33], GF MAE [35, 81, 91]
	User gender	MAD NDCG [15], MAD precision [15]	DIF [33]
GoodReads	Popularity		PopQ@1 [61]
	Author gender	DP [56], EED [56], frequency per group [29], exposure per group [29]	gender tendency [17], AWRF [56, 57], BDV [18], FAIR, EED, DP, EUR, RUR, IAA, EEL [56], $\Delta_{\text{diff}}, \Delta_{\text{abs}}, \Delta_{\text{sq}}, \Delta_{\text{KL}}$ [29]
Goodbooks	Blockbuster	frequency per group [83]	

Table 6. Bias types, metrics, and corresponding paper references per dataset.

Just Accepted

interactions until all users have interacted with at least k books and all books have been interacted with by at least k users). This practice raises questions on whether the conclusions of each study hold outside the limits of their experimentation or can be directly compared with the conclusions of a different study on the same dataset.

Additionally, the datasets include very little information on the users and authors. Specifically, bias towards user characteristics has only been studied on Book-Crossing. Author gender has been studied on multiple datasets, but the enrichment is not necessarily consistent, as various sources have been used to discern gender, namely

VIAF [17], Wikidata [18] and Genderize.io [65]. Notably, outside of author features, the only book characteristic that has been considered is publisher location.

Table 6 shows for each dataset which types of bias have been measured, as well as based on which metrics. Even though some papers have studied the same type of bias on a dataset, they mostly use different metrics. Therefore, it is mostly not possible to directly compare the studies unless they explicitly use each other's methods as one of the baseline comparisons. We will discuss cases where the same metrics have been used across studies.

Popularity bias on Book-Crossing. Daniil et al. [14], Kowald and Lacic [31], Naghiaei et al. [45] measure popularity-recommendation frequency correlation based on algorithms trained on Book-Crossing. Daniil et al. [14], Naghiaei et al. [45] measure $\%\Delta GAP$, yet these studies are explicitly reproductions of each other. Daniil et al. [14] conclude that the differences in evaluation strategies among the studies is the driver behind the divergence in conclusions, despite the use of the same metric and dataset.

Additionally, Balasubramanian et al. [6], Wei et al. [80], Zhang et al. [89, 90] measure long-tail HR and long-tail NDCG on Book-Crossing. These studies use each other's methods as baselines in chronological order, as they all attempt to augment the information space of the long-tail items to empower them.

User age bias on Book-Crossing. Liu et al. [35], Wei and He [81], Zhang et al. [91] measure MAE-based group fairness with user age being the sensitive characteristic on Book-Crossing, and they are all concerned with meta-learning. In this case, Liu et al. [35] and Zhang et al. [91] compare their respective frameworks with *CLOVER*, the framework proposed by Wei and He [81] as a baseline. However, the two studies do not report the exact same numbers for *CLOVER*'s GF MAE (presumably due to experimental design differences), so comparing between them directly is not feasible.

Activity level bias on Book-Crossing. Multiple studies measure or mitigate activity level bias on Book-Crossing by comparing algorithm performance on groups of users with different activity levels [44, 54, 84, 90]. However, among other differences, the studies divide the users into groups differently between them: Rahmani et al. [54] compare the top 5% most active users with the rest, Yang et al. [84] compare the top 20% most active users with the rest, Zhang et al. [90] rank the users based on activity level and divide them into 4 parts, and Moody and Glass [44] compare the users with activity level above the median with the rest. We cannot conclude which mitigation method ([54, 84, 90]) is more successful in addressing activity level bias.

Publicly available datasets with user-item interactions are invaluable for recommender systems research, as they facilitate comparison across methods and support reproducibility. However, our analysis shows that, in the context of bias in book recommendation, their potential to support cumulative knowledge is currently limited. Even when the same dataset and the same bias type are studied, differences in preprocessing, user or item grouping strategies, metric choice, and evaluation protocols often prevent direct comparison across studies. As 4 out of 10 studies included in this survey do not grant access to a public repository with their code and experiments (see Figure 3b in the Appendix), addressing these challenges requires more explicit reporting of dataset preprocessing, clearer operationalization of bias-related concepts, and greater transparency around experimental setups.

Just Accepted

9 Discussion

9.1 General observations

There has been relatively little research specifically addressing bias in book recommender systems compared to the extensive work done on bias in other domains, for example, music recommendation. At first glance, the limited research is surprising given that bias is a well-known and documented issue in the book market in general [32]. One explanation is that book reading is considered a relatively niche habit compared to other media, and there hasn't been as strong an emphasis on personalization for the sake of increasing consumption. In recommender

Table 7. Coverage of research dimensions. Papers are classified by general bias type, bias subject, specific bias type, and research action according to the framework described in Section 5. Empty cells highlight critical research gaps.

Bias Type	Subject	Specific Type	Measurement	Mitigation
Interaction-based	Item	Popularity bias	4	13
		Blockbuster bias	1	0
		Subtotal	*5*	*13*
	User	Activity level bias	2	7
		Popularity affinity bias	4	3
		Subtotal	*6*	*10*
	Joint	Popularity + Activity	1	4
		Popularity + Affinity	3	2
		Subtotal	*4*	*6*
Attribute-based	Item	Author nationality bias	1	0
		Author gender bias	2	4
		Publisher location bias	0	1
		Subtotal	*3*	*5*
	User	User age bias	1	4
		User gender bias	1	1
		Subtotal	*2*	*5*
	Joint	User + item attributes	0	0
		Subtotal	*0*	*0*

systems, research is often driven by streaming companies. Users consume music, series, and movies faster, which might inherently require further automation and personalization and hence lead to bias propagation.

As seen in table 7, most of the papers that we include in this survey study interaction-based bias, especially popularity. There exists varied work on popularity bias and its effect on users and items, with a notable emphasis on mitigation while also ensuring sufficient accuracy. Given the fact that the book market is often assumed to follow a long-tail distribution, this is not surprising. This observation aligns with Dinnissen and Bauer [16] who found popularity bias to be a prevalent topic in fair music recommendation literature (see Section 3).

Another observation is the ubiquity of incidental and somewhat superficial studies on bias in book recommender systems. Many of these studies seem to treat bias as a generic machine learning-related issue without critically engaging with the book domain. Researchers often use book datasets simply because they are available rather than because they are concerned with the specific implications of bias in literary recommendation. In most of these papers, the book dataset was one of many. Book recommendation results are sometimes treated as an afterthought: in some cases, book-related results are hidden in appendices, while in others, they are omitted altogether. This raises questions about why book recommendation research is sidelined. We hypothesize that due to the book datasets' sparsity, the results are sometimes too inconclusive to be included in the main body of the paper and discussed in detail. As a result, the field lacks focus and continuity. It also becomes harder for researchers, especially outside of the field of recommender systems, to find material on bias in book recommendation and draw conclusions.

To add to that, much of the existing research relies on the same few datasets, primarily American ones, which fails to account for global reading habits and cultural differences. Additionally, book recommendation studies often include very few external features, making it difficult to explore biases related to other relevant characteristics, both for users and books. For users, the only external features tested are gender and age, presumably because

Just Accepted

these are the ones available in the datasets. For books, a few more characteristics are tested; however, the majority of papers are concerned with author gender. In some cases, the studies use Linked Open Data to obtain this information, and gender might be easier to discern than other sensitive characteristics. This observation points to a well-known issue in quantitative research: sometimes we measure what is available rather than what is important.

9.2 Gaps & future directions

Table 7 makes the imbalances concrete and reveals specific gaps. Interaction-based bias accounts for the bulk of research. For example, popularity bias has been studied in 17 out of the 40 papers examined in this survey. The only author characteristic that has received substantial attention is gender. While publishers are a very important stakeholder in the book market, only bias towards their location has been explored. Attribute-based bias remains largely unexplored. Notably, no paper simultaneously addresses bias affecting both users and items based on external attributes despite recent emphasis in recommender systems research on multistakeholder solutions.

One gap in the mitigation of bias in book recommendation is effective comparison between the various methods developed, which would determine which is the 'best' among them. While most of the mitigation papers compare their method to some baselines, the emphasis is on the method rather than the effect on the book dataset result. As discussed in Section 8, the studies included in this survey mostly use different metrics and evaluation protocols, even when measuring the same type of bias based on the same dataset. An extensive comparative study that tests the different mitigation methods for a set of book datasets and given different bias metrics would shed light on what the state-of-the-art is when it comes to the book domain specifically.

User-book interaction datasets, as they are now, are limited. There is much progress to be made on improving them, in the sense of enriching them with publicly available information. Ekstrand and Kluver [17] take advantage of Linked Open Data sources to create a valuable dataset that includes author gender, which is used by some of the other studies that we survey. Daniil et al. [13] follow a similar approach to supplement a smaller dataset with author country of citizenship. While publicly available information is not complete and is known to be biased in its own right, it would be a step towards the right direction. At the same time, book themes as a bias subject have barely been studied. Extracting information on the themes of a book by textually analyzing the description might be even more accessible than information on the authors' personal characteristics. Future research on bias in book recommendation could explore this avenue.

9.3 Recommendations for interdisciplinarity

The field of bias in book recommendation would benefit from connection between theoretical background and practical studies. There is insight to be gained from interdisciplinary research, especially when interacting with a field as rich as LIS. Recommender systems studies can rely on existing LIS frameworks that describe just and inclusive library services [28, 37, 72, 86] and design metrics for bias and inclusivity for each component of the book ecosystem accordingly. Grounding on theory can assist with interdisciplinary communication and ensure overall progress.

A crucial step that can support the aforementioned goal is engagement with real library or commercial literary recommenders. In simulations, researchers are limited to making assumptions regarding both how a recommendation is done and also what is a relevant type of bias in the given use case. Practitioners in the book domain have a more sophisticated understanding of what bias entails and can support researchers in their pursuit. Additionally, since bias mitigation for many studies means to simply reduce bias compared to a baseline, in an actual use case stakeholders can give input on what an 'ideal' system would look like (e.g., appropriate level of representation for every group of authors) [73]. Context specificity is crucial here; book reading is very much dependent on the cultural context, even today that many of us exist in the same online world. Books are still

Just Accepted

produced and recommended in decentralized ways. By only studying the same few American datasets, researchers are restricted to a one-dimensional view of bias and can only make narrow conclusions.

The differences in terminologies and research methods used by LIS and computer science professionals are a significant hurdle in creating fair and effective recommendation systems. The lack of integration between these fields hampers the development of systems that align with the core values of diversity, accessibility, and fairness, which libraries uphold. An essential step forward is greater collaboration between LIS professionals and computer scientists to establish shared terminology and research practices. Interdisciplinary projects can support the design of recommender systems that take bias into account in a targeted and situated way. Without such intentional safeguards, these systems risk perpetuating existing biases and further marginalizing underrepresented voices.

10 Conclusion

In this paper, we conducted a literature review on bias in book recommender systems across computer science venues. We gathered and analyzed 40 papers from various perspectives. We found that, despite the prevalence of book datasets with ratings in recommender systems research, studies often focus more on the 'bias' part rather than the 'book' part. Accordingly, we proposed a set of future directions based on gaps in literature, as well as recommendations for interdisciplinary practices. We hope that by collecting and mapping recent research, this paper can serve as a stepping stone for researchers, computer scientists or otherwise, who wish to contribute to the field of bias in book recommendation.

Just Accepted

References

[1] Himan Abdollahpouri and Robin Burke. 2021. Multistakeholder recommender systems. In *Recommender systems handbook*. Springer, 647–677.

[2] Anne-Mette Bech Albrechtslund. 2020. Amazon, Kindle, and Goodreads: implications for literary consumption in the digital age. *Consumption Markets & Culture* 23, 6 (2020), 553–568.

[3] Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz. 2018. A survey of book recommender systems. *Journal of Intelligent Information Systems* 51 (2018), 139–160.

[4] American Library Association. 2021. Access to Library Resources and Services. https://www.ala.org/advocacy/intfreedom/access Accessed: 26 March 2025.

[5] Khalid Anwar, Jamshed Siddiqui, and Shahab Saquib Sohail. 2020. Machine learning-based book recommender system: a survey and new perspectives. *International Journal of Intelligent Information and Database Systems* 13, 2-4 (2020), 231–248.

[6] Keshav Balasubramanian, Abdulla Alshabanah, Elan Markowitz, Greg Ver Steeg, and Murali Annavaram. 2024. Biased User History Synthesis for Personalized Long-Tail Item Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 189–199. doi:10.1145/3640457.3688141

[7] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[8] Jill Cirasella. 2021. Ideology, Policy, and Practice: Structural Barriers to Collections Diversity in Research and College Libraries. *College & Research Libraries* 82, 6 (2021), 865–878. https://crl.acrl.org/index.php/crl/article/view/25342/33226

[9] Doris Hargrett Clack. 1995. Subject Access to African American Studies Resources in Online Catalogs: Issues and Answers. *Cataloging & Classification Quarterly* 19, 2 (1995), 49–66. doi:10.1300/J104v19n02_04

[10] Catherine Nicole Coleman. 2020. Managing Bias When Library Collections Become Data. *International Journal of Librarianship* 5, 1 (2020), 8–19. doi:10.23974/ijol.2020.vol5.1.162

[11] Martina Currenti. 2023. TikTok as a Marketing Tool in the Hands of Publishers. *Logos* 34, 1 (2023), 24–37.

[12] Sarah Park Dahlen. 2020. "We need diverse books": Diversity, activism, and children's literature. *Literary cultures and twenty-first-century childhoods* (2020), 83–108.

[13] Savvina Daniil, Mirjam Cuper, Cynthia Liem, Jacco van Ossenbruggen, and Laura Hollink. 2022. Hidden Author Bias in Book Recommendation. *arXiv preprint arXiv:2209.00371* (2022).

[14] Savvina Daniil, Mirjam Cuper, Cynthia C. S. Liem, Jacco van Ossenbruggen, and Laura Hollink. 2024. Reproducing Popularity Bias in Recommendation: The Effect of Evaluation Strategies. *ACM Trans. Recomm. Syst.* 2, 1, Article 5 (March 2024), 39 pages. doi:10.1145/3637066

[15] Yashar Deldjoo, Alejandro Bellogin, and Tommaso Di Noia. 2021. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Information Processing & Management* 58, 5 (2021), 102662. doi:10.1016/j.ipm.2021.102662

[16] Karlijn Dinnissen and Christine Bauer. 2022. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data* 5 (2022), 913608.
[17] Michael D Ekstrand and Daniel Kluver. 2021. Exploring author gender in book rating and recommendation. *User Modeling & User-Adapted Interaction* 31, 3 (2021).
[18] Stavroula Eleftherakis, Georgia Koutrika, and Sihem Amer-Yahia. 2024. Optimizing Neighborhoods for Fair Top-N Recommendation. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 57–66. doi:10.1145/3627043.3659539
[19] Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* 6, 1 (2023), 3.
[20] J. Folta. 2025. Fable's AI-generated end-of-year reading summaries veered into bigotry. https://lithub.com/fables-ai-generated-end-of-year-reading-summaries-veered-into-bigotry/ Accessed: 26 March 2025.
[21] National Endowment for the Arts. 2023. Arts Participation Patterns in 2022 Highlights from the Survey of Public Participation in the Arts. (Oct 2023). https://www.arts.gov/sites/default/files/2022-SPPA-final.pdf
[22] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. doi:10.1145/230538.230561
[23] Giovanni Gabbolini, Edoardo D'Amico, Cesare Bernardis, and Paolo Cremonesi. 2021. On the instability of embeddings for recommender systems: the case of matrix factorization. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (Virtual Event, Republic of Korea) *(SAC '21)*. Association for Computing Machinery, New York, NY, USA, 1363–1370. doi:10.1145/3412841.3442011
[24] Cynthia R. Greenlee. 2025. Fable AI's New Model Is More Inclusive, But Still Struggles with Racism. https://www.nytimes.com/2025/01/03/us/fable-ai-books-racism.html Accessed: 26 March 2025.
[25] Asela Gunawardana, Guy Shani, and Sivan Yogev. 2012. Evaluating recommender systems. In *Recommender systems handbook*. Springer, 547–601.
[26] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. 2022. Provider fairness across continents in collaborative recommender systems. *Information Processing & Management* 59, 1 (2022), 102719. doi:10.1016/j.ipm.2021.102719
[27] L. Jahnke, K. Tanaka, and C. Palazzolo. 2022. Ideology, policy, and practice: Structural barriers to collections diversity in research and college libraries. *College & Research Libraries* 83, 2 (2022), 166–184. doi:10.5860/crl.83.2.166
[28] Andrea Jamison. 2023. *Decentering whiteness in libraries: a framework for inclusive collection management practices*. Rowman & Littlefield.
[29] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1065–1075. doi:10.1145/3442381.3450080
[30] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction* 34, 5 (2024), 1777–1834.
[31] Dominik Kowald and Emanuel Lacic. 2022. Popularity bias in collaborative filtering-based multimedia recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–11.
[32] Tin Cheuk Leung and Koleman S Strumpf. 2022. Discrimination in the Publishing Industry? *Available at SSRN 3660289* (2022).
[33] Cheng-Te Li, Cheng Hsu, and Yang Zhang. 2022. FairSR: Fairness-aware Sequential Recommendation through Multi-Task Learning with Preference Graph Embeddings. *ACM Trans. Intell. Syst. Technol.* 13, 1, Article 16 (Feb. 2022), 21 pages. doi:10.1145/3495163
[34] Chen Lin, Dugang Liu, Hanghang Tong, and Yanghua Xiao. 2022. Spiral of Silence and Its Application in Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering* 34, 6 (2022), 2934–2947. doi:10.1109/TKDE.2020.3013973
[35] Shenghao Liu, Yu Zhang, Lingzhi Yi, Xianjun Deng, Laurence T. Yang, and Bang Wang. 2024. Dual-Side Adversarial Learning Based Fair Recommendation for Sensitive Attribute Filtering. *ACM Trans. Knowl. Discov. Data* 18, 7, Article 165 (June 2024), 20 pages. doi:10.1145/3648683
[36] Yang Liu, Alan Medlar, and Dorota Glowacka. 2023. What We Evaluate When We Evaluate Recommender Systems: Understanding Recommender Systems' Performance using Item Response Theory. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 658–670. doi:10.1145/3604915.3608809
[37] M Elena Lopez, Bharat Mehra, and Margaret Capse. 2023. An exploratory social justice framework to develop public library services with underserved families. *Public Library Quarterly* 42, 6 (2023), 576–601.
[38] Dr. Prakashbhai N. Makwana. 2024. The Dewey Decimal System and Traditional Libraries. *International Journal of Science and Research (IJSR)* 13, 6 (June 2024). doi:10.21275/SR24530155734
[39] Kay Mathiesen. 2015. Informational justice: A conceptual framework for social justice in library and information services. *Library trends* 64, 2 (2015), 198–225.
[40] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. 1097–1101.
[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

Unaccepted

[42] Joanna Misztal-Radecka and Bipin Indurkhya. 2021. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management* 58, 3 (2021), 102519. doi:10.1016/j.ipm.2021.102519
[43] Joanna Misztal-Radecka and Bipin Indurkhya. 2023. A bias detection tree approach for detecting disparities in a recommendation model's errors. *User Modeling and User-Adapted Interaction* 33, 1 (2023), 43–79.
[44] Jennifer Moody and David H. Glass. 2016. A Novel Classification Framework for Evaluating Individual and Aggregate Diversity in Top-N Recommendations. *ACM Trans. Intell. Syst. Technol.* 7, 3, Article 42 (Feb. 2016), 21 pages. doi:10.1145/2700491
[45] Mohammadmehdi Naghiaei, Hossein A Rahmani, and Mahdi Dehghan. 2022. The unfairness of popularity bias in book recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 69–81.
[46] Yesha Naik and Barry Trott. 2012. Finding Good Reads on Goodreads: Readers Take RA into Their Own Hands. *Reference & User Services Quarterly* 51, 4 (2012), 319–323. http://www.jstor.org/stable/refuseserq.51.4.319
[47] Library & Information Science Education Network. 2024. Exploring AI-Based Recommendation Systems in Libraries. https://www.lisedunetwork.com/exploring-ai-based-recommendation-systems-in-libraries/ Accessed: 25 March, 2025.
[48] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.
[49] Hiroki Okamura, Keisuke Maeda, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2024. Flexibly manipulating popularity bias for tackling trade-offs in recommendation. *Information Processing & Management* 61, 2 (2024), 103606. doi:10.1016/j.ipm.2023.103606
[50] Hope Olson. 2002. *The Power to Name: Locating the Limits of Subject Representation in Libraries*. Kluwer Academic, Dordrecht.
[51] Yesid Ospitia-Medina, Sandra Baldassarri, Cecilia Sanz, and José Ramón Beltrán. 2022. Music Recommender Systems: A Review Centered on Biases. *Advances in Speech and Music Technology: Computational Aspects and Applications* (2022), 71–90.
[52] OverDrive, Inc. 2025. Libraries Break Digital Lending Records in 2024 with Over 739 Million Checkouts. https://company.overdrive.com/2025/01/27/libraries-break-digital-lending-records-in-2024-with-over-739-million-checkouts/ Accessed: 01 April 2025.
[53] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021). arXiv:https://www.bmj.com/content/372/bmj.n71.full.pdf doi:10.1136/bmj.n71
[54] Hossein A. Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. 2022. Experiments on Generalizability of User-Oriented Fairness in Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2755–2764. doi:10.1145/3477495.3531718
[55] Hossein A. Rahmani, Mohammadmehdi Naghiaei, and Yashar Deldjoo. 2024. A Personalized Framework for Consumer and Producer Group Fairness Optimization in Recommender Systems. *ACM Trans. Recomm. Syst.* 2, 3, Article 19 (June 2024), 24 pages. doi:10.1145/3651167
[56] Amifa Raj and Michael D. Ekstrand. 2022. 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* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 726–736. doi:10.1145/3477495.3532018
[57] Amifa Raj and Michael D Ekstrand. 2024. Towards Optimizing Ranking in Grid-Layout for Provider-Side Fairness. In *European Conference on Information Retrieval*. Springer, 90–105.
[58] S. R. Ranganathan. 1931. *The Five Laws of Library Science*. Madras Library Association, Madras, India.
[59] Alex Reece. 2024. Frisco Library Sees Record-Breaking Patronage During First Fiscal Year Open, Data Shows. https://communityimpact.com/dallas-fort-worth/frisco/government/2024/11/13/frisco-library-sees-record-breaking-patronage-during-first-fiscal-year-open-data-shows/ Accessed: 01 April 2025.
[60] Kara Reuter. 2024. Making Magic Happen: Building and Launching a Reader's Advisory Kiosk. *Information Technology and Libraries* 43, 2 (2024).
[61] Wondo Rhee, Sung Min Cho, and Bongwon Suh. 2022. Countering Popularity Bias by Regularizing Score Differences. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 145–155. doi:10.1145/3523227.3546757
[62] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2021. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook* (2021), 1–35.
[63] Jonathan H Rystrøm. 2024. An Offer You Cannot Refuse? Trends in the Coercive Impact of Amazon Book Recommendations. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–15.
[64] Joyce G Saricks. 2005. *Readers' advisory service in the public library*. American Library Association.
[65] Shrikant Saxena and Shweta Jain. 2024. Exploring and mitigating gender bias in book recommender systems with explicit feedback. *Journal of Intelligent Information Systems* 62, 5 (2024), 1325–1346.

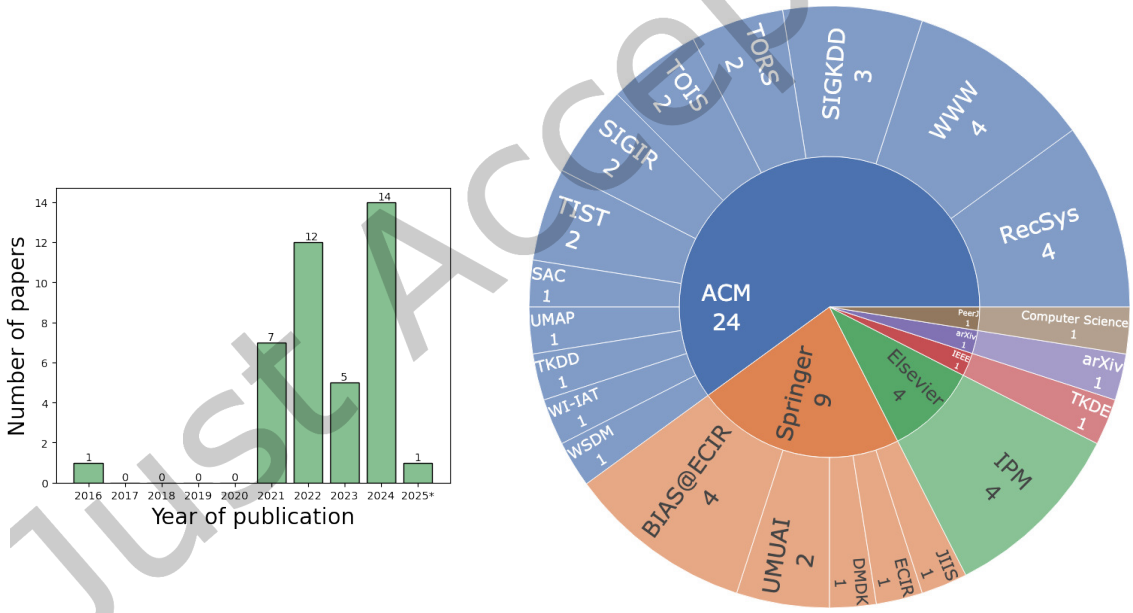
Just Accepted

[66] Lijuan Shen and Liping Jiang. 2024. Eliminating bias: enhancing children's book recommendation using a hybrid model of graph convolutional networks and neural matrix factorization. *PeerJ Computer Science* 10 (2024), e1858.
[67] Ayşe Sirel. 2021. Reflection of Paradigm Change in Information Technology to Library Architecture: The Helsinki Oodi Library. *Rigas Tehniskas Universitates Zinatniskie Raksti* 17, 1 (2021), 123–135.
[68] R. Sivasankari, S. Suriya, S. Sindhu, J. Shyamala Devi, and J. Dhilipan. 2024. AI-Powered Recommendation Systems and Resource Discovery for Library Management. In *Advances in Library and Information Science Applications of Artificial Intelligence in Libraries*. IGI Global, 223–244. https://www.igi-global.com/chapter/ai-powered-recommendation-systems-and-resource-discovery-for-library-management/346535 Accessed: 22 March 2025.
[69] Shuai Tang and Xiaofeng Zhang. 2022. CADPP: An Effective Approach to Recommend Attentive and Diverse Long-tail Items. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Melbourne, VIC, Australia) *(WI-IAT '21)*. Association for Computing Machinery, New York, NY, USA, 218–225. doi:10.1145/3486622.3493961
[70] Beau Steenken Tarica LaBossiere, Endia Paige. 2024. Keeping Up With: Bias. American Library Association. https://www.ala.org/acrl/publications/keeping_up_with/bias Accessed: 2024-04-04.
[71] Riku Togashi, Mayu Otani, and Shin'ichi Satoh. 2021. Alleviating Cold-Start Problems in Recommendation through Pseudo-Labelling over Knowledge Graph. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 931–939. doi:10.1145/3437963.3441773
[72] University of Colorado Boulder. 2025. Anti-Racist Collections Review Workbook. https://libguides.colorado.edu/anti-racist-collections-review-acquisitions/workbook2?utm_source=chatgpt.com Accessed: 23 March, 2025.
[73] Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. 2024. Diversity of what? On the different conceptualizations of diversity in recommender systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 573–584.
[74] Simon Wakeling, Paul Clough, Barbara Sen, and Lynn Silipigni Connaway. 2012. "Readers who borrowed this also borrowed...": recommender systems in UK libraries. *Library Hi Tech* 30, 1 (2012), 134–150.
[75] Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 86–94. doi:10.1145/3240323.3240369
[76] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2605–2610. doi:10.18653/V1/P19-1248
[77] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–43.
[78] Zhidan Wang, Lixin Zou, Chenliang Li, Shuaiqiang Wang, Xu Chen, Dawei Yin, and Weidong Liu. 2024. Toward Bias-Agnostic Recommender Systems: A Universal Generative Framework. *ACM Trans. Inf. Syst.* 42, 6, Article 142 (June 2024), 30 pages. doi:10.1145/3655617
[79] Katie Warburton, Charles Kemp, Yang Xu, and Lea Frermann. 2023. Quantifying Bias in Library Classification Systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/94b7d9zr
[80] Chunyu Wei, Jian Liang, Di Liu, Zehui Dai, Mang Li, and Fei Wang. 2023. Meta Graph Learning for Long-tail Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 2512–2522. doi:10.1145/3580305.3599428
[81] Tianxin Wei and Jingrui He. 2022. Comprehensive Fair Meta-learned Recommender System. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 1989–1999. doi:10.1145/3534678.3539269
[82] Lanling Xu, Zihan Lin, Jinpeng Wang, Sheng Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Promoting Two-sided Fairness with Adaptive Weights for Providers and Customers in Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 918–923. doi:10.1145/3640457.3688169
[83] Emre Yalcin. 2022. Exploring potential biases towards blockbuster items in ranking-based recommendations. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2033–2073.
[84] Hao Yang, Xian Wu, Zhaopeng Qiu, Yefeng Zheng, and Xu Chen. 2024. Distributional Fairness-aware Recommendation. *ACM Trans. Inf. Syst.* 42, 5, Article 131 (April 2024), 28 pages. doi:10.1145/3652854
[85] Wenhao Yang, Yingchun Jian, Yibo Wang, Shiyin Lu, Lei Shen, Bing Wang, Haihong Tang, and Lijun Zhang. 2024. Not All Embeddings are Created Equal: Towards Robust Cross-domain Recommendation via Contrastive Learning. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3195–3206. doi:10.1145/3589334.3645357
[86] Nasim Yazdani. 2024. A Pedagogy of Care in Academic Libraries: A Framework to Increase Underrepresented Communities' Sense of Belonging and Engagement. *Journal of the Australian Library and Information Association* 73, 4 (2024), 505–528.
[87] Zygmunt Zajac. 2017. Goodbooks-10k: a new dataset for book recommendations. http://fastml.com/goodbooks-10k. *FastML* (2017).

Just Accepted

[88] Kuanyi Zhang, Min Xie, Yi Zhang, and Haixian Zhang. 2023. Utilizing Implicit Feedback for User Mainstreaminess Evaluation and Bias Detection in Recommender Systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 42–58.
[89] Yin Zhang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, and Ed H. Chi. 2021. A Model of Two Tales: Dual Transfer Learning Framework for Improved Long-tail Item Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2220–2231. doi:10.1145/3442381.3450086
[90] Yin Zhang, Ruoxi Wang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, James Caverlee, and Ed H. Chi. 2023. Empowering Long-tail Item Recommendation through Cross Decoupling Network (CDN). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 5608–5617. doi:10.1145/3580305.3599814
[91] Zheng Zhang, Qi Liu, Zirui Hu, Yi Zhan, Zhenya Huang, Weibo Gao, and Qingyang Mao. 2024. Enhancing Fairness in Meta-learned User Modeling via Adaptive Sampling. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3241–3252. doi:10.1145/3589334.3645369
[92] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. 22–32.
[93] Erion Çano and Maurizio Morisio. 2015. Characterization of public datasets for Recommender Systems. In *2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*. 249–257. doi:10.1109/RTSI.2015.7325106
[94] Seda Nur Şanlı and Ayşe Sirel. 2022. The Effects of Developing Information Technologies on 21st Century Library Architecture. In *Proceedings of the International Conference of Contemporary Affairs in Architecture and Urbanism (ICCAUA)*, Vol. 5. 12–26.

A Appendix

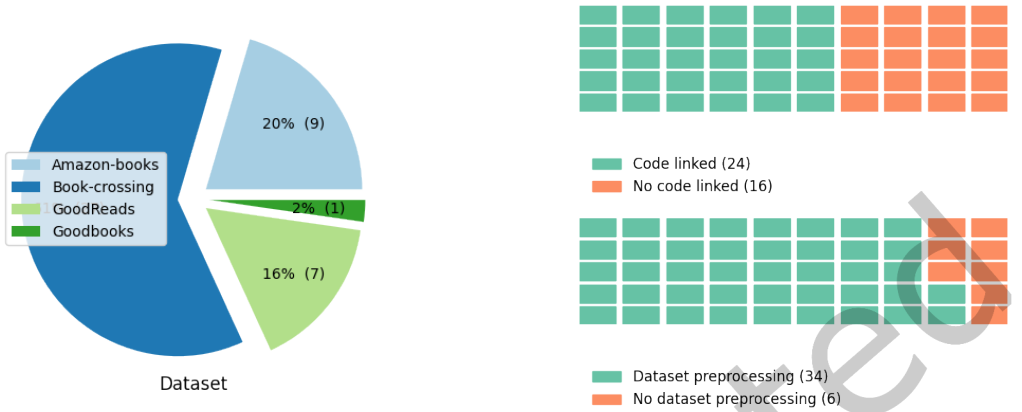


(a) Year of publication (b) Venue and publisher

Fig. 2. Papers per publication year, venue, and publisher. *We conducted the survey in 2024 but one of the papers has official copyright in 2025.

Figures 2, 3 and 4 show basic statistics about the papers. According to figure 2a, the big bulk of computer science research on bias in book recommender systems is within the last few years. Figure 2b shows the paper distribution among publishers and venues. Unsurprisingly, most of the papers come from ACM publications. Otherwise, the

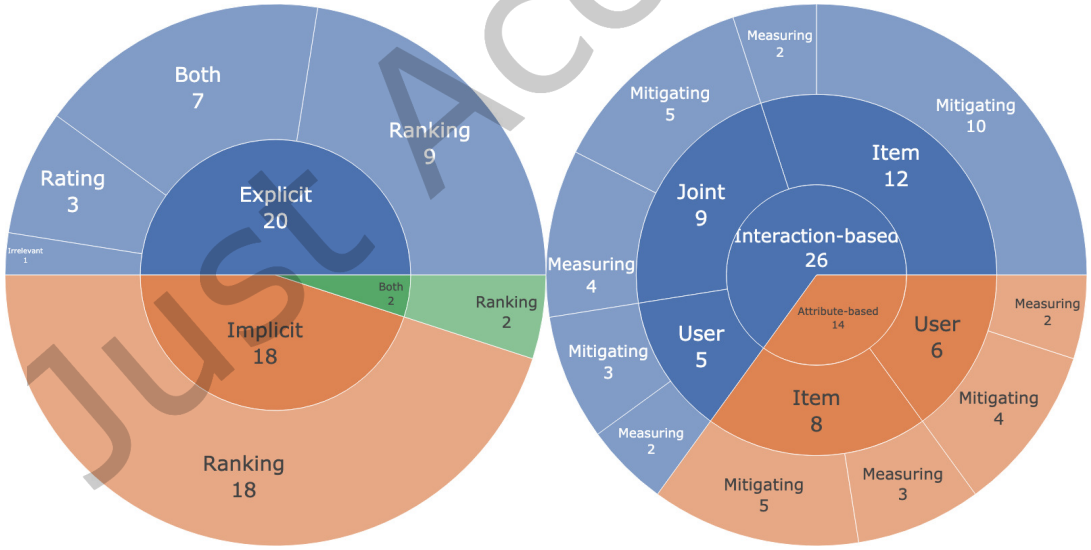
Just Accepted



(a) Dataset used

(b) Code availability and dataset preprocessing

Fig. 3. Papers per dataset, code inclusion, and dataset preprocessing. Note that Goodreads and Goodbooks stem from the same source.



(a) Prediction task and type of feedback

(b) Bias type, bias perspective and action taken

Fig. 4. Papers per prediction task, type of feedback, type of bias, and action taken

Just Accepted

papers we included are scattered across conferences and journals. In figure 3a we see that *Book-Crossing* is used in more than half of the papers, with *Goodreads* and *Amazon-books* used in the majority of the other papers. The results across papers are not necessarily comparable even if they use the same dataset, since, as figure 3b shows, the majority of the papers train their models in a preprocessed version of the given dataset (e.g., k-core filtering). According to the same figure, a bit more than half of the papers provide a link to a repository where they include their code. Figure 4a shows that roughly the same amount of papers use explicit (i.e., ratings) and implicit (i.e., interactions) feedback as the basis of training a model. Regardless of the type of features, most of the papers analyze bias given a ranking task. This result is in line with recent trends in recommender systems which favor ranking rather than rating prediction. Most papers study interaction-based bias, especially from an item perspective, as figure 4b shows. Across bias types and perspectives there is a balance between measurement and mitigation, with the exception of interaction-based bias towards items where most papers present a mitigation strategy.

Received 11 June 2025; revised 31 August 2026; accepted 4 October 2026

Just Accepted