

Like a Skilled DJ - an Expert Study on News Recommendations Beyond Accuracy

Thomas E. Kolb^{1,*,\dagger}, Irina Nalis^{1,\dagger} and Julia Neidhardt¹

¹Christian Doppler Laboratory for Recommender Systems, TU Wien, Vienna, Austria

Abstract

In the past, recommender systems were primarily focused on optimizing accuracy. However, in recent years, there has been an increasing awareness that considerations beyond accuracy are necessary. The definition of what constitutes a good recommendation is a crucial issue. The most precise prediction may not always be the recommendation that satisfies the user best. This study offers a comprehensive investigation into the present advancements within the realm of beyond-accuracy measurements, especially the metrics diversity, serendipity, and novelty. Collaborative efforts between algorithmic models and domain experts can enrich recommendation quality, particularly in labeling and categorizing content. To address this, we present a study conducted by experts in the news domain. This study provides new insights into the multifaceted nature of this challenge. Employing an interdisciplinary approach, we underscore the significance of constructing a system that revolves around the user. Recent discussions about algorithmic content filtering and its societal implications underscore the importance of maintaining human involvement in the decision-making loop.

Keywords

recommender systems, beyond-accuracy measures, domain-expert study

1. Introduction

Recommender systems have traditionally focused on accuracy, aiming to predict how users rate items based on their past preferences. However, in recent years, there has been a growing recognition that assessing recommender systems based solely on accuracy is insufficient. Beyond-accuracy measures, such as novelty, diversity, serendipity, and coverage, have emerged as crucial dimensions for evaluating recommender systems and attract an increasing number of studies [1]. News recommender systems face several challenges and potential problems that impact their ability to promote diversity, novelty, and serendipity. To avoid the problem of filter bubbles [2] requires designing recommender systems that prioritize presenting a variety of viewpoints and topics to users, rather than solely relying on personalized recommendations based on past behavior.


11th International Workshop on News Recommendation and Analytics in conjunction with ACM RecSys 2023, September 18–22 2023, Singapore.


*Corresponding author.

^{\dagger}These authors contributed equally.

✉ thomas.kolb@tuwien.ac.at (T. E. Kolb); irina.nalis-neuner@tuwien.ac.at (I. Nalis); julia.neidhardt@tuwien.ac.at (J. Neidhardt)

ORCID 0000-0002-2340-0854 (T. E. Kolb); 0000-0001-7101-3229 (I. Nalis); 0000-0001-7184-1841 (J. Neidhardt)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Another research gap can be found in methodological approaches in the limited integration of user perspectives and real-world evaluations in recommender system studies. According to a survey of research papers on the performance of recommender systems, the vast majority of the studies evaluated algorithmic models exclusively through offline trials [3]. Few papers combined offline experiments with user research. This reveals a research gap in the studies of recommender systems' scant incorporation of user opinions and actual evaluations.

Integrating user feedback, such as explicit ratings or explicit indications of interest in specific topics or news sources, can enhance personalization while maintaining a balance with diverse and serendipitous recommendations. Exploring how different recommendation strategies affect user engagement and satisfaction is crucial. Developing algorithms that explicitly optimize for diversity, novelty, and serendipity involves considering not only the relevance and popularity of articles but also the diversity of perspectives, coverage of niche topics, and unexpected content that may pique user interest. This is where our interdisciplinary research that combines insights into user experience, and social sciences can contribute to a comprehensive understanding of the complexities and trade-offs involved in designing diverse and engaging news recommendations. For instance, Choi et al. (2022) [4] propose a novel news recommendation model that aims to provide personalized recommendations by considering a user's individual interest rather than relying solely on popular articles. Winecoff et al. (2019) [5] emphasize the importance of integrating psychological concepts into recommender systems (RS) by addressing the limitations of commonly used similarity functions and algorithm validations. Their research highlights the need to develop similarity functions that align with human cognition and to rigorously evaluate their performance through methodologically sound user testing.

In the design of recommender systems, it is crucial to consider users' psychology and incorporate domain expertise to ensure the systems meet users' needs and preferences [5, 6]. To address the limitations of traditional recommendation techniques and facilitate a more comprehensive evaluation, we propose an approach that investigates beyond accuracy measures with domain experts [5, 6]. In this workshop submission, we introduce the results of a labeling study with experts with profound domain knowledge, as they work as editors at the news outlet that has been researched. This approach enables a more systematic study of beyond-accuracy measures in the specific context of news domains.

Therefore, our study also answers calls to integrate the user perspective into the development of beyond-accuracy measures. For instance, some authors [7, 8] underline the necessity to closely observe the process in which the individual is confronted with the recommendation and whether for instance one is at the beginning of a decision-making process or already decided. Depending on the decision phase, the recommender system should consider entirely different sets of items to cater to these perspectives. Moreover, specifics of the news domain and user needs for information, entertainment [8] need to be considered.

Within our study we aim to extract multi-facet feature sets that capture different aspects of user preferences and behavior toward reading recommendations in the news domain in reference to the beyond-accuracy paradigm e.g., [9]. Therefore, editors of a news outlet were invited to participate in a labeling study on ideal reading recommendations from the perspective of diversity, novelty, and serendipity. The research design for this study integrates domain expertise to improve automated categorization, therefore it is built around an expert survey and labeling study to evaluate reading recommendations in recommender systems. By involving

relevant stakeholders, in our case editors of the news outlet that is being investigated, this study aims to overcome the limitations of previous approaches and contribute to the development of recommender systems that align with users' needs and values. Moreover, it seeks to harvest domain expertise that expands a mere user-centric approach, which is described in human-in-the-loop approaches, yet oftentimes comes at the cost of overlooking domain expertise [10]. The integration of domain expertise is crucial in building models when labeling texts with categories requires specialized knowledge and only limited annotated data are available [6]. Moreover, as demonstrated by Han et al. (2022) [6] collaborative efforts with domain experts can improve automated categorization by leveraging their understanding of categories and their confidence in annotation. Hence, our study provides novel insights into the design and development process.

This submission particularly explores the concept of serendipity in recommender systems. Serendipity refers to surprising or unexpected discoveries that are still relevant to users [1, 11]. Although serendipity has attracted considerable research interest, designing for serendipity remains a challenge [1]. The current interpretation of serendipity as a narrow evaluation metric for algorithmic performance limits our understanding and hinders the ability to design for serendipity. According to Smets et al. (2022) [1], serendipity ought to be viewed as a user experience rather than just an offline evaluation statistic like novelty or diversity. As a result, a user-centric analysis is beneficial to any attempt to research serendipity in recommender systems. The following topics are being addressed:

- Content analysis and model performance: Evaluation of ranking and content analysis regarding the news article resort diversity.
- Emerging characteristics of beyond accuracy measures: Which characteristics of a reading recommendation are ideally present?
- User behavior and domain-specific preferences: What does an ideal reading recommendation look like?

Our research aligns with the broader field of Digital Humanism [12], which emphasizes the importance of striving for beyond-accuracy objectives and fairness in software programs and algorithms [13]. By focusing on beyond-accuracy measures, we aim to enhance the quality and usefulness of recommender systems while promoting fairness and user-centric design principles.

In the following sections, we will discuss the concepts of diversity, serendipity, and novelty, highlighting their importance in evaluating recommender systems. We will also present our experimental design for evaluating the influence of different features on serendipitous encounters. This work represents a step towards a more integrated and user-centric approach to beyond-accuracy measures in recommender systems.

2. Related Work

2.1. Fairness and Bias

The relevance of an item depends on various factors, including the consumer's goals, situational context, and the specific purpose of the recommender system from different stakeholders'

viewpoints. The current focus on optimizing accuracy measures using historical data may not adequately capture the true value and relevance of recommended items. Therefore, bias is a field that currently attracts increasing awareness as fairness towards the representation of items but also towards user preferences and stakeholder perspectives (e.g., news outlets) need to be better represented. However, a recent review [14] criticizes the fragmented and inconsistent nature of existing studies on bias in recommender systems, calling for a systematic survey and taxonomy to organize research on recommendation debiasing. Existing research has focused on addressing this issue by increasing the coverage of long-tail items. To further advance the development of metrics of news recommenders, Smets et al. (2022) [1] emphasize the importance of considering multiple stakeholders in the design of news recommenders, revealing that their development involves a negotiation process among actors beyond just users. Their findings call for an expanded framework that accounts for preconditions, product owners, and indirect stakeholder involvement, offering a more comprehensive understanding of news recommenders.

2.2. Beyond Accuracy Paradigm

Beyond accuracy measures User behavior and decision-making are significantly influenced by recommender systems, which highlights the necessity for ethically sound designs that uphold democratic principles. The significance of making sure that these platforms, which enable users to interact with the vast amount of information online, uphold the values of the cultures and people they are utilized [15]. Although many researchers have attempted to evaluate their methodologies in recent years using criteria other than accuracy (e.g., innovation, diversity, serendipity, or coverage), there are still a number of significant problems that need to be resolved [16]. Usually, in static settings, relatively basic user models are taken into account, and beyond-accuracy solutions are not tailored to the requirements or preferences of particular users or groups of users, nor are they adjusted to different domains. Therefore, a rising number of people are calling for these technologies to be evaluated for more than just accuracy (e.g., [9]). Under the umbrella concept of beyond accuracy, several aspects are frequently discussed in order to prevent potentially harmful effects from personalized recommendations that have been shown to lead to filter bubbles [2]. Frequently discussed are diversity, novelty, and serendipity [17]. Notwithstanding the necessary critique and call for beyond accuracy-measures, recommender systems should continue to focus on accuracy since it helps to lower prediction mistakes. For instance, it appears vital in many instances to prevent inaccurate recommendations because, among other things, this could affect how well the system is viewed [18]. Therefore, it is crucial to develop measurements that allow for a sensible trade-off between accuracy and other criteria for good recommendations.

Diversity Diversity in news recommendation is a topic of significant research interest and has spurred research on the challenge of remaining relevant to user interests, hence avoiding content that is too diversified or irrelevant to their preferences while yet offering sufficient variety, e.g. introducing users to new topics and categories. In order to keep users interested, an NRS must strike a balance between remaining relevant to their interests, hence avoiding content that is too diversified or irrelevant to their preferences while yet offering sufficient

variety, e.g. introducing users to new topics and categories. Raza and Ding (2023) [19] present a deep neural network, that meets the needs of users to obtain information in topics in which they have shown interest before, but that goes beyond accuracy metrics. Lee and Lee (2022) [20] explored the role of perceived personalization and news diversity in news recommendation services with a focus on the efficacy of news diversity and trust in NRS. They looked at how users' intentions for remaining around were affected by perceived personalization and news diversity. Through the mediation of trust, user enjoyment, and perceived utility, they discovered in their study that perceived personalization significantly affected continuance intention.

Novelty The term novelty often refers to whether something is new. When recommendations include products or topics the user was previously unfamiliar with, they are seen as being more helpful [21]. Users, however, can differ in terms of their needs for novelty or variety due to differences in their personalities and interests. In the research on recommender systems, novelty is often described in two conceptually distinct ways [17]. The first method takes into account whether a person is familiar with a certain item. Clearly, it's challenging to capture this. Furthermore, novelty is oftentimes defined as the antithesis of popularity, hence it could be applied as a measure to counterbalance popularity bias [22]. Moreover, novelty is closely related to serendipity, together both possible beyond-accuracy measures come with the challenge to find a good balance between relevance and positive surprise.

Serendipity The concept of serendipity is discussed by Smets et al. (2022) [1] from the perspective of its potential to help to mitigate popularity bias and to increase the usefulness of recommendations by enabling better discoverability. In their research in line with the beyond accuracy paradigm on diversity, coverage and serendipity, Smets et al. (2022) [1] present a feature repository, highlight the importance of integrating metadata, user interface, and information access in facilitating serendipitous experiences. In addition, evidence for the potential of serendipity has been investigated by Niu and Al-Doulat (2021) [23] regarding the use of surprise in improving user satisfaction and inspiring curiosity in recommender systems in the health domain. Ziarani and Ravanmehr (2021) [24] present a systematic literature review on serendipity in recommender systems, exploring various aspects and approaches. Chen et al. (2021) [25] examine the values of user exploration in recommender systems, emphasizing the importance of serendipity in addressing the cold-start and filter-bubble problems. Furthermore, Abdollahpouri et al. (2021) [13] focus on news recommender systems and discuss the direction toward the next generation of these systems. Another issue, particularly relevant in designing for serendipity is to be found in the cold start problem. A problem, that could be addressed via a model proposed by Xu et al. (2022) [26] in which a self-serendipity recommender system to address the cold-start problem and provide diverse but relevant recommendations is presented.

3. Editor Study

This collaborative study with Falter Verlagsgesellschaft m.b.H.¹ (FALTER) focuses on investigating the needs for diversity, serendipity, and novelty in recommender systems. The research

¹<https://www.falter.at/>

aims to understand the variations of these needs within the news domain. It further explores how bias co-evolves with various outcome measures, including beyond-accuracy objectives. Specifically conducted within the news domain, the study examines the weekly newspaper FALTER, which publishes a wide range of news items both in print and online formats. By analyzing the content, meta-data, and additional news formats, the research aims to enhance the recommendation process, aligning it more effectively with the preferences and expectations of users.

An expert-in-the-loop approach was chosen to answer the raised questions. This is especially important as recent research has shown that often multiple different stakeholders are involved in such a system [27]. In addition, a survey on their personal perception of the importance of diverse, novel, and serendipitous reading recommendations was issued. Moreover, the editors were invited to answer an open-ended question on what an ideal reading recommendation would look like. In total twelve editors participated in the study. The editors were chosen in a way to represent the whole news medium by covering all important areas (feuilleton, politics, city-life, nature, economy, etc.).

3.1. Data

The data corpus provided by FALTER contains more than 100.000 news articles. Due to FALTER being a weekly news paper with a focus on investigative journalism these corpus consists of high quality texts. However, it is important to note that the texts are exclusively in German. FALTER employs a monetization model based on subscriptions, which results in the majority of their articles being concealed behind a paywall. To enhance the reproducibility of this approach we have provided access to the articles utilized for this study within a GitHub repository². The repository grants access to both content and metadata of 875 news articles. These articles were utilized in either of two ways: as candidate items or as baseline recommendations for the scope of this study.

3.2. Candidate Item Selection & Baseline Recommendations

A representative set of candidate items was created based on the 15 most recent significant stories of each of the participating editors. It is important to note that columns were excluded from this selection process. This process was supported by a domain expert from FALTER. The resulting list consists of 12 (editors) x 15 (candidate items) = 180 articles³. Pyserini [28] is utilized to generate recommendations for each of the 168 candidate articles. A traditional lexical model (BM-25) is employed to capture the essence of Lucene-based tools, such as Elastic Search⁴, which are widely employed in the industry. The recommendations are generated based on the corpus presented beforehand. Recommendations are considered valid if recommended news article was published within the past 365 days, using a snapshot of the corpus dated 2022-12-12. The initial three paragraphs of the query items serve as the search query parameter for the recommendation process, employing the default parameters of the BM-25 algorithm within the

²<https://github.com/ThomasEKolb/an-expert-study-on-news-recommendations-beyond-accuracy>

³Actually, a total of 168 articles were chosen because the editors did not consistently have 15 significant articles each.

⁴<https://www.elastic.co/elasticsearch/>

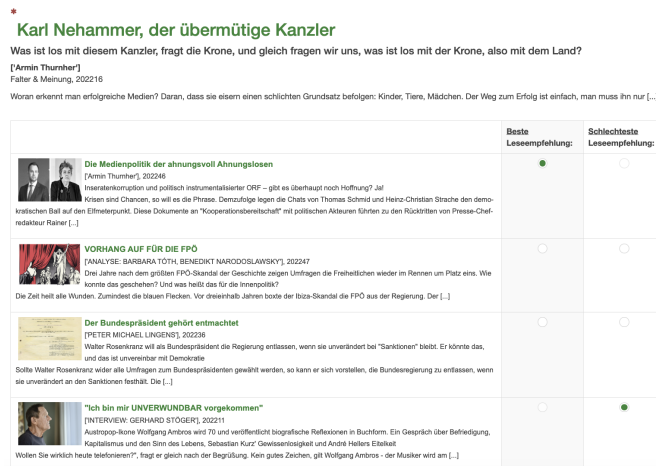


Figure 1: Editor study - BWS

Pyserini framework. A set of 15 items is recommended for each candidate article, resulting in a total of 2520 recommended items.

3.3. Survey Implementation

In addition to asking the editors about their preferences towards the beyond-accuracy metrics like diversity, novelty, and serendipity as also about their view of a good reading recommendation, each editor was requested to rank 15 recommendations for 15 of their own articles. Typically, assigning a ranking to a collection of 15 items proves to be challenging. Thus, we used a ranking technique known as best-worst scaling (BWS), which increases the agreement among annotators compared with alternative methodologies [29]. The tuples for the survey were created by using a script⁵ provided by Kiritchenko & Mohammad alongside their publication [30]. The parameters were set as following: “factor: 2”, “4 items per tuple”, and a total of “100 iterations”. Figure 1 shows one of the surveys that was generated based on the provided tuples. At the top, the query article is displayed, featuring its headline, subtitle, and a concise text excerpt. The headline is linked to the full version of the article. Directly beneath the query article, a list presents four suggested articles (= one tuple created following the BWS methodology). The editor has the option to designate one article as the “best” choice and another as the “worst” option for a good reading recommendations in relation to the query article. As survey tool LimeSurvey⁶ was used. For each editor a dedicated survey was programmatically created. LimeSurvey offers great flexibility in configuring an interface for the BWS approach.

Once all editors completed their survey, the scores were calculated based on the collected annotations using counts analysis. This variation of BWS computes scores for each item by subtracting the percentage of times it was chosen as the worst from the percentage of times it was chosen as the best, resulting in a score ranging from -1 to +1. This allows to create a more

⁵<https://saifmohammad.com/WebPages/BestWorst.html>

⁶<https://www.limesurvey.org>

robust and reliable article ranking for the given query articles. The outcomes of the various stages are available in the repository linked to this work. To identify the quality of the labeled data the split-half reliability was calculated for each editor.

3.4. Analysis

By utilizing ClayRs [31], a Python framework for evaluation, we compared the correlation between BM-25 based recommendations and the outcomes of the editor study. The framework's evaluation module⁷, facilitates the use of diverse metrics to analyze the results. Furthermore, this framework is specifically crafted to enhance the replicability of the research conducted. Besides the correlation metrics, the comparison of the two article rankings involves an examination of Jaccard similarity (intersection over union) among the article resort. This analysis enables the exploration of the degree of similarity in the article resort present within the top positions of both lists.

We presented the expert raters with an open-ended question on their individual perception of an ideal reading recommendation. In addition, with a description of diversity, novelty, and serendipity, and asked to rank for themselves their individual preferences for either of these beyond-accuracy measures for ideal reading recommendations. By applying a multi-method approach of a labeling study in combination with a survey, we were able to mitigate several shortcomings of research on recommender systems as described by [5].

4. Results

One key takeaway is the relatively high prioritizing of serendipity in comparison with the two other beyond-accuracy measures. In the expert's study, a component of the survey was included to solicit opinions from editors on the best reading suggestions. Numerous editors stressed the value of serendipity and the investigation of fresh knowledge and ideas in their written comments. They reported a need for book suggestions that would broaden their knowledge, pique their curiosity, and provide them with a compelling reason to keep reading. Additionally, in order to assess beyond-accuracy metrics, the editors were required to rank the significance of a reading recommendation's diversity, novelty, and serendipity on a separate basis. This study emphasizes how crucial it is to include serendipity as a fundamental component of recommender systems in the news domain. While accuracy is unquestionably important, offering suggestions that go beyond mere accuracy and encourage chance meetings can increase user happiness and engagement. The ideal reading recommendation should be like a talented DJ who creates a mix that surprises and engages the listener, as one editor so eloquently put it. Recommender systems can more effectively achieve the targeted results of expanding knowledge, arousing interest, and introducing users to cutting-edge concepts and material by incorporating the editors' insights and giving serendipity priority. The editors' answers emphasize the value of serendipity in the context of news suggestions, highlighting its usefulness for fostering an interesting and educational reading experience.

⁷<https://swapuniba.github.io/ClayRS/evaluation/introduction/>

Table 1

BM-25 vs. BWS expert labeling results - Spearman correlation (SC)

	SC@3	SC@5	SC@10	SC@15
all articles (mean)	0.208	0.221	0.235	0.196

Table 2

Jaccard similarity (JS) over all ranked list pairs (JS) based on the article ressorts

	JS@3	JS@5	JS@10	JS@15
mean	0.626	0.681	0.798	1.0
std	0.302	0.257	0.195	0.0

According to the findings presented by Kiritchenko and Mohammad [30], the split-half reliability of the BWS conducted with the editors was assessed. The average Spearman correlation coefficient, computed for each query article and its corresponding recommendations, is 0.785, with a standard deviation of 0.092. These results strongly indicate that the editors exhibited a high level of consistency during the labeling task. Another crucial aspect is the evaluation of the BM-25 based recommendations and the comparison of their ranking with the article ranking created by the editors during the labeling task. Table 1 highlights that there is certain low positive correlation between the two lists. This highlights how important such labeled datasets are. It is not enough to just use “another” recommendation algorithm and apply it on a dataset. The outcomes from calculating the mean of the Jaccard similarity⁸ and the corresponding standard deviation across all pairs of ranked lists emphasize this observation. There is a certain ressort wise overlap which increases if more items within the list are considered by beginning from the top three until the full list of 15 items.

5. Discussion

This study goes beyond the conventional approaches in recommender systems by raising the importance of incorporating individual-level factors and user characteristics into the recommendation process. Unlike previous works that mainly focused on independent individuals without considering social context or domain specific expertise, this research acknowledges the domain knowledge of news editors in the investigation of news reading recommendations. The findings underscore the effectiveness of integrating an expert study with BWS and the LimeSurvey platform for efficiently acquiring domain knowledge. To overcome the limitations of previous studies, we integrated forced choice and Best-Worst-Scaling methods in our study, to address the limitations of traditional single-item approaches and provide a more robust and context-aware assessment of users’ judgments of reading recommendations.

⁸List pairs with missing ressort data were excluded from the Jaccard similarity calculation.

6. Future Work

Future studies could deepen these results by drawing upon psychological theories and empirical studies, help to understand why individuals have certain preferences and how these preferences are associated with contextual information, personality, and demographic characteristics. The insights into the editors' perspective on the beyond-accuracy measurements raise the importance of further investigating serendipity. Furthermore, considering this as a multi-stakeholder issue, it is essential to incorporate the audience's viewpoint to perform a comprehensive comparison of diverse opinions regarding these metrics. For instance, via A/B tests, which could be designed as controlled field experiments. Moreover, performing topic extraction on the news articles presents an additional opportunity for enhancing the depth of understanding within the labeled data.

Acknowledgments

This research is supported by the Christian Doppler Research Association (CDG).

References

- [1] A. Smets, L. Michiels, T. Bogers, L. Björneborn, Serendipity in recommender systems beyond the algorithm: A feature repository and experimental design, in: 16th ACM Conference on Recommender Systems. CEUR Workshop Proceedings, 2022, pp. 44–66.
- [2] L. Michiels, J. Leysen, A. Smets, B. Goethals, What are filter bubbles really? a review of the conceptual and empirical work, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 274–279.
- [3] D. Jannach, C. Bauer, Escaping the mcnamara fallacy: towards more impactful recommender systems research, *AI Magazine* 41 (2020) 79–95.
- [4] S. Choi, H. Kim, M. Gim, Do not read the same news! enhancing diversity and personalization of news recommendation, in: Companion Proceedings of the Web Conference 2022, 2022, pp. 1211–1215.
- [5] A. A. Winecoff, F. Brasoveanu, B. Casavant, P. Washabaugh, M. Graham, Users in the loop: a psychologically-informed approach to similar item retrieval, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 52–59.
- [6] K. Han, R. Rezapour, K. S. Nakamura, D. Devkota, D. C. Miller, J. Diesner, An expert-in-the-loop method for domain-specific document categorization based on small training data, *Journal of the Association for Information Science and Technology* 74 (2022) 669 – 684.
- [7] M. Jesse, D. Jannach, Digital nudging with recommender systems: Survey and future directions, *Computers in Human Behavior Reports* 3 (2021) 100052.
- [8] D. Jannach, M. Quadrona, P. Cremonesi, Session-based recommender systems, in: *Recommender Systems Handbook*, Springer, 2022, pp. 301–334.
- [9] D. Jannach, Evaluating conversational recommender systems: A landscape of research, *Artificial Intelligence Review* 56 (2023) 2365–2400.

- [10] N. Sambasivan, R. Veeraraghavan, The deskilling of domain expertise in ai development, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–14.
- [11] L. Björneborn, Three key affordances for serendipity: Toward a framework connecting environmental and personal factors in serendipitous encounters, *Journal of documentation* (2017).
- [12] H. Werthner, E. Prem, E. A. Lee, C. Ghezzi, *Perspectives on Digital Humanism*, Springer Nature, 2022.
- [13] H. Abdollahpouri, E. C. Malthouse, J. A. Konstan, B. Mobasher, J. Gilbert, Toward the next generation of news recommender systems, in: Companion Proceedings of the Web Conference 2021, 2021, pp. 402–406.
- [14] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, *ACM Transactions on Information Systems* 41 (2023) 1–39.
- [15] J. Stray, A. Halevy, P. Assar, D. Hadfield-Menell, C. Boutilier, A. Ashar, L. Beattie, M. Ekstrand, C. Leibowicz, C. M. Sehat, et al., Building human values into recommender systems: An interdisciplinary synthesis, *arXiv preprint arXiv:2207.10192* (2022).
- [16] D. Jannach, P. Pu, F. Ricci, M. Zanker, Recommender systems: Trends and frontiers, *AI Magazine* 43 (2022) 145–150.
- [17] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems, *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7 (2016) 1–42.
- [18] F. Fouss, E. Fernandes, A closer-to-reality model for comparing relevant dimensions of recommender systems, with application to novelty, *Inf.* 12 (2021) 500.
- [19] S. Raza, C. Ding, Relevancy and diversity in news recommendations *, in: CEUR WORKSHOP PROCEEDINGS, volume 3411, 2023, pp. 6–15.
- [20] S. Y. Lee, S. W. Lee, Normative or effective? the role of news diversity and trust in news recommendation services, *International Journal of Human–Computer Interaction* (2022) 1–14.
- [21] B. Alhijawi, A. A. Awajan, S. Fraihat, Survey on the objectives of recommender systems: Measures, solutions, evaluation methodology, and new perspectives, *ACM Computing Surveys* 55 (2022) 1 – 38.
- [22] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The connection between popularity bias, calibration, and fairness in recommendation, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 726–731.
- [23] X. Niu, A. Al-Doulat, Luckyfind: Leveraging surprise to improve user satisfaction and inspire curiosity in a recommender system, in: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, 2021, pp. 163–172.
- [24] R. J. Ziarani, R. Ravanmehr, Serendipity in recommender systems: a systematic literature review, *Journal of Computer Science and Technology* 36 (2021) 375–396.
- [25] M. Chen, Y. Wang, C. Xu, Y. Le, M. Sharma, L. Richardson, S.-L. Wu, E. Chi, Values of user exploration in recommender systems, in: Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 85–95.
- [26] Y. Xu, E. Wang, Y. Yang, H. Xiong, GS²-rs: A generative approach for alleviating cold start

and filter bubbles in recommender systems, *IEEE Transactions on Knowledge and Data Engineering* (2023).

- [27] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Multistakeholder recommendation: Survey and research directions, *User Modeling and User-Adapted Interaction* 30 (2020) 127–158. URL: <https://doi.org/10.1007/s11257-019-09256-1>. doi:10.1007/s11257-019-09256-1.
- [28] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations, in: *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, 2021, pp. 2356–2362.
- [29] S. Kiritchenko, S. Mohammad, Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 465–470. URL: <https://aclanthology.org/P17-2074>. doi:10.18653/v1/P17-2074.
- [30] S. Kiritchenko, S. M. Mohammad, Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling, in: *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California, 2016.
- [31] P. Lops, C. Musto, M. Polignano, Semantics-aware content representations for reproducible recommender systems (score), in: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 354–356. URL: <https://doi.org/10.1145/3503252.3533723>. doi:10.1145/3503252.3533723.