Disentangling the Performance Puzzle of Multimodal-aware Recommender Systems

Daniele Malitesta[1,*], Giandomenico Cornacchia[1,*], Claudio Pomo[1] and Tommaso Di Noia[1]

[1]*Politecnico di Bari, via Orabona, 4, 70126 Bari, Italy*

Abstract

In domains such as fashion, music, food, and micro-video recommendation, items' representation can be suitably enhanced through multimodal side information (extracted from images, texts, or audio). Multimodal-aware recommender systems (MRSs) refer to the family of RSs which integrates extracted multimodal items' features into the classical recommendation pipeline to learn users' fine-grained taste towards these aspects, thus boosting the recommendation accuracy. As a general trend, recent multimodal recommendation approaches tend to share similar strategy patterns with few variations on the central theme. This does not allow, in most cases, to easily interpret their accuracy recommendation improvements, also because these models are usually evaluated under different implementations. To bridge such an evaluation gap, in this paper, we propose one of the first benchmark analyses on the performance of MRSs by comparing five popular and recent approaches on widely-adopted recommendation datasets. After introducing the helpful background notions and formulations for a multimodal-aware recommendation, we first study the models' accuracy in a unified experimental framework and hyper-parameter setting. Then, differently from the existing works, we provide a new evaluation perspective by assessing their recommendation performance under the lens of novelty and diversity of recommendations. Besides confirming some of the observations from the related literature, results shed light on unexpected findings, which show how a careful hyper-parameter tuning can make shallow and less recent approaches quite competitive against the state-of-the-art ones.

Keywords

Recommender Systems, Multimodality, Evaluation

1. Introduction

Online platforms for e-commerce (e.g., Amazon), media streaming (e.g., Netflix), social networks (e.g., Instagram), and traveling (e.g., Booking) currently host a large amount of digital data, such as images, texts, and videos. Recommender systems (RSs), which work under the hood of such platforms to learn user-item preference patterns and promote a personalized user experience, have started to leverage such content to enhance the quality of their recommendations.

In this respect, the literature [1] has introduced a novel family of RSs, called multimodal-aware recommender systems (MRSs). Indeed, it has been widely demonstrated that in specific scenarios

2nd Edition of EvalRS: a Rounded Evaluation of Recommender Systems, August 6 - August 10, 2023, Long Beach, CA, USA

*Corresponding authors: Daniele Malitesta and Giandomenico Cornacchia.

daniele.malitesta@poliba.it (D. Malitesta); giandomenico.cornacchia@poliba.it (G. Cornacchia); claudio.pomo@poliba.it (C. Pomo); tommaso.dinoia@poliba.it (T. Di Noia)

© 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)

CEUR-WS.org/Vol-3450/paper2.pdf

such as fashion [2], music [3], food [4], and micro-video [5] recommendation, MRSs may benefit from *multimodal* side information to tackle (i) the sparsity of the user-item matrix, (ii) the difficulty of learning preference patterns for users and items with few recorded interactions (i.e., cold-start scenario), and (ii) the inexplicable nature of users' actions (e.g., clicks, views) which are implicit and may not be easily profiled.

By leveraging pre-trained deep learning models for image classification [6], text sentiment analysis [7], and audio classification [8], MRSs inject high-level multimodal features of items extracted through such deep networks into the recommendation pipeline to augment items' embedded representations, thus learning finer-grained users' preferences. Starting from the pioneer work by He and McAuley [9], which integrated only visual features of product images into BPR-MF [10], several approaches have later adopted similar solutions, with many noteworthy attempts optionally introducing the textual and audio modalities to learn more refined users and items latent representations [1, 11, 12, 13, 14].

Despite their indisputable success as recognized in the recent literature, mainly thanks to their recommendation accuracy improvements to the existing baselines, some performance evaluation concerns still raise. Indeed, as most of these methods follow similar strategy patterns with few variations on the main theme, it can be challenging to unveil which technique is actually providing the most impactful contribution to the recommendation performance. Additionally, most of such approaches are trained and evaluated under different implementations, which come with their own data splitting/sampling, hyper-parameter searching, and evaluation protocols.

To address such an evaluation gap in the literature, our work aims to provide the first extensive benchmarking setting for the multimodal-aware recommendation. Specifically, our contributions are threefold:

- We provide a unified framework to benchmark five state-of-the-art multimodal-aware recommender systems (i.e., VBPR [9], MMGCN [1], MGAT [11], GRCN [12], and LATTICE [13]) on three popular recommendation datasets from the Amazon catalog [15] (i.e., Office, Toys, and Clothing).

- We run extensive hyper-parameter explorations to fine-tune all tested models under the same settings for a fair comparison. While confirming some findings from the existing literature, results also show how careful hyper-parameter tuning can make even shallow approaches (e.g., VBPR) quite competitive against more complex and recent ones (e.g., GRCN).

- In addition to assessing recommendation accuracy, we complement the evaluation through an analysis of the novelty [16, 17] and diversity [18] of the produced lists of recommendation. To the best of our knowledge, this work is the first to perform this analysis in the domain of multimodal-aware recommendation.

The rest of the paper is organized as follows. First, we depict the related literature on multimodal-aware recommendation and novelty/diversity in recommendation (Section 2). Then, in Section 3, we provide the useful background notions about the (multimodal-aware) recommendation problem, and report the main formulations of the tested MRSs for our study. Afterwards, we present our proposed analysis in Section 4, where we report details about the adopted experimental settings to conduct our investigation. Finally, we show and discuss about

the obtained results in Section 5, and define the directions to follow in future directions of this analysis in Section 6. To foster the reproducibility of our benchmark, we share the code and datasets adopted in this work: https://github.com/sisinflab/MultiModal-Eval.

2. Related Work

This section provides an outlook on multimodal-aware recommendation and focuses on the evaluation of recommendation algorithms under the lens of novelty/diversity.

2.1. Multimodal-aware recommendation

Independently on the domain and task (e.g., fashion [19, 20, 21], music [22, 3, 23], food [4, 24, 25], and micro-video [1, 26, 27] recommendation) the common rationale in multimodal-aware recommendation is about augmenting users and/or items representation through their *multimodal* profiles.

Indeed, multimodal learning [28, 29] has demonstrated to be beneficial by tackling known issues in the recommendation community such as data sparsity and cold-start [9, 3, 30]. Additionally, multimodal content may help to unveil the possible intents behind implicit user-item interactions [31] through attention mechanisms [20, 32, 11, 33, 34] for the sake of explainability.

Given the increasing impact graph neural networks [35, 36] have had on recommendation [37, 38], several solutions introduce multimodality on nodes of the user-item graphs (but also knowledge graphs [39]) and refine such representations through the message-passing schema. After some initial attempts which work by simply injecting the multimodal item features into the graph-based recommendation pipeline [40], more advanced techniques propose to learn separate graph representations for each modality [1] and disentangle the users' preferences at modality level [11, 12, 14]. Finally, recent approaches are trained to uncover multimodal structural differences among items in the catalog [13, 41, 42].

Despite showing superior accuracy recommendation performance, we recognize a critical issue in the training and evaluation of such approaches. Concretely, each of these models usually come with their own implementations, meaning that they perform custom data splitting/sampling, hyper-parameter searching, and evaluation protocols. To this end, we provide a unified experimental framework to train and test five of the most popular and recent multimodal-aware recommender systems (i.e., VBPR [9], MMGCN [1], MGAT [11], GRCN [12], and LATTICE [13], refer to Table 1) under the same shared settings.

2.2. Novelty and diversity in recommendation

User experience plays a crucial role in recommendation platforms, as highlighted by several academic studies [55, 56, 57]. Such works emphasize that suggesting interesting lists of items satisfies users and encourages them to remain loyal to the platform, ultimately leading to increased profits [58]. To ensure a good user experience, the recommended items must be nontrivial, diverse, possibly unexpected [57, 59], fair [60, 61], and explainable [61, 62]. However, designing dedicated models for recommendation systems presents significant challenges, mainly because evaluating them requires conducting user studies.

Table 1
An overview on the selected multimodal-aware recommender systems, along with their publication venue and year, and a non-exhaustive set of papers where they are used as baselines.

Models	Venue	Baseline in
VBPR [9]	AAAI 2016	[20, 43, 32, 11, 13, 44]
MMGCN [1]	MM 2019	[39, 45, 46, 21, 47, 48]
MGAT [11]	IPM 2020	[49, 50]
GRCN [12]	MM 2020	[51, 13, 44, 47, 48, 52]
LATTICE [13]	MM 2021	[47, 48, 53, 14, 54]

Consequently, researchers have invested substantial effort in exploring beyond-accuracy dimensions within the field of recommendation systems over the past two decades [18, 17, 63]. These dimensions refer to aspects beyond the traditional accuracy metric, aiming to improve the overall user experience.

While user experience has been a crucial aspect when evaluating multimodal-aware intelligent systems [64, 65, 66] for years, in recommendation it has gained attention only recently [67, 68]. As for multimodal-aware recommendation, most of the research efforts have focused on emphasizing the advantages of multimodal recommendation models in addressing the cold start user problem [9, 3, 69]. However, to the best of our knowledge, there is a lack of recent scientific literature that explicitly considers the impact of multimodality on user experience in terms of novelty and diversity of the produced recommendations. To this end, our work stands first and foremost as an attempt to bridge such an evaluation gap.

3. Background

This section first provides notations and formulations about the general recommendation problem. Then, we introduce the multimodal-aware recommendation problem by describing our study's five selected multimodal baselines.

3.1. Recommendation problem

Let us denote with $u \in \mathcal{U}$, $i \in \mathcal{I}$, and $r \in \mathcal{R}$ a user, an item, and their recorded interaction (if any), respectively. Given $\rho(\cdot)$ as the preference score prediction function for each user-item pair, a recommender system (RS) aims to build, for each user, a top-n list of items maximizing the following posterior probability ($prob$):

$$\hat{\Theta}_\rho = \arg\max_{\mathbf{\Theta}_\rho} prob(\mathbf{\Theta}_\rho | \, \mathcal{U}, \mathcal{I}, \mathcal{R}), \tag{1}$$

where $\mathbf{\Theta}_\rho = [\theta_\rho^{(0)}, \theta_\rho^{(1)}, \ldots, \theta_\rho^{(|\mathcal{W}_\rho|-1)}]$ is the vector collecting all trainable weights for the inference function $\rho(\cdot)$, $\mathcal{W}_\rho = \{\theta_\rho^{(0)}, \theta_\rho^{(1)}, \ldots, \theta_\rho^{(|\mathcal{W}_\rho|-1)}\}$ is the set of such weights, and $|\mathcal{W}_\rho|$ its cardinality. For instance, in the case of latent factor models (e.g., matrix factorization [10]), the set of trainable weights $\mathcal{W}_\rho$ consists, in the minimal setting, of the user and item embeddings.

3.2. Multimodal-aware recommendation problem

Let us denote with $m \in \mathcal{M}$ a single modality to profile an item or the user's preference towards that modality. Generally speaking, m is one of $\{visual, textual, audio\}$ meaning that, for instance, a product image and description involve the $visual$ and $textual$ modalities, respectively, a song involves the $audio$ modality, and a video involves the $visual, textual, audio$ modalities.

By exploiting deep neural networks already pre-trained for other tasks (e.g., image classification, text sentiment analysis, and audio classification), the necessary initial step is to extract high-level features from the multimodal input data and inject them into the recommendation algorithm. Then, such additional multimodal embeddings may be either trained or fixed to provide a supervision signal for the model training in the end-to-end downstream task. Finally, the obtained multimodal representations could be fused into a single representation to refine the output of the preference prediction score function $\rho(\cdot)$. On such basis, we complement the vector collecting all trainable weights introduced in the previous section with an additional vector $\mathbf{\Theta}_m = [\theta_m^{(0)}, \theta_m^{(1)}, \ldots, \theta_m^{(|\mathcal{W}_m|-1)}]$ which collects all trainable weights referring to the modality $m \in \mathcal{M}$, where $\mathcal{W}_m = \{\theta_m^{(0)}, \theta_m^{(1)}, \ldots, \theta_m^{(|\mathcal{W}_m|-1)}\}$.

In the following, we summarize the main strategies implemented in each of the five selected multimodal-aware recommendation systems.

3.2.1. VBPR

Visual-bayesian personalized ranking [9] (i.e., VBPR) is among the first attempts to bring visual-aware side information (e.g., high-level visual features extracted from product images [70, 71]) into the BPR-MF recommendation algorithm. In addition to the user and item *collaborative* embeddings, whose inner product estimates the interaction score, the authors introduce two *visual* user and item embeddings, where the latter is the high-level features extracted from product images through a pre-trained convolutional neural network. The two inner products coming from the *collaborative* and *visual* embeddings are summed to obtain the final prediction score. Even though VBPR is originally designed as a single-modality recommendation approach, we follow [13] and introduce additional user and item embeddings for each further modality in the same manner as the visual one.

3.2.2. MMGCN

Multimodal graph convolution network [1] (i.e., MMGCN) brings multimodality to graph-based recommendation. Specifically, the authors propose to learn a separate graph convolutional network for each considered modality, thus resulting in three user and item representations accounting for users' different attitudes towards each modality. Finally, all modality embeddings for both users and items are combined through element-wise addition, and the preference prediction score is obtained via inner product.

3.2.3. MGAT

Multimodal graph attention network [11] (i.e., MGAT) is a slight variation to the MMGCN framework, where the graph convolutional layer is augmented with the gate and attention mechanisms. The obtained multimodal representations are averaged, resulting in a single representation for users and items, which is eventually exploited for the score prediction.

3.2.4. GRCN

Graph-refined convolutional network [12] (i.e., GRCN) adopts the information conveyed by modalities to refine the entries of the adjacency matrix. Indeed, given the implicit nature of the user-item interactions in the bipartite graph, the idea is to find (and prune) the edges which may not correspond to the actual preferences of each user. The learned multimodal user and item representations are combined through concatenation to obtain a final representation for the preference score prediction.

3.2.5. LATTICE

Latent structure mining method for multimodal recommendation [13] (i.e., LATTICE) builds an item-item similarity graph for each modality and refines such a structure via graph structure learning. The updated adjacency matrices are combined through a weighted sum to give a different importance weight to each modality. The overall adjacency matrix is then used to refine item embeddings through a graph convolutional network. Finally, the learned item embeddings may represent any backbone based on user and item latent factors (e.g., BPR-MF).

4. Proposed analysis

This section describes our proposed analysis for multimodal-aware recommendation. First, we report on the adopted datasets, along with details about the extraction of multimodal features. Then, we introduce and formalize the set of evaluation metrics accounting for accuracy, novelty, and diversity of recommendation. Finally, we outline the details about reproducibility for our work, by providing information about dataset splitting and filtering strategies, and the hyper-parameter search.

4.1. Datasets

In our study, we conduct extended experiments on three popular review datasets from the Amazon catalog [15] to better generalize the insight derived from our analysis. The categories are: Office Products (**Office**), (b) Toys & Games (**Toys**), and (c) Clothing, Shoes & Jewelry (**Clothing**). Each dataset consists of both visual and textual modalities, where the former are made available by McAuley et al. [15]. Thus, in our analysis, we utilize the already pre-extracted 4,096-dimensional visual features which are publicly available[1]. For the textual modality, by following [13], we aggregate the item's title, descriptions, categories, and brand, thereby

[1]https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html.

Table 2
Statistics of the tested datasets.

Datasets	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{R}\rvert$	Sparsity (%)
Office	4,905	2,420	53,258	99.5513
Toys	19,412	11,924	167,597	99.9276
Clothing	39,387	23,033	278,677	99.9693

generating textual embeddings. In this regard, we leverage sentence transformers [72, 13], acquiring 1,024-dimensional sentence embeddings. Additional dataset information can be found in Table 2.

4.2. Evaluation metrics

The choice of appropriate evaluation metrics plays a crucial role in assessing the effectiveness of recommendation systems. In this work, and differently from the existing literature on multimodal-aware recommendation, we take into account metrics measuring the accuracy, along with the novelty and diversity of recommendation. The metrics are listed hereinafter. Note that all metrics are formulated in a way higher values stand for better performance.

4.2.1. Accuracy

In this study, we adopt the recall, normalized discounted cumulative gain, and precision as metrics for evaluating the *accuracy* performance calculated on top-k recommendation lists.
Recall. The recall measures the ability of the system to retrieve relevant items from the recommendation list, emphasizing the importance of comprehensive coverage with respect to the list of user interactions [73]:

$$\text{Recall@}k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\text{Rel}_u \text{ @}k|}{|\text{Rel}_u|}. \tag{2}$$

The term Rel_u indicates the set of relevant items for user u, while Rel_u @k is the set of relevant recommended items in the top-k list.
nDCG. The normalized discount cumulative gain (nDCG) takes into accounts both relevance and ranking position of recommended items, considering the varying degrees of relevance:

$$\text{nDCG@}k = \frac{1}{|\mathcal{U}|} \sum_{u} \frac{\text{DCG}_u\text{@}k}{\text{IDCG}_u\text{@}k}, \tag{3}$$

where $\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_{u,i}}-1}{\log_2(i+1)}$ quantifies the cumulative gain of relevance scores through the recommended list, with $rel_{u,i} \in \text{Rel}_u$, and IDCG represents the cumulative gain of relevance scores for a perfect (i.e., ideal) recommender system.
Precision. The precision (dubbed as "Prec" in the following) focuses on the quality of the recommendations by quantifying the proportion of relevant items among the recommended

ones, highlighting the system's ability to provide accurate suggestions:

$$\text{Prec@}k = \frac{1}{|U|} \sum_{u \in \mathcal{U}} \frac{|\text{Rel}_u \text{ @}k|}{k}. \tag{4}$$

Indeed, precision and recall are strictly related since both depend on the user relevance of the recommendation list, with the former normalized on recommender list cardinality and the latter on the user relevance list cardinality.

4.2.2. Novelty and diversity

To comprehensively evaluate the quality of recommendation, we also employ popular metrics to quantify the recommendation *novelty* and *diversity*. The selected metrics are the Expected Free Discovery (EFD) and Gini index, which are complemented by a measure of the *coverage* provided by the recommendation lists.

EFD. The EFD, proposed by Vargas and Castells [17], is a novelty metric that utilizes the *inverse collection frequency*. It quantifies the effectiveness of an algorithm in recommending relevant long-tail items.

$$\begin{aligned} \text{EFD@}k = C \sum_{i_k \in R} \text{disc}(k) P\left(\text{Rel}_u \text{ @}k \mid i_k, u\right) \cdot \\ \cdot \left(-\log_2 p(i \mid \text{ seen}, \theta)\right). \end{aligned} \tag{5}$$

Gini Index. The Gini index (dubbed as "Gini" in the following) quantifies the disparity in item popularity among users. It allows to measure if certain items are consistently favored by a large portion of users at the detriment of other ones. In this work, we use its normalized version according to which a high Gini index indicates a more balanced distribution of item recommendations, implying that a wide variety of items are being suggested to users. On the other hand, a low Gini index suggests that a few popular items are dominating the recommendations, leading to a less diverse recommendation experience [74]. Its formulation is:

$$\text{Gini@}k = 1 - \left(\frac{\sum_{i=1}^{|\mathcal{I}|} (2i - |\mathcal{I}| - 1) P|_{@k}(i)}{|\mathcal{I}| \sum_{i=1}^{|\mathcal{I}|} P|_{@k}(i)} \right), \tag{6}$$

where $P|_{@k}(i)$ is the popularity of each item i, in the top-k users recommended list, indexed in non-decreasing order (i.e., $P|_{@k}(i) \leq P|_{@k}(i+1)$).

Item Coverage. The item coverage (dubbed as "iCov" in the following) provides insights on the coverage (item-side) measured in the recommendation lists. A higher item coverage indicates that a larger portion of the item space is being explored and recommended to users, suggesting a broader coverage of user preferences and a potentially more comprehensive recommendation experience. Specifically, we have:

$$\text{iCov@}k = \frac{|\bigcup_u \hat{\mathcal{I}}_u @k|}{|\mathcal{I}_{train}|}, \tag{7}$$

where $\hat{\mathcal{I}}_u @k$ represent the list of top-k recommended items for a user u.

Table 3
Set of explored and fixed hyper-parameters for our study.

Models	Hyper-parameters	Values
All	*epochs*	200
	batch_size	1024
	seed	123
VBPR	*lr*	{1e-2, 1e-3, 1e-4, 3e-5, 1e-5}
	factors	64
	comb_mod	concat
	l_w	{1e-5, 1e-2}
MMGCN	*lr*	{1e-2, 1e-3, 1e-4, 3e-5, 1e-5}
	num_layers	3
	factors	64
	factors_multimod	(256, None)
	aggregation	mean
	concatenation	False
	has_id	True
	latent_factors	64
	l_w	{1e-2, 1e-5}
MGAT	*lr*	{1e-2, 1e-3, 1e-4, 3e-5, 1e-5}
	num_layers	2
	factors	64
	factors_multimod	(256, 100)
GRCN	*lr*	{1e-2, 1e-3, 1e-4, 3e-5, 1e-5}
	num_layers	2
	num_routings	3
	factors	64
	factors_multimod	128
	aggregation	add
	weight_mode	confid
	fusion_mode	concat
LATTICE	*lr*	{1e-2, 5e-3, 1e-3, 5e-4, 1e-4}
	n_layers	1
	n_ui_layers	2
	top_k	20
	l_m	0.7
	factors_multim	64

4.3. Reproducibility

To ensure reproducibility, we provide detailed information about the dataset preprocessing and splitting, models' tuning and evaluation.

The datasets have been filtered following the p-core strategy with $p = 5$. Then, we split the dataset with 80%/20% train-test hold-out strategy. For the hyper-parameter tuning phase, we remove the 50% of the test set and use it to validate the results on Recall@20. For all models, we fix the maximum number of epochs to 200 and select the model weights according to the epoch providing the best results on the validation set.

The complete set of hyper-parameters is reported in Table 3. The code for the entire pipeline (which is implemented in Elliot [75]) can be found at this link https://github.com/sisinflab/MultiModal-Eval.

5. Results and Discussion

In this section, we seek to answer the following research questions (i.e., RQs):

RQ1. *What is the accuracy performance of multimodal-aware recommender systems and is it aligned with the findings from the existing literature?* Section 5.1 aims to investigate the recommendation performance in terms of *accuracy* (i.e., Recall, nDCG, and Prec).

RQ2. *What is the recommendation performance of such models in terms of novelty and diversity of the produced lists of recommendation?* Section 5.2 unveils important insights in terms of novelty and diversity of recommendation (i.e., iCov, Gini, and EFD).

5.1. Accuracy performance (RQ1)

The results of the *accuracy* metrics analysis is reported in Table 4. As a general remark, we notice how the results are quite standard across the different datasets.

Overall, LATTICE is the best model, showing its superior performance across all the datasets and the metrics, while VBPR is the second-to-best model. Surprisingly, complex and recent approaches such as MMGCN, MGAT, and GRCN do not outperform a shallow and classic model such as VBPR. Conversely, LATTICE's results are aligned with the findings from the literature.

From a *dataset-wise* analysis, the highest metrics-performance variation between LATTICE and VBPR is observable for Toys and Clothing while it is limited in Office. Indeed, we should point out that Toys and Clothing have three times and four times the interactions of Office, respectively, but they are much sparser. This aspect highlights how LATTICE manages to recommend more accurate items despite the high dataset sparsity.

From a *metric-wise* analysis, LATTICE, compared to VBPR, correctly predicts relevant items (i.e., high Recall) with a higher probability to be in a top positions of the recommendation lists (i.e., nDCG). The same trend is not observable in the Recall/Prec metric pair, but this is explainable considering that the latter formulation is normalized as the number of recommended items increases. Thus, it can result in a lower performance variation between LATTICE and VBPR at the increase of k.

From a *model-wise* analysis, we notice how MMGCN has better performance on Toys while showing the lowest performance at the increase of interactions number and sparsity. GRCN has an opposite trend compared to MMGCN, boosting its performance with highly sparse data. MGAT performs in the middle of MMGCN and GRCN with no remarkable note.

Summary. *Accuracy results demonstrate that, with the only exception of LATTICE (whose trend is almost aligned with the existing literature) all other approaches' performance is heavily influenced by the hyper-parameter exploration and dataset characteristics. Indeed, even shallow models (e.g., VBPR) show competitive if not superior accuracy measures compared to more recent and complex solutions (e.g., MMGCN, GRCN).*

5.2. Novelty/diversity performance (RQ2)

Table 5 summarizes the results of the *novelty* and *diversity* metrics analysis. Overall, we observe that some trends are quite aligned with findings from the accuracy evaluation, but also that some other ones show deviations which we carefully describe and explain in the following.

Table 4
Accuracy results of the tested baselines when considering the top-10, top-20, and top-50 recommendation lists. **Boldface** and underline stand for best and second-to-best results on each dataset/metric pair, respectively.

Datasets	Models	$k = 10$			$k = 20$			$k = 50$		
		Recall	**nDCG**	**Prec**	**Recall**	**nDCG**	**Prec**	**Recall**	**nDCG**	**Prec**
Office	VBPR	0.0652	0.0419	0.0164	0.1025	0.0533	0.0133	0.1774	0.0721	0.0095
	MMGCN	0.0455	0.0300	0.0124	0.0798	0.0405	0.0109	0.1575	0.0598	0.0084
	MGAT	0.0427	0.0277	0.0119	0.0745	0.0377	0.0102	0.1450	0.0552	0.0079
	GRCN	0.0393	0.0253	0.0118	0.0667	0.0339	0.0099	0.1250	0.0488	0.0075
	LATTICE	**0.0664**	**0.0449**	**0.0173**	**0.1029**	**0.0566**	**0.0137**	**0.1780**	**0.0751**	**0.0096**
Toys	VBPR	0.0710	0.0458	0.0131	0.1006	0.0545	0.0096	0.1523	0.0667	0.0061
	MMGCN	0.0256	0.0150	0.0052	0.0426	0.0200	0.0044	0.0785	0.0285	0.0033
	MGAT	0.0375	0.0230	0.0072	0.0595	0.0294	0.0059	0.1039	0.0398	0.0043
	GRCN	0.0554	0.0354	0.0108	0.0831	0.0436	0.0083	0.1355	0.0559	0.0056
	LATTICE	**0.0805**	**0.0512**	**0.0148**	**0.1165**	**0.0617**	**0.0110**	**0.1771**	**0.0759**	**0.0069**
Clothing	VBPR	0.0339	0.0181	0.0034	0.0529	0.0229	0.0027	0.0847	0.0292	0.0017
	MMGCN	0.0227	0.0119	0.0023	0.0348	0.0150	0.0018	0.0609	0.0201	0.0012
	MGAT	0.0244	0.0127	0.0025	0.0384	0.0162	0.0019	0.0699	0.0225	0.0014
	GRCN	0.0319	0.0164	0.0032	0.0496	0.0209	0.0025	0.0858	0.0281	0.0017
	LATTICE	**0.0502**	**0.0275**	**0.0051**	**0.0744**	**0.0336**	**0.0038**	**0.1186**	**0.0425**	**0.0024**

On the one hand, in terms of recommendation *novelty* (i.e., EFD), we can see that LATTICE is the best model, with VBPR being the second-to-best approach in each dataset and for different settings of k. Indeed, this is a further demonstration on how accuracy and novelty of recommendation may be highly correlated [76, 77, 78], also in multimodal-aware recommendation.

On the other hand, when considering the *diversity* (i.e., Gini) and *coverage* (i.e., iCov) metrics, we notice some trends deviation to the accuracy performance. Specifically, we see how GRCN is the best model in all settings. This suggests that this approach may be (un)purposely giving up on the accuracy to promote a wider set of items from the catalog, with a corresponding positive effect on the system serendipity. Indeed, while its accuracy performance is not the best one, its diversity and coverage metrics outperform all other models almost on every dataset, even reaching 100% of covered items at $k = 50$. A much more impressive trend is recognizable for Gini, which is higher than the second-to-best model. On a dataset level, it is worth pointing out that, even with more sparse datasets, the GRCN constantly reaches a high iCov and Gini measures. The second-to-best model in terms of diversity is VBPR. Notwithstanding its high accuracy, VBPR settles once again as a compelling model in terms of diversity and coverage.

Summary. *While novelty results are almost aligned with the accuracy trends observed in RQ1, the diversity/coverage measures depict a different scenario. In this respect, GRCN seems to be the approach providing the most diversified item recommendations but at the expense of the accuracy, while VBPR manages to reach a more balanced performance among all metrics.*

Table 5
Novelty and diversity results of the tested baselines when considering the top-10, top-20, and top-50 recommendation lists. **Boldface** and underline stand for best and second-to-best results on each dataset/metric pair, respectively.

Datasets	Models	$k = 10$			$k = 20$			$k = 50$		
		EFD	**Gini**	**iCov** (%)	**EFD**	**Gini**	**iCov** (%)	**EFD**	**Gini**	**iCov** (%)
Office	VBPR	0.1753	0.3634	93.83	0.1479	0.396	10.23	0.1115	0.4413	99.59
	MMGCN	0.1140	0.0128	3.07	0.1027	0.0231	4.64	0.0845	0.0546	10.23
	MGAT	0.1079	0.0132	5.14	0.0963	0.0241	8.12	0.0792	0.0575	17.23
	GRCN	0.1215	**0.4587**	**99.01**	0.1051	**0.4892**	**99.79**	0.0829	**0.5286**	**100**
	LATTICE	**0.1827**	0.2128	87.86	**0.1513**	0.2652	95.90	**0.1125**	0.3414	99.30
Toys	VBPR	0.1948	0.2645	84.90	0.1527	0.3011	92.82	0.1051	0.3585	97.85
	MMGCN	0.0648	0.0989	37.87	0.0570	0.1450	52.51	0.0455	0.2296	72.88
	MGAT	0.0929	0.1036	40.95	0.0796	0.1439	55.71	0.0612	0.2183	76.24
	GRCN	0.1604	**0.3954**	**92.66**	0.1298	**0.4329**	**97.73**	0.0932	**0.4864**	**99.73**
	LATTICE	**0.2090**	0.1656	73.80	**0.1665**	0.2026	86.58	**0.1151**	0.2662	95.94
Clothing	VBPR	0.0502	0.2437	83.40	0.0413	0.2791	92.33	0.0291	0.3344	98.00
	MMGCN	0.0292	0.0136	7.58	0.0240	0.0236	12.44	0.0182	0.0493	23.34
	MGAT	0.0315	0.0201	11.05	0.0263	0.0326	17.36	0.0205	0.0622	30.90
	GRCN	0.0481	**0.3990**	**93.37**	0.0397	**0.4368**	**97.77**	0.0293	**0.4929**	**99.73**
	LATTICE	**0.0738**	0.1022	58.49	**0.0589**	0.1384	76.20	**0.0413**	0.2037	93.23

6. Conclusion and Future Work

In this work, we provide one of the first benchmarking study on multimodal-aware recommender systems through *accuracy*, *novelty*, and *diversity* recommendation measures. By implementing a unified evaluation framework where we train and test five state-of-the-art multimodal recommender systems on three popular datasets, we set the basis to a fair and complete comparison setting. In terms of accuracy, the observed results demonstrate how a careful hyper-parameter exploration can lead shallow multimodal approaches (e.g., VBPR) to be competitive to more recent solutions; on the contrary, other recent techniques such as LATTICE show to be consistently outperforming the other baselines (as reported in the related literature). When measuring novelty and diversity of recommendation, GRCN seems to be a strong baseline for the diversification of the recommendation lists, but VBPR is the solution reaching the most balanced accuracy, novelty, and diversity performance.

The current benchmarking study paves the way to deeper analyses about the specific influence of each component of the multimodal recommendation pipeline on the overall performance, for instance when considering the different impact of each modality (e.g., visual and textual). We plan to run more extensive experimental settings considering: (i) additional datasets and baselines, (ii) deeper hyper-parameter explorations, and (iii) recommendation metrics accounting for bias and fairness.

Acknowledgment

This work was partially supported by the following projects: Secure Safe Apulia, MISE CUP: I14E20000020001 CTEMT - Casa delle Tecnologie Emergenti Comune di Matera, CT_FINCONS_II, CT_FINCONS_III, OVS Fashion Retail Reloaded, LUTECH DIGITALE 4.0.

References

[1] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T. Chua, MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video, in: ACM Multimedia, ACM, 2019, pp. 1437–1445.

[2] X. Han, Z. Wu, Y. Jiang, L. S. Davis, Learning fashion compatibility with bidirectional lstms, in: ACM Multimedia, ACM, 2017.

[3] S. Oramas, O. Nieto, M. Sordo, X. Serra, A deep multimodal approach for cold-start music recommendation, in: DLRS@RecSys, ACM, 2017, pp. 32–37.

[4] W. Min, S. Jiang, R. C. Jain, Food recommendation: Framework, existing solutions, and challenges, IEEE Trans. Multim. 22 (2020) 2659–2671.

[5] Z. Yi, X. Wang, I. Ounis, C. MacDonald, Multi-modal graph contrastive learning for micro-video recommendation, in: SIGIR, ACM, 2022, pp. 1807–1811.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, IEEE Computer Society, 2016, pp. 770–778.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 3980–3990.

[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. W. Wilson, CNN architectures for large-scale audio classification, in: ICASSP, IEEE, 2017, pp. 131–135.

[9] R. He, J. J. McAuley, VBPR: visual bayesian personalized ranking from implicit feedback, in: AAAI, AAAI Press, 2016, pp. 144–150.

[10] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: bayesian personalized ranking from implicit feedback, in: UAI, 2009.

[11] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, T. Chua, MGAT: multimodal graph attention network for recommendation, Inf. Process. Manag. 57 (2020) 102277.

[12] Y. Wei, X. Wang, L. Nie, X. He, T. Chua, Graph-refined convolutional network for multimedia recommendation with implicit feedback, in: ACM Multimedia, ACM, 2020, pp. 3541–3549.

[13] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: ACM Multimedia, ACM, 2021, pp. 3872–3880.

[14] T. Kim, Y. Lee, K. Shin, S. Kim, MARIO: modality-aware attention and modality-preserving decoders for multimedia recommendation, in: CIKM, ACM, 2022, pp. 993–1002.

[15] J. J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: SIGIR, ACM, 2015, pp. 43–52.

[16] S. Vargas, Novelty and diversity enhancement and evaluation in recommender systems and information retrieval, in: SIGIR, ACM, 2014, p. 1281.

[17] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: RecSys, ACM, 2011, pp. 109–116.
[18] G. Shani, A. Gunawardana, Evaluating recommendation systems, in: Recommender Systems Handbook, Springer, 2011, pp. 257–297.
[19] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, B. Zhao, POG: personalized outfit generation for fashion recommendation at alibaba ifashion, in: KDD, ACM, 2019.
[20] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: SIGIR, ACM, 2019, pp. 765–774.
[21] H. Zhan, J. Lin, K. E. Ak, B. Shi, L. Duan, A. C. Kot, $aˆ3$-fkg: Attentive attribute-aware fashion knowledge graph for outfit preference prediction, IEEE Trans. Multim. 24 (2022) 819–831.
[22] Z. Cheng, J. Shen, S. C. H. Hoi, On effective personalized music retrieval by exploring online user behaviors, in: SIGIR, ACM, 2016, pp. 125–134.
[23] K. Vaswani, Y. Agrawal, V. Alluri, Multimodal fusion based attentive networks for sequential music recommendation, in: BigMM, IEEE, 2021, pp. 25–32.
[24] Z. Lei, A. U. Haq, A. Zeb, M. Suzauddola, D. Zhang, Is the suggested food your desired?: Multi-modal recipe recommendation with demand-based knowledge graph, Expert Syst. Appl. 186 (2021) 115708.
[25] W. Wang, L. Duan, H. Jiang, P. Jing, X. Song, L. Nie, Market2dish: Health-aware food recommendation, ACM Trans. Multim. Comput. Commun. Appl. 17 (2021) 33:1–33:19.
[26] X. Chen, D. Liu, Z. Xiong, Z. Zha, Learning and fusing multiple user interest representations for micro-video and movie recommendations, IEEE Trans. Multim. 23 (2021) 484–496.
[27] D. Cai, S. Qian, Q. Fang, C. Xu, Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation, IEEE Trans. Multim. 24 (2022) 805–818.
[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: ICML, Omnipress, 2011, pp. 689–696.
[29] T. Baltrusaitis, C. Ahuja, L. Morency, Challenges and applications in multimodal machine learning, in: The Handbook of Multimodal-Multisensor Interfaces, Volume 2 (2), Association for Computing Machinery, 2018, pp. 17–48.
[30] D. Verma, K. Gulati, V. Goel, R. R. Shah, Fashionist: Personalising outfit recommendation for cold-start scenarios, in: ACM Multimedia, ACM, 2020, pp. 4527–4529.
[31] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, T. Chua, Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention, in: SIGIR, ACM, 2017, pp. 335–344.
[32] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, M. S. Kankanhalli, User diverse preference modeling by multimodal attentive metric learning, in: ACM Multimedia, ACM, 2019, pp. 1526–1534.
[33] J. Li, X. Xu, W. Yu, F. Shen, Z. Cao, K. Zuo, H. T. Shen, Hybrid fusion with intra- and cross-modality attention for image-recipe retrieval, in: SIGIR, ACM, 2021, pp. 244–254.
[34] Y. Deldjoo, T. D. Noia, D. Malitesta, F. A. Merra, Leveraging content-style item representation for visual recommendation, in: ECIR (2), volume 13186 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 84–92.

[35] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Networks 20 (2009) 61–80.
[36] M. M. Bronstein, J. Bruna, T. Cohen, P. Velickovic, Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, CoRR abs/2104.13478 (2021).
[37] X. Wang, X. He, M. Wang, F. Feng, T. Chua, Neural graph collaborative filtering, in: SIGIR, ACM, 2019, pp. 165–174.
[38] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering graph convolution network for recommendation, in: SIGIR, ACM, 2020, pp. 639–648.
[39] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, in: CIKM, ACM, 2020, pp. 1405–1414.
[40] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec, Graph convolutional neural networks for web-scale recommender systems, in: KDD, ACM, 2018, pp. 974–983.
[41] Y. Liu, S. Yang, C. Lei, G. Wang, H. Tang, J. Zhang, A. Sun, C. Miao, Pre-training graph transformer with multimodal side information for recommendation, in: ACM Multimedia, ACM, 2021, pp. 2853–2861.
[42] Z. Liu, Y. Ma, M. Schubert, Y. Ouyang, Z. Xiong, Multi-modal contrastive pre-training for recommendation, in: ICMR, ACM, 2022, pp. 99–108.
[43] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, M. S. Kankanhalli, MMALFM: explainable recommendation by leveraging reviews and images, ACM Trans. Inf. Syst. 37 (2019) 16:1–16:28.
[44] K. Liu, F. Xue, D. Guo, L. Wu, S. Li, R. Hong, MEGCF: multimodal entity graph collaborative filtering for personalized recommendation, ACM Trans. Inf. Syst. 41 (2023) 30:1–30:27.
[45] Y. Wei, X. Wang, X. He, L. Nie, Y. Rui, T. Chua, Hierarchical user intent graph network for multimedia recommendation, IEEE Trans. Multim. 24 (2022) 2701–2712.
[46] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, L. Nie, Dualgnn: Dual graph neural network for multimedia recommendation, IEEE Trans. Multim. 25 (2023) 1074–1084.
[47] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: WWW, ACM, 2023, pp. 845–854.
[48] W. Wei, C. Huang, L. Xia, C. Zhang, Multi-modal self-supervised learning for recommendation, in: WWW, ACM, 2023, pp. 790–800.
[49] K. Liu, F. Xue, S. Li, S. Sang, R. Hong, Multimodal hierarchical graph collaborative filtering for multimedia-based recommendation, IEEE Transactions on Computational Social Systems (2022) 1–12. doi:10.1109/TCSS.2022.3226862.
[50] Q. Du, L. Yu, H. Li, N. Ou, X. Gong, J. Xiang, M^3rec: Cross-modal context enhanced micro-video recommendation with mutual information maximization, in: ICME, IEEE, 2022, pp. 1–6.
[51] F. Lei, Z. Cao, Y. Yang, Y. Ding, C. Zhang, Learning the user's deeper preferences for multi-modal recommendation systems, ACM Trans. Multimedia Comput. Commun. Appl. 19 (2023). URL: https://doi.org/10.1145/3573010. doi:10.1145/3573010.
[52] K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, R. Hong, Multimodal graph contrastive learning for multimedia-based recommendation, IEEE Transactions on Multimedia (2023) 1–13. doi:10.1109/TMM.2023.3251108.
[53] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, L. Wang, Latent structures mining with contrastive modality fusion for multimedia recommendation, CoRR abs/2111.00678 (2021).

[54] Z. Mu, Y. Zhuang, J. Tan, J. Xiao, S. Tang, Learning hybrid behavior patterns for multimedia recommendation, in: ACM Multimedia, ACM, 2022, pp. 376–384.
[55] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems, ACM Trans. Interact. Intell. Syst. 7 (2017) 2:1–2:42.
[56] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, What recommenders recommend: an analysis of recommendation biases and possible countermeasures, User Model. User Adapt. Interact. 25 (2015) 427–491.
[57] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma, How good your recommender system is? A survey on evaluations in recommendation, Int. J. Mach. Learn. Cybern. 10 (2019) 813–831.
[58] S. Vrijenhoek, M. Kaya, N. Metoui, J. Möller, D. Odijk, N. Helberger, Recommenders with a mission: Assessing diversity in news recommendations, in: CHIIR, ACM, 2021, pp. 173–183.
[59] A. Gharahighehi, C. Vens, Diversification in session-based news recommender systems, Personal and Ubiquitous Computing (2021). URL: https://doi.org/10.1007/s00779-021-01606-4. doi:10.1007/s00779-021-01606-4.
[60] G. Cornacchia, V. W. Anelli, G. M. Biancofiore, F. Narducci, C. Pomo, A. Ragone, E. D. Sciascio, Auditing fairness under unawareness through counterfactual reasoning, Inf. Process. Manag. 60 (2023) 103224.
[61] G. Cornacchia, F. Narducci, A. Ragone, A general model for fair and explainable recommendation in the loan domain (short paper), in: KaRS/ComplexRec@RecSys, volume 2960 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.
[62] G. Cornacchia, F. M. Donini, F. Narducci, C. Pomo, A. Ragone, Explanation in multi-stakeholder recommendation for enterprise decision support systems, in: CAiSE Workshops, volume 423 of *Lecture Notes in Business Information Processing*, Springer, 2021, pp. 39–47.
[63] C. Zhai, W. W. Cohen, J. D. Lafferty, Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, in: SIGIR, ACM, 2003, pp. 10–17.
[64] K. Jokinen, T. Hurtig, User expectations and real experience on a multimodal interactive system, in: INTERSPEECH, ISCA, 2006.
[65] A. N. M. Perrin, H. Xu, E. Kroupi, M. Rerábek, T. Ebrahimi, Multimodal dataset for assessment of quality of experience in immersive multimedia, in: ACM Multimedia, ACM, 2015, pp. 1007–1010.
[66] Y. Zhang, H. Tan, Effects of multimodal warning types on driver's task performance, physiological data and user experience, in: HCI (12), volume 12773 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 304–315.
[67] P. J. Chia, J. Tagliabue, F. Bianchi, C. He, B. Ko, Beyond ndcg: Behavioral testing of recommender systems with reclist, WWW '22 Companion, Association for Computing Machinery, New York, NY, USA, 2022, p. 99–104. URL: https://doi.org/10.1145/3487553.3524215. doi:10.1145/3487553.3524215.
[68] F. Bianchi, P. J. Chia, C. Greco, C. Pomo, G. de Souza P. Moreira, D. Eynard, F. Husain, J. Tagliabue, Evalrs 2023. well-rounded recommender systems for real-world deployments, CoRR abs/2304.07145 (2023).
[69] X. Pan, Y. Chen, C. Tian, Z. Lin, J. Wang, H. Hu, W. X. Zhao, Multimodal meta-learning

for cold-start sequential recommendation, in: CIKM, ACM, 2022, pp. 3421–3430.
[70] Y. Deldjoo, T. D. Noia, D. Malitesta, F. A. Merra, A study on the relative importance of convolutional neural networks in visually-aware recommender systems, in: CVPR Workshops, Computer Vision Foundation / IEEE, 2021, pp. 3961–3967.
[71] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia, V-elliot: Design, evaluate and tune visual recommender systems, in: RecSys, ACM, 2021, pp. 768–771.
[72] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Conference on Empirical Methods in Natural Language Processing, 2019.
[73] R. A. Baeza-Yates, B. A. Ribeiro-Neto, Modern Information Retrieval - the concepts and technology behind search, Second edition, 2011.
[74] W. Sun, S. Khenissi, O. Nasraoui, P. Shafto, Debiasing the human-recommender system feedback loop in collaborative filtering, in: WWW (Companion Volume), ACM, 2019, pp. 645–651.
[75] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: SIGIR, ACM, 2021, pp. 2405–2414.
[76] V. W. Anelli, A. Bellogín, T. D. Noia, D. Jannach, C. Pomo, Top-n recommendation algorithms: A quest for the state-of-the-art, in: UMAP, ACM, 2022, pp. 121–131.
[77] V. W. Anelli, Y. Deldjoo, T. D. Noia, E. D. Sciascio, A. Ferrara, D. Malitesta, C. Pomo, How neighborhood exploration influences novelty and diversity in graph collaborative filtering, in: MORS@RecSys, volume 3268 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.
[78] V. W. Anelli, Y. Deldjoo, T. D. Noia, D. Malitesta, V. Paparella, C. Pomo, Auditing consumer- and producer-fairness in graph collaborative filtering, in: ECIR (1), volume 13980 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 33–48.