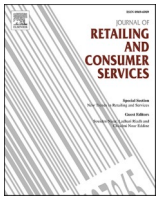
Journal of Retailing and Consumer Services 68 (2022) 103003

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

journal homepage: www.elsevier.com/locate/jretconser



Economic corollaries of personalized recommendations

Mir Majid Molaie, Wonjae Lee *

Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea

ARTICLE INFO

Keywords:
Artificial intelligence
Consumer purchase diversity
E-commerce
Field experiment
Machine learning
Neural networks
Personalization
Recommender systems
Sale diversity

ABSTRACT

The impact of recommendation systems (RSs) on the diversity of consumption is not transparent or well understood. Available studies, whether experimental or theoretical, show inconsistent and even opposite results, which manifests as debate in the literature. In this paper, we investigate the impact of two main recommender systems, neural collaborative filtering and deep content filtering, on sales diversity via a randomized field experiment. Our results confirm the capability of recommender engines in increasing or decreasing aggregate sales diversity. Nonetheless, they amplify homogenization and reduce individual-level consumption diversity. In conclusion, our research reconciles seemingly contradict previous findings and illustrates that the design of the RS is the decisive factor in homogenizing or diversifying product sales.

1. Introduction

Recommender systems are a subgroup of information filtering technologies and are applied to handle the issue of information overload (Schreiner et al., 2019). These systems discover the preferences and interests of users by refining a great amount of dynamically generated data based upon their interactions with items. Subsequently, they predict the willingness of a user to purchase/consume a particular item.

Despite the omnipresence of recommender systems and algorithmic content curation, there are few studies that examine their societal and economic outcomes. Most research efforts attempt to address the technological aspects of recommendation systems and to improve the accuracy of the matchmaking process, whereas papers that seek to elucidate the byproducts of recommender systems at the market level are thin on the ground. In particular, there is a consensus view in the literature that personalized recommendations generate more engagement and more sales (Adomavicius et al., 2018), but the difficulty of gaining access to appropriate research settings and the complexity of personalization algorithms have led to protracted controversy over the distribution of sales and the consumption diversity of users.

One school of thought believes that because a pernicious feedback loop lays the foundation of recommender systems, these systems are inherently biased and decrease sales diversity as a result of recommending blockbusters and well-known products, a pattern reflecting what is termed the "Matthew effect" (Lee and Hosanagar, 2019; Li, 2021). Conversely, others speculate that recommender systems flatten sales distributions by reducing the search cost and exposing users to a large variety of products (both unpopular or popular)—a phenomenon called the Internet's "long tail" (Brynjolfsson et al., 2011; Holtz et al., 2020; Donnelly et al., 2021). This theory explains a new marketing strategy provoked by the Internet in general, and recommender systems in particular, which benefits from low-volume sales of hard-to-find products and boosts niche or artisan buying habits (Hoskins, 2020). To exacerbate the matter, there is no agreement with regard to the impact of recommendation systems on individual-level diversity either. Individual-level diversity predominantly shapes the user experience and has important ramifications for internet companies and online platforms. Research has shown that higher individual-level diversity is desirable (Anderson et al., 2020); however, the relationship between recommender systems and this issue remains unclear. Pariser (2011) coined the term "filter bubble," referring to how online personalization can effectively isolate users and confine their consumption to content that is homogenous within but diverse across users. A different view is proposed by Hosanagar et al. (2014) and Möller et al. (2018), who demonstrate empirically that recommendation systems promote the idea of a global village by virtue of increasing individual-level diversity and commonality among users.

In spite of the significant implications of diversity for users and firms, debate continues in the literature. Although both popular sides of the argument present persuasive results and empirical evidence to underpin their viewpoints, the most studies regard recommender systems as a "black box" and do not consider possible variants of these systems. Given

* Corresponding author.
E-mail addresses: majid@kaist.ac.kr (M.M. Molaie), wnjlee@kaist.ac.kr (W. Lee).

https://doi.org/10.1016/j.jretconser.2022.103003
Received 15 December 2021; Received in revised form 28 March 2022; Accepted 3 April 2022
Available online 13 May 2022

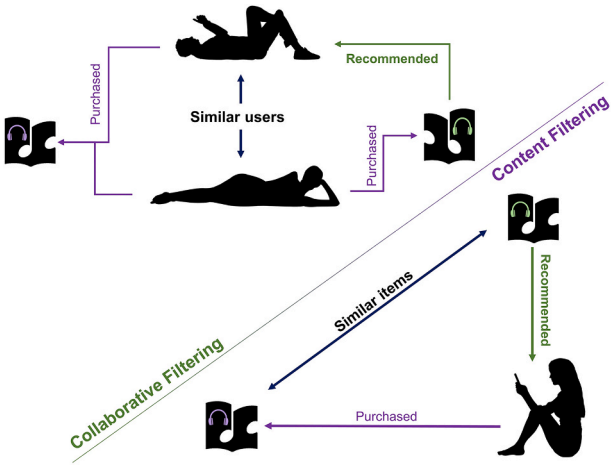


Fig. 1. A Schematic of collaborative filtering and content-based filtering. Pictures were purchased from https://www.123rf.com.

the fact that recommender systems apply different matchmaking approaches, the outcomes at the market level may not be similar.

There are various approaches that can be used when designing recommendation systems, among which collaborative filtering is most popular. This family of algorithms follows the theory of winner-take-all and has shown very notable performance capabilities; however, these algorithms cannot recommend new items or even items with a limited number of interactions (Chinchanachokchai et al., 2021). Collaborative filtering does not take into account the attributes of items or what is termed "metadata" and is associated with the cold start problem. To overcome these limitations, it is common to combine one of these algorithms with content-based filtering (Srivastava et al., 2020). Content-based filtering even demonstrates performance superior to that of collaborative filtering in some applications, such as recommendations of webpages, mobile apps, publications, and news (Jannach and Jugovac, 2019). The underlying mechanism in content-based filtering is the long tail theory, which holds that people have different tastes and that the utilization of product side features as well as implicit and explicit user feedback will result in a higher satisfaction rate. Fig. 1 schematically shows the difference between content-based and collaborative filterings. The integration of these two techniques is termed a hybrid recommendation system, and along with corresponding modifications, these make up the most commonly used algorithms in a range of business domains.

Consequently, the authors believe there are three main sources of these contretemps in the literature. First, the majority of relevant publications employ lab experiments or are based on simulations of archival data, which renders causal implications weaker. In addition, archival data is often confounded by a previously deployed personalization technique (Chaney et al., 2018). Second, experimental studies thus far evaluate either one type of recommender system (collaborative filtering or content-based) or a hybrid algorithm, thereby making the generalization very difficult. Third, most existing approaches for measuring diversity at the individual level are simple and do not account for the similarity between products.

To reconcile the aforementioned contrasting views, in this paper we report a randomized field experiment focusing on an online audiobook app that provides purchase and streaming services to consumers. The experiment consists of one control group and two treatment groups, with neural collaborative filtering (NCF) and deep content filtering (DCF) applied to provide personalized suggestions in the treatment groups, while users in the control group received randomly sampled popular products.

Based on the Gini coefficient, we find that collaborative filtering decreases aggregate-level diversity. In contrast, content-based filtering increases aggregate-level diversity. Upon a further analysis of sales using the Jensen–Shannon divergence metric, we show both algorithms have homogenizing effects on user behavior or the collection of products users consume. This impact is stronger in the group treated by content-based filtering. We conclude personalized recommendations pave the way for a frictionless exploration and raise consumers' propensity to buy through increasing homogeneity, albeit detrimental for some users.

These findings shed light on the impact of different personalization strategies on aggregate and individual diversity levels and help us reconcile different viewpoints in the literature. Notably, we show that recommender systems have the ability to fashion the consumption behavior of users into any particular form. Collaborative and content-based filterings can generate significantly distinct sale distributions; therefore, it can be safely deduced that the extent to which these methods are combined determines their impact at the market level. The authors believe that this research can make a valuable contribution to the information systems, economics, and computational social science communities.

2. Prior works

Methodologically, this paper builds on an emerging field of research that attempts to understand the societal and economic impacts of personalization techniques by field experiments. Here, we present recently published studies that are most relevant to ours.

Anderson et al. (2020), by using a dataset from Spotify, highlight the importance of diversity in the recommendation content, finding that users with more diverse consumption patterns are more likely to subscribe. Claussen et al. (2019) implement a field experiment on a news organization, and reports that algorithmic recommendations, when supplied with enough data, can outperform human editors in terms of user engagement but can also create information bubbles around readers. As a result, individual-level diversity is decreased. Holtz et al. (2020), via a field experiment on Spotify's podcast app, document a similar diversity reduction at the individual level, whereas aggregate diversity increases. They also demonstrate that the existence of a personalization algorithm can affect content sought by users organically from other sections of the app not controlled during the experiment. The algorithms employed in all of the above-mentioned papers are hybrid recommendation systems. Donnelly et al. (2021) implement a field experiment on a retailer website and shows that collaborative filtering

increases the aggregate diversity. In contrast, Lee and Hosanagar (2019) report that collaborative filtering causes a slight lift in individual-level diversity but decreases aggregate diversity significantly. Li et al. (2021), in an experiment on an online book retailer, present findings broadly aligned with those of Lee and Hosanagar (2019). To the best of our knowledge, ours is the first experimental research to date on how content-based and collaborative filtering, in the same setting, can affect sales diversity.

This paper also affords concern about new advancements in the development of recommendation systems that employ deep learning-based recommender engines. Artificial neural networks, owing to their stellar performance capabilities, have recently been the subject of considerable interest in many research fields and have been applied in numerous information retrieval studies (Zhang et al., 2019; Chen et al., 2021). Also, the prediction of consumer behavior is an interesting task (Lombardi et al., 2013; Panniello and Gorgoglione, 2016), where deep learning models have been used (Kim et al., 2021). However, the majority of relevant empirical papers utilize traditional matchmaking paradigms, e.g., WRMF and TF-IDF. Accordingly, whether deep learning-based personalization algorithms would yield similar or different results at the market level relative to the outcomes of traditional methods is an interesting question.

3. Research setting

In order to evaluate the causal impacts of personalized recommendations on users' interactions and sales distributions, we partnered with an online media (e-book and audiobook) streamer and used their platform as our setting. We only focused on the audiobooks section for the experiments, referring to recorded versions of books. Although audiobook recommendation is a novel domain, this choice is a sound decision in the interest of generality and interpretability. The business model studied here is one-off purchase, which is prevalent in e-commerce, and the results are thus generalizable to any other retailer that uses a similar model. It is also interpretable as the major attributes of the item, or the item overall, can be represented as metadata. Furthermore, audiobooks are gaining popularity owing to their easy-to-use and creative contents, and they are considered as the fastest growing section of publishing and entertainment, as the global market share of audiobooks grows by 20–25 percent annually (Stewart et al., 2019).

To minimize the impact of the experiment on the company's business and alleviate concerns about the influence of other sections of their commercial app, we developed a new mobile application. Considering the central role of smartphone ecosystem in today's shopping experience, the exercise of an app for the field experiment is another interesting point of the current paper, as we are witnessing e-commerce is incrementally becoming "mobile", and this channel is dominating online retailing industry (Lee et al., 2019; Verkijika et al., 2021). The app for the experiment had three main pages: (1) a home page that recommended new audiobooks and is the default page when opening the app (Fig. 2), (2) a library and wish-list page that lists audiobooks purchased and lists items the user desires to purchase in the future, and (3) a discover page that lists audiobooks based on categories and authors. There is also a search engine on this page that allows users to search for and explore all available items using keywords.

Due to the host company's future plans, we invited users who had interacted with at least 20 e-books in the main app to install the new app for a two-week experiment. The idea here is to infer their taste in audiobooks from their interaction history with e-books. The interaction history includes purchase, adding to wish-list and reading the demo, all of which are indirect reflections of users' preferences and can be considered as implicit feedbacks. Since only up to 18% of e-books has been converted into audiobooks, we prioritized users whose e-book interaction history has a higher number of audiobook counterparts. As an incentive, they were offered notable discounts on all items upon joining. At the end, 3094 individual users installed the app and participated in the experiment. We matched their e-book interaction history with available audiobooks, and on average there were 4.35 matches that

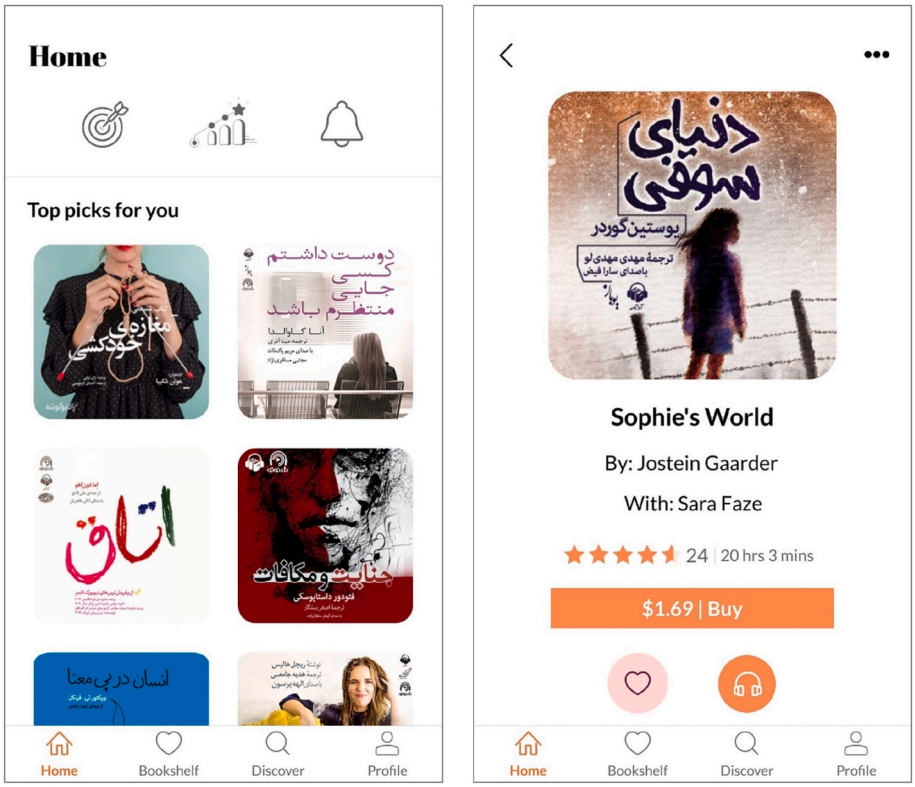


Fig. 2. The home page of the app, presenting recommendations under the title of "Top picks for you". Figure on the right exemplifies a purchase page.

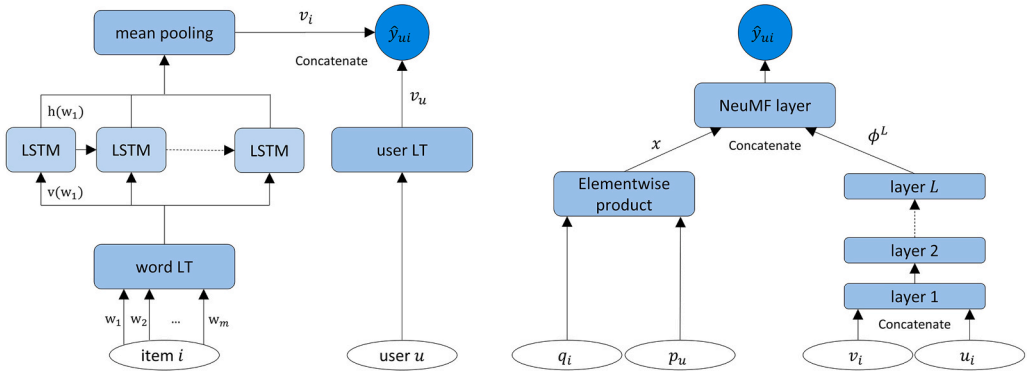


Fig. 3. Architecture of the used neural networks (DCF on left and NCF on right).

would be used to generate recommendations. When users signed up and started using the app, they were randomly assigned to the control or to one of two treatment groups. Of these users, 1029 (33.25%) and 1033 (33.38%) were in the NCF and DCF treatment groups respectively, whereas 1032 (33.35%) were assigned to the control group. The treatment in this context is the presence of the personalized contents on the homepage of the app. On the other hand, in the control group, this recommendation panel is filled with non-personalized content—a combination of random and popularity-based recommendations. In fact, a small proportion of products have historically dominated many markets such as book publishing (Yucesoy et al., 2018) and motion picture industry (Kumar et al., 2014). The Pareto principle is often used by managers and economists to describe this consumption pattern (Brynjolfsson et al., 2011). This aphorism asserts that a large quantity of sales (e.g. 80%) are generated by a few bestsellers (e.g. 20% of products). We follow this principle to promote a traditional consumption pattern in the control group in order to establish a realistic baseline to which the impact of personalized recommendations could be compared. Accordingly, 80% of recommendations in the control group are randomly chosen from 20% of the most interacted audiobooks in the dataset, and the remaining recommendations come from the rest of the dataset. This approach imitates real-life business practices as 80% of recommendations includes bestsellers as well as trending products, and 20% of them represents hand-picked titles such as 'editors' choices.'

As mentioned before, both personalization engines are based on neural network models with implicit feedback. Implicit feedback was utilized to train the models owing to their high abundance and availability. Recommendations based on explicit data were the prime focus of early research activities in this field, especially when the goal was rating prediction. However, at present, given that the task of suggesting a short list of items to consumers is more practical, investigators pay more attention to implicit feedbacks (Chen et al., 2021). To formulate the problem of learning from implicit data, let Y be a user-item interaction matrix with M users and N items:

$$y_{ui} = \begin{cases} 1, & \text{an interaction (user } u, \text{ item } i) \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$$

Y is a binary matrix, where an entry with a value of 1 merely shows the existence of some interaction between item i and user u, not necessarily an indication of a preference. This, however, can be regarded as a noisy reflection of the user's interest. Likewise, a value of 0 does not represent disfavor or dislike, meaning instead that user i has not viewed item u, as this value is treated as an unobserved or missing entry. Therefore, it stands to reason that implicit data is associated with a lack of negative feedback. In this class of recommendation systems, the problem is to estimate the scores of missing data in the interaction matrix Y. Here, we briefly introduce the matchmaking procedures used and depict the architectures of the recommendation systems adopted in this research.

The recommendation system in the first treatment group is a neural network-based collaborative filtering system, abbreviated as NCF. In this recently developed type of collaborative filtering (He et al., 2017), the inner product, which is the main factor when modeling the interaction between users and items in matrix factorization (MF), is replaced with a multi-layer perceptron (MLP). The key idea here is the capability of MLP to learn more abstract features of user-item interactions by applying a non-linear kernel on embedding features, while the inner product uses linear multiplication, which may not sufficiently learn the intricate structure of the interaction data. Collaborative filtering based on MLP can be defined as follows:

$$z^{(1)} = \varphi_1(\mathbf{u}_u, \mathbf{v}_i) = [\mathbf{u}_u, \mathbf{v}_i]$$

$$\varphi^{(2)}(z^{(1)}) = \alpha^1(\mathbf{W}^{(2)} z^{(1)} + \mathbf{b}^{(2)})$$

$$\cdots\cdots$$

$$\varphi^{(L)}(z^{(L-1)}) = \alpha^L(\mathbf{W}^{(L)} z^{(L-1)} + \mathbf{b}^{(L)})$$

$$\hat{y}_{ui} = \alpha(\mathbf{h}^\top \varphi^{(L)}(z^{(L-1)}))$$

In the above expressions, $\mathbf{W}$, $\mathbf{h}$, $\mathbf{b}$, φ, z and α denote the weight matrix, the weight of the output layer, the bias vector, the mapping function, the output of the corresponding layer and the activation function, respectively. In addition, $\mathbf{u}_u$ and $\mathbf{v}_i$ are user and item embeddings, which are the output of a fully connected layer. Furthermore, it has been proposed that MLP and generalized matrix factorization (GMF) can mutually reinforce each other and that fusing the two may result in a better model. GMF is considered as a generic neural network-based form of matrix factorization for which the input is the element wise product of item and user latent factors. It is defined as shown below.

$$\mathbf{x} = \varphi^1(\mathbf{p}_u, \mathbf{q}_i) = \mathbf{p}_u \odot \mathbf{q}_i$$

$$\hat{y}_{ui} = \alpha(\mathbf{h}^\top \mathbf{x})$$

In the above formulation, $\mathbf{p}_u$ and $\mathbf{q}_i$ are user and item embeddings, similar to those in the previous equation. NCF is the resultant model after combining MLP and GMF; the corresponding prediction layer can be formulated as follows:

$$\hat{y}_{ui} = \sigma(\mathbf{h}^\top [\mathbf{x}, \varphi^{(L)}(z^{(L-1)})])$$

To fuse the results of MLP and GMF, NeuMF concatenates, instead of simple summation, the second-to-last layers of the two networks (Fig. 3) to generate a feature vector that can be passed to the ensuing layers. Subsequently, the outputs are projected with $\mathbf{h}$, and afterwards, a logistic activation function.

The second treatment group corresponds to a pure content-based recommender which adopts deep neural networks to learn the embeddings of both users and items jointly. It is therefore referred to as deep content filtering (DCF). Unlike NCF, item textual descriptions are taken into account when learning the embeddings by long short-term memory

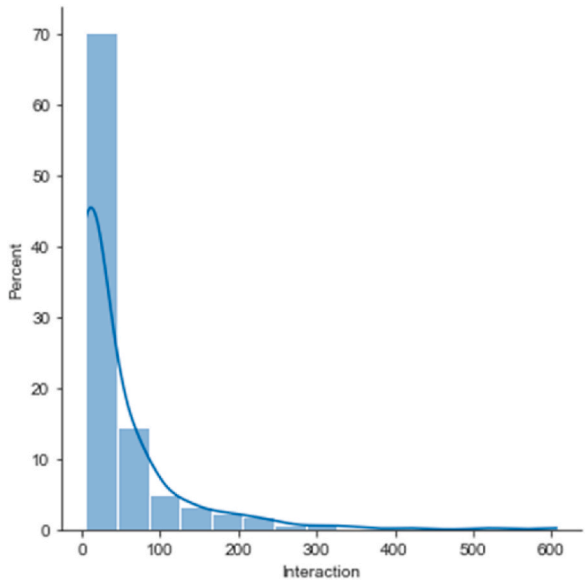


Fig. 4. KDE of items in the dataset.

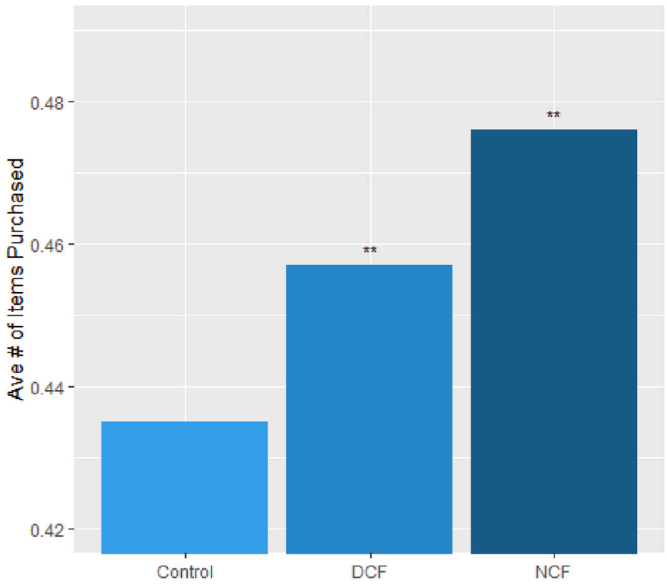


Fig. 5. Average number of items purchased in different treatment arms.

(LSTM). The textual content of each item is a sequence of terms of an arbitrary length, and given the remarkable performance of LSTM in generating latent representations of sequences, the core of this model consists of LSTM units. This technique approaches the recommendation problem as a question answering (QA) scenario, where the user preferences—the description of items with which the user has interacted—can be considered as a question and the task of recommending items to a user is similar to finding the most proper items (answer) based on their descriptions. In DCF, first a dense vector representation of users ($\mathbf{v}_u$) and word descriptions ($\mathbf{v}_w$) are generated by the user and item lookup tables using weight matrices specific for users ($\mathbf{W}_u$) and items ($\mathbf{W}_w$). Then, word representations ($\mathbf{v}_w$) go through an RNN network with LSTM units and mean pooling in order to obtain the item embeddings ($\mathbf{v}_i$). At the end, both embeddings are concatenated, and the following equation expresses the aforementioned operations:

$$\hat{y}_{ui} = \sigma(\mathbf{W}[\mathbf{v}_u\ ,\ \mathbf{v}_i] + \mathbf{b})$$

More details on DCF can be found in the literature (Suglia et al., 2017). To carry out the top-N recommendation task, the recommender should produce a list of items ordered by relevance regarding the user profile. Therefore, both probabilistic models have a logistic regression layer that predicts score $s(u,i)$ or the likelihood that user u would like item i and sorts item in a decreasing order.

The recommender models were trained with preexisting user-audiobook interaction data. The original data has 792 audiobooks but is highly sparse and many users have only one interaction. Therefore, we employed a subset of the data that contains 336 audiobooks, and 15547 interaction records of 2796 users. Fig. 4 illustrates the histogram and kernel density estimation (KDE) of the data. For training and offline assessment, we adopted a commonly used method in the literature, the *leave-one-out* evaluation (Rendle et al., 2012), and applied *Hit Ratio* (HR) to judge the performance of the ranked list (Lee et al., 2011). The training was started with the hyperparameters reported in the original papers, and we then tuned them to improve their predictive accuracy.

According to the result of top-10 evaluation, NCF achieves an accuracy of 0.539, while the accuracy of DCF reaches 0.4821. Neural collaborative filtering comparably outperforms deep content filtering and shows an improvement of 11.8%. Whether features different than simple textual descriptions such as audio and narrator characteristics can elevate the performance of DCF is an interesting question for future research in the audiobook recommendation.

4. Results and discussion

Technically speaking, the outcomes when employing personalization techniques can be divided into two categories: the first-order impact, which describes the more immediate influence of personalization on consumers' consideration sets, product sales and business revenue, and second-order impact, which tells of the more complicated consequence of personalization on the consumption diversity of users and overall sales distributions of firms. Although the first-order impact has been well studied and documented (Baier et al., 2010; Liaukonyte, 2021), we briefly present the impact of personalization on the sale volume to later discuss the connection between sale and diversity.

Fig. 5 shows the average number of items purchased in each treatment arm during the field experiment. Consumers exposed to NCF purchase 9.45% more items compared to users in the control group, whereas DCF generates 5.02% more sales. Both treatments are successful in introducing fitting products and increasing conversion; however, NCF considerably outperforms DCF similar to the offline evaluation. Our results confirm the previous findings in the literature about the potential of personalized recommendations in capturing users' interest.

Empirical studies in psychology and marketing show when consumers face many options wherefrom to choose, they ignore the majority of choices and narrow their attention to a smaller collection, which is called "consideration set" (Hauser, 2014). Recently, Li et al. (2021) find that personalized recommendations largely mediates the positive economic effects through increasing the size of consumers' consideration set. Here, we approach this problem from a different angle. Our conjecture is not only the size but also the diversity of the consideration set is important.

As mentioned before, the second-order impact comprises aggregate and individual-level diversities. We examine the concentration of market share or the issue of aggregate-level diversity using the Gini coefficient. The Gini coefficient has been proven to be a powerful measure of distributional inequality and has been used to examine income inequality and wealth distribution, among other aspects, and was recently used in sales diversity as well (Brynjolfsson et al., 2011). The Gini coefficient is calculated based on the Lorenz curve. Fig. 6 presents the Lorenz curves and changes in the aggregate sales diversity of the control and treatment groups, and Fig. 7 illustrates the percentage of sales generated by each product in descending order, known as the long

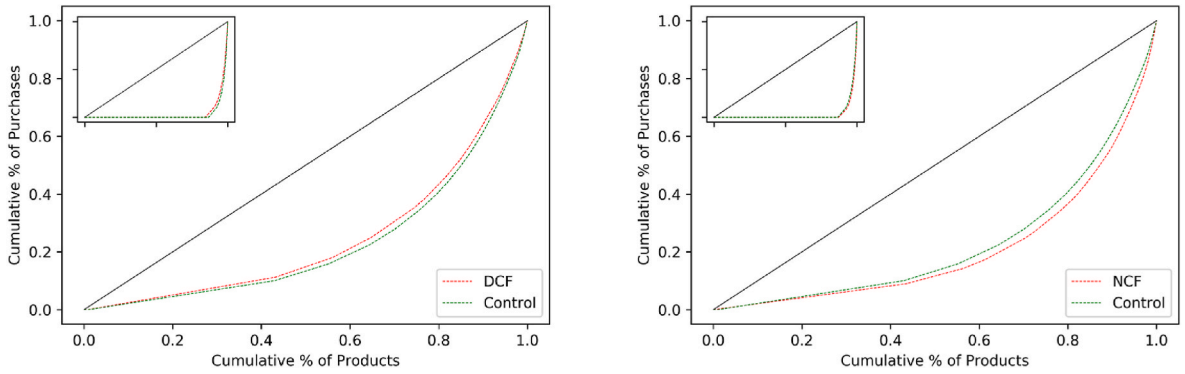


Fig. 6. Purchase Lorenz curves for DCF and NCF (the inset Lorenz curves are for all products).

tail curve. A steeper curve shows that a greater share of sales comes from popular items and that diversity is low.

We also measure the individual-level diversity, which indicates the miscellany of contents users consume. As noted in the introduction, we believe that a consistent notion of diversity should be employed that not only captures the frequency of the products but also the similarity between them. Therefore, we use an information-theory based metric called Jensen–Shannon divergence (JSD) (Virosztek, 2021) to quantify the difference (or similarity) between audiobooks.

JSD is an intuitive idea that measures the distance between symbolic sequences, especially where their frequency distributions have heavy tails. Assuming that p and q are the probability distribution vectors of two audiobooks, the JSD between them is calculated as (Gerlach et al., 2016):

$$D(p,q) = H\left(\frac{p+q}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(q),$$

where $H(p) = -\sum_i p_i log p_i$.

D is between 0 and 1, where 0 means two audiobooks have precisely the same frequencies of symbols and 1 indicates they are the most dissimilar; i.e., they do not share a single word. After tokenizing the raw text of the audiobooks, we calculate $D_{i,j}$ between each pair (i,j) of books. We then quantify the individual-level diversity by averaging the distances between the books purchased by a user.

The Gini coefficients and the individual average diversity levels of the different groups are illustrated in Fig. 8 (left and right, respectively). Because the Gini coefficient is an aggregate statistical measure, we use the permutation test technique (Good, 2013) to calculate the null distribution and determine the p-value. If the impact of the treatment is significant, its bar is labeled by p-value stars according to the magnitude of significance (* 10%, ** 5%, and *** 1%).

According to Figs. 6 and 7, the treatment arms generate significantly different results on diversity compared to a world without personalization. NCF decreases aggregate diversity [$Gini_{control} = 0.727$ *vs* $Gini_{DCF} = 0.765$, $p < 0.1$]. NCF limits the variety of products sold and causes users to interact with less diverse products in the aggregate. In other words, NCF is regarded as a subjective technique that makes recommendations based on the collective and social behavior of consumers, thereby augmenting the exploration of more popular or top-selling items. Table 1 presents the impact of treatments on the market share of top N products of the control group. It indicates whether recommender systems increase or decrease the sale percentage of big hits or well-known audiobooks in the control group. According to the table, the use of NCF is associated with an increase in the market share of top 20 audiobooks of the control group; however, this shift is stronger for top 10 titles. Furthermore, a close look at Fig. 7 reveals that NCF decreases the number of unique products sold by 5.6%. Taken together, the results confirm the existence of concentration bias in collaborative filtering, which has been reported in numerous studies, and there is growing interest in understanding and debiasing these methods (Morik et al., 2020).

On the other hand, DCF enhances the diversity at the aggregate level [$Gini_{control} = 0.727$ *vs* $Gini_{DCF} = 0.662$, $p < 0.05$]. This personalization method serves as a tool that can reduce sales concentrations, in contrast to NCF. Considering that DCF is no longer influenced by the past sales of products, we do not see the same dynamic cycle as found with NCF. Fig. 7 shows that DCF increases the number unique products sold by 16.8%, and guides consumers to the long tail of unknown products, as it predicts user preferences solely based on their profiles and item descriptions. It successfully promotes the visibility and sales of niche

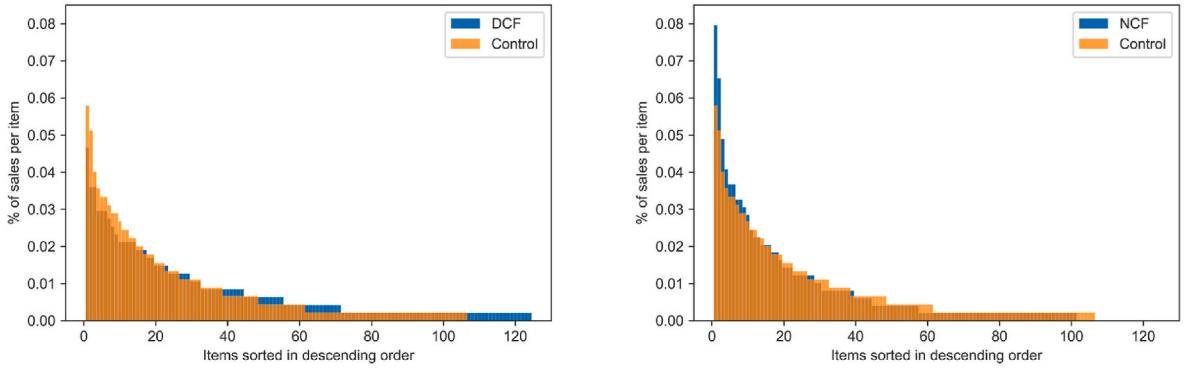


Fig. 7. The percentage of sales per items in each treatment arm of the experiment.

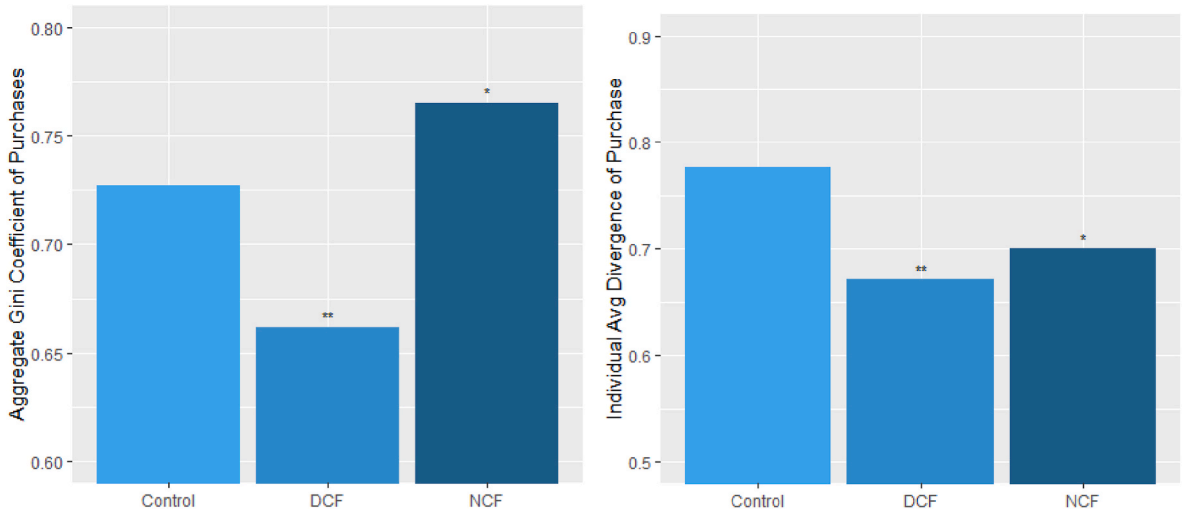


Fig. 8. Aggregate diversity according to Gini coefficients (left) and individual diversity outcomes by JSD (right).

Table 1
The shift in the market share of top N audiobooks of the control group caused by treatments.[a].

	Treatment market share – control market share	
	Top 10 items	Top 20 items
Neural collaborative filtering	0.0325	0.0156
Deep content filtering	−0.086	−0.093

[a] It should be noted that top N items in the control group are not necessarily top N in NCF and DCF.

products. Less popular items lose their share on the market under the influence of NCF, whereas DCF did not show this effect; rather, it helps niche products attract more attention. Our findings regarding aggregate diversity are compatible with most recent findings in the literature, and we unearth evidence for both sides of the argument. These results indicate that recommender systems can skew sales distributions towards hits or niches. These results emphasize that the basic design of the algorithm is of great importance, as different ways of constructing or modeling the preferences and attributes can yield wholly different outcomes in terms of sales diversity.

As shown in Fig. 8 (right), when the contents and similarity of items are considered, both algorithms homogenize users' purchases and decrease individual-level diversity. DCF considers the similarity between the user profile and the item description; therefore, it stands to reason that this class of algorithmic curations can limit consumption diversity by users [$M_{control} = 0.776$, $SD_{control} = 0.124$ *vs* $M_{DCF} = 0.671$, $SD_{DCF} = 0.143$, $p < 0.05$]. However, the reduction of individual-diversity in the NCF group is surprising given that it also decreases aggregate diversity [$M_{control} = 0.776$, $SD_{control} = 0.124$ *vs* $M_{NCF} = 0.70$, $SD_{NCF} = 0.162$, $p < 0.1$]. This shows that the impact of a recommender system on homogenization may not be linked to the overall distribution of sales. The most persuasive explanation for this phenomenon is that NCF helps consumers explore more products, whereas in the end, similar users may arrive at buying similar collections of audiobooks. In fact, this is the idea behind collaborative filtering—a user receives the finest recommendations from someone who has similar preferences (Fig. 1). NCF boosts homogenization along learned latent factors and creates virtual cliques, meaning that it can fracture global village into tribes, a phenomenon that was long recognized by its authors (Resnick et al., 1994).

Homogenization indicates that the models are elucidating user preferences and learning from the data, and when the utility increases, users would likely experience homogenization, meaning that it is not inherently unpleasant. However, substantial evidence exists that homogenization can have detrimental effects on consumers' propensity to buy, churn rates, and lifetime value (Anderson et al., 2020). In the current study, NCF raises the purchase volume by 9.45%, while the increase is 5.02% for DCF. The lower individual-level diversity in the DCF group corresponds directly to a noteworthy drop in the performance, which manifests the over-exploitation of user preferences by the model. DCF causes superfluous homogeneity and provides consumers with a less diverse consideration set. Furthermore, homogenization has grave social consequences and systemic effects on societies and thus deserves more attention in future studies (Milano et al., 2020).

To bridge the gap in the literature, we compare our findings with those reported recently. Looking closely at the results above, we find that DCF creates a more equal distribution and amplifies the homogenization of user behavior, thereby having a balkanization effect (Konstan et al., 2012). This function of DCF is reminiscent of the hybrid algorithms used in several recent studies (Claussen et al., 2019; Holtz et al., 2020; Donnelly et al., 2021). In Donnelly et al. (2021), the authors mention that the employed algorithm undertakes modified collaborative filtering while taking into account product metadata (e.g., color) and user-specific tastes. Hence, it can be considered as a hybrid algorithm. This shows even a slight combination of item metadata and collaborative filtering can alter the overall sales distribution. Notably, Hosanagar et al. (2014), by employing a hybrid algorithm that is 90% content-based, show a similar finding with regard to aggregate diversity, but they find that a hybrid algorithm increases individual consumption diversity. This contrast can be justified by the different measurement of diversity used. They only evaluate the number of items purchased and do not examine the similarity between them.

The outcome of the treatment group exposed to NCF, with regard to aggregate diversity, is also compatible with recent findings in which collaborative filtering has been applied (Lee and Hosanagar, 2019; Li et al., 2021). However, our finding appears to be in flat contradiction of Lee and Hosanagar (2019) with regard to individual-level diversity. Again, they use the Gini coefficient to quantify individual-level diversity, which is neutral as regards the content of the products. In fact, when the quantity of items bought by a consumer increases, the Gini coefficient calculated over all products decreases. The Gini coefficient is a balance-only diversity measure and is therefore appropriate for computing aggregate diversity but not reliable for estimating homogenization when items belong to different categories (Morales et al., 2021). Considering this shortcoming of the Gini coefficient, Claussen et al. (2019) and Holtz et al. (2020) used the Hirschman-Herfindahl Index (HHI) and Shannon entropy, respectively, to account for topic and

category differences of items. There is also a study involving a simulation, the results of which are in good agreement with those here (Chaney et al., 2018). That study shows that personalization achieves higher consumption/engagement levels, but at the cost of pushing users into consumption bubbles.

The randomized field experiment reported here has interesting managerial implications. Most importantly, it confirms the notable influence of recommendation systems on sales dispersion. Retailers need to define their marketing strategies in concert with the personalization algorithms on their online platforms. For example, firms generally have a desire to increase sale/engagement so they usually adopt collaborative filtering. NCF maximizes revenue but diminishes the visibility of items with limited historical data, so the introduction of a new product becomes challenging, and it not only negatively affects the long-term income but also enfeebles their clients who produce low-budget, niche items. On the other hand, when they are interested in exposing their customer to the back-catalog titles, they can benefit from content-based methods. DCF can fathom unique taste of users, as the higher aggregate-level diversity is accompanied with a higher average sale number compared to the control but it comes at the cost of lower individual-level diversity. Our results prove the existence of a trade-off between sale and diversity, and the extent to which these algorithms are combined can determine the impact at the market level.

5. Future research directions

Our study directly speaks to the debate in the literature and highlights the corollaries of two main personalization techniques at the market level. Here, we suggest a number of directions to be explored. First, with the advent of the Internet and digitalization, the "spoken word is now as powerful as the written word.[1]" One can listen to audiobooks and become educated while doing other tasks. This medium deserves more attention from the RS community in the future. Second, the long-term impact of recommender systems could also be an interesting extension for the current work, as multiple updates of recommendations based on user feedback might yield different results. Third, it would be interesting to study the effects of other types of personalization (e.g., social networks and graph-based methods). Fourth, this study is a part of a larger discussion. We present the potential and ability of different algorithms to reshape sales distributions, but which strategy is better: blockbuster (Elberse, 2013) or long tail (Anderson, 2006)?

6. Conclusion

Given the current circumstances around the world, online retailers and web-based service providers are gaining in popularity. Recommender systems have been demonstrated to be a powerful tool for these businesses, and their significant value is clear. Therefore, it is of utmost importance to understand the way these intermediaries influence the behaviors of users and how they impact the visibility and sales of different products, from best-selling to niche items. In order to address this issue, we carried out a randomized field experiment employing two most common recommender systems in the context of e-commerce. The analysis of our results shows that both treatments enhance sales; however, collaborative filtering creates concentration bias. In other words, it reinforces the sales of already best-sellers, whereas content-based recommender flattens the distribution of sales and expose users to niche items. Our study reveals that both algorithms decrease individual-level diversity compared to a world without personalization, and homogenization is an inevitable corollary of personalized recommendations. In light of our results, marketing strategists can benefit from a combination of these two matchmaking approaches and find an optimal point that best suits their needs.

[1] Jordan Peterson.

Acknowledgement

This work was supported by BK21 Plus Postgraduate Organization for Content Science.

References

Adomavicius, G., Bockstedt, J.C., Curley, S.P., Zhang, J., 2018. Effects of online recommendations on consumers' willingness to pay. Inf. Syst. Res. 29 (1), 84–102.
Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., Lalmas, M., 2020. Algorithmic effects on the diversity of consumption on Spotify. In: Proceedings of the Web Conference 2020, pp. 2155–2165.
Anderson, C., 2006. The Long Tail: Why the Future of Business Is Selling Less of More. Hachette Books.
Baier, D., Stüber, E., 2010. Acceptance of recommendations to buy in online retailing. J. Retailing Consum. Serv. 17 (3), 173–180.
Brynjolfsson, E., Hu, Y., Simester, D., 2011. Goodbye pareto principle, hello long tail: the effect of search costs on the concentration of product sales. Manag. Sci. 57 (8), 1373–1386.
Chaney, A.J., Stewart, B.M., Engelhardt, B.E., 2018, September. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 224–232.
Chen, S.S., Choubey, B., Singh, V., 2021. A neural network based price sensitive recommender model to predict customer choices based on price effect. J. Retailing Consum. Serv. 61, 102573.
Chinchanachokchai, S., Thontirawong, P., Chinchanachokchai, P., 2021. A tale of two recommender systems: the moderating role of consumer expertise on artificial intelligence based product recommendations. J. Retailing Consum. Serv. 61, 102528.
Claussen, J., Peukert, C., Sen, A., 2019. The Editor vs. The Algorithm: Targeting, Data and Externalities in Online News. Data and Externalities in Online News.
Donnelly, R., Kanodia, A., Morozov, I., 2021. The Long Tail Effect of Personalized Rankings. *Available at SSRN 3649342*.
Elberse, A., 2013. Blockbusters: Why Big Hits–And Big Risks–Are the Future of the Entertainment Business. Faber & Faber.
Gerlach, M., Font-Clos, F., Altmann, E.G., 2016. Similarity of symbol frequency distributions with heavy tails. Phys. Rev. X 6 (2), 021009.
Good, P., 2013. Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media.
Hauser, J.R., 2014. Consideration-set heuristics. J. Bus. Res. 67 (8), 1688–1699.
He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S., 2017. Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web, pp. 173–182.
Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., Aral, S., 2020, July. The engagement-diversity connection: evidence from a field experiment on spotify. In: Proceedings of the 21st ACM Conference on Economics and Computation, pp. 75–76.
Hosanagar, K., Fleder, D., Lee, D., Buja, A., 2014. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. Manag. Sci. 60 (4), 805–823.
Hoskins, J.D., 2020. The evolving role of hit and niche products in brick-and-mortar retail category assortment planning: a large-scale empirical investigation of US consumer packaged goods. J. Retailing Consum. Serv. 57, 102234.
Jannach, D., Jugovac, M., 2019. Measuring the business value of recommender systems. ACM Transact. Manag. Info. Sys.(TMIS) 10 (4), 1–23.
Kim, J., Ji, H., Oh, S., Hwang, S., Park, E., del Pobil, A.P., 2021. A deep hybrid learning model for customer repurchase behavior. J. Retailing Consum. Serv. 59, 102381.
Konstan, J.A., Riedl, J., 2012. Recommender systems: from algorithms to user experience. User Model. User-Adapted Interact. 22 (1), 101–123.
Kumar, A., Smith, M.D., Telang, R., 2014. Information discovery and the long tail of motion picture content. MIS Q. 38 (4), 1057–1078.
Lee, D., Hosanagar, K., 2019. How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. Inf. Syst. Res. 30 (1), 239–259.
Lee, S., Song, S.I., Kahng, M., Lee, D., Lee, S.G., 2011. Random walk based entity ranking on graph for multidimensional recommendation. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 93–100.
Lee, Y., Kim, H.Y., 2019. Consumer need for mobile app atmospherics and its relationships to shopper responses. J. Retailing Consum. Serv. 51, 437–442.
Li, X., Grahl, J., Hinz, O., 2021. How do recommender systems lead to consumer purchases? A causal mediation analysis of a field experiment. Inf. Syst. Res.
Liaukonyte, J., 2021. Personalized and Social Commerce. *Available at SSRN 3846888*.
Lombardi, S., Gorgoglione, M., Panniello, U., 2013. The effect of context on misclassification costs in e-commerce applications. Expert Syst. Appl. 40 (13), 5219–5227.
Milano, S., Taddeo, M., Floridi, L., 2020. Recommender systems and their ethical challenges. AI Soc. 35 (4), 957–967.
Möller, J., Trilling, D., Helberger, N., van Es, B., 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. Inf. Commun. Soc. 21 (7), 959–977.
Morales, P.R., Lamarche-Perrin, R., Fournier-S'Niehotta, R., Poulain, R., Tabourier, L., Tarissan, F., 2021. Measuring diversity in heterogeneous information networks. Theor. Comput. Sci. 859, 80–115.

Morik, M., Singh, A., Hong, J., Joachims, T., 2020. Controlling fairness and bias in dynamic learning-to-rank. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 429–438.
Panniello, U., Gorgoglione, M., Tuzhilin, A., 2016. Research note—in CARSs we trust: how context-aware recommendations affect customers' trust and other business performance measures of recommender systems. Inf. Syst. Res. 27 (1), 182–196.
Pariser, E., 2011. The Filter Bubble: what the Internet Is Hiding from You. Penguin, UK.
Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L., 2012. BPR: Bayesian Personalized Ranking from Implicit Feedback. *arXiv preprint arXiv:1205.2618*.
Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186.
Schreiner, T., Rese, A., Baier, D., 2019. Multichannel personalization: identifying consumer preferences for product recommendations in advertisements across different media channels. J. Retailing Consum. Serv. 48, 87–99.
Srivastava, A., Bala, P.K., Kumar, B., 2020. New perspectives on gray sheep behavior in E-commerce recommendations. J. Retailing Consum. Serv. 53, 101764.
Stewrt, D., Casey, M., Wigginton, C., 09 December 2019. The Ears Have it: the Rise of Audiobooks and Podcasting. Retrieved from. https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2020/rise-of-audiobooks-podcast-industry.html.
Suglia, A., Greco, C., Musto, C., De Gemmis, M., Lops, P., Semeraro, G., 2017. A deep architecture for content-based recommendations exploiting recurrent neural networks. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 202–211.
Verkijika, S.F., Neneh, B.N., 2021. Standing up for or against: a text-mining study on the recommendation of mobile payment apps. J. Retailing Consum. Serv. 63, 102743.
Virosztek, D., 2021. The metric property of the quantum Jensen-Shannon divergence. Adv. Math. 380, 107595.
Yucesoy, B., Wang, X., Huang, J., Barabási, A.L., 2018. Success in books: a big data approach to bestsellers. EPJ Data Sci. 7, 1–25.
Zhang, S., Yao, L., Sun, A., Tay, Y., 2019. Deep learning based recommender system: a survey and new perspectives. ACM Comput. Surv. 52 (1), 1–38.