



On Practical Diversified Recommendation with Controllable Category Diversity Framework

Tao Zhang
selous.zt@alibaba-inc.com
Alibaba Group
Hangzhou, Zhejiang, China

Luwei Yang*
luwei.ylw@alibaba-inc.com
Alibaba Group
Hangzhou, Zhejiang, China

Zhibo Xiao
xiaozhibo.xzb@alibaba-inc.com
Alibaba Group
Hangzhou, Zhejiang, China

Wen Jiang
wen.jiangw@alibaba-inc.com
Alibaba Group
Hangzhou, Zhejiang, China

Wei Ning
wei.ningw@alibaba-inc.com
Alibaba Group
Hangzhou, Zhejiang, China

ABSTRACT

Recommender systems have made significant strides in various industries, primarily driven by extensive efforts to enhance recommendation accuracy. However, this pursuit of accuracy has inadvertently given rise to echo chamber/filter bubble effects. Especially in industry, it could impair user's experiences and prevent user from accessing a wider range of items. One of the solutions is to take diversity into account. However, most of existing works focus on user's explicit preferences, while rarely exploring user's non-interaction preferences. These neglected non-interaction preferences are especially important for broadening user's interests in alleviating echo chamber/filter bubble effects. Therefore, in this paper, we first define diversity as two distinct definitions, i.e., user-explicit diversity (U-diversity) and user-item non-interaction diversity (N-diversity) based on user historical behaviors. Then, we propose a succinct and effective method, named as Controllable Category Diversity Framework (CCDF) to achieve both high U-diversity and N-diversity simultaneously. Specifically, CCDF consists of two stages, User-Category Matching and Constrained Item Matching. The User-Category Matching utilizes the DeepU2C model and a combined loss to capture user's preferences in categories, and then selects the top- K categories with a controllable parameter K . These top- K categories will be used as trigger information in Constrained Item Matching. Offline experimental results show that our proposed DeepU2C outperforms state-of-the-art diversity-oriented methods, especially on N-diversity task. The whole framework is validated in a real-world production environment by conducting online A/B testing. The improved conversion rate and diversity metrics demonstrate the superiority of our proposed framework in industrial applications. Further analysis supports the complementary effects between recommendation and search that diversified

recommendation is able to effectively help users to discover new needs, and then inspire them to refine their demands in search.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Personalization**; **Information retrieval diversity**.

KEYWORDS

Recommender systems, Controllable category diversity, Diversification, User-category matching

ACM Reference Format:

Tao Zhang, Luwei Yang, Zhibo Xiao, Wen Jiang, and Wei Ning. 2024. On Practical Diversified Recommendation with Controllable Category Diversity Framework. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3589335.3648323>

1 INTRODUCTION

Recommender systems (RS) have achieved great success in many industries, e.g., e-commerce [16], online news [22], and multimedia content [5]. Pursuing the accuracy of recommendation based on user attributes and historical preferences has attracted tremendous attention [6, 19, 24, 29, 30]. However, when accuracy is the only target, echo chamber/filter bubble effects will inevitably be introduced in recommendation [9, 17, 25]. It means user's interests will be reinforced by exposed items which are recommended by learning from user's historical preferences. As a result, the user will be isolated in this small set of items from his/her interests. This phenomenon will hurt user's experience and prevent user from accessing more diverse merchandise in e-commerce platforms or videos in video sharing websites.

To alleviate the echo chamber/filter bubble effects, diversity in recommender systems has attracted a growing research interest [23]. The diversified recommendation is proven to help users to broaden their interests and improve their experiences [14]. Some post-processing techniques, Maximal Marginal Relevance (MMR) [3] and Determinantal Point Process (DPP) [21], focus on reshuffling and re-ranking the retrieved candidates to promote diversity. However, these methods are applied after candidate generation stage. In order to consider diversity in candidate generation stage, some Graph Neural Network based methods are proposed [11, 25, 26, 28].

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW '24 Companion, May 13–17, 2024, Singapore, Singapore
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0172-6/24/05...\$15.00
<https://doi.org/10.1145/3589335.3648323>

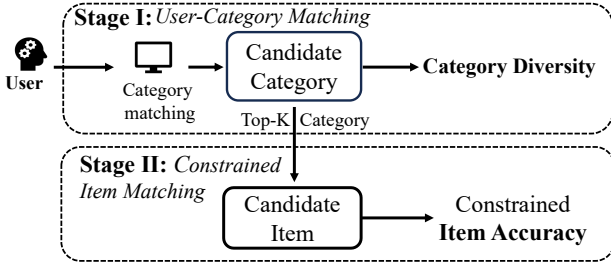


Figure 1: The proposed Controllable Category Diversity Framework (CCDF), which is divided into two stages. Top: User-Category Matching. Bottom: Constrained Item Matching

They aim to expand the candidate pool by leveraging graph structures or content-based similarities, which expands the range of items available for recommendation. However, all these works evaluate diversity based on the number of categories, ignoring the relationship between users and categories. Even within diversified recommendations, there exists the possibility that the presented categories are still within user’s historical interactions (U-diversity). The key to alleviating the echo chamber/filter bubble effects is to provide users with new choices (N-diversity), rather than just diversifying among user’s existing interests. The detailed definitions of U-diversity and N-diversity can be found in Section 3.1.

Therefore, in this paper, we propose a succinct and effective method, called Controllable Category Diversity Framework (CCDF), which is able to explicitly control category diversity and achieve both high U-diversity and N-diversity. The pipeline of the framework is shown in Figure 1. This framework is divided into two stages. The first stage is User-Category Matching, which is responsible for selecting top-K categories that the user may be interested in. The second stage is Constrained Item Matching, where the emphasis lies on accurately retrieving items within the constraints of the top-K categories. This process divides the responsibilities of category diversity and item accuracy into two different stages, User-Category Matching and Constrained Item Matching respectively, which equips CCDF with the controllability on category diversity by parameter K . If K is small, the retrieved items will be more related to user’s historical preferences. If K is large, the retrieved items will be more diversified. As a result, the user can access more diversified and novel items with larger K , thereby alleviating echo chamber/filter bubble effects.

Specifically, we novelly formalize User-Category Matching as a next category prediction task and construct the DeepU2C model, which employs User Net, Category Net and Wide Layer, to predict the next category label based on user historical behaviors. Moreover, we propose a combined loss function to further improve the performance. We evaluate our DeepU2C model on U-diversity and N-diversity tasks to show its effectiveness in category diversity modeling. To further validate the effectiveness of the CCDF, we conduct online experiments on DeepU2C model in conjunction with Constrained Item Matching which utilizes posterior weighted score. The online A/B testing results demonstrate significant improvements in both conversion rate and diversity metrics, confirming the effectiveness of our proposed CCDF in industrial applications. Further analysis supports the complementary effects [27] between recommendation and search that diversified recommendation is

able to effectively help users to discover new needs, and then inspire them to refine their demands in search.

Our main contributions are summarized as follows:

- We innovatively define the diversity as two distinct definitions, user-explicit diversity (U-diversity) and user-item non-interaction diversity (N-diversity), to evaluate the effectiveness of our method more comprehensively.
- We propose a succinct and effective two-stage Controllable Category Diversity Framework (CCDF), which is able to control category diversity explicitly and achieve both high U-diversity and N-diversity simultaneously.
- We formalize User-Category Matching as a next category prediction task, and then utilize the DeepU2C model and a combined loss to capture user’s preferences in categories. These categories are used as trigger information for Constrained Item Matching.
- We compare our method on offline real-world industrial datasets with state-of-the-art methods. We also implement it in a real-world industrial production environment for A/B testing. The results validate the superiority of our method. The emerging complementary effects between recommendation and search are observed.

2 RELATED WORK

2.1 Accuracy in Recommendation

As one of the primary objectives of recommendation, accuracy has been studied extensively since it was proposed. The goal is to accurately recommend items based on user’s preferences or interests. In industry, the recommendation system architecture is usually a funnel architecture. The first funnel is match module (aka candidate generation module) which is responsible for retrieving thousands of items from a large candidate pool. Collaborative Filtering [16] is one of the most representative match techniques which has been widely applied in industry. In recent years, deep learning techniques have been utilized to generating candidates. Deep candidate generation model [8] applies deep learning techniques to effectively assimilate many signals and model their interactions. Some followed works, e.g., MIND [15] and ComiRec [4], introduce effective modifications to the user and item deep learning networks and negative sampling techniques to improve the accuracy of user and item matching. The second funnel is rank module which is used to rank the retrieved items elaborately. There are numerous research papers on rank module. Some of them are applied successfully in industry, such as WDL [6], DeepFM [10], DIN [30], DIEN [29] and DMIN [24]. Almost all of the methods in match and rank modules emphasize on the accuracy of recommended items based on user’s preferences or interests. Nevertheless, a sole emphasis on accuracy may inadvertently result in a dearth of diversity in recommendation which may cause users to fall into echo chamber/filter bubble effects [9, 25].

2.2 Diversity in Recommendation

Diversity is introduced to solve the echo chamber/filter bubble effects in order to make the recommended results more plentiful. Most earlier methods achieve diversity by re-ranking the recommendation list based on the user-item relevance scores and some diversity metrics. Maximum Marginal Relevance (MMR) [3] is one of the

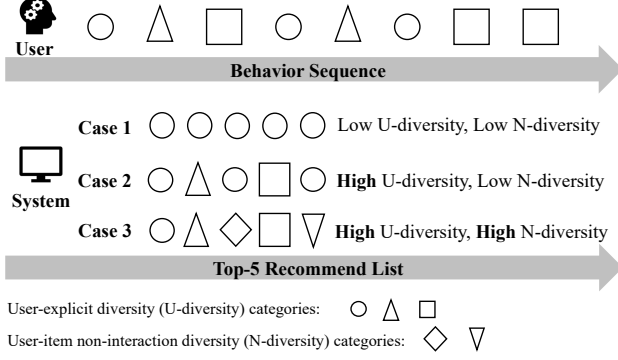


Figure 2: Illustration of U-diversity and N-diversity. Different shapes refer to different categories. Top: user historical behavior. Middle: three top-5 recommendation cases. Bottom: U-diversity categories and N-diversity categories.

most representative diversity methods in this field, which considers a linear combination of the relevance and diversity scores. It greedily selects the item if it has both high relevance score to the user and high diverse score to previously selected items. Determinantal Point Process (DPP) [21] directly models the dissimilarity among items in a set-wise way instead of pair-wise way which has been used in MMR. A unified model is used to capture the global correlations among items. In addition to the diversity approach on re-ranking stage, a number of graph-based methodologies [25, 26, 28] have been proposed in match module to directly generate diverse items. These techniques create a heterogeneous graph by conceptualizing users and items as nodes and abstracting behavioral interactions as edges. The intuitive sense that the graph-based methods are able to improve the diversity is that higher order neighbors of a user contain more diverse items. DGCN [28] selects neighbor nodes based on the inverse category frequency, and employs category-boosted negative sampling and adversarial learning in obtaining node embeddings. DDGraph [26] utilizes dynamic user-item graph construction to select diverse items dynamically. A quantile progressive candidate selection scheme is used. DGRc [25] applies a submodular neighbor selection for diverse item acquisition, and a loss reweighting method to focus on long-tail items. However, these solutions have yet to consider user-item non-interaction diversity explicitly, which is essential for alleviating the echo chamber/filter bubble effects.

Category is used for grouping items. In e-commerce industry, the category groups products by considering their shared characteristics, e.g., apparel, food, home appliances, etc. In movie industry, category could be genre such as action, drama, and musical. There are a few of research on improving diversified recommendations by employing category information. Choi and Han [7] use the category correlation to improve diversity in collaborative filtering. DGCN [28] applies category information in the category-boosted negative sampling scheme to obtain similar but negative items. It will help the model better to distinguish user's preferences within a category. Matroid Constraints [1] relies on category similarity to measure the resemblance between candidate items and existing ones. Besides, Hu and Pu [13] propose an organization-based interface by grouping recommendations with categories. The results

Table 1: The overall notations of our work.

Notation	Description
u	a user
i	an item
c	a category
\mathcal{U}	the set of users
\mathcal{I}	the set of items
\mathcal{I}_c	the set of items belongs to category c
\mathcal{C}	the set of categories
\mathcal{I}_u	the list of items interacted by user u
\mathcal{P}_u	the basic profiles of user u
\mathcal{F}_{uc}	the crossing features between u and c
y	label, either positive or negative
s_{uc}	the number of interactions between u and c
$C(\cdot)$	a mapping function for item to its category
c_{neg}	negative samples from user non-interacted category set
c_{nei}	neighbor samples from user interacted category set
n_{neg}	number of negative samples
n_{nei}	number of neighbor samples
\mathcal{G}	a heterogeneous graph of user-category interactions
\mathcal{V}, \mathcal{E}	the set of vertexes and edges in \mathcal{G}

show that the organization interface effectively increases user's perceived diversity of recommendations and further enhances their satisfaction. Nevertheless, these works only treat categories as auxiliary information. Instead, we will use category to redefine diversity into U-diversity and N-diversity. Our proposed method CCDF directly models user's preferences in categories to obtain controllable diversity in order to achieve both high U-diversity and N-diversity.

3 METHOD

3.1 Preliminaries

3.1.1 Diversity. Motivated by [26], we innovatively define diversity as two distinct definitions, user-explicit diversity (U-diversity) and user-item non-interaction diversity (N-diversity). Specifically, U-diversity measures the extent to which the categories in recommended items overlap with the user's historical categories. It represents the diversity in the set of recommended items that align with the user's previous interactions or expressed categories. N-diversity, also known as novelty, captures the degree to which the categories in recommended items differ from user historical behaviors or interests. It focuses on introducing users to new and unfamiliar yet relevant categories, expanding their horizons, and encouraging exploration beyond user's existing preferences. Figure 2 shows three cases to illustrate the concepts of U-diversity and N-diversity in the context of e-commerce. Case 1 recommendation list only contains one category products which results in both low U-diversity and N-diversity. While case 2 recommendation list contains multiple categories, it fails to extend user's existing preferences. Thus it has high U-diversity but low N-diversity. Case 3 recommendation list not only has a good category converge on user's existing preferences, but also introduces novel items beyond user's known category preferences. As a result, it has both high U-diversity and N-diversity. N-diversity is a key factor in alleviating echo chamber/filter bubble effects. Achieving both high U-diversity

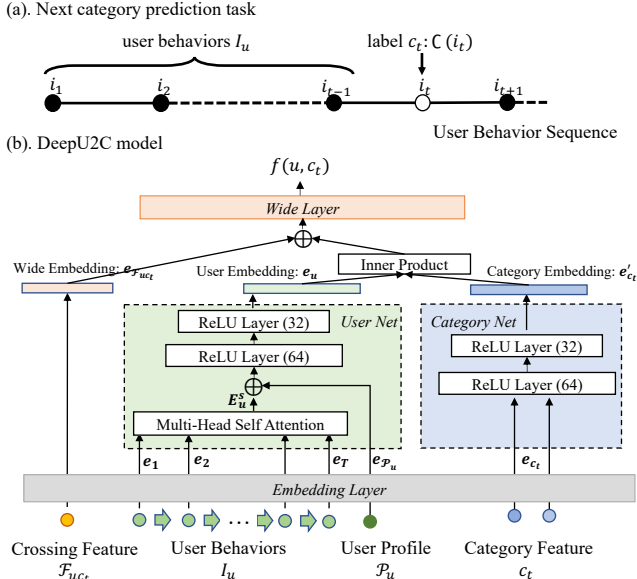


Figure 3: The User-Category Matching. (a): definition of next category prediction task. (b): DeepU2C model architecture.

and N-diversity will make real-world e-commerce recommender systems more effective and healthy. Thus, we will evaluate the performance of our framework and state-of-the-art methods on both U-diversity and N-diversity tasks.

3.1.2 Definitions. Let $u \in \mathcal{U}$ denote the user and $i \in \mathcal{I}$ denote the item. $C(\cdot)$ represents a mapping function that maps each item to its category, and $c \in \mathcal{C}$ denotes the category. An instance is represented by a tuple $\langle I_u, \mathcal{P}_u, \mathcal{F}_{uc_t}, c_t, s_{uc_t}, y \rangle$, where I_u is the user historical behaviors which are the items interacted by user u . \mathcal{P}_u represents the basic profiles of user u , e.g., gender, age and so on. \mathcal{F}_{uc_t} represents the crossing features between u and c_t . c_t is the target category, and s_{uc_t} is the number of interactions between user u and category c_t . y represents the label, either positive or negative. Usually, if the user has interacted items with c_t , y is positive. Negative examples will be sampled from user non-interacted category set, which is denoted as c_{neg} . The overall notations are summarized in Table 1.

3.2 The Proposed Framework

As shown in Figure 1, our proposed Controllable Category Diversity Framework (CCDF) is structured into two stages. The first stage, User-Category Matching, is responsible for identifying a curated set of top- K categories that align with the user's potential interests. The degree of control over category diversity is determined by the value of K . The second stage, Constrained Item Matching, focuses on retrieval of items within the top- K categories chosen in the first stage.

The advantage of our CCDF is that, it is able to explicitly control the diversity level of the recommended list by controlling parameter K . With a larger K , the retrieved items will be more N-diversified, which helps users discover new items, thereby alleviating echo chamber/filter bubble effects.

3.3 Stage I: User-Category Matching

The goal of User-Category Matching in CCDF is to capture user's preferences in categories, and subsequently select the top- K categories from the category candidate set that align with their interests to customize and control the category diversity level in recommendation list. We model this stage as a next category prediction task, which aims to predict what the next category the user is most likely to click. When the target category c_t exists in the user historical behaviors, i.e., $c_t \in C(I_u)$, it will be U-diversity next category prediction task, otherwise it will be N-diversity next category prediction task. We will evaluate our method and state-of-the-art baselines on these tasks.

3.3.1 Next Category Prediction Task. As shown in Figure 3 (a), we first obtain user historical behavior items in the form of a sequence in chronological order. Then we randomly choose an item from this sequence, whose category c_t will be selected as the label. The user historical behaviors before this label will be selected as I_u . The main objective of next category prediction task is to maximize the probability of user u will click the item with category c_t .

3.3.2 DeepU2C Model. The basic idea of the model architecture is borrowed from Wide&Deep network [6]. The motivation originates from its superior performance in memorization and generalization, which is quite suitable for improving U-diversity and N-diversity next category prediction tasks simultaneously. As shown in Figure 3 (b), the model mainly consists of *Embedding Layer*, *User Net*, *Category Net*, and *Wide Layer*.

There are four groups of input features, *crossing features* \mathcal{F}_{uc_t} , *user historical behaviors* I_u , *user profile* \mathcal{P}_u , and *target category* c_t . Each feature is normally encoded into a high-dimensional one-hot vector and is further transformed into low-dimensional dense features by the *Embedding Layer*. Thus *crossing features*, *user historical behaviors*, *user profile* and *target category* are transformed as $e_{\mathcal{F}_{uc_t}}$, E_{I_u} , $e_{\mathcal{P}_u}$ and e_{c_t} respectively. Note that, $E_{I_u} = \{e_1, e_2, \dots, e_T\} \in \mathbb{R}^{T \times d_{model}}$, where T represents the length of user historical behaviors and d_{model} is the dimension of item embedding e_i .

The *User Net* extracts user interests $e_u \in \mathbb{R}^d$ from user historical behaviors E_{I_u} and user profile features $e_{\mathcal{P}_u}$. Multi-Head Self-Attention (MHSA) [20] is used here to refine the behavior's item representation. Given the user historical behaviors E_{I_u} as input, the MHSA outputs $E_u^s \in \mathbb{R}^{T \times d}$, which is formulated as:

$$\begin{aligned} E_u^s &= \text{MHSA}(E_{I_u}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_H) \mathbf{W}^O, \\ \text{head}_h &= \text{Attention}(E_{I_u} \mathbf{W}_h^Q, E_{I_u} \mathbf{W}_h^K, E_{I_u} \mathbf{W}_h^V) \\ &= \text{Softmax}\left(\frac{E_{I_u} \mathbf{W}_h^Q \cdot (E_{I_u} \mathbf{W}_h^K)^T}{\sqrt{d_h}}\right) \cdot E_{I_u} \mathbf{W}_h^V, \end{aligned} \quad (1)$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{d_{model} \times d_h}$ are projection matrices of the h -th head for query, key and value respectively, The H is the number of heads and $\mathbf{W}^O \in \mathbb{R}^{Hd_h \times d}$ is a linear matrix. Finally, a feed-forward network, consisting two fully-connected layers and ReLU activation [2], is adopted to fusing E_u^s and user profile features $e_{\mathcal{P}_u}$,

$$e_u = \text{ReLU}\left(\text{ReLU}((e_{\mathcal{P}_u} \oplus \text{sum}(E_u^s)) \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2)\right), \quad (2)$$

where $\text{sum}(\cdot)$ is a sum-pooling operator and \oplus means concatenation, $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{b}_1, \mathbf{b}_2$ are learnable parameters of the feed-forward network.

The *Category Net* is responsible for learning the latent embedding $\mathbf{e}'_{c_t} \in \mathbb{R}^d$ of target category c_t . Here, we adopt two fully-connected layers with ReLU activation [2] to obtain a more expressive embedding for category,

$$\mathbf{e}'_{c_t} = \text{ReLU}(\text{ReLU}(\mathbf{e}_{c_t} \mathbf{W}_3 + \mathbf{b}_3) \mathbf{W}_4 + \mathbf{b}_4), \quad (3)$$

where $\mathbf{W}_3, \mathbf{W}_4$ and $\mathbf{b}_3, \mathbf{b}_4$ are learnable parameters. It should be noted that these two deep networks, *User Net* and *Category Net*, are beneficial to the N-diversity next category prediction task due to their superior generalization performance.

The *Wide Layer* is a generalized linear layer with the form $\text{Wide}(\mathcal{F}) = \mathbf{w}^T \mathcal{F} + b$ where \mathcal{F} consists of the crossing feature's embedding $\mathbf{e}_{\mathcal{F}_{uc_t}}$ and the dot product of user embedding and category embedding which are outputs from *User Net* and *Category Net* respectively. The *Wide Layer* has a better memorization capability which is suitable to improving the performance in U-diversity next category prediction task. In summary, after input $\langle \mathcal{F}_{uc_t}, \mathcal{I}_u, \mathcal{P}_u, c_t \rangle$, we get the final user-category score:

$$\begin{aligned} f(u, c_t) &= \mathbf{w}^T \cdot (\mathbf{e}_{\mathcal{F}_{uc_t}} \oplus \text{Deep}(\mathcal{I}_u, \mathcal{P}_u, c_t)) + b \\ &= \mathbf{w}^T \cdot (\mathbf{e}_{\mathcal{F}_{uc_t}} \oplus \langle \mathbf{e}_u, \mathbf{e}'_{c_t} \rangle) + b, \end{aligned} \quad (4)$$

where *Deep* contains *Embedding Layer*, *User Net* and *Category Net*, \mathbf{w} and b are learnable parameters of *Wide Layer*, and its output is a scalar.

3.3.3 Combined Loss Function. To improve the model performance, we employ a combined loss function consisting of cross entropy loss and triplet loss functions. The probability that a user u will interact with an item with category c_t is calculated as:

$$\Pr(c_t | \mathcal{I}_u, \mathcal{P}_u, \mathcal{F}_{uc_t}) = \frac{\exp(f(u, c_t))}{\sum_{c_j \in \mathcal{C}} \exp(f(u, c_j))}, \quad (5)$$

then we define the cross entropy loss L_c as follows:

$$L_c(u, c_t) = -\log(\Pr(c_t | \mathcal{I}_u, \mathcal{P}_u, \mathcal{F}_{uc_t})). \quad (6)$$

The sum operator of Equation 5 is computationally expensive. Thus, the sampled softmax technique [8] with n_{neg} negative categories generated by random sampling from user non-interacted category set is adopted to train our model more time-efficiently.

In addition, we introduce the triplet loss [12] to enhance the ranking capability of our model. A heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is created based on user-category interactions, where $\mathcal{V} = \mathcal{U} \cup \mathcal{C}$. There is an edge $e_{u,c} \in \mathcal{E}$ if u has interacted with an item with category c . Then, for each positive sample data $\langle u, c_t \rangle$, we sample n_{nei} neighbor categories from u as c_{nei} . Ultimately, we calculate the triplet loss L_t as follows:

$$\begin{aligned} L_t(u, c_t, c_{\text{nei}}) &= \frac{1}{n_{\text{nei}}} \sum_{c_j \in c_{\text{nei}}} \max(\mathbf{sign}(s_{uc_t} - s_{uc_j}) \\ &\quad \cdot (f(u, c_j) - f(u, c_t)) + m, 0), \end{aligned} \quad (7)$$

where the margin m is a hyper-parameter in triplet loss. $\mathbf{sign}(s) = 1$ when $s > 0$, otherwise $\mathbf{sign}(s) = -1$. Suppose we have $\langle u, c_j \rangle$ where $s_{uc_t} > s_{uc_j}$, this loss function will steer our model's output to

Table 2: Statistics of the offline datasets.

	Dataset	Users	Categories	Interactions
offline	Taobao	50,000	3,266	2,628,956
	Alibaba.com	50,000	4,360	3,419,462

ensure that $f(u, c_t)$ is greater than $f(u, c_j)$ by a margin m , thereby improving the ranking ability of our model. Note that s_{uc_t} is the number of interactions between user u and category c_t .

The overall loss function is a weighted sum of cross entropy and triplet loss functions,

$$L = \sum_{(u, c_t) \in \mathcal{D}} L_c(u, c_t) + \lambda \cdot L_t(u, c_t, c_{\text{nei}}), \quad (8)$$

\mathcal{D} is the collection of training samples, and λ is a hyper-parameter.

3.4 Stage II: Constrained Item Matching

The objective of Constrained Item Matching in CCDF is to retrieve items accurately within the constraints of the top- K categories which are output from previous stage. A popular mechanism in industry, posterior weighted score, is used here. We calculate the real posterior click-through rate (CTR), conversion rate (CVR), and conversion purchase rate (CPR) for each item. The final score of each item is obtained by a weighted sum of CTR, CVR, and CPR, with weights 1, 10, and 100 respectively. These items are grouped by categories and the top- N items of each category are retained. The above process is formulated as

$$\begin{aligned} \text{score}_i &= 1 \times \text{CTR}_i + 10 \times \text{CVR}_i + 100 \times \text{CPR}_i, \\ \mathcal{I}_c^N &= \underset{i \in \mathcal{I}_c}{\text{argmax}}(\text{score}_i, N), \end{aligned} \quad (9)$$

where $\text{argmax}(\text{score}_i, N)$ returns the top- N items according to above calculated score. The top- K categories in the first stage serve as triggers to retrieve items.

4 EXPERIMENTS

In this section, we perform offline experiments and online A/B testing to validate the effectiveness of CCDF. We evaluate the category diversity performance of our DeepU2C model by comparing it with these existing state-of-the-art diversity-oriented methods on offline real-world datasets. We also show the superiority of the whole CCDF framework by conducting real-world online A/B testing.

4.1 Offline Experimental Setup

4.1.1 Datasets. We evaluate the effectiveness of our method on category diversity performance with two real-world datasets. One of them is from an existing public e-commerce dataset, which is named as Taobao¹. The other is collected from our real-world industrial e-commerce website, which is referred as Alibaba.com. Considering the evaluation of original large datasets is very time-consuming, we sample the top 50,000 active users according to interactions for both datasets. We designate logs from the penultimate day as the validation set, logs from the last day as the test set and the rest logs as the training set. The statistics of these two datasets are shown in Table 2. The user historical behaviors are truncated with the

¹<https://tianchi.aliyun.com/dataset/649>

Table 3: Offline experimental results on real-world datasets. The bold number is the best result, while the underlined number is the second best result.

Alibaba.com				Taobao			
Method	HR@5	HR@15	HR@30	HR@5	HR@15	HR@30	
U^*	Statistics	<u>0.601</u>	0.718	1.000	0.500	<u>0.755</u>	1.000
	LightGCN	0.523	<u>0.733</u>	0.833	<u>0.525</u>	0.731	0.829
	DGCN	0.449	0.674	0.790	0.448	0.669	0.783
	DGRec	0.375	0.577	0.720	0.390	0.587	0.701
	DeepU2C	0.688	0.810	<u>0.843</u>	0.781	0.843	<u>0.875</u>
N^*	Statistics	0.000	0.000	0.000	0.000	0.000	0.000
	LightGCN	0.088	0.228	0.368	0.061	0.149	0.255
	DGCN	<u>0.107</u>	0.241	0.372	0.068	0.156	0.264
	DGRec	0.098	<u>0.256</u>	<u>0.380</u>	<u>0.085</u>	<u>0.168</u>	<u>0.278</u>
	DeepU2C	0.245	0.458	0.610	0.248	0.429	0.580

* U means U-diversity task, N means N-diversity task.

length of 20 (i.e., $|I_u| = 20$) for offline experiments and 30 for online A/B testing respectively. This truncation scheme is a common and reasonable practice for balancing complexity and accuracy.

4.1.2 Baselines. We compare our DeepU2C model with the following state-of-the-art methods.

- **Statistics:** A naive statistical method obtaining the preferred categories by considering the user historical behavior times-tamp and number of interactions with categories.
- **LightGCN** [11]: A GCN-based model but removes the transformation matrix, non-linear activation, and self-loop.
- **DGCN** [28]: A GNN-based diversified recommendation method with rebalanced neighbor discovering and category-boosted negative sampling.
- **DGRec** [25]: A variation of GNN-based diversified recommendation method with submodular neighbor selection and layer attention.

4.1.3 Evaluation Metric. Hit Ratio (HR) is used to measure the category accuracy of next category prediction task, defined as:

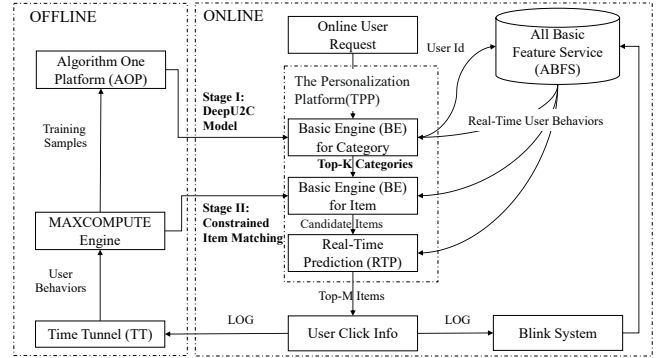
$$HR@K = \frac{\sum_{(u, c_t) \in \mathcal{D}_{test}} \sigma(\text{target category } c_t \text{ in top-}K)}{|\mathcal{D}_{test}|}, \quad (10)$$

where $\sigma(\cdot)$ denotes indicator function and \mathcal{D}_{test} denotes the test set consisting of pairs of user and target category (u, c_t) . To evaluate all methods comprehensively, we divide \mathcal{D}_{test} into two parts for U- and N-diversity tasks based on whether $c_t \in C(I_u)$ or not.

4.1.4 Parameter Setting. For parameter setting, we tune all methods using the validation set. Adam [18] is used as optimizer with a mini-batch size of 100. The learning rate is fixed as 0.001. The embedding sizes of input features, E_{I_u} , e_{c_t} , $e_{\mathcal{F}_{uc_t}}$, and e_{p_u} , are set to 64, 32, 8, and 8, respectively. The number of heads H used in MHSA is set to 8, and the fully-connected layers in *User Net* and *Category Net* employ hidden layers of size 64 and 32. We randomly sample 500 (i.e., $n_{neg} = 500$) negative categories and 5 (i.e., $n_{nei} = 5$) neighbor categories for each positive sample. Other hyper-parameters m and λ in the loss function are tuned by Grid Search with $[0.2, 0.3, 0.4, 0.5]$ and $[0.1, 0.5, 1.0, 1.5, 2.0]$ respectively. To achieve the best results, the hyper-parameter m is set as 0.4 and λ is set as 1.0. We stop training until convergence within 5 epochs.

Table 4: Offline ablation experimental results on Alibaba.com.

Method		HR@5	HR@15	HR@30
U	DeepU2C	0.688	0.810	0.843
	w/o triplet loss	0.645	0.782	0.810
	w/o wide	0.483	0.602	0.668
	w/o deep	0.534	0.689	0.788
N	DeepU2C	0.245	0.458	0.610
	w/o triplet loss	0.221	0.387	0.552
	w/o wide	0.233	0.435	0.578
	w/o deep	0.000	0.000	0.000

**Figure 4: The online implementation architecture of our proposed CCDF.**

4.2 Offline Experimental Results

The offline experimental results are shown in Table 3. Regarding U-diversity task, where the next category has occurred in user historical behaviors, our proposed DeepU2C achieves the best values of HR@5 and HR@15 for both Alibaba.com and Taobao datasets. LightGCN and Statistics are strong competitors. For the N-diversity task, where the next category does not occur in user historical behaviors, DGCN and DGRec obtain relatively better performance compared to their performance in U-diversity task. The reason is that they try to aggregate diverse categories in the graph. Nevertheless, our proposed method achieves 0.245, 0.458, and 0.610 for HR@5, HR@15, and HR@30 on Alibaba.com respectively, ranking first place among various competitors. Similar results are observed on Taobao dataset. These results show the better generalization performance of the DeepU2C model, especially for non-interacted categories. One of the reasons can be attributed to the effectiveness of deep networks including *User Net* and *Category Net*, and the elaborated combined loss function.

The ablation experimental results on Alibaba.com are shown in Table 4. We observe that *Wide Layer* greatly improves the performance of DeepU2C on U-diversity task due to their excellent memorization capability. For example, after being equipped with the *Wide Layer*, the value of HR@5 is increased from 0.483 to 0.688 on U-diversity task. For N-diversity task, deep networks in *User Net* and *Category Net* which have a better generalization capability play a crucial role. Thus, the values of HR@5, HR@15 and HR@30 are dropped to zero after removing them. While the *Wide Layer* has little effect on these metrics. Note that the triplet loss has positive effects on both two tasks.



Figure 5: Two real-world cases from online traffic. The upper part is user historical behaviors in chronological order, while the lower part is the recommended categories from CCDF. U-diversity categories are in blue and N-diversity categories are in red.

4.3 Online Implementation Architecture

We have deployed our proposed CCDF in a real-world production environment to do online A/B testing. Specifically, CCDF is deployed in Alibaba.com homepage recommendation scene, *Just for You*, by leveraging several algorithm platforms in Alibaba Group. Figure 4 illustrates the online architecture of our proposed CCDF. There are two main parts in the architecture, online and offline parts.

The online part is responsible for generating the final top- M items that will be exposed to the end user. To be more specific, the online user request is served by The Personalization Platform (TPP) where the match and rank modules are inside. The DeepU2C model is deployed in Basic Engine (BE) for Category. To make the online serving more time-efficient, the category embeddings from *Category Net* are calculated offline beforehand. We only calculate the remaining parts in real-time. In practice, the response time is usually less than 2ms for thousands of candidate categories. The Constrained Item Matching is deployed in BE for Item. It accepts the top- K categories for item retrieval. Note that this match module has also deployed other candidate generation methods, such as item-based Collaborative Filtering method and deep candidate generation models. The next module is Real-Time Prediction (RTP), which is responsible for calculating each candidate item's score elaborately. Then the final top- M items will be exposed to the end user. All Basic Feature Service (ABFS) is utilized here to return necessary user features, such as user profile features, real-time user historical behaviors, etc.

For the offline part, the user logs are recorded by Time Tunnel (TT). A big data platform, called MAXCOMPUTE, is used to process the logs in order to prepare training samples for training DeepU2C model. Algorithm One Platform (AOP) will consume these training samples to train the DeepU2C model. Once the training is finished, the model will be pushed to BE for Category for online serving. Note that, the posterior weighted score for each item in Constrained

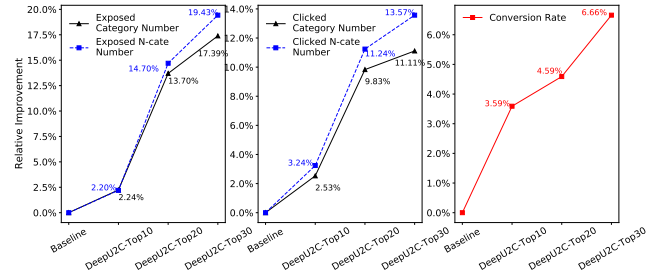


Figure 6: The improvements of CCDF with different top- K parameters on online A/B testing compared with the baseline. K is the controllable parameter for diversity which is set as 10, 20, and 30.

Item Matching is calculated in MAXCOMPUTE. Then \mathcal{I}_c^N is implemented as an inverted index in BE for Item. N is set as 300 for online experiments.

4.4 Online A/B Testing

Usually, the efforts to deploy a new model in the production environment are not trivial, thus we compare our proposed DeepU2C, which has obtained the best performance in offline experiments, with existing online baseline Statistics method for two weeks. Both of them are employed with the same Constrained Item Matching, i.e., posterior weighted score. To make comparisons fairly, all of other parts are untouched. For training online version of DeepU2C, we collect the 90-day of real online user interactions from our platform, which has millions of users and thousands of categories.

4.4.1 Numerical Results. The A/B testing results are presented in Figure 6. When $K = 10$, the exposed category number and clicked category number are increased by 2.24% and 2.53% respectively. Furthermore, the corresponding conversion rate is increased by 3.59%, which proves the assumption that diversity is able to improve user's experiences and increase website revenues. When we tweak the K to 20 and further to 30, the exposed category number is increased by 13.70% and 17.39%, the clicked category number is

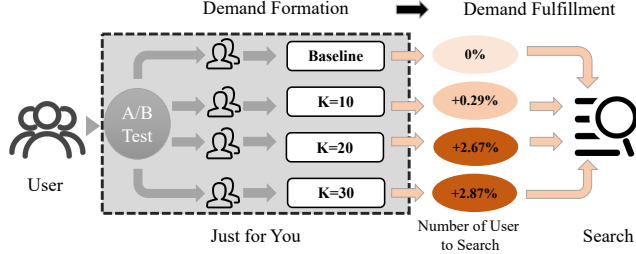


Figure 7: The demonstration of user flow from homepage recommendation scene, Just for You, to search.

increased by 9.83% and 11.11%. The conversion rate is increased by 4.59% and 6.66%. Meanwhile, it is important to report the exposed and clicked N-category number which are used to denote the N-diversity. These values indicate how many categories that have not occurred in user historical behaviors are exposed and clicked. Under K in settings 10, 20, and 30, the exposed N-category number and clicked N-category number are increased by 2.20%, 14.70%, 19.43% and 3.24%, 11.24%, 13.57% respectively. All these improvements are statistically significant using an unpaired t-test. These findings provide clear validation of the superiority of our proposed CCDF. With an increase in K , not only does the diversity level improve, but also there is a significant enhancement in conversion rate. The continuous improvements of N-diversified metrics also indicate that our framework is able to help user to obtain new products and discover new needs effectively, and then alleviate the echo chamber/filtering bubble effects of our system. Notably, the retrieval of N-diversified items, which are neglected by other candidate generation methods in match module, is the key reason why our framework is able to achieve huge improvements on conversion rate online.

4.4.2 Case Study. We present two real-world cases from online traffic in Figure 5. User historical behaviors are listed in the upper part, while the recommended categories from CCDF are shown in the lower part. For case A, the user has expressed strong interests in *Lip Gloss* by clicking products (1), (2), and (3), and *Face Mask* by clicking products (5) and (7). Meanwhile, this user has some sparse interests, *Women's T-Shirts* and *Sunglasses*. Our proposed CCDF successfully predicts the user's preferences for these four categories, which would result in a high U-diversity. Furthermore, when the K is increased, we find that more N-diversity categories have been presented. These categories are relevant to the user's preferences to some extent, which broadens user's interests. For example, the new categories *Full Strip Eyelashes*, *Women's Perfume*, *UV Gel* and *Mascara Brushes* are extended from user original interests *Lip Gloss* and *Face Mask*. Besides categories in cosmetics, there is a high probability that a lady would like to view products in *Casual Dresses*, *Handbags*, *Earrings*, and *Necklace*. The same conclusions can also be observed in case B. *Smartphone* is extended to *Data Cables* and *Business laptops*, *Living Room Chairs* is extended to *Living Room Sofas* and *Dining Tables*, etc. Both cases have verified the effectiveness of CCDF, which is able to achieve high U-diversity and N-diversity by tweaking the value of K .

4.4.3 The Impacts on User Search. These online experiments on homepage recommendation scene, *Just for You*, also have positive impacts on user's search activities. Figure 7 shows the user flow

Table 5: Experimental results of users from Just for You to search on online A/B testing.

Method	Number of User to Search	Conversion Category Number in Search	Conversion Rate in Search
Baseline	0%	0%	0%
K=10	+0.29%	+0.97%	+2.28%
K=20	+2.67%	+2.28%	+3.57%
K=30	+2.87%	+4.01%	+5.54%

from *Just for You* to search under different experimental settings. Table 5 presents the experimental results of users from *Just for You* to search. We observe that when the value of K is increased, the diversity-level of recommendation list is increased (as presented in Section 4.4.1), then more users in recommendation scenes are willing to go to search. The number of users from *Just for You* to search is increased by 0.29%, 2.67%, and 2.87% with K setting as 10, 20, and 30 respectively. One of the reasons is that the diversified recommendations help user to access more items, and in turn encourage user to go to search to refine their needs. The improved conversion category number and conversion rate in search further validate the promotional effect of diversified recommendations on user search. More specifically, in search, the conversion category number of these users are increased by 0.97%, 2.28%, and 4.01% respectively, and the conversion rate of these users are improved by 2.28%, 3.57%, and 5.54% respectively. The above experimental results support the existence of *complementary effects* [27] between recommendation and search. Some customers may form new demands while browsing diverse results in recommendations (demand formation) and then they actively initiate a search to meet their needs more efficiently (demand fulfillment).

5 CONCLUSION

In this paper, we innovatively define the diversity as two distinct definitions, user-explicit diversity (U-diversity) and user-item non-interaction diversity (N-diversity) based on user historical behaviors. Then, we propose a succinct and effective Controllable Category Diversity Framework (CCDF), which is able to explicitly control category diversity to achieve both high U-diversity and N-diversity. There are two stages in CCDF, User-Category Matching and Constrained Item Matching. We formalize User-Category Matching as a next category prediction task and construct a DeepU2C model which incorporates User Net, Category Net, Wide Layer, and a combined loss. This two-stage framework divides the responsibilities of category diversity and item accuracy into two different stages, which equips CCDF with the controllability on category diversity by parameter K . The larger the parameter K , the more N-diversified items will be retrieved. As a result, it is beneficial to alleviate echo chamber/filter bubble effects. Offline experimental results on real-world datasets and online A/B testing results validate the superiority of our framework. The complementary effects between recommendation and search have been observed in our online A/B testing.

In the future, we will strive to improve the accuracy of DeepU2C model, especially for N-diversity task in stage I. Then we will try to model user's preferences for items under constraint of specific category through deep networks to improve the accuracy of stage II.

REFERENCES

- [1] Zeinab Abbassi, Vahab S Mirrokni, and Mayur Thakur. 2013. Diversity maximization under matroid constraints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 32–40.
- [2] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).
- [3] Jaime G. Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- [4] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2942–2951.
- [5] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- [6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishikesh Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ipsir, et al. 2016. Wide & deep learning for recommender systems. In *the 1st workshop on deep learning for recommender systems*. 7–10.
- [7] Sang-Min Choi and Yo-Sub Han. 2010. A content recommendation system based on category correlations. In *2010 Fifth International Multi-conference on Computing in the Global Information Technology*. IEEE, 66–70.
- [8] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.
- [9] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*.
- [10] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.
- [11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*.
- [12] Elad Hoffer and Nir Ailon. 2015. Deep Metric Learning Using Triplet Network. In *Similarity-Based Pattern Recognition*. Springer International Publishing, Cham, 84–92.
- [13] Rong Hu and Pearl Pu. 2011. Helping users perceive recommendation diversity.. In *DiveRS@ RecSys*. 43–50.
- [14] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems – A survey. *Knowledge-Based Systems* 123 (2017), 154–162.
- [15] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2615–2623.
- [16] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Comput.* (2003).
- [17] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.
- [18] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the Convergence of Adam and Beyond. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net.
- [19] Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. 2021. Sparse-Interest Network for Sequential Recommendation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*. ACM, 598–606.
- [20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [21] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical Diversified Recommendations on YouTube with Determinantal Point Processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.
- [22] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- [23] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Fernando Diaz, and Xue Liu. 2022. A Survey of Diversification Techniques in Search and Recommendation. *arXiv preprint arXiv:2212.14464* (2022).
- [24] Zhibo Xiao, Luwei Yang, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. 2020. Deep multi-interest network for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.
- [25] Liangwei Yang, Shengjie Wang, Yunzhe Tao, Jiankai Sun, Xiaolong Liu, Taiqing Wang, and Philip S. Yu. 2023. DGRN: Graph Neural Network for Recommendation with Diversified Embedding Generation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*.
- [26] Rui Ye, Yuqing Hou, Te Lei, Yunxing Zhang, Qing Zhang, Jiale Guo, Huaiwen Wu, and Hengliang Luo. 2021. Dynamic graph construction for improving diversity of recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*.
- [27] Zhe Yuan, AJ Chen, Y Wang, and T Sun. 2022. *How Recommendation Affects Customer Search: A Field Experiment*. Technical Report. Working Paper.
- [28] Yu Zheng, Chen Gao, Liang Chen, Depeng Jin, and Yong Li. 2021. DGCN: Diversified recommendation with graph convolutional networks. In *Proceedings of the Web Conference*.
- [29] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5941–5948.
- [30] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.