Full-stage Diversified Recommendation: Large-scale Online Experiments in Short-video Platform

Nian Li*
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China

Yunzhu Pan*
University of Electronic Science and Technology of China
Chengdu, China

Chen Gao†
Depeng Jin
Tsinghua University
Beijing, China

Qingmin Liao
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China

ABSTRACT

The recommender systems on online platforms assist users in finding personalized information, yet this also leads to the issue of limited diversity, potentially giving rise to societal issues such as filter bubbles. Despite significant progress in diversified recommendation algorithms, they have not been extensively experimented with and evaluated for effectiveness in large-scale, full-stage industrial recommender systems. Specifically, industrial recommenders usually consist of three stages of matching, ranking, and re-ranking, in which specific characteristics lead to critical challenges for promoting both recommendation diversity and user engagement. First, user interests are partially observed due to only relevance maximization. Second, item-side feature-aware bias causes imbalanced recommendations. Last, the impact of diversity perception on user engagement stresses the necessity of explicit diversity modeling. To address these challenges in industrial systems, in this work, we deploy several existing diversified algorithms in a real-world short-video platform, including exploration-exploitation, feature-aware debiasing, and diversity optimization. We conduct large-scale online A/B testing for evaluation via online metrics of user engagement and recommendation diversity. Performance improvement across full stages demonstrates the effectiveness of these simple solutions. From comparing performance across different stages and algorithms, we identify that the ranking stage is the most suitable for real-world deployment, and the combination of debiasing and diversity optimization is a promising direction in terms of diversified recommendations. This work provides experiential guidance for the large-scale deployment of diversified algorithms and the construction of a more inclusive platform on the Web.

CCS CONCEPTS

• **Information systems → Information systems applications**.

*Contribute equally to this work.
†Corresponding author (chgao96@tsinghua.edu.cn).

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05.
https://doi.org/10.1145/3589334.3648144

KEYWORDS

Diversified Recommendation; Full-stage Recommender System; Online Experiments

ACM Reference Format:
Nian Li, Yunzhu Pan, Chen Gao, Depeng Jin, and Qingmin Liao. 2024. Full-stage Diversified Recommendation: Large-scale Online Experiments in Short-video Platform. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589334.3648144

1 INTRODUCTION

As an information filter, the recommender system plays an indispensable role in various platforms on the Web, such as video, takeout, and music services, *etc.* Users are exposed to tailored items for both their satisfactory experience and the platform's commercial profit. Therefore, optimizing for *relevance* to provide items best matching user interests has always been the most essential objective for recommendations. However, only pursuing relevance may lead to users' homogeneous consumption and further cause societal issues concerning consumers, information providers, and recommender practitioners, like filter bubble [35], echo chamber [43], and information cocoon [29]. To address it, many works have concentrated on diversified recommendations in recent years, which improves item diversity exposed to users while maintaining accuracy. Most approaches [5, 25, 30, 45] improve the diversity by pre-defined strategies or explicitly joint modeling with relevance. Besides, more advanced models have been adopted for diversified recommendations, such as graph neural networks (GNN) [53, 57].

Despite recent progress in diversified algorithms for recommendations, most works verify model effectiveness by the trade-off between ranking metrics (*e.g.*, Recall, AUC) and diversity metrics (*e.g.*, Coverage, ILAD) on collected offline datasets [31, 53, 57]. Some works also present relatively simple online results of a single diversified algorithm in real-world systems [46]. However, online performance is still unexplored extensively and compared between **multiple diversified algorithms** when deployed in **full-stage** recommender systems of real-world applications. To be more specific, in the scenario of industrial applications, the number of items to be processed ranges from millions to billions. Therefore, the recommender system is usually divided into multiple stages, including matching, ranking (coarse and fine ranking), and re-ranking [16].

Each stage has specific characteristics, such as the scale of candidate items, objectives when serving the whole recommender, *etc.* On the one hand, along with these characteristic stages, diversified recommendations face three critical challenges, including partially observed user interests, item-side feature-aware bias, and users' diversity perceptions. On the other hand, when deployed in real-world applications, the online effects of existing diversified algorithms on user engagement (how positive or negative the user feedback is) and exposure diversity may be not guaranteed. The performance can also differ in terms of deployment across full stages. For example, it will be questionable whether deploying the popular maximal marginal relevance (MMR) [5] in the ranking stage improves the diversity of final exposed items to users.

In this work, we deploy existing diversified algorithms of exploration-exploitation, feature-aware debiasing, and diversity optimization to address the aforementioned challenges in industrial systems. To extensively evaluate their online effects, we deploy these algorithms across the full-stage recommender system of a short-video platform with over 100 million daily active users. The deployment is summarized as follows.

- **Exploration and exploitation (E&E) for capturing complete user interests.** Due to the maximization of relevance, the normal recommendation model tends to constantly exploit items that users have given positive feedback, thus only capturing partial user interests [29]. E&E algorithms aim to balance the exploitation of items with known utility and the exploration of items with uncertain utility for better user engagement and diversity [14, 27, 38]. That is to say, the recommendation includes more diverse and new items that users are satisfied with. In this work, we adopt bandit-based and hard-strategy-based algorithms for exploration, leveraging the content-based categorical system in the video platform (see Section 3 for details). Roughly speaking, we identify specific categories to be explored or exploited according to estimated utility, from which videos are further selected. Regarding the deployment, we choose the matching and ranking stages, of which the item pool is relatively large.
- **Debiasing algorithm for mitigating item-side feature-aware bias.** Since users only give positive feedback to items with specific features, ranking results by the relevance of model prediction are usually of low diversity. Take the video platform as an example, if film videos are much more popular among users than religious ones, the recommendation will consist of more film videos, bringing a certain homogeneity. This phenomenon can be seen as item-side feature-aware bias [59], where 'feature' means item feature such as video category. Therefore, we attempt to debias by adjusting model predictions. Specifically, videos are ranked top when the model predicts that they will obtain much better feedback than average levels of all the videos of the same category. The debiasing algorithm is deployed in the ranking stage.
- **Diversity optimization for users' diverse perception.** Many existing works have demonstrated that the perception of low diversity or repeated recommendations can harm user engagement [26, 51]. For example, users can give negative feedback if the instant impression of exposed items is homogeneous, even if they may be interested in them. Therefore, ranking models are proposed to optimize the combination of relevance and diversity [5, 7, 22]. We choose to model diversity perception in the form of the sliding window with the advanced work SSD [22], which can imitate users' realistic experience in the scenario of the short video. The deployment lies in fine ranking and re-ranking stages, whose candidate videos range from tens to hundreds. In this way, the computational complexity is acceptable in the end phase of online service.

We conduct large-scale experiments of A/B testing involving over ten million users. The performance is evaluated by the relative change of important online metrics based on the comparison between experimental and base groups. We collect metrics from three aspects, including user interaction, recommendation diversity, and users' dwell time in the platform. Improvement in most results demonstrates the effectiveness of the deployed algorithms. From performance comparisons across full stages, we conclude that the fine ranking stage is the most suitable for diversified recommendations. As for algorithms, debiasing and diversity optimization perform better, and their effective combination may be a promising direction for diversified recommendations.

To summarize, the main contributions of this work are as follows,

- We investigate the full-stage effects of multiple diversified algorithms in the real-world recommender system with online experiments.
- We deploy the aforementioned algorithms and conduct large-scale A/B experiments across the full stages in the recommender system of a real-world short-video platform. There is up to 4.645% improvement in terms of important online metrics. Empirical results can contribute to the online deployment of diversified recommendation algorithms and the mitigation of societal issues on the Web.

2 FULL-STAGE RECOMMENDER SYSTEM

Due to the huge amount of items in the real-world application, online recommender systems are usually deployed as the architecture of multiple stages, including matching, ranking, and re-ranking [16]. Each stage tackles items of different magnitudes from the upstream stage. There are specific characteristics of objective, model design, and data input across these stages. Therefore, existing diversified algorithms have distinct designs when serving at different stages. The overview of the full-stage recommender system is shown in Figure 1.

2.1 Matching

The input of this first stage is millions or even billions of items. For the real-time response to user feedback in online services, the matching must retrieve items roughly meeting user interests as efficiently as possible. There are usually multiple matching channels to filter a broad scope of items as the candidate pool for the following stage.

Since the efficiency of selecting items from large-scale candidates is an essential problem, diversified algorithms proposed in the matching stage are usually based on collaborative filtering (CF). Some works improved recommendation diversity through predefined strategies. For example, Kwon *et al.* [25] identified four types of users and adopted four filtering algorithms independently

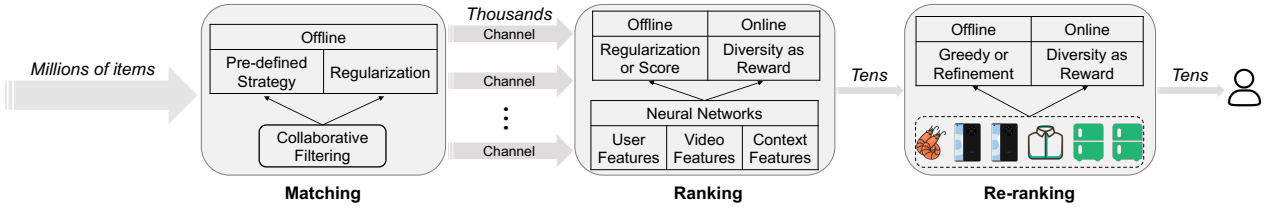


Figure 1: The full-stage pipeline of online recommender system in industrial applications. Each stage has specific designs for diversified algorithms when serving the whole system.

to select recommendation items, which were further merged to increase the diversity. Cheng *et al.* [10] pre-selected relevant and diverse items for each user to train a model of parameterized matrix factorization (MF) coupled with a structural SVM. Besides, there have been diversified models based on GNNs, in which neighborhood or negative sampling is enhanced for more information propagation of the disadvantaged items [52, 57]. Diversity modeling in an explicit way is another direction. Specifically, diversity is defined based on item relationships and further serves as regularization in the loss function of the recommendation model. For instance, Wasilewski and Hurley [45] proposed to combine MF loss and diversity regularization for the optimization of users' and items' embeddings. These works are mainly evaluated on offline datasets with observed user-item interactions because online experiments require handling large-scale items.

2.2 Ranking

This stage aims to rank items that users are interested in at top positions by *prediction scores* of the ranking model, *i.e.*, users' interaction probability under various behaviors. However, considering the requirements of low computational complexity and latency, the stage is usually further divided into two stages, **Coarse Ranking** and **Fine Ranking**. Generally speaking, due to the large scale of the candidate pool from the matching stage, the function of coarse ranking is to relieve computational pressure in the next fine ranking.

In this stage, more complicated models (*e.g.*, attention mechanism [37], recurrent neural network (RNN) [11]) and richer features (user, item, and context features) will be incorporated. On the one hand, diversity can also serve as the regularization term in the loss function for RNN-based sequential recommendations [9]. On the other hand, diversity can be combined with prediction scores for item selection. For example, Li *et al.* [30] proposed to calculate an item's diversity score for the generation of a recommendation list based on the occurrences of its category among already determined items. In addition to offline methods, some online diversified algorithms have also been developed for recommendations, which can collect user feedback continuously. Specifically, diversity is fused in the reward function for the action generation in multi-armed bandit (MBA) [14, 27] or RL framework [42, 56], which model invariant and variant user interests respectively [48].

2.3 Re-ranking

This stage processes tens of items from the output of the fine ranking. The major goal is to refine the recommendation list of several consecutive items for accuracy maximization since the list-wise context can influence user feedback as a whole [32]. Besides, various goals beyond accuracy are also taken into consideration for better user experience, such as diversity, fairness [12], serendipity [24], and even necessary operation objectives for the business. These measurements are based on the relationship among items of the recommendation list, which is approximately the exposure shown on the screens of user devices.

In the re-ranking, diversity is usually optimized along with relevance. In a greedy way, some works propose to generate the recommendation list by adding the item one by one, such as maximal marginal relevance (MMR) [5], determinantal point process (DPP) [7]. Furthermore, diversity can be defined as the form of a submodular objective function [2, 39]. In a refinement way, the existing recommendation list output from the ranking stage will be modified for diversity improvement. For example, Ziegler *et al.* [62] merged two lists independently ranked by relevance and diversity scores to generate a more diverse list. Yu *et al.* [54] proposed to swap the item contributing least to the list diversity with another one with the highest relevance score in the remaining items. As for online methods for diversified recommendations, there is no clear difference between the ranking and re-ranking stages. In other words, both MAB-based and RL-based algorithms can be adopted.

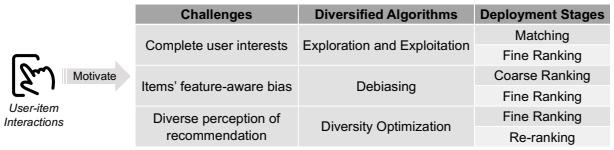
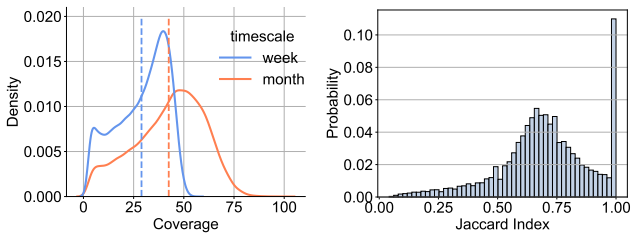


Figure 2: Overview of motivations, challenges, deployed diversified algorithms, and corresponding deployment stages.

3 DIVERSIFIED ALGORITHM

The overview of our proposed diversified algorithms is shown in Figure 2. Three algorithms are corresponding to critical challenges in diversified recommendations, *i.e.*, capturing users' complete interests, mitigating item-side feature-aware bias, and improving users' diversity perceptions. As the motivation, each challenge is further illustrated by the diversity analysis of user interactions in the short-video platform. The algorithms are deployed in different stages of the online recommender system, considering the complexity of the online system, we will only describe primary ideas.



(a) Coverage PDF (b) Histogram of Jaccard Index

Figure 3: (a) PDF of users' exposed videos for one week and one month. The vertical dotted lines represent the average of distributions. (b) The histogram of Jaccard Index between weekly and monthly exposure categories for each user.

The utilized dataset consists of a random sample of about 66,000 users' interaction records for one month, including two aspects of data for analysis,

- **User behavior.** In this platform, user behaviors denote various feedback to the videos, such as watching time, like, comment, collect, *etc.*
- **Video category.** This is the unified classification based on video content. Each video is marked with hierarchical categories of three levels by human or deep-learning algorithms. For example, a video of playing the drum kit will be tagged as 'Music', 'Musical instrument', and 'Drum kit'.

3.1 Exploration and Exploitation

3.1.1 Motivation. We first illustrate that user interests the recommender captures are incomplete and confined to limited video content. Figure 3 (a) shows the probability density function (PDF) of level-2 categorical coverage of users' exposed videos in the first week and the whole month, respectively. Obviously, the coverage is far smaller than the theoretical maximum (more than 300), demonstrating that the exposed videos cover a very small part of the whole platform content. In order to further investigate whether the recommender can explore new videos for a longer time, we calculate the categorical overlap between weekly and monthly individual exposure. Figure 3 (b) shows the histogram of Jaccard Index for each user. There is a clear peak around 1.0, and the rest concentrates around 0.7. Therefore, there is a limited increase when the exposure timescale extends from one week to one month, which means that the ability to explore more content for users needs improvement. In summary, the recommender captures user interests incompletely and most users consistently watch homogeneous videos with limited content.

3.1.2 Algorithm. E&E algorithm is based on the content-based categorical system in the platform. The general idea is to estimate category utility through user feedback, from which videos are further selected. Since exploring new content among tens of videos is actually meaningless, we choose the matching and ranking stage for deployment. Considering their difference in scales of candidate videos and available data, we deploy two algorithms individually.

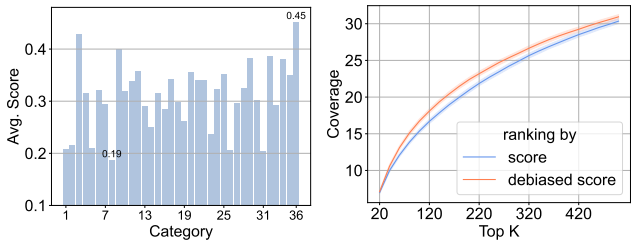
- **Matching.** We adopt a standard setting of multi-armed bernoulli bandit [6], where an arm denotes a category. Parameters in the likelihood function, *i.e.*, the cumulative reward of each category, are estimated by thompson sampling [6]. Specifically, each exposed video is associated with a category, and corresponding user feedback is collected to update the parameters of the category (arm). Top categories are selected based on the sampling θ_c from a beta distribution,

$$\theta_c \sim \textbf{Beta}(S_c + \alpha, F_c + \beta), \tag{1}$$

  where α and β are prior parameters. S_c and F_c denote the number of positive (*e.g.*, comment) and negative feedback (*e.g.*, watching less than 3 seconds) for category c. Finally, videos of these categories are sampled from a premium pool, in which candidate videos are of high quality and widely acclaimed. This serves as an additional matching channel, in which a certain proportion of videos must be ranked top in the subsequent stage of coarse ranking.
- **Fine Ranking.** The strategy consists of two parts, suppressing the most consumed categories and boosting novel ones. Specifically, we first aggregate the total watching time of each category in recent consumption and select the top N_a categories as a set $\mathcal{A}$. After ranking for video candidates by predicted scores from the base model, we further select categories of top N_b videos as a set $\mathcal{B}$. In this way, $\mathcal{A} \cup \mathcal{B}$ denotes the content that users are most interested in recently. We then randomly choose one category in $\mathcal{A} \cup \mathcal{B}$ to suppress, which is to reduce the weights (to be multiplied by prediction scores) of videos with this category in the ranking. In order to avoid a negative impact on ranking results based on model prediction, this operation only takes effect with the probability of p_s. As for exploring novel content, boosting categories not in $\mathcal{A} \cup \mathcal{B}$ directly will be risky, since user interests are not taken into consideration. Therefore, we additionally predict the categorical probability distribution of users' next consumption, and select top N_c categories as a set $\mathcal{C}$. Finally, one category in $\mathcal{C} \setminus (\mathcal{A} \cup \mathcal{B})$ is randomly chosen to increase its ranking weight. Similarly, this takes effect with the probability of p_b.

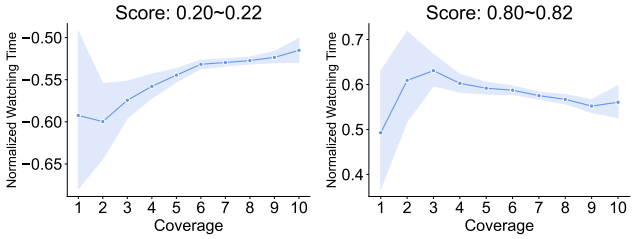
3.2 Debiasing in Ranking

3.2.1 Motivation. As stated in the introduction, there is usually a great item-side feature-aware bias of consumption propensity. In the video platform, this means that videos with different content can obtain different levels of user feedback. We calculate average prediction scores from base recommendation model for each level-1 category and show the distribution in Figure 4 (a). The maximum and minimum prediction scores among categories can differ by a factor of two to three, demonstrating the existence of item-side bias. Therefore, ranking videos by the scores will lead to massive exposure of categories obtaining high-level feedback from most users, which further brings homogeneous consumption. To verify the necessity of debiasing, we select top K videos with ranking scores or debiased scores (see the next subsection for details) for each user and calculate the level-2 categorical coverage. Figure 4 (b) shows the coverage under different K. Obviously, there are significant and steady coverage improvements when ranking by debiased scores, which means a more diverse candidate generation.



(a) Category Score (b) Coverage *w.r.t.* K

Figure 4: (a) Average scores of videos predicted by the ranking model for each level-1 category. (b) Coverage of top K videos when individually ranking for each user, by prediction scores and debiased scores respectively. The line shade represents 95% confidence interval.



(a) Low Relevance (b) High Relevance

Figure 5: Window-wise normalized watching time under different coverage, controlling for low and high average relevance. The line shade represents 95% confidence interval.

3.2.2 Algorithm. The primary idea of debiasing is to adjust prediction scores of videos by the propensity of user consumption in the platform [19]. To be more specific, the propensity is represented by the empirical probability of various user behaviors for each video, which takes the entire platform as the statistical caliber. For example, it will be 0.4 if one video is consumed by 400 users under certain behavior with 1000 exposures. In terms of debiasing, we adjust the prediction score s_{v_i} of video v_i with category c as follows,

$$\hat{s}_{v_i} = \frac{s_{v_i}}{\sum_{v_j \in \mathcal{V}_c} p_{v_j} / |\mathcal{V}_c|}, \tag{2}$$

where $\mathcal{V}_c$ denotes all the videos with category c to be ranked, and p_{v_i} is the propensity score. This formula indicates that the video will be ranked top when the user has a much higher probability to consume it than the general users, who have been exposed to videos of the same category.

For the deployment, we focus on the ranking stage, including **Coarse Ranking** and **Fine Ranking**. Besides, as shown in Figure 4 (b), the coverage improvement by ranking with debiased scores is trivial when the length of the video list is less than 40. Therefore, we exclude the re-ranking stage which only handles tens of candidate videos.

3.3 Diversity Optimization

3.3.1 Motivation. The perception of content repeatability (*i.e.*, low diversity) has been proven to influence user engagement significantly in many platforms, such as LinkedIn [26], TikTok [33], WeChat [51]. Therefore, direct optimization for diversity is as important and necessary to user engagement as relevance.

Here we verify this impact in the short-video platform. Specifically, the exposure sequence for each user is divided by multiple windows with a size of 10. We calculate the diversity, relevance, and user engagement of each window. For diversity, level-2 categorical coverage is utilized. For relevance, we estimate it as average prediction scores from the model. For user engagement, although the watching time is a good metric, it's greatly biased by video duration [58]. Therefore, we choose to further normalize it as $(w-\mu)/\sigma$, where w is the watching time. μ and σ denote the mean and standard deviation of all the watching time for all the videos with the same duration as the current video, respectively. As an important confounder, we control relevance at a certain level and show two representative relationships between window diversity and user engagement in Figure 5. It's obvious that the impact is significant, regardless of low or high relevance. Large fluctuations around coverage of $1 \sim 2$ show that users may give negative feedback when the diversity is extremely low. In addition, too much high diversity will also harm user experience sometimes.

3.3.2 Algorithm. Many post-processing approaches aim to maximize both relevance and diversity of the ranking list. In this way, exposed items will match user interests and provide users with diverse perceptions simultaneously. Two of the most popular post-processing methods are Maximal Marginal Relevance (MMR) [5] and Determinantal Point Process (DPP) [7]. However, as early efforts, the effectiveness of MMR is limited and the computation complexity of DPP is very high. Inspired by [22], we choose to maximize both relevance and diversity in the form of a sliding window. The advantages are as follows,

- **Realistic modeling of users' diversity perceptions.** In the utilized short-video platform, users consume videos with the sliding action, thus the perception is window-wise, *i.e.*, several consecutive videos. Therefore, diversity calculation for multiple sliding windows in an video sequence can imitate the process of user consumption and model realistic diversity perceptions.
- **Low computational complexity.** The computational complexity for DPP objective, *i.e.*, the occurrence probability of the chosen items, is $O(N^2 d)$, where N is the number of all the candidate items, and d is the dimension of item embedding. In contrast, it will be reduced to $O(NTd)$ when leveraging the trick of modified Gram-Schmidt orthogonalization, where T is the length of the recommendation list. Generally speaking, we have $T \ll N$ in the online ranking stage.

For each video sequence with length T, sliding through it with the fixed size w will generate L windows, where $L = \max(1, T - l + 1)$ when the sliding step is 1. Stacking these L windows and embedding each video will define a *trajectory tensor* $\mathbf{X} \in \mathbb{R}^{L \times l \times d}$. In this way, the diversity for the video sequence is represented by the volume of $\mathbf{X}$, which means that more diverse videos will span larger latent space in embeddings. Finally, the objective to be optimized combines both items' prediction scores and the diversity,

formulated as follows,

$$\max_{\{v_1,v_2,\cdots,v_T\}\subset\mathcal{V}} \sum_{i=1}^{T} s_{v_i} + \prod_{\sigma_{ijk}>0} \sigma_{ijk}, \tag{3}$$

where $\mathcal{V}$ is the set of candidate videos, and σ_{ijk} is the singular value of tensor $\mathbf{X}$.

Considering real-time requirements for online services, we focus on the stage of **Fine Ranking** and **Re-ranking**, where the number of videos to rank ranges from tens to hundreds.

4 ONLINE EXPERIMENTS

To evaluate the diversified algorithms introduced in Section 3, we deploy them in the full-stage recommender system of a short-video platform. For the experimental results, we concentrate on the online performance comparison from both stage and algorithm aspects.

4.1 Deployment and Evaluation Metrics

4.1.1 Deployment. All the experiments are conducted in large-scale A/B testing. Algorithms are deployed in randomly aligned experimental groups isolated from normal (base) groups. The experiments account for 10% or 20% of online traffic and involve over ten million users, lasting for 3 ∼ 13 days. The deployment details are summarized as follows.

- **Exploration and Exploitation.** For the matching, we set $\alpha = \beta = 0$. For the fine ranking, we set $N_a = 6, N_b = 6, N_c = 10$ and $p_s = 25\%, p_b = 15\%$.
- **Feature-aware Debiasing.** Although there are multiple behaviors to be predicted, we choose to leverage the scores for long view (see next subsection for definition) to rank videos for simplicity. In other words, s_{v_i} denotes the user's long-view probability for video v_i predicted by the ranking model.
- **Diversity Optimization.** In order to model interaction and content information simultaneously, we measure the similarity between videos with embeddings from both the ranking model and multimedia understanding. Specifically, we correspondingly define two trajectory tensors and combine their volumes together as overall diversity. For the window size, we set $l = 10$.

4.1.2 Evaluation Metrics. We leverage the relative improvement of important online metrics during the experiments to evaluate the algorithms. Specifically, interaction and diversity metrics are taken into consideration, representing the real-time user engagement and the content richness of individual exposed videos, respectively. Besides, we also report two of the most important platform metrics from the perspective of commercial business.

Interaction Metrics

- **Exposure** denotes the total number of exposed videos shown on the screens of user devices.
- **Like** denotes the total number of videos users have liked.
- **Comment** denotes the total number of videos users have commented on.
- **Long view** denotes the total number of videos that $w \geq \min(d, 18)$ and $d \geq 3$, where w is user's watching time and d is the video duration.

Diversity Metrics

- **Concentration** measures the timely repeatability of several consecutive videos. For each user, an exposure sequence is divided as some windows in the order of exposed time. In this way, the concentration of a window is calculated as $N - C$, where N is the fixed window size, and C is the number of unique level-1 categories in this window. For a certain experimental or base group during the experiments, the concentration is averaged over the sequential windows across all the users.
- **Coverage** measures the overall richness of exposed videos over a longer period. Formally, it's defined as the number of level-2 categories covered by the exposed videos. Similarly, the coverage is also averaged over all the users.

Platform Metrics

- **Sum. Time** denotes the dwell time in the platform during the experiments summed over all the users.
- **Avg. Time** denotes the dwell time in the platform during the experiments averaged over all the users.

4.2 Performance Comparison

Table 1 shows the A/B performance, *i.e.*, the relative improvement of experimental groups compared with base groups, for all the algorithms. We compare the performance between two different stages for each deployed algorithm and conclude distinct findings.

4.2.1 Exploration and Exploitation.

- **E&E algorithm performs better in the stage of fine ranking.** Across most interaction and diversity metrics, the algorithm gains up to 0.186% profit or obtains comparable results. In addition, the deployment in the fine ranking gains higher profit consistently compared with that in the matching stage, except for the coverage. Generally speaking, this can be explained by the fact that candidate videos contained in the fine ranking match user interests more accurately. In contrast, the additional channel in the matching stage is more likely to explore videos out of user acceptance, although the exposed videos cover richer content for a longer period. Therefore, compared to the matching stage, users give better feedback, and corresponding interaction metrics obtain greater improvement in the fine ranking.
- **E&E algorithm in the stage of fine ranking assists in exploring new video content that users are interested in.** To verify the ability of exploration, we further define **Novelty** and **Serendipity** metrics based on the number of new level-2 categories shown in the current window but not the past window of the individual video sequence, under specific user feedback. For novelty, it usually denotes the recommendation of new content, thus all the exposed videos are considered to constitute the sequence. For serendipity, it usually denotes new content interested users, thus only videos of effective consumption[1] are considered. In this way, the algorithm in fine ranking obtains improvement of 0.278% for novelty and 0.312% for serendipity respectively, capturing more complete user interests.

4.2.2 Feature-aware Debiasing.

[1]Defined as the user giving any of the following feedback to the video: long view, complete view ($w \geq d$), like, comment, like some comments, download, collect, share, or follow the author.

Table 1: A/B performance comparison of the diversified algorithms across different recommender stages. ↑ (↓) denotes that the higher (lower) value means better diversity, and the value is the improvement in percentage. Better results of one algorithm in different stages are marked in bold.

Algorithm	Stage	Interaction				Diversity		Platform	
		Exposure	Like	Comment	Long view	Concentration ↓	Coverage ↑	Sum. Time	Avg. Time
E&E	Matching	+0.110	-0.154	-0.001	+0.024	-1.091	**+1.272**	-0.011	-0.026
	Fine Ranking	**+0.186**	**-0.007**	**+0.173**	**+0.059**	**-1.610**	+0.765	**+0.003**	**-0.017**
Debiasing	Coarse Ranking	-0.227	-1.072	-1.299	+0.082	-0.861	+2.759	+0.117	+0.100
	Fine Ranking	**-0.110**	**+0.423**	**-0.429**	**+0.175**	**-3.338**	**+3.360**	**+0.167**	**+0.141**
Diversity Optimization	Fine Ranking	**+1.306**	**+0.732**	-0.819	**+0.917**	**-4.645**	-0.181	+0.077	+0.016
	Re-ranking	+0.069	+0.423	**-0.255**	-0.219	-2.309	**-0.059**	**+0.097**	**+0.077**

- **Debiasing algorithm in the stage of fine ranking mitigates the trade-off between accuracy and diversity more effectively.** The deployment in the fine ranking gains two positive profits (0.423% for like and 0.175% for long view) out of four interaction metrics, and also gains improvement larger than 3.3% for two diversity metrics, which are significant benefits in terms of online experiments. The largest up to 0.167% profits of platform metrics among all the experiments further show that it captures user interests more accurately. In addition, the performance exceeds that of coarse ranking for all the metrics, demonstrating its advantage again.
- **User consumption becomes richer in terms of categories by debiasing.** Since top videos in the ranking will be homogeneous due to the aforementioned bias, we further verify whether users consume richer video content. Specifically, we focus on videos of effective consumption, approximately representing top-ranking ones, and calculate the corresponding coverage of level-1 categories averaged over users. The relative improvements are +0.099% and +0.414% for the stage of coarse and fine ranking, respectively. This demonstrates the effectiveness of debiasing for users' richer consumption. In other words, top-ranking videos are more balanced, thus feature-aware bias is moderated to some extent.

4.2.3 Diversity Optimization.

- **Performance is comparable between the deployment in the stage of fine ranking and re-ranking.** Across all eight metrics, the performance of the two stages beat each other in half of them. This is reasonable since the two stages are close in the recommender procedure, and the video output from the fine ranking is the input of the re-ranking. Therefore, the overall degree of matching user interests is similar for videos to rank. Nevertheless, due to the longer length of the ranking list, the deployment in the fine ranking requires more time-consuming (model inference, service call, *etc.*) than that in the re-ranking (relative change of +0.155% *v.s.* -0.002% compared with the base group).

Table 2: A/B percentage improvement of SSD with respect to users' sessional diversity metrics. ↑ (↓) denotes that the higher (lower) value means better diversity.

Stage	Session Coverage ↑		Session Bad Case ↓	
	level-1	level-2	level-1	level-2
Fine Ranking	+0.956	+0.537	-3.809	-1.549
Re-ranking	+0.618	+0.412	-3.118	-0.490

- **Diversity optimization provides more diverse perceptions of real-time recommendation for users.** In order to investigate whether recommended video content is richer in a short-term period, we further define diversity metrics within one session[2]. The metrics include **Session Coverage** and **Session Bad Case** based on both level-1 and level-2 categories, where the latter denotes the exposure of extremely low diversity that may harm user experience, such as consecutive videos with the same category. The results are shown in Table 2, and all the metrics obtain positive profit across two stages, especially for bad cases. This demonstrates that users have more diverse and less homogeneous perceptions of each concentrated usage.

4.2.4 Comparison across Full-stage. We further compare performance across full-stage and conclude from the following two perspectives.

Deployment stage. The deployment in the fine ranking obtains the overall highest profit across all the diversified algorithms. In the upstream of this stage, there are too many noisy candidate videos out of user interest. In the downstream of this stage, the room for diversity improvement is limited since the videos left only cover users' major interests, which are usually homogeneous. Therefore, fine ranking is the most suitable stage for improving user engagement and recommendation diversity simultaneously.

Diversified algorithm. Debiasing has the most significant influence on recommendation diversity (up to 3.360%) as well as the dwell time in the platform (up to 1.167%). This simple yet effective algorithm addresses the necessity of modeling diversity, rather than ranking by only prediction scores. However, take the debiasing in the coarse ranking as an example, implicit modeling may worsen

[2]Two sessions have an interval larger than five minutes.

interaction metrics, *i.e.*, harm user engagement. In contrast, explicit diversity optimization can gain the highest up to 1.306% profit.

In summary, fusing explicit diversity modeling with the debiasing of users' propensity to different video content can be a promising direction for diversified recommendations.

5 RELATED WORK

5.1 Diversified Recommendation

As an additional objective beyond accuracy, diversity is incorporated in recommendations for extending the scope of user consumption. Generally speaking, diversity is taken into consideration by pre-defined strategies or explicit modeling. Specifically, strategies are usually designed to fetch items of more diverse categories. For example, Kwon *et al.* [25] selected top items in four types classified by users' purchasing intentions independently, which are further merged for heterogeneous recommendations. Zheng *et al.* [57] proposed to sample diverse items based on their categories in GNNs learning. The sampling includes disadvantaged neighbors in the process of embedding aggregation and negative items of the same category with positive ones. Some works follow this idea to develop GNN-based models for diversified recommendation [52, 53].

Many works model diversity based on item relationships in an explicit way. For the in-processing, *i.e.*, during the model designs and training, diversity serves as the regularization term in loss function or score for ranking, which is usually combined with relevance score [47]. Wasilewski and Hurley [45] proposed to train the recommendation model with both relevance loss and diversity regularization, where the latter is the negative of intra-list average distance (ILAD) [55] based on item similarity. Following this approach, Chen *et al.* [9] defined the diversity term as users' interests in different item categories. In terms of ranking scores, Li *et al.* [30] proposed to select items based on the summation of relevance and diversity scores. For the post-processing, the diversity of the recommendation list is maximized jointly with the relevance, which is defined based on fixed prediction scores. The two most representative methods are MMR [5] and DPP [7], which model the diversity with local pair-wise similarity and global correlation among items, respectively. In the scenario of multiple recommendation lists, *i.e.*, sliding windows, Huang *et al.* [22] proposed to define the diversity with time-series decomposition.

The methods above can be classified as the offline setting, where models are evaluated and trained based on static user-item interactions. For the online setting, RL models the diversity with the fusion in reward functions [14, 27] or E&E strategies [56]. Some works also present relatively simple online results of a single diversified algorithm in real-world systems [46].

Different from these works, we extensively investigate online performance and compare them between multiple diversified algorithms in a full-stage recommender system of real-world application.

5.2 Full-stage Recommender Systems

There are some works to jointly optimize for better recommendation across the matching and ranking stages. Since independent modeling in each stage may be suboptimal, Ma *et al.* [34] explicitly considered the ranking model while training the matching (*a.k.a.*, candidate generation) model based on off-policy learning. Exploration strategies are also investigated in such a two-stage system, where LinUCB [28] is deployed individually [20]. Many works focused on the advances of single stage, for example, the frameworks developed for the matching stage were dedicated to the efficiency of large-scale systems [3, 15, 60] or fairness [44, 60] for the latter recommendation. The ranking stage is the most studied, and the basic framework to predict the probability of user-item interaction, with the input of various user and item feature embeddings [13, 17, 18, 40]. Besides, in the deployment of the industrial system, very long item sequences are considered for modeling long-term user interests [4, 8, 37]. In terms of the re-ranking, the item list is re-ordered for better accuracy, diversity, fairness, *etc.* For accurate predictions of user feedback, many neural models based on recurrent networks [1, 61] or self-attention networks [21, 36] have been proposed in recent years. For the diverse recommendation list, both non-learning [5, 7] and learning [49, 50] based approaches are studied for the maximization of both relevance and diversity. Similarly, fairness is optimized for fair exposure of different items with specific features [23, 41].

In this work, we extensively compare the difference in the online effects of diversified algorithms when deployed across the full stages in the recommender system.

6 CONCLUSIONS

In this work, we deploy several simple yet effective algorithms to address critical challenges in diversified recommendations. We conduct large-scale online experiments in a short-video platform to investigate their full-stage effects on user engagement and recommendation diversity. From performance comparison in terms of important online metrics, we identify the fine ranking as the most suitable stage for real-world deployment. Besides, the combination of debiasing and diversity optimization can be a promising direction for future advances in diversified algorithms. Finally, we observe a co-growth trend of new users' long-term engagement and recommendation diversity, further demonstrating the effectiveness of our deployed algorithms. This work can provide beneficial experiences in diversified recommendations for researchers from both academia and industry. As for future works, we will deploy more advanced diversified algorithms to verify our findings. For more precise online evaluations, we will further leverage direct diversity-relevant questionnaires sent to users.

7 ETHICAL CONSIDERATIONS

The issue of diversity in recommender systems is closely related to social issues such as filter bubbles and information cocoons [29]. Enhancing user engagement and consumption diversity on the Web contributes to the development of more inclusive and active online platforms, and also helps to prevent users from inadvertently falling into communities filled with only homogeneous content or opinions.

8 ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under 62272262 and 72342032, and the National Key Research and Development Program of China under 2022YFB3104702.

REFERENCES

[1] Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 135–144.

[2] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. 2015. Optimal greedy diversity for recommendation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[3] Fedor Borisyuk, Krishnaram Kenthapadi, David Stein, and Bo Zhao. 2016. CaSMoS: A framework for learning candidate selection models over structured queries and documents. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 441–450.

[4] Yue Cao, XiaoJiang Zhou, Jiaqi Feng, Peihao Huang, Yao Xiao, Dayao Chen, and Sheng Chen. 2022. Sampling Is All You Need on Modeling Long-Term User Behaviors for CTR Prediction. *arXiv preprint arXiv:2205.10249* (2022).

[5] Jaime 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*. 335–336.

[6] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems* 24 (2011).

[7] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2017. Improving the diversity of top-N recommendation via determinantal point process. In *Large Scale Recommendation Systems Workshop*.

[8] Qiwei Chen, Changhua Pei, Shanshan Lv, Chao Li, Junfeng Ge, and Wenwu Ou. 2021. End-to-End User Behavior Retrieval in Click-Through RatePrediction Model. *arXiv preprint arXiv:2108.04468* (2021).

[9] Wanyu Chen, Pengjie Ren, Fei Cai, Fei Sun, and Maarten de Rijke. 2020. Improving end-to-end sequential recommendations with intent-aware diversification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 175–184.

[10] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to recommend accurate and diverse items. In *Proceedings of the 26th international conference on World Wide Web*. 183–192.

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[12] Enyan Dai and Suhang Wang. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 680–688.

[13] Jingtao Ding, Yuhan Quan, Xiangnan He, Yong Li, and Depeng Jin. 2019. Reinforced Negative Sampling for Recommendation with Exposure Data.. In *IJCAI*. Macao, 2230–2236.

[14] Qinxu Ding, Yong Liu, Chunyan Miao, Fei Cheng, and Haihong Tang. 2021. A hybrid bandit framework for diversified recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4036–4044.

[15] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. 2018. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference*. 1775–1784.

[16] Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. 2022. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1623–1625.

[17] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[18] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[19] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003), 1161–1189.

[20] Jiri Hron, Karl Krauth, Michael I Jordan, and Niki Kilbertus. 2020. Exploration in two-stage recommender systems. *arXiv preprint arXiv:2009.08956* (2020).

[21] Jinhong Huang, Yang Li, Shan Sun, Bufeng Zhang, and Jin Huang. 2020. Personalized flight itinerary ranking at fliggy. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2541–2548.

[22] Yanhua Huang, Weikun Wang, Lei Zhang, and Ruiwen Xu. 2021. Sliding Spectrum Decomposition for Diversified Recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3041–3049.

[23] Chen Karako and Putra Manggala. 2018. Using image fairness representations in diversity-based re-ranking for recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 23–28.

[24] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A survey of serendipity in recommender systems. *Knowledge-Based Systems* 111 (2016), 180–192.

[25] Hyokmin Kwon, Jaeho Han, and Kyungsik Han. 2020. ART (Attractive Recommendation Tailor) How the Diversity of Product Recommendations Affects Customer Purchase Preference in Fashion Industry?. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2573–2580.

[26] Pei Lee, Laks VS Lakshmanan, Mitul Tiwari, and Sam Shah. 2014. Modeling impression discounting in large-scale recommender systems. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1837–1846.

[27] Chang Li, Haoyun Feng, and Maarten de Rijke. 2020. Cascading hybrid bandits: Online learning to rank for relevance and diversity. In *Fourteenth ACM Conference on Recommender Systems*. 33–42.

[28] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.

[29] Nian Li, Chen Gao, Jinghua Piao, Xin Huang, Aizhen Yue, Liang Zhou, Qingmin Liao, and Yong Li. 2022. An Exploratory Study of Information Cocoon on Short-form Video Platform. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4178–4182.

[30] Shuang Li, Yuezhi Zhou, Di Zhang, Yaoxue Zhang, and Xiang Lan. 2017. Learning to diversify recommendations based on matrix factorization. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE, 68–74.

[31] Zihan Lin, Hui Wang, Jingshu Mao, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and Ji-Rong Wen. 2022. Feature-aware Diversified Re-ranking with Disentangled Representations for Relevant Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3327–3335.

[32] Weiwen Liu, Yunjia Xi, Jiarui Qin, Fei Sun, Bo Chen, Weinan Zhang, Rui Zhang, and Ruiming Tang. 2022. Neural Re-ranking in Multi-stage Recommender Systems: A Review. *arXiv preprint arXiv:2202.06602* (2022).

[33] Yunfei Lu, Peng Cui, Linyun Yu, Lei Li, and Wenwu Zhu. 2022. Uncovering the Heterogeneous Effects of Preference Diversity on User Activeness: A Dynamic Mixture Model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3458–3467.

[34] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H Chi. 2020. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020*. 463–473.

[35] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* penguin UK.

[36] Changhua Pei, Yi Zhang, Yongfeng Zhang, Fei Sun, Xiao Lin, Hanxiao Sun, Jian Wu, Peng Jiang, Junfeng Ge, Wenwu Ou, et al. 2019. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM conference on recommender systems*. 3–11.

[37] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2685–2692.

[38] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. 2014. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, 461–469.

[39] Lijing Qin and Xiaoyan Zhu. 2013. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.

[40] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[41] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *Advances in Neural Information Processing Systems* 32 (2019).

[42] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2022. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 957–965.

[43] Cass R Sunstein. 2009. *Going to extremes: How like minds unite and divide.* Oxford University Press.

[44] Lequn Wang and Thorsten Joachims. 2022. Fairness in the First Stage of Two-Stage Recommender Systems. *arXiv preprint arXiv:2205.15436* (2022).

[45] Jacek Wasilewski and Neil Hurley. 2016. Incorporating diversity in a learning to rank recommender system. In *The twenty-ninth international flairs conference*.

[46] 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*. 2165–2173.

[47] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Fernando Diaz, and Xue Liu. 2022. A Survey of Diversification Techniques in Search and Recommendation. *arXiv e-prints* (2022), arXiv–2212.

[48] Qiong Wu, Yong Liu, Chunyan Miao, Yin Zhao, Lu Guan, and Haihong Tang. 2019. Recent advances in diversified recommendation. *arXiv preprint arXiv:1905.06589*

(2019).

[49] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling document novelty with neural tensor network for search result diversification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 395–404.

[50] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, Wei Zeng, and Xueqi Cheng. 2017. Adapting Markov decision process for search result diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 535–544.

[51] Ruobing Xie, Cheng Ling, Shaoliang Zhang, Feng Xia, and Leyu Lin. 2022. Multi-granularity Fatigue in Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4595–4599.

[52] Liangwei Yang, Shengjie Wang, Yunzhe Tao, Jiankai Sun, Xiaolong Liu, Philip S Yu, and Taiqing Wang. 2022. DGRec: Graph Neural Network for Recommendation with Diversified Embedding Generation. *arXiv preprint arXiv:2211.10486* (2022).

[53] 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 *Fifteenth ACM Conference on Recommender Systems*. 651–655.

[54] Cong Yu, Laks Lakshmanan, and Sihem Amer-Yahia. 2009. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th international conference on extending database technology: Advances in database technology*. 368–378.

[55] Mi Zhang and Neil Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM conference on Recommender systems*. 123–130.

[56] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 world wide web conference*. 167–176.

[57] 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 2021*. 401–412.

[58] Yu Zheng, Chen Gao, Jingtao Ding, Lingling Yi, Depeng Jin, Yong Li, and Meng Wang. 2022. DVR: Micro-Video Recommendation Optimizing Watch-Time-Gain under Duration Bias. In *Proceedings of the 30th ACM International Conference on Multimedia*. 334–345.

[59] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*. 2980–2991.

[60] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3985–3995.

[61] Tao Zhuang, Wenwu Ou, and Zhirong Wang. 2018. Globally optimized mutual influence aware ranking in e-commerce search. *arXiv preprint arXiv:1805.08524* (2018).

[62] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. 22–32.