Surrogate for Long-Term User Experience in Recommender Systems

Yuyan Wang
Google Research, Brain Team
yuyanw@google.com

Mohit Sharma
Google
mohitsharma@google.com

Can Xu
Google
canxu@google.com

Sriraj Badam
Google
srirajdutt@google.com

Qian Sun
Google
qians@google.com

Lee Richardson
Google
leerich@google.com

Lisa Chung
Google
lchung@google.com

Ed H. Chi
Google Research, Brain Team
edchi@google.com

Minmin Chen
Google Research, Brain Team
minminc@google.com

ABSTRACT

Over the years we have seen recommender systems shifting focus from optimizing short-term engagement toward improving long-term user experience on the platforms. While defining good long-term user experience is still an active research area [30, 32], we focus on one specific aspect of improved long-term user experience here, which is user revisiting the platform. These long term outcomes however are much harder to optimize due to the sparsity in observing these events and low signal-to-noise ratio (weak connection) between these long-term outcomes and a single recommendation. To address these challenges, we propose to establish the association between these long-term outcomes and a set of more immediate term user behavior signals that can serve as surrogates for optimization.

To this end, we conduct a large-scale study of user behavior logs on one of the largest industrial recommendation platforms serving billions of users. We study a broad set of sequential user behavior patterns and standardize a procedure to pinpoint the subset that has strong predictive power of the change in users' long-term visiting frequency. Specifically, they are predictive of users' increased visiting to the platform in 5 months among the group of users with the same visiting frequency to begin with. We validate the identified subset of user behaviors by incorporating them as reward surrogates for long-term user experience in a reinforcement learning (RL) based recommender. Results from multiple live experiments on the industrial recommendation platform demonstrate the effectiveness of the proposed set of surrogates in improving long-term user experience.

CCS CONCEPTS

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

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539073

KEYWORDS

Recommender Systems, Long-Term User Experience, Sequential User Behavior, Reward Surrogate

ACM Reference Format:
Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H. Chi, and Minmin Chen. 2022. Surrogate for Long-Term User Experience in Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539073

1 INTRODUCTION

Recommender systems are becoming an integral part of our daily life by facilitating information acquisition and decision-making in retail [10, 46], media [6, 15], travel [16, 18], food [54], news [33, 69] and social platforms [55, 60]. Recommendation algorithms that focus on users' immediate responses such as clicks and likes have gained immense success over the years [15, 65, 68]. However, it has become increasingly clear that over-indexing on short-term engagement can lead to undesirable recommendations, such as clickbait contents or pigeon-holing effects which hurt long-term user experience [9, 38, 67]. Recognizing the drawback of over-emphasizing the short-term metrics, algorithm designers resort to optimizing other objectives that are more aligned with the long-term user experience on recommendation platforms. As an example, Wu et al. [58] argued the long-term goal of a recommender system is to not only satisfy the user's needs in the current session, but also to see them come back to the platform more often in the future.

Optimizing long-term user experience is however challenging as the desired long-term outcome is much sparser, noisier and naturally manifests over a much longer horizon than the short-term engagement signals. A natural question one would ask is: Are there any alternative objectives that are predictive of the long-term outcome, while being easier to optimize? An objective that is easier to optimize should have stronger connection to the recommendation. For example, the effect of a recommendation on a user's behavior in the current session (e.g. clicks, likes) is much easier to establish than in the future sessions (e.g. returning to the platform in one month). Meanwhile, we also need to ensure that optimizing the alternatives will lead to improved long-term user experience.

The effect of a recommendation on a user's long-term experience is manifested through the aggregation of their sequential medium-term behaviors. Motivated by this, we start with identifying medium-term user behaviors as alternatives and study their associations with the long-term user experience, in particular changes in visiting frequencies. On an industrial recommendation platform serving billions of users, we analyze various evolving behavior patterns of the users who visited the site with increasing frequency over a 5-month time period. Specifically, we identify a subset of sequential consumption patterns, diversity of the consumed contents being one of them, that are associated with increased visiting frequencies in the long term. These behavior patterns have impressive predictive power in differentiating users' future long-term experience. In particular, they are predictive of which users will revisit the platform more often in 5 months among a group of users with the same visiting frequency to begin with.

To validate the effectiveness of leveraging the identified sequential user behavior patterns in optimizing long-term user experience, we experiment with incorporating them as reward surrogates in an RL-based recommender. The idea of using surrogate outcomes as a proxy for long-term outcomes was first proposed in medicine and biostatistics [24, 42, 56]. Recent literature in causal inference showed that even when the long-term outcome is available, using surrogate outcomes improves the efficiency in estimating treatment effects [4, 62]. In our case, the proposed reward surrogate enables the system to nudge the users toward the desired user behavior patterns for an improved long-term experience. Live experiments show the effectiveness of these behavior patterns as surrogates.

To summarize, the contributions of our work include:

- **Measurements:** We propose a set of metrics to capture users' sequential and temporal consumption patterns.
- **Analytical insights:** We identify a set of user behavior patterns that are associated with users' long-term experience.
- **Surrogate selection:** We standardize the procedure of selecting user behavior patterns as surrogates for long-term user experience, based on robust predictive modeling.
- **Algorithmic Improvements:** We validate the efficacy of the surrogates in optimizing long-term user experience in an RL recommender.

2 RELATED WORK

Understanding user behavior on recommender systems. There is a large body of work on understanding user behavior in recommender systems from various fields, ranging from human-computer interaction [28, 35, 45, 59], marketing [3, 48, 64] and information retrieval [2, 31, 40, 52, 61]. For example, Knijnenburg et al. [28], Xiao and Benbasat [59] provided insights into the mechanisms underlying the user experience in recommender systems. Structural and probabilistic models were proposed to learn the evolving user preferences [3, 40, 67]. Content-wise, Goel et al. [17] studied the distribution of user interests in long-tail and niche contents.

User behavior on recommender systems is a combined effect of user preference, algorithmic recommendation and other confounding factors such as personality characteristics. Anderson et al. [2] studied the algorithmic effects of recommendations on the content diversity of users' consumption. Villermet et al. [52] proposed to disentangle human and algorithmic behavior in online music consumption to better determine the effect of recommendation. Hansen et al. [19], Zhou et al. [70] studied consumption patterns on video and music streaming platforms as part of the feedback loop of the recommendation algorithms and user behaviors. Karumur et al. [26], Xiao and Benbasat [59] showed the effect of personal and situational characteristics on user behaviors on recommender systems. Another line of work is around building simulation or conducting field experiments to understand user behaviors while controlling for potential confounding [9, 21, 64].

While there has been extensive work on analyzing user behavior on recommender platforms, limited work has been done on analyzing the sequential and evolving aspect of user behaviors toward understanding their long-term experience on the platform, which is the main contribution of our work.

Optimizing long-term user experience in recommender systems. Improving long-term user experience usually entails optimizing a noisy, sparse and delayed signal. There are efforts that aim at optimizing this signal directly by imposing additional assumptions on the environment and the users. For example, Wu et al. [58] assumed a stationary distribution of user preference and recommendation candidates and optimizes for user return directly; Zhang et al. [66] proposed to leverage a counterfactual estimate of the delayed reward to avoid waiting on long-term labels. Both works adopted the bandit framework and assumed that the recommender does not change users states or alter their interests, which however are often violated in real world recommendation products.

RL for recommender systems has achieved notable success in the past few years [5, 11, 14, 22, 34, 44, 69], one reason being that they naturally account for the shifts in user states due to the recommendations. There has been work on improving longer term user experience by extending the planning horizon of RL-based recommenders. For example, Zheng et al. [69] introduced a user activeness score which is used as a reward for future returns; Ji et al. [23] proposed to predict the future long-term value of the fresh items; Zou et al. [73] developed a hierarchical LSTM to model complex user behaviors with the reward containing delayed metrics such as return time. These works either propose to advance user behavior modeling to better predict the long-term behavior yet still suffer from the delay in collecting the reward [69], or rely on existing input features to impute the long-term user behavior [23, 73]. Our work differs in that we propose to establish the connection between long-term user experience and a set of sequential user behaviors that may not be captured by the existing system. We then leverage the insights derived from these analyses by incorporating those behaviors as reward surrogates in RL-based recommenders.

Another line of research on improving long-term user experience is around avoiding myopic recommendations and attending to users' long-term interests [8, 38, 39]. This includes a better exploration-exploitation trade-off [13, 51], diversity-focused recommendations [1, 2, 71, 72] and distribution-aware recommendations [27, 49, 67]. Sequential recommendation algorithms [20, 25, 50] were also proposed to adapt to users' changing interests.

Surrogate outcomes. A critical challenge in estimating long-term treatment effects in clinical trials is that long-term outcomes are

often observed with a delay of months or even years [4]. The idea of surrogate outcomes was first proposed to address this challenge [42]. It was also shown that even when the long-term outcome is observed, using surrogate outcomes can improve the efficiency in estimating treatment effects [4] and policy evaluation [62].

The notion of surrogate has also been studied in RL literature when the true reward is noisy or costly to be observed [12, 53, 63]. A common approach is to hand-engineer a dense reward by leveraging domain knowledge which serves as a reward surrogate to guide the policy learning. Such approaches have theoretical guarantees under some assumptions [37] and have been successful in multiple RL application scenarios [29, 43]. However, we have not seen much work on leveraging domain knowledge to design reward surrogates for recommender systems. Our work aims at bridging this gap and identifying interpretable user behavior patterns as surrogates to improve the long-term user experience in these systems.

3 MEASUREMENTS

As we will see shortly, improved long-term user experience such as increased visiting frequency to the platform, is a sparse and noisy signal that manifests over a time horizon of weeks or even months, making it difficult to optimize directly. This motivates us to search for medium-term user behavior signals as surrogate objectives. To be used as a surrogate, the user behavior signals should 1) exhibit stronger and cleaner association to the recommendation than the long-term outcomes for easier optimization, and 2) connect to the long-term outcomes such that optimizing the surrogate leads to improved long-term user experience.

In this section, we define a set of candidate metrics for measuring user behavior patterns on recommendation platforms over medium-term horizons. In the next section, we present the analysis on the evolution of these user behaviors over time and their associations with long-term user experience.

We first introduce the notion of a topic cluster, which will be used in defining diversity-related user behaviors below. Similar to [13], we define the topic clusters for each item by: 1) taking the item co-occurrence matrix, where entry (i, j) counts the number of times item i and j were consumed by the same user consecutively; 2) performing matrix factorization to generate one embedding for each item; 3) using k-means to cluster the learned embeddings into 10K clusters; 4) assigning the *top 3* nearest clusters to each item.

We use S to denote a user's consumption history over a certain time period, where S may contain repeated items if the user consumes a content more than once.

3.1 Diversity

Diversity [36, 41, 47] measures the broadness of the set of contents that a user engages with. We propose the following three measurements of diversity.

3.1.1 Ratio-based Diversity. We define the *ratio-based diversity* as the proportion of unique topic clusters in S:

$$D_{\text{ratio}}(S) = \frac{|\text{unique topic clusters in } S|}{|S|}, \tag{1}$$

where $|\cdot|$ returns the size of the collection, i.e. $|S|$ is the number of items (including repeated ones) the user consumed.

3.1.2 Distribution-based Diversity. A user's consumption history can be viewed as a distribution over topics. Let N_i be the number of consumed items from topic i, the *entropy-based diversity* is:

$$D_{\text{entropy}}(S) = -\sum_i \hat{p}_i log(\hat{p}_i), \tag{2}$$

where $\hat{p}_i = N_i / \sum_i N_i$ is the proportion of items from topic i in S.

One caveat of using entropy as the diversity measure is that it naturally grows with the cardinality of S. For example, a uniform distribution on two topic clusters has entropy $-log(1/2) \approx 0.69$, while a uniform distribution on five has entropy $-log(1/5) \approx 1.61$. In other words, the same shape of topic distribution (e.g. uniform) on different supports of S will lead to different entropy measures. To better capture the shape of the distribution regardless of the support, we propose to use the negative of the Kullback–Leibler (KL) divergence between the topic distribution P and the uniform distribution P_u on the *same* support as the diversity measure:

$$\begin{aligned} D_{\text{KL}}(S) &= -KL(P||P_u) = -\sum_i \hat{p}_i log(\hat{p}_i/(1/C)) \\ &= D_{\text{entropy}}(S) - log(C), \end{aligned} \tag{3}$$

where $C := |i : \hat{p}_i > 0|$ is the number of clusters the user has consumed in S. We name this measure $D_{\text{KL}}(S)$ as *KL-divergence diversity*. $D_{\text{KL}}(S)$ measures the concentration of the user's interests, i.e. how different is their interest distribution from a uniform distribution. With this normalization, a uniform distribution on a set S of two topic clusters vs. five will both have $D_{\text{KL}}(S) = 0$.

3.2 Repeated Consumption

When a user is particularly interested in a certain item on the platform, she may interact with it repeatedly. We measure the user's repeated consumption behavior from a user consumption history S as the ratio of *repeated consumption* from S:

$$R_{\text{repeatedCons}}(S) = \frac{|\{s \in S : \text{s has been consumed before}\}|}{|S|}. \tag{4}$$

As an example, for a user with $S = \{s_1, s_1, s_2, s_1, s_3, s_4, s_3\}$, her re-consumption ratio is $R_{reconsumption}(S) = 3/7$ as there are two repeated consumption of s_1 and one of s_3.

3.3 High-Quality Consumption

The consumption time measures the engagement level of an item that the click-based signals do not capture. As the contents on the platform can be of various lengths, we measure the quality of a consumption based on the completion ratio, i.e. the ratio between the consumption time and the total length of the item. A *high-quality consumption* is defined as having greater than $X\%$ completion ratio or greater than Y consumption time[1]. *High-quality consumption ratio* is the proportion of consumption that are high-quality:

$$R_{\text{highQualCons}}(S) = \frac{|\{s \in S : \text{s is a high-quality consumption}\}|}{|S|}. \tag{5}$$

[1]We omitted the value for X, Y here and Z below for business-compliance reasons.

3.4 Persistent Topics

The fact that a user consumes an item from a certain topic cluster does not necessarily mean that she is interested in that topic. It could be a transient or transactional interest (e.g. looking up some content for a family member or performing a one-time task) instead. This motivates us to define a metric to capture the user's true interest by only looking at topic clusters that she has repeatedly consumed. We define *persistent topic ratio* as the portion of topic clusters with more than Z number of consumption from the user's history S:

$$R_{\text{persistentTopic}}(S) = \frac{|\text{unique topic clusters with more than Z consumption in } S|}{|\text{unique topic clusters in } S|}. \quad (6)$$

3.5 Page-Specific Revisits

A recommendation platform usually provides multiple surfaces/pages for users to interact with. Given the same amount of time spent on the platform, users may vary in their interaction patterns with different pages on the platform. To study users' revisiting behavior, we look at the time it takes for a user to come back and revisit each surface/page on the platform. The surfaces/pages we look at are:
(1) *Home page*: the first destination of the recommendation website;
(2) *Search page*: the page for users to search for a particular content;
(3) *Consumption page*: the page for users to consume an item.

We define the *page-specific revisit time* as:

$$T_{\text{revisit}}(S, \text{Page}) = Avg(\text{Time between two consecutive } \textit{high-quality consumptions} \text{ from Page}), \quad (7)$$

where Page $\in$ {Home, Search, Consump}. We focus on visits with at least one high-quality consumption during that visit.

4 ANALYSIS

In this section, we present the analysis of the sequential behavior patterns in Section 3 and their relationship with long-term user experience, on a real world content recommendation platform.

4.1 Data

We study one of the largest industrial recommendation platforms serving billions of users, and analyze the user visiting logs over a 20-week period [2]. Users on recommendation platforms are heterogeneous in their visiting frequency. Some visit the platform occasionally, while others regularly. We call a user *low-frequency* if they visit less than A days[3] over a 14-day window, and consistently behave like that over at least two 14-day windows. We call a user *high-frequency* if they visit more than B days over a 14-day window, and consistently behave like that over at least two 14-day windows.

We divide the 20-week period into 10 time buckets of 2 weeks (14 days). For every user, we compute their behavior pattern statistics defined in Section 3 over each bucket, and analyze their temporal patterns across the whole period. We study the users who started as low-frequency users in the beginning of the analysis period. Some of them increased their returning frequency and became high-frequency users at the end of the analysis period, while others remained as low-frequency users. As their experience on the platform improves, we imagine users will return to the platform more often. In other words, the improved long-term user experience is manifested as an increase in users' long-term visiting frequency.

[2]We performed the same analysis over two different 5-month periods and results were consistent.
[3]We omitted the value for A and B here for business-compliance reasons.

4.2 Analysis Results

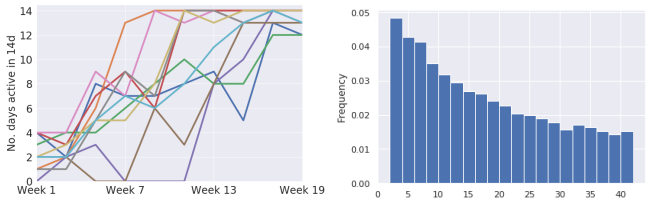
Here we present the results on the user behavior patterns in Section 3, and their association with improved long-term user experience manifested through an increase in long-term visiting frequency.

4.2.1 Statistics on Improved Long-Term User Experience. We first present some descriptive statistics on the pattern of increased visiting frequency as an improved long-term user experience.

Sparsity. Among the 2 million users who started as low-frequency users, only about 2.3% of them became high-frequency users at the end of the 5 months, making the changes in visiting frequency an extremely sparse signal as a measure for long-term user experience.

Heterogeneity. Among the low-frequency users who transitioned to high-frequency users in the same time period, their paths to becoming high-frequency users are very different. Fig. 1a illustrates the transition patterns of 10 randomly sampled low-frequency users who became high-frequency in 5 months. We see that some low-frequency users transitioned to high-frequency directly in less than a month, while others gradually increased their visiting frequency over a much longer horizon, with ups and downs in between and eventually became high-frequency. The heterogeneity of these revisiting patterns results in a very low signal-to-noise ratio for long-term user experience if attempting to model and optimize directly.

Long time horizon. We also summarize the time it takes for an average low-frequency user to become high-frequency over time. We extend the time period of the analysis to 10 months (40 weeks) and Figure 1b shows the distribution of the time it takes for a low-frequency user to become high-frequency. The average time horizon is 15.32 weeks and the median is 14 weeks, suggesting that an improved user experience is a long and gradual process.



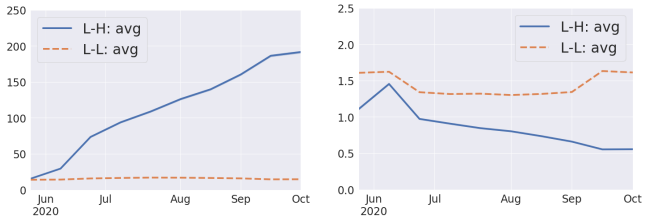
(a) A sample of 10 low-frequency users who became high-frequency in 5 months. **(b)** Distribution of time (in weeks) it takes for for increased visiting frequencies.

Figure 1: Descriptive plots for visiting frequencies over time.

In summary, long-term user experience is a sparse, noisy and delayed signal that is hard to optimize directly. Next, we look at medium-term user behavior patterns proposed in Section 3 as surrogates. To understand their relationship with the long-term user experience, we compare between the low-frequency users who became high-frequency at the end of the 5-month period (denoted as **'L-H'**) against those who remained low-frequency (**'L-L'**).

4.2.2 Sequential Consumption Diversity Patterns. We first study the association between increased visiting frequency and sequential diversity patterns: When the users visit the platform more frequently in the long term, do they gradually consume a larger set of topics, or do they develop a more concentrated interest over time?

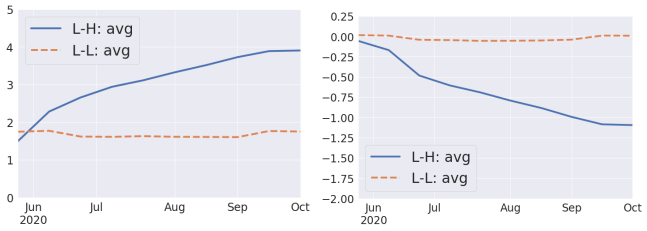
Ratio-Based Diversity Patterns. Figure 2a shows that, not surprisingly, when low-frequency users transition to high-frequency users, they gradually consume more topics. However, when we look at ratio-based diversity $D_{\text{ratio}}(S)$ as defined in Eq. (1), we see a reverse trend in that the users had a lower ratio-based diversity as they increase their visiting frequency (Figure 2b)[4].



(a) Number of topic clusters. **(b)** Ratio-based diversity $D_{\text{ratio}}(S)$.

Figure 2: Patterns for number of topics and ratio-based diversity.

Distribution-Based Diversity Patterns. Another set of diversity metrics proposed in Eq. (2) and (3) measures the user behavior patterns as a distribution over the topic clusters. Figure 3a shows that as the low-frequency users increased their visiting frequency, the entropy of the topics they consume is also *increasing*. However, their KL-divergence diversity $D_{\text{KL}}(S)$ is *decreasing* over time, suggesting that their interest distribution is moving further away from a uniform distribution. These observations corroborate our findings above, in that when users visits more often, they 1) consume *a more diverse set* of topics as measured by topic counts and entropy-based diversity metrics, and 2) also form *a more concentrated interest* around a subset of topics as indicated by the decrease in the ratio-based and KL-divergence based diversity metrics.



(a) Entropy diversity $D_{\text{entropy}}(S)$. **(b)** KL-divergence diversity $D_{\text{KL}}(S)$.

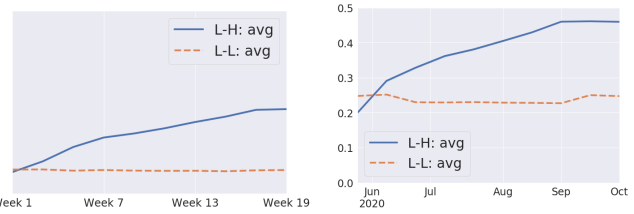
Figure 3: Patterns for distribution-based diversity.

The first part of our findings is aligned with [2] which stated that generalist users (i.e. users with a broader interest measured by a lower similarity score from their past history) are much more likely to remain on a music streaming platform than specialist users. The second part of our finding, which to our knowledge has not been discussed before, further characterizes that users develop more concentrated and consistent interests on a subset of topics as they stay and revisit the platform more often.

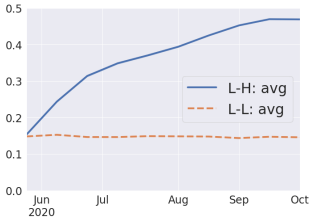
[4]Note that all the figures in this paper include the standard errors of the metrics, but some are hardly visible as the standard error of the average statistics are very small with very large sample sizes ($\sim O(1/\sqrt{n})$ where n is the sample size, according to central limit theorem).

4.2.3 Sequential Consumption Quality Patterns. We examine repeated consumption, high-quality consumption, and persistent topics in Section 3.2-3.4 with increased visiting frequency over time.

Repeated Consumption and High-Quality Consumption. We see that as the users visit the platform more often, a larger portion of their consumption history were on items that they have consumed before (Figure 4a); and a larger proportion were high-quality consumption (Figure 4b). In addition, users more than doubled their high-quality consumption ratio (from below 20.0% to 46.0%), when transitioning to high-frequency users (Figure 5).



(a) Repeated consumption ratio $R_{\text{repeatedCons}}(S)$. **(b)** High-quality consumption ratio $R_{\text{highQualCons}}(S)$.



(c) Persistent topic ratio $R_{\text{persistentTopic}}(S)$.

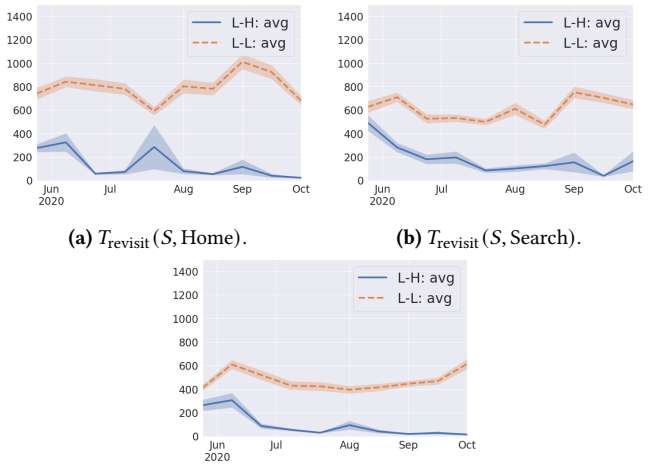
Figure 4: Patterns for repeated consumption, high-quality consumption and persistent topics.

The significant increase in repeated item consumption is surprising. It however validates our observation in Section 4.2.2 that the users develop more concentrated interests as they revisit the platform more often.

Persistent Topics. Figure 4c shows that an increased proportion of persistent topics is associated with improved long-term user experience. The persistent topic ratio more than tripled (from 15.3% to 46.9% in Fig. 4c) when an average low-frequency user transitioned to a high-frequency user over time, indicating that more contents they consumed are indicative of their true interests as opposed to transient and transactional interests.

4.2.4 Sequential Page Revisit Patterns. We examine if an improved long-term user experience entails more frequent visits to a specific page/surface. To control for confounding caused by the amount of consumption, we look at a subset of low-frequency users who consumed the same number of items in the beginning. Figure 5 shows

the average time it takes for a user to come back and have a *high-quality consumption* on Home Page, Search Page, and Consumption Page respectively. We see that as the low-frequency users became high-frequency, they visited all the pages more frequently (decreasing trend of the blue curves in Fig. 5). Among them, the revisit time to the homepage shows the biggest difference (especially in the beginning) between the low-frequency users who became high-frequency in the end ('L-H') and those who did not ('L-L')(Figure 5a). This suggests an association that low-frequency users who visit the Home Page more often are more likely to become high-frequency users in the long term.



(a) $T_{\text{revisit}}(S, \text{Home})$. **(b)** $T_{\text{revisit}}(S, \text{Search})$.

(c) $T_{\text{revisit}}(S, \text{Consumption})$.

Figure 5: Average time (in hours) it takes between two consecutive page visits with high-quality consumption, on low-frequency users who consumed the same number of items in the first 4 weeks.

In summary, as low-frequency users gradually transition to high-frequency, they consume contents from a broader set of topics (Figure 2a and 3a); meanwhile, they also develop more concentrated interests with a sharper topic distribution (Figure 3b) and engage with some topics more often than the rest (Figure 4c). In addition, they generate more repeated consumptions and high-quality consumptions (Figure 4a and 4b), and visit specific pages of the recommendation platform more often than others (Figure 5).

5 SURROGATE SELECTION

Now we describe the procedure to identify user behavior patterns that strongly associate with the long-term user experience, specifically the transition from low-frequency to high-frequency users.

We use the same data and aggregation method as described in 4.1, and use the user behavior patterns defined in Section 3 as features toward predicting long-term user visiting frequency changes. We segment the user behavior log of 20 weeks into 10 buckets of 14 days each, and extract the users who were low-frequency in the *2nd* bucket. The features we include in the predictive modeling are: (1) their behavior patterns in the *2nd* bucket (when all users were low-frequency), and (2) the *difference* in their behavior patterns between the two time buckets. This captures the sequential/evolving aspect of the behavior patterns, denoted with superscript 'diff' in the results below, while controlling for the possible confounding introduced by visiting frequency. The response is a binary indicator on whether a low-frequency user became high-frequency at the end of the 5-month study period.

We choose *random forests* [7] as the model class, as it offers us 1) *flexibility* in handling extremely imbalanced data; 2) *interpretability* so one can easily analyze feature importance[5] to pinpoint the subset of user behavior patterns with the strongest associations; and 3) *robustness* to different user distributions to reduce potential spurious correlation between the user behavior patterns and the long-term outcome. For 3), if a user behavior pattern has high feature importance across different user distributions (e.g. created by bootstrapping, i.e., sampling users with replacement), then it is more likely to be causal to the response, therefore more likely to be an effective surrogate for the long-term objective, i.e. visiting frequency increase over time.

A single run of the random forest model with 200 trees and max tree depth 5 shows that the proposed set of user behavior patterns reaches test AUC 0.691 (training AUC 0.697). The number is quite impressive considering that we are only using the behavior patterns from the beginning (1st and 2nd 2-week buckets) to predict what will happen 5 months later, among a group of similar users to begin with. To refine the predictive modeling and control for possible confounding introduced by the amount of activities, we propose *stratified modeling* by slicing the users into segments according to a fine-grained activity measure, which is the number of consumed items in the first bucket.

Table 1 summarizes the feature importance results for each slice and overall, where the cutoff-points are picked so that each segment has roughly the same number of users. Full results on the feature importance scores for all features can be found in Appendix. We see that entropy-based diversity $D_{\text{entropy}}(S)$ and average time between homepage visits $T_{\text{revisit}}(S, \text{Home})$ consistently appear as the top features in predicting visiting frequency increases across different user segments. We therefore choose them as the *surrogates for long-term user experience.*

6 EXPERIMENTS

We conduct a series of live A/B experiments on the same industrial recommendation platform to verify that optimizing these surrogates indeed leads to improved long term user experience. Experiments are run on a REINFORCE [57] recommender [11] in an RL setting. We would like to point out that our proposed approach also applies to supervised learning based recommenders, as it serves as a general methodology to replace the optimization objective with surrogate objectives, regardless of the optimization framework.

6.1 Background: a REINFORCE Recommender

Chen et al. [11] formulates the recommendation problem as a Markov Decision Process (MDP) over $(\mathcal{S}, \mathcal{A}, \mathbf{P}, R, \gamma)$, where $\mathcal{S}$ is the state space representing the users' interest and context, $\mathcal{A}$ is the discrete action space and $a \in \mathcal{A}$ a recommended item. $\mathbf{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow$

[5]Here the importance is measured by the average decrease in Gini impurity, i.e. $\sum_c P(c) \cdot (1 - P(c))$ where $P(c)$ is the probability of class c in the training data.

# Consump	AUC	Top 5 important features
$(0, c_1]$	0.57	$T_{\text{revisit}}(\text{Home})$, D_{entropy}, $T_{\text{revisit}}(\text{Consump})$, D_{KL}, $T^{\text{diff}}_{\text{revisit}}(\text{Home})$
$(c_1, c_2]$	0.62	$T_{\text{revisit}}(\text{Home})$, D_{entropy}, $D^{\text{diff}}_{\text{entropy}}$ $R^{\text{diff}}_{\text{repeatedCons}}$, $T_{\text{revisit}}(\text{Search})$
$(c_2, c_3]$	0.63	$T_{\text{revisit}}(\text{Home})$, D_{entropy}, $T^{\text{diff}}_{\text{revisit}}(\text{Search})$, $D^{\text{diff}}_{\text{KL}}$, $D^{\text{diff}}_{\text{entropy}}$
$(c_3, c_4]$	0.62	$T_{\text{revisit}}(\text{Home})$, $T^{\text{diff}}_{\text{revisit}}(\text{Home})$, $D^{\text{diff}}_{\text{entropy}}$, D_{KL}, $D^{\text{diff}}_{\text{KL}}$
$(c_4, c_5]$	0.58	$T_{\text{revisit}}(\text{Home})$, $R^{\text{diff}}_{\text{repeatedCons}}$, $T^{\text{diff}}_{\text{revisit}}(\text{Home})$, $T^{\text{diff}}_{\text{revisit}}(\text{Consump})$, $D^{\text{diff}}_{\text{entropy}}$
(c_5, ∞)	0.61	$T_{\text{revisit}}(\text{Home})$, $T^{\text{diff}}_{\text{revisit}}(\text{Home})$, $T^{\text{diff}}_{\text{revisit}}(\text{Consump})$, $D^{\text{diff}}_{\text{entropy}}$, $R^{\text{diff}}_{\text{persistentTopic}}$
Overall	0.61	$T_{\text{revisit}}(\text{Home})$, $T^{\text{diff}}_{\text{revisit}}(\text{Home})$, D_{entropy}, $D^{\text{diff}}_{\text{KL}}$, $D^{\text{diff}}_{\text{entropy}}$

Table 1: Feature importance results sliced by number of consumption in the first 14 days (1st column). We omitted the actual values for $0 < c_1 < c_2 < c_3 < c_4 < c_5$ for business-compliance reasons.

$\mathbb{R}$ is the state transition probability and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, with $r(s, a)$ as the immediate reward (user feedback) of action a under state s. $\gamma < 1$ is the discounting factor for future rewards. Given the latent user state s_t, a softmax policy over the item corpus $\mathcal{A}$ is parameterized by θ as:

$$\pi_\theta(a|s_t) = \frac{\exp(\mathbf{u}_{s_t}^\top \mathbf{v}_a)}{\sum_{a' \in \mathcal{A}} \exp(\mathbf{u}_{s_t}^\top \mathbf{v}_{a'})}, \quad \forall a \in \mathcal{A}, \tag{8}$$

Here $\mathbf{u}_{s_t}$ is the latent user state and $\mathbf{v}_a$ is the embedding for item a, and both are learned together with other variables in the network. At time t, the agent acts according to $\pi_\theta(a|s_t)$. The policy parameters θ are learned using REINFORCE [57] to maximize the cumulative reward over all trajectories. In an offline batch learning setting, importance sampling is applied to correct for the off-policy distribution shift between the learned policy $\pi_\theta(\cdot|s)$ and behavior policy $\beta(\cdot|s)$ that generates the trajectories. This results in a gradient estimate for θ as

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \sum_{s_t \sim d_t^\beta(s), a_t \sim \beta(\cdot|s_t)} \left[\frac{\pi_\theta(a_t|s_t)}{\beta(a_t|s_t)} R_t(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right], \tag{9}$$

where $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$ is the discounted future return, and $d_t^\beta(s)$ is the state visitation probability under β. We incorporate the proposed user behavior patterns into the learning objective of the agent by *reshaping* R_t with the identified surrogates.

6.2 Consumption Diversity as Surrogate

6.2.1 Reward Formulation. As shown in Section 4 and 5, the entropy-based diversity $D_{\text{entropy}}(S)$ is highly predictive of increased visiting frequency over time. Denoting $R_t^o(s_t, a_t)$ as the original reward used in the REINFORCE recommender, we reshape the reward by

$$R_t(s_t, a_t) = R_t^o(s_t, a_t) \cdot \exp\left[m \left(D_{\text{entropy}}(S_t) - D_{\text{entropy}}(S_{t-1}) \right) \right], \tag{10}$$

where $D_{\text{entropy}}(S_t)$ measures the consumption diversity over a 2-week window *including* the current consumption a_t, and $D_{\text{entropy}}(S_{t-1})$ the consumption diversity over the same window but *excluding* the current consumption. The reward formulation with diversity as the surrogate is intuitive: When the recommended item is consumed[6] and it increases the diversity of the user's consumption history (i.e. $D_{\text{entropy}}(S_t) - D_{\text{entropy}}(S_{t-1}) > 0$), we assign a higher reward than its original value by applying a multiplier that's greater than 1. Otherwise a multiplier lower than 1 is applied (i.e. $D_{\text{entropy}}(S_t) - D_{\text{entropy}}(S_{t-1}) < 0$). A scaling factor $m > 0$ controls the strength of the surrogate reward. In our experiments, we find that $m = 5$ works well without the need for aggressive tuning. $D_{\text{entropy}}(S_{t-1})$ is also concatenated into the user state u_{s_t} for the model to better learn and adapt to the proposed reward changes.

We choose multiplicative design for the reward surrogate as opposed to an additive one with the following reason: If the original reward for an action $R_t^o(s_t, a_t)$ is large, indicating that the user enjoys an item, then it gets amplified even more with the multiplier if the action also increases consumption diversity.

6.2.2 Results. Figure 6 summarizes the live A/B experiments where we report the percentage improvements of the metrics over the baseline REINFORCE algorithm *over time*. Figure 6a and 6b shows the movements in a top-line metric capturing user overall enjoyment on the platform, and a proxy metric for user's long-term visiting frequency, respectively. We see that the reward surrogate model not only improves the top-line metric and user retention significantly, but it also exhibits a strong learning effect over the course of the experiment. This indicates that the proposed change enables users to *continuously* discover and consume more diverse contents toward an improved user experience in the long term. We also see growing differences in the number of topic clusters consumed (Figure 6c), which confirms the effect of the reward surrogate.

6.3 Homepage Revisits as Reward Surrogate

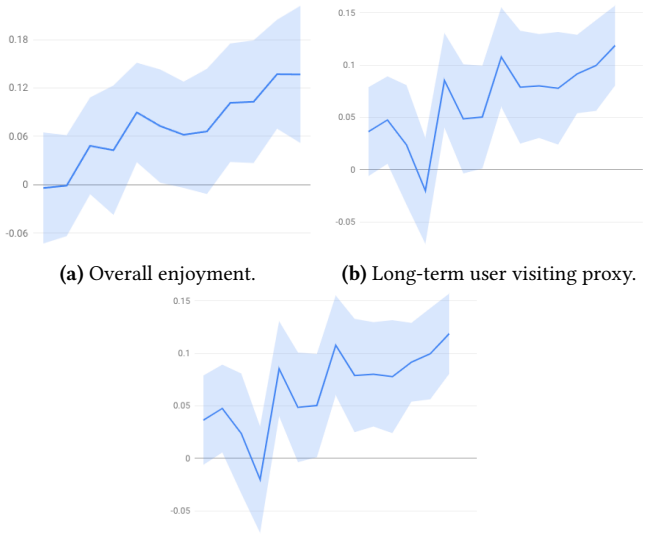
6.3.1 Reward Formulation. Based on the findings in Section 4 and 5, we propose to leverage homepage visits as reward surrogates:

$$R_t(s_t, a_t) = R_t^o(s_t, a_t) \cdot \left[(1 + c\mathbb{1}(T_{\text{revisit}}(S, \text{Home}) < T_0) \right], \tag{11}$$

where the recommended item will receive a boost with multiplier c if the user consumes the item and comes back to the homepage with a high-quality consumption within the next T_0 time. $c > 0$ is a tuning parameter controlling the strength of the reward surrogate. We tuned c among $\{5, 10, 20\}$ and $c = 10$ performed the best.

6.3.2 Results. Compared with the baseline, we see an overall increase in user visiting frequency (Figure 7a), which mainly comes from the low-frequency users (Figure 7b). Figure 7c shows a significant increase in the number of homepage visits, which confirms the effect of the proposed reward surrogate in nudging the users toward the desired behavior patterns (i.e. visit homepage more often). We also see an increase in the number of satisfied consumptions (Figure 7d), which indicates a more satisfactory long-term user experience.

[6]Note that when the recommended item a_t is not consumed, then $R_t(s_t, a_t) = 0$ and the proposed reward surrogate will have no effect.

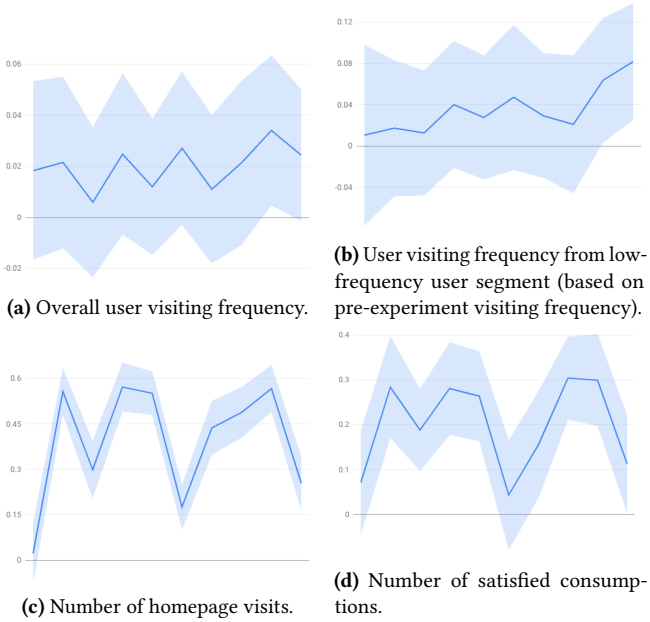


(a) Overall enjoyment.

(b) Long-term user visiting proxy.

(c) Number of topics consumed.

Figure 6: Entropy-based diversity as reward surrogate; Results are shown as percentage difference in metric values against baseline.



(a) Overall user visiting frequency.

(b) User visiting frequency from low-frequency user segment (based on pre-experiment visiting frequency).

(c) Number of homepage visits.

(d) Number of satisfied consumptions.

Figure 7: Homepage revisits T_{revisit} as reward surrogate.

7 CONCLUSION

In this work, we provide analytical insights and algorithmic improvements for optimizing long-term user experience on recommender systems. To tackle the challenges of long-term user experience being a noisy, sparse and delayed signal, we propose to establish the association between the long-term objective and a set of medium-term user behavior signals that can serve as surrogate objectives. Specifically, we define and identify a set of medium-term user behavior patterns that are predictive of the changes in user's visiting frequency to the platform, which includes consumption diversity, repeated consumption behavior and curating persistent topics/interests etc. We then validate the efficacy of those behavior patterns by incorporating them as reward surrogates in an RL-based recommender system. Live experiment results on an industrial recommendation platform shows the effectiveness of these surrogates in improving long-term user experience, which is manifested through increased user visiting frequency. Our work provides practical guidance for algorithm designers to identify and leverage interpretable user behavior patterns as surrogates for optimizing long-term user experience in recommender systems.

REFERENCES

[1] Gediminas Adomavicius and YoungOk Kwon. 2011. Maximizing aggregate recommendation diversity: A graph-theoretic approach. In *Proc. of the 1st International Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011)*. Citeseer, 3–10.

[2] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *The Web Conf 2020*. 2155–2165.

[3] Asim Ansari, Yang Li, and Jonathan Z Zhang. 2018. Probabilistic topic model for hybrid recommender systems: A stochastic variational Bayesian approach. *Marketing Science* 37, 6 (2018), 987–1008.

[4] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.* Technical Report. National Bureau of Economic Research.

[5] Xueying Bai, Jian Guan, and Hongning Wang. 2019. Model-Based Reinforcement Learning with Adversarial Training for Online Recommendation. *arXiv preprint arXiv:1911.03845* (2019).

[6] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, NY, USA., 35.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*. 160–169.

[9] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Recsys' 18*. 224–232.

[10] Thomas Chatzidimitris, Damianos Gavalas, Vlasios Kasapakis, Charalampos Konstantopoulos, Damianos Kypriadis, Grammati Pantziou, and Christos Zaroliagis. 2020. A location history-aware recommender system for smart retail environments. *Personal and Ubiquitous Computing* (2020), 1–12.

[11] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *WSDM 2019*. 456–464.

[12] Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. 2019. Surrogate objectives for batch policy optimization in one-step decision making. (2019).

[13] Minmin Chen, Yuyan Wang, Can Xu, Ya Le, Mohit Sharma, Lee Richardson, Su-Lin Wu, and Ed Chi. 2021. Values of User Exploration in Recommender Systems. In *Recsys' 21*. 85–95.

[14] Shi-Yong Chen, Yang Yu, Qing Da, Jun Tan, Hai-Kuan Huang, and Hai-Hong Tang. 2018. Stabilizing reinforcement learning in dynamic environment with application to online recommendation. In *KDD' 18*. 1187–1196.

[15] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Recsys' 16*. 191–198.

[16] Daniel R Fesenmaier, Karl W Wöber, and Hannes Werthner. 2006. *Destination recommendation systems: Behavioral foundations and applications.* Cabi.

[17] Sharad Goel, Andrei Broder, Evgeniy Gabrilovich, and Bo Pang. 2010. Anatomy of the long tail: ordinary people with extraordinary tastes. In *WSDM 2010*. 201–210.

[18] Mihajlo Grbovic and Haibin Cheng. 2018. Real-time personalization using embeddings for search ranking at airbnb. In *KDD' 18*. 311–320.

[19] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting consumption towards diverse content on music streaming platforms. In *WSDM 2021*. 238–246.

[20] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 191–200.

[21] David Holtz, Ben Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. 2020. The engagement-diversity connection: Evidence from a field

experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 75–76.
[22] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. 2019. SlateQ: A tractable decomposition for reinforcement learning with recommendation sets. (2019).
[23] Luo Ji, Qin Qi, Bingqing Han, and Hongxia Yang. 2021. Reinforcement Learning to Optimize Lifetime Value in Cold-Start Recommendation. *arXiv preprint arXiv:2108.09141* (2021).
[24] Marshall M Joffe and Tom Greene. 2009. Related causal frameworks for surrogate outcomes. *Biometrics* 65, 2 (2009), 530–538.
[25] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM' 18*. IEEE, 197–206.
[26] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. 2018. Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers* 20, 6 (2018), 1241–1265.
[27] Mesut Kaya and Derek Bridge. 2019. A comparison of calibrated and intent-aware recommendations. In *Recsys' 19*. 151–159.
[28] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 441–504.
[29] Jens Kober, J Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274.
[30] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction* 22, 1 (2012), 101–123.
[31] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *KDD' 09*. 447–456.
[32] Elisabeth Lex, Dominik Kowald, Paul Seitlinger, Thi Ngoc Trang Tran, Alexander Felfernig, Markus Schedl, et al. 2021. Psychology-informed recommender systems. *Foundations and Trends® in Information Retrieval* 15, 2 (2021), 134–242.
[33] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 31–40.
[34] Bogdan Mazoure, Paul Mineiro, Pavithra Srinath, Reza Sharifi Sedeh, Doina Precup, and Adith Swaminathan. 2021. Improving Long-Term Metrics in Recommendation Systems using Short-Horizon Offline RL. *arXiv preprint arXiv:2106.00589* (2021).
[35] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Making recommendations better: an analytic model for human-recommender interaction. In *CHI'06 extended abstracts on Human factors in computing systems*. 1103–1108.
[36] Klaus Nehring and Clemens Puppe. 2002. A theory of diversity. *Econometrica* 70, 3 (2002), 1155–1198.
[37] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. 278–287.
[38] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW' 14*. 677–686.
[39] Zachary A Pardos and Weijie Jiang. 2020. Designing for serendipity in a university course recommendation system. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 350–359.
[40] Francesco Sanna Passino, Lucas Maystre, Dmitrii Moor, Ashton Anderson, and Mounia Lalmas. 2021. Where To Next? A Dynamic Model of User Preferences. In *The 2021 World Wide Web Conference, WWW*. 19–23.
[41] GP Patil and Charles Taillie. 1982. Diversity as a concept and its measurement. *Journal of the American statistical Association* 77, 379 (1982), 548–561.
[42] Ross L Prentice. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine* 8, 4 (1989), 431–440.
[43] Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences* 3, 6 (1999), 233–242.
[44] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. 2005. An MDP-based recommender system. *Journal of Machine Learning Research* 6, 9 (2005).
[45] Donghee Shin. 2020. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior* 109 (2020), 106344.
[46] Brent Smith and Greg Linden. 2017. Two decades of recommender systems at Amazon. com. *Ieee internet computing* 21, 3 (2017), 12–18.
[47] Barry Smyth and Paul McClave. 2001. Similarity vs. diversity. In *International conference on case-based reasoning*. Springer, 347–361.
[48] Yicheng Song, Nachiketa Sahoo, and Elie Ofek. 2019. When and how to diversify—a multicategory utility model for personalized content recommendation. *Management Science* 65, 8 (2019), 3737–3757.
[49] Harald Steck. 2018. Calibrated recommendations. In *Recsys' 19*. 154–162.
[50] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM' 19*. 1441–1450.
[51] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2017. # exploration: A study of count-based exploration for deep reinforcement learning. In *31st Conference on Neural Information Processing Systems (NIPS)*, Vol. 30. 1–18.
[52] Quentin Villermet, Jérémie Poiroux, Manuel Moussallam, Thomas Louail, and Camille Roth. 2021. Follow the guides: disentangling human and algorithmic curation in online music consumption. In *Recsys' 21*. 380–389.
[53] Jingkang Wang, Yang Liu, and Bo Li. 2020. Reinforcement learning with perturbed rewards. In *AAAI*, Vol. 34. 6202–6209.
[54] Y Wang, Y Ning, I Liu, and XX Zhang. 2018. Food discovery with uber eats: Recommending for the marketplace. (2018).
[55] Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang. 2014. Friendbook: a semantic-based friend recommendation system for social networks. *IEEE transactions on mobile computing* 14, 3 (2014), 538–551.
[56] Christopher J Weir and Rosalind J Walley. 2006. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in medicine* 25, 2 (2006), 183–203.
[57] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.
[58] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1927–1936.
[59] Bo Xiao and Izak Benbasat. 2007. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS quarterly* (2007), 137–209.
[60] Xing Xie. 2010. Potential friend recommendation in online social network. In *2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*. IEEE, 831–835.
[61] Xuhai Xu, Ahmed Hassan Awadallah, Susan T. Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite, and Farnaz Jahanbakhsh. 2020. Understanding user behavior for document recommendation. In *theWebConf 2020*. 3012–3018.
[62] Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. 2020. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835* (2020).
[63] Gregory Yauney and Pratik Shah. 2018. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. In *Machine Learning for Healthcare Conference*. PMLR, 161–226.
[64] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research* 31, 1 (2020), 76–101.
[65] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.
[66] Xiao Zhang, Haonan Jia, Hanjing Su, Wenhan Wang, Jun Xu, and Ji-Rong Wen. 2021. Counterfactual Reward Modification for Streaming Recommendation with Delayed Feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 41–50.
[67] Xing Zhao, Ziwei Zhu, and James Caverlee. 2021. Rabbit Holes and Taste Distortion: Distribution-Aware Recommendation with Evolving Interests. In *The Web Conf 2021*. 888–899.
[68] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *Recsys' 19*. 43–51.
[69] 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.
[70] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. 2010. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 404–410.
[71] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *PNAS* 107, 10 (2010), 4511–4515.
[72] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *WWW' 05*. 22–32.
[73] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *KDD' 19*. 2810–2818.

A APPENDIX

A.1 Full results on surrogate selection

Feature	Number of consumptions in 14d						Avg
	$(0, c_1]$	$(c_1, c_2]$	$(c_2, c_3]$	$(c_3, c_4]$	$(c_4, c_5]$	(c_5, ∞)	
T_{revisit}(Home)	0.186	0.135	0.113	0.129	0.116	0.099	0.13
$T^{\text{diff}}_{\text{revisit}}$(Home)	0.068	0.04	0.057	0.089	0.094	0.081	0.072
D_{entropy}	0.072	0.1	0.082	0.05	0.061	0.041	0.068
$D^{\text{diff}}_{\text{KL}}$	0.061	0.059	0.066	0.063	0.074	0.052	0.062
$D^{\text{diff}}_{\text{entropy}}$	0.042	0.067	0.061	0.066	0.054	0.06	0.058
D_{KL}	0.068	0.065	0.058	0.064	0.036	0.055	0.058
T_{revisit}(Consump)	0.069	0.059	0.044	0.059	0.058	0.043	0.055
$R^{\text{diff}}_{\text{repeatedCons}}$	0.03	0.067	0.046	0.043	0.066	0.061	0.052
$R^{\text{diff}}_{\text{highQualCons}}$	0.046	0.042	0.053	0.056	0.043	0.064	0.051
T_{revisit}(Search)	0.064	0.066	0.039	0.036	0.033	0.067	0.051
$T^{\text{diff}}_{\text{revisit}}$(Consump)	0.037	0.017	0.044	0.062	0.058	0.066	0.047
$R^{\text{diff}}_{\text{persistentTopic}}$	0.031	0.04	0.054	0.045	0.046	0.062	0.046
$R_{\text{highQualCons}}$	0.04	0.043	0.045	0.06	0.031	0.05	0.045
$D^{\text{diff}}_{\text{ratio}}$	0.04	0.036	0.043	0.039	0.062	0.044	0.044
$T^{\text{diff}}_{\text{revisit}}$(Search)	0.028	0.046	0.067	0.034	0.034	0.042	0.042
D_{ratio}	0.037	0.032	0.044	0.03	0.035	0.029	0.035
$R_{\text{repeatedCons}}$	0.027	0.033	0.028	0.021	0.04	0.025	0.029
$R_{\text{persistentTopic}}$	0.028	0.023	0.027	0.027	0.033	0.035	0.029
$N^{\text{diff}}_{\text{active}}$	0.012	0.015	0.019	0.015	0.019	0.012	0.015
N_{active}	0.014	0.014	0.01	0.011	0.009	0.011	0.012

Table 2: Feature importance results for the stratified surrogate selection model, where the feature importance score for every fine-grained user activity group is reported. The rows are ranked by the average importance score across all user slices. We drop the dependency on S for the user activity patterns defined in Section 3 for ease of notation.