Investigating Listeners' Responses to Divergent Recommendations

Rishabh Mehrotra
rishabhm@spotify.com
Spotify, London
London, UK

Chirag Shah
University of Washington
Seattle, US
chirags@uw.edu

Benjamin Carterette
benjaminc@spotify.com
Spotify, NY
New York, US

ABSTRACT

Recommender systems offer great opportunity not only for users to discover new content, but also for the providers of that content to find new audience, followers, and fans. Users often come to a recommender system with certain expectations about what it will recommend to them, and a recommender system that is optimized for creating opportunities for content creators may provide recommendations that are very different from what a user is expecting. We hypothesize that some users' expectations have a much wider range of acceptability than others, and users with more "receptivity" to subversion of their expectations are likely to accept such divergence in the recommended content. Understanding users' responses to such recommendations is vital to platforms that need to serve multiple stakeholders. In this work we investigate logged behavioral responses of users of an audio streaming platform to recommendations that deviate from their expectation, or "divergent" recommendations. We present three classes of listener response to divergent recommendations that can be identified in interaction logs with the aim of predicting which users can be targeted for future divergent recommendations. We derive a number of user characteristics based on user's music consumption which we think are predictive of user's receptivity, train models to predict receptivity of these users, and run a live A/B test to validate our approach by correlating with engagement.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

User receptivity, Recommender Systems

ACM Reference Format:
Rishabh Mehrotra, Chirag Shah, and Benjamin Carterette. 2020. Investigating Listeners' Responses to Divergent Recommendations. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20), September 22–26, 2020, Virtual Event, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3383313.3418482

1 INTRODUCTION

Recommender systems offer great opportunity not only for users to discover new content, but also for the providers of that content to find new customers, followers, and fans. But most user-centric recommender systems are optimized solely for user-centric objectives, including clicks [18], dwell time [17], session length time [7], streaming time, and conversion [14], among others. Platform ecosystems have recently witnessed explosive growth by facilitating efficient interactions between multiple stakeholders, including for example buyers and retailers (Amazon), guests and hosts (AirBnb), riders and drivers (Uber), and listeners and artists (Spotify). Recommender systems powering such multi-stakeholder platforms could provide even more value by additionally optimizing their models for different stakeholder objectives, including exposure, fairness, diversity, promotion, and revenue metrics [11, 13].

While helpful overall, optimizing a model for objectives that do not directly serve the user might distort recommendations, in ways which might decrease user satisfaction. But each individual user is different, and each has their own expectations of what they will see when they visit their favorite recommendation platform. Not all users will appreciate deviance from their expectation, and some may be more sensitive to the degree of deviation than others. If we can detect which users are more likely to become frustrated or dissatisfied with their recommendations, and quantify whether or not a user is receptive to such distortions, we can give ourselves more leeway to explore and serve recommendations based on other stakeholder objectives.

In this work we investigate user behavior in a large audio streaming platform. We refer to recommendations that deviate from expectation as *divergent*, and provide a system that gives divergent music recommendations to listeners. We define three notions of user receptivity based on user behavioral response to these recommendations in terms proxy measures of satisfaction and effort, specifically based on: (i) drop in engagement; (ii) increase in effort; and (iii) return rate of users. All three of these are attempting to capture different aspects of user *receptivity* to recommendations that diverge from their expectation. These notions of receptivity is integral to the work and part of our contribution. Furthermore, we identify features that can be used to predict in which class a user may be given a context (e.g. nostalgia, currency), and train models to predict degree to which their behavior will change given distorted recommendations. We contend that our experimental insights derived from live AB test data from over 20 million listeners have implications on the design and impact of recommendation systems powering online multi-stakeholder platforms and motivates future work, which we briefly discuss at the end of the paper.

RecSys '20, September 22–26, 2020, Virtual Event, Brazil
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7583-2/20/09.
https://doi.org/10.1145/3383313.3418482

2 RELATED WORK

Recent advances in optimization and evaluation of recommender systems have resulted in putting forth a number of scenarios wherein the recommendations surfaced to users are intentionally divergent from the ones that would achieve high relevance or accuracy. Greedily selecting relevant items creates filter bubble style issues, and recent research has proposed ways of introducing diversity in recommendations [3, 5, 10, 16]. Beyond diversity, another key factor contributing to divergent recommendations being surfaced in front of users is the presence of conflicting optimization objective for the recommender model. Recent advancements in multi-stakeholder platforms have necessitated the need to train such models for a number of , often conflicting, objectives [1, 11, 12]. The presence of non-user centric objectives results in serving recommendations which are not necessarily aimed at piquing user interest, but to ensure performance on other stakeholder metrics (e.g. fairness and equality of exposure to suppliers, platform revenue, et cetera). Finally, intentional randomized distortions in surfaced content play a key role in developing unbiased counterfactual evaluation frameworks, which enable efficient large scale experimentation [4, 8, 9].

We believe a user's receptivity is an intervening factor when we want to see the impacts of introducing distortions in recommendations to their engagement. In this context, Sun et al. [15] define receptivity τ as $\tau = 1 - \frac{R-A+1}{2}$. Here, $R \in [0, 1]$ is the resistance of user u_i to object o_j and $A \in [0, 1]$ is the acceptance of u_i to o_j. For our work, the object will be a recommendation or a recommendation strategy. Related to our notion of receptivity, recent work has also investigated measuring consumer sensitivity to advertising, by trialing different level of audio advertising interrupting their streaming sessions and assessed long-run demand effects [6].

3 UNDERSTANDING PREDICTORS OF RECEPTIVITY

Our goal is to define the notion of user receptivity to divergence in recommendations, and understand its impact on key behavioral aspects of user engagement. In this section, we begin by presenting a motivating example and define the three notions of receptivity. Further, we present user behavior characteristics which could help us predict user receptivity.

3.1 Motivating example

Suppose a user visits the home screen of an entertainment site with an underlying recommendation system using contextual bandits or similar technologies. Because the recommendations are contextual, this user may be accustomed to recommendations changing for opaque reasons – for example, if they visit the home screen multiple times in a day and see slightly different recommendations each time. To what extent can we vary recommendations while maintaining the user's satisfaction with the home screen and not losing their trust? This motivates our idea of *receptivity.*

Because satisfaction is often not directly observable, this notion of receptivity is captured in the way the user changes their interaction patterns in response to changes—the *proxies* to satisfaction that we use to assess changes to algorithmic recommendations. Users who change their behavior very little in response to substantial distortions in recommendations could be thought of as very receptive. Users who abandon the home screen entirely in favor of other ways to access content may be thought of as having little receptivity. Others may put in more effort up to a point, until they find they are putting in too much effort for their satisfaction and give up on the home screen.

Finally, even if some users find themselves drifting away from the home screen if the recommendations become too variable or they require too much effort, users themselves may engage in some form of *explore/exploit* with the system and eventually come back to the home screen if either variability in recommendations or effort required decrease to a level the user feels more comfortable with.

3.2 Defining user receptivity

The above examples motivate definitions of user "receptivity" to varying recommendations that are fully in terms of changes in user behavior. To be more specific, we quantify "variability in recommendations" using two states: one is the "normal" state that users are accustomed to, that which comes from a contextual bandit approach that performs some exploration. The recommendations surfaced in this state are derived from a model optimized for a user-centric objective aimed at increasing user delight. The second is a fully-randomized home screen wherein the recommendations are significantly distorted by randomly selecting content to surface to user, from a set of pre-filtered candidates.

To measure satisfaction or engagement, we use *reach*, that is, the amount of content a user has interacted with. To measure effort, we use *depth*, that is, how deep the user is willing to look in the recommendations to find something desirable. Then we can define user receptivity in terms of changes in their engagement and effort between the fully-randomized condition and the "normal" recommendations.

3.2.1 *Definition 1: Engagement-centric receptivity.*

Our first definition is concerned with observing whether there is a drop in user engagement when the user is exposed to random recommendations. Essentially, we say the user is receptive if and only if their proportional decrease in engagement between the two conditions is less than some threshold.

$$\tau_u = 1 \Leftrightarrow \frac{|o_u - o'_u|}{o_u} \leq \rho$$

Here τ_u is a binary variable representing user receptivity, o_u is user engagement with the regular home screen, o'_u is user engagement with the fully-randomized home screen, and ρ is a threshold.

3.2.2 *Definition 2: Effort-centric receptivity.*

Our second definition relates receptivity to both engagement and effort. In this case, we say the user is most receptive if, between the two conditions, their engagement does not drop and their effort does not increase. The user is receptive if their engagement does not drop, but they are putting in more effort to achieve the same level of engagement.

If the user's engagement drops between the two conditions but they are putting in effort, they may exhibit some receptivity but are probably on a path to frustration. And if a user's engagement drops and they are not putting in any additional effort, they are simply not receptive to the randomized/divergent recommendations.

Thus we can say:

$$\tau = \begin{cases} 2 & \frac{|o_u - o'_u|}{o_u} < \rho \text{ and } \frac{|e_u - e'_u|}{e_u} < \epsilon \\ 1 & \frac{|o_u - o'_u|}{o_u} < \rho \text{ and } \frac{|e_u - e'_u|}{e_u} \geq \epsilon \\ 0 & \frac{|o_u - o'_u|}{o_u} \geq \rho \text{ and } \frac{|e_u - e'_u|}{e_u} < \epsilon \\ -1 & \frac{|o_u - o'_u|}{o_u} \geq \rho \text{ and } \frac{|e_u - e'_u|}{e_u} \geq \epsilon \end{cases}$$

Here, e_u represents effort put into the regular home screen and e'_u represents effort put into the fully randomized home screen. The threshold ϵ is on relative difference in effort.

3.2.3 Definition 3: Emendation-centric receptivity. Finally, we can view the two conditions in time sequence. A user that reduces engagement with the home screen when switched in conditions and then returns to regular engagement when switched back is less receptive than one that maintains the same level of engagement throughout.

Cost and Risk of Variations in Recommendations: It is important to note that these three definitions present an increasing level of annoyance users might have with variations in the surfaced recommendations. These varied notions of receptivity equips system designers to ascertain the level of user receptivity they are willing to risk in order to push other recommendation objectives.

3.3 User characteristics

An important question we aim to address in this work is: What types of users exhibit greater receptivity? We leverage large scale historic user interaction information to identify and quantify key user characteristics, which might be predictive of user receptivity and response to variability in served recommendations, including: age, gender, auditacity, currency, popularity score, skip rate, nostalgia, number of streams, number of artists streamed, 30 day engagement, 30 day discovery engagement, 30 day user-collection engagement.

The six categories of features cover demographic information, measures derived from listening behavior, and diversity of historically consumed content. Effort and engagement features are computed using the past 30 day interaction data. Finally, we derive features based on personality traits of users, using the notions of propensity to listen to music with characteristics such as how current it is, nostalgia, popularity, and auditacity.

Nostalgia describes to what extent a user listens to music that makes them feel nostalgic. A user with low nostalgia never listens to nostalgic music. A user with high nostalgia frequently listens to their generation's music. Currency scores users by how regularly they listen to brand new music. Auditacity scores users by how much skipping they do, i.e. how often the user skips content vs how often they allow the content to play without skipping.

4 PREDICTING USER RECEPTIVITY

Based on the defined notions of user receptivity and user characteristics, we aim at predicting response of users based on the distortions in the recommended content. We train predictive models on historic user interaction data, and conduct a live AB test to validate our findings.

4.1 Live AB Test

We conduct a large scale randomized live AB test on a music streaming platform and use user interaction data before and during the test period to make inference about user's receptivity. Specifically, for a period of one month, instead of surfacing recommendations drawn from a model trained on user satisfaction, we randomly select playlists to show to the users on the homescreen. Users can scroll left and right, and up and down, to see more playlists. We define satisfaction and effort in terms of their interactions with playlists and how deep they scroll.

Our dataset consists of these users, their interactions in the month before the randomization (when they saw "normal" recommendations), the month during the randomization, and the month after. We computed all features in Table **??**, for over 20 million users and 100 million sessions.

4.2 Predictive Models

We train a simple XGBoost model based on the extracted user interaction features. Specifically, for each user session we extract the features identified in Table **??** and assign three receptivity labels based on the three notions of receptivity defined in Section 3.2. The three labels help us define three separate prediction task, each predicting one specific notion of receptivity from among the three outlined before.

Additionally, we implemented three different baselines for predicting whether a user is receptive or not: (i) *Random baseline*, wherei we labeled each user with 'receptive' or 'non-receptive' randomly; (ii) *Engagement-based predictor*, wherein users with system engagement above the median for the platform were labeled them 'receptive', otherwise 'non-receptive', and (iii) *Diversity-based predictor*, wherein if a user's openness to artist diversity is above the mean for the platform, we labeled him with 'receptive', otherwise with 'non-receptive'.

4.3 Heterogeneity in User Response to Random Recommendations

We begin by investigating how users respond to random distortions in the served recommendations. For each user, we compute the proportional drop in engagement levels during the random recommendations period compared to the engagement in a normal recommendation period, i.e., $\frac{|d_u - d'_u|}{d_u}$ where d_u is the number of days in the normal recommendation month the user streamed tracks from their regular home screen, while d'_u is corresponding engagement number for the duration of fully-randomized home screen. A positive number indicates drop in engagement resulting due to variations in the served recommendations. In Figure 1 we plot the distribution of these proportional drop in engagement for three user segments: low use users (active 0-5 days a month), medium use users (active 15-20 days a month) and heavy use users (active 25+ days a month), based on user's activity levels in the past month.

We observe stark differences in how users respond to random recommendations being served to them. Heavy use users experience the biggest dip in engagement levels, with most users witnessing a drop of over 25% engagement. Medium and low use users do

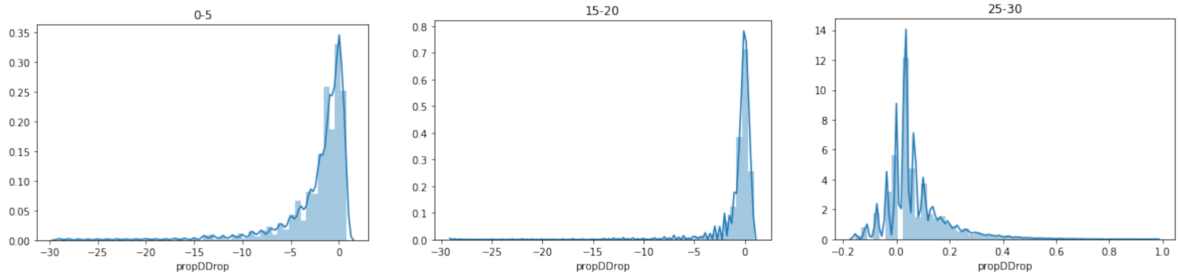

Figure 1: Distribution of dip in user engagement between the random recommendations period and normal recommendation period. Left: Low use users, active 0-5 days a month; Middle: medium use users, active 15-20 days a month; Right: high use users, active 25+ days a month.

not witness such significant drop in engagement. Low use users were anyway not actively engaging with the home screen, so either they remain unaffected by randomized recommendations, or their engagement increases. Even a slight increase in engagement activity for low use users would result in high proportional increases.

These results demonstrate the presence of user level heterogeneity in how users respond to distortions in their recommendations. Further, the risk of random variations in recommended content is higher for heavy use users; while being significantly lower for low use users, since they anyway were not actively engaging with recommended content before.

4.4 How well can we predict user receptivity?

Given the heterogeneity in user response to distortions in their recommendations, we next investigate how predictive is user's receptivity. Table 1 compares the performance of the simple baselines and the prediction model trained on the user features identified earlier. We observe that randomly labeling a user as 'receptive' or not does as we would suspect – giving us 50% accuracy – no better than a coin-toss. The other two baselines, however, do yield better and more reasonable accuracies. This indicates that engagement and diversity are two important factors that contribute to user receptivity.

However, it is important to note that diversity of consumption is not strongly predictive of user receptivity. User's consumption diversity is often hypothesized to be related to user's acceptability of varied recommendations [2]. Indeed, a user with a higher consumption diversity is often expected to be acceptable to recommendations departing their taste. However, we find strong evidence that this trait is not very predictive of user's receptivity and acceptability of distortions in the served recommendations. It also highlights that not any random diversity is good diversity, and user's who prefer diversity prefer personalized diversity.

We observe significant boost in receptivity prediction upon training a model with nuanced user features, with over 10% gains across all metrics. We also observe that it is easier to predict engagement receptivity, than to predict effort or emendation receptivity, with engagement accuracy reaching 70%. We posit two confounding factors: (i) predicting effort and stickiness of users is harder problem in general, and (ii) the features used are better suited for engagement receptivity task, and we need additional explanatory features for the other receptivity tasks.

4.4.1 Identifying Key Predictors of User receptivity. Figure 2 shows key predictors for the three notions of user receptivity. We note that engagement appears to be a good predictor for the first definition of receptivity; skip rate and other effort proxies such as number of artists listened to appear to be good predictors for the second definition. Good features for the third definition include listening to more current music and propensity for discovery, both features that intuitively make sense as things that would drive a user to return to the home screen after having abandoned it previously, particularly if the home screen is showing new releases personalized to the user.

4.5 Leveraging User receptivity to Predict Future Engagement

The findings and results so far highlight our ability to understand and predict user receptivity. In this section, we demonstrate the usefulness of user receptivity by quantifying how predictive it is of future engagement with the platform. For each user, we consider the first-X days of the randomized recommendation period and using information from the past month of normal recommendation and the first-X days of distorted recommendations, we investigate how predictive user behavior is for future engagement with the homescreen. Figure **??** presents the correlation plots for the three notions of receptivity with future engagement (during the next 4 weeks) with the platform.

We observe that effort based receptivity is more predictive of future engagement, slightly more than engagement based receptivity and significantly more than emendation receptivity. Further, we observe a stark difference across days: engagement receptivity is predictive of short term, however effort receptivity is more correlated with longer term future engagement with the platform. One possible interpretation indicates that soliciting high effort from the user over a sustained period of time leads to decrease in user delight, and might result in dis-engagement with the platform.

Overall, while the correlation scores are far from perfect, they are statistically significant and therefore a useful indicator of user receptivity. These results offer preliminary indications that user receptivity is indeed a useful trait to quantify and consider in decision making and recommendation strategy development process.

5 CONCLUSION & IMPLICATIONS

In this work, we hypothesized that users are differently receptive to divergence in their recommendations, and proposed three different notion of user receptivity. We conducted an online test wherein

	Engagement receptivity				Effort receptivity				Emendation receptivity			
Method	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Random	0.50	0.5	0.5	0.5	0.50	0.5	0.5	0.5	0.50	0.5	0.5	0.5
Diversity based	0.58	0.58	0.55	0.55	0.54	0.56	0.56	0.55	0.53	0.54	0.54	0.54
Engagement based	0.63	0.65	0.63	0.63	0.56	0.57	0.57	0.56	0.55	0.56	0.56	0.55
Proposed	0.70	0.72	0.70	0.70	0.617	0.62	0.63	0.62	0.603	0.60	0.60	0.60

Table 1: Classifier performance for each of our three definitions of receptivity.

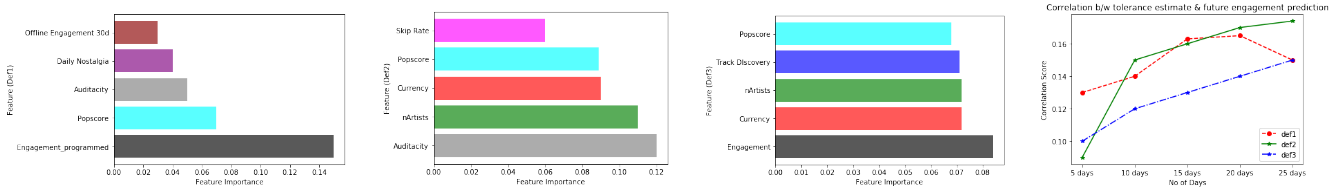

Figure 2: Top 5 features by importance, for predicting the three notions of user receptivity: engagement receptivity (left), effort receptivity (center-left), emendation receptivity(center-right). Right: Correlation between receptivity estimates and future engagement. Def1, def2 & def3 refer to engagement centric, effort centric and emandation centric definitions of receptivity.

we purposefully distorted recommendations served to users and leverage user's behavioral signals around engagement and effort to predict how receptive a user is. We observe that different users are receptive to different extent, and that user receptivity is predictable based on certain user characteristics.

The wider implication of this work is that we can leverage a deeper understanding of users' responses to changes in their recommendations to provide recommendations that have greater aggregate value to everyone with a stake in the platform: not just users, but content providers and platform owners as well. Recommendation strategies can be designed that predict which users are better for targeting divergent recommendations and that include an understanding of how users are likely to react to those recommendations. Finally, our work has implications for the impact of collecting randomized data. Such data is very useful for offline testing and training, so understanding how users will react to it, and particularly how it might induce negative responses in users, will enable system developers to better adjust the scope of randomized data collection tests.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Recommender systems as multistakeholder environments. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 347–348.

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

[3] Pablo Castells, Neil J Hurley, and Saul Vargas. 2015. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*. Springer, 881–918.

[4] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.

[5] Nina Hagemann, Michael P O'Mahony, and Barry Smyth. 2018. Module advisor: a hybrid recommender system for elective module exploration. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 498–499.

[6] Jason Huang, David Reiley, and Nick Riabov. 2018. Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio. *Available at SSRN 3166676* (2018).

[7] Dietmar Jannach and Malte Ludewig. 2017. When recurrent neural networks meet the neighborhood for session-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 306–310.

[8] Thorsten Joachims and Adith Swaminathan. 2016. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 1199–1201.

[9] Thorsten Joachims, Adith Swaminathan, Yves Raimond, Olivier Koch, and Flavian Vasile. 2018. REVEAL 2018: offline evaluation for recommender systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 514–515.

[10] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-Based Systems* 123 (2017), 154–162.

[11] Rishabh Mehrotra and Benjamin Carterette. 2019. Recommendations in a marketplace. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 580–581.

[12] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.

[13] Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. 2020. Bandit based Optimization of Multiple Objectives on a Music Streaming Platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3224–3233.

[14] Lili Shan, Lei Lin, and Chengjie Sun. 2018. Combined Regression and Tripletwise Learning for Conversion Rate Prediction in Real-Time Bidding Advertising. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 115–123.

[15] Lifeng Sun, Xiaoyan Wang, Zhi Wang, Hong Zhao, and Wenwu Zhu. 2017. Social-aware video recommendation for online social groups. *IEEE Transactions on Multimedia* 19, 3 (2017), 609–618.

[16] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 109–116.

[17] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*. 113–120.

[18] 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 *Proceedings of the 13th ACM Conference on Recommender Systems*. 43–51.