Analyzing user engagement in news application considering popularity diversity and content diversity

Atom Sonoda1 · Yoshifumi Seki2 · Fujio Toriumi3

Received: 7 July 2021 / Accepted: 17 August 2022 / Published online: 8 September 2022
© The Author(s) 2022

Abstract

The diversity of recommender systems has been well analyzed, but the impact of their diversity on user engagement is less understood. Our study is the first to attempt to analyze the relationship between diversity and user engagement in the news domain. In this study, we introduce the notion of popularity diversity, propose metrics for it, and analyze user behavior on popular news applications in terms of content diversity and popularity diversity, the impact of which we find to be closely related to user activity. We also find that users who use these services for longer periods have not only higher content diversity and popularity diversity but also a tendency to increase diversity as the week progresses. Users with low content diversity had a 224% greater withdrawal rate than those with high content diversity, and users with low popularity diversity had a 112% greater withdrawal rate than those with high popularity diversity. Although many studies have examined changes in diversity in recommender systems, we notice this trend is affected by user withdrawal. Finally, we confirm that popularity diversity and content diversity explain user withdrawal and predict user behavior. Our study reveals the relationship between diversity and engagement in the news domain and introduces the impact of popularity bias per user as a metric of diversity.

Keywords Diversity · User behavior analysis · Recommender systems · News media

Introduction

Diversity is a crucial metric of recommender systems, especially for user experience [16]. A lack of diversity in recommender systems can cause problematic filter bubbles, which are phenomena of online personalization that effectively
isolate people from a diversity of viewpoints or content [19]. Recommender systems are also related to echo chambers: information silos that reinforce biased ideas by only exposing users to information and communities they already prefer. It is believed that the loss of diversity in the information people receive leads to a leaning toward the political parties they support and triggers extremist views. The spread of recommender systems has made their diversity important for both user experiences and social issues.

We must therefore develop better news recommender systems and consider how best to use their diversity, as the news domain is more closely related to social issues than other domains are (e.g., commerce, cinema, and music), making news diversity proportionately more important. Although this importance is well recognized, the impact of diversity on news services and users remains unclear. Because news services are businesses that have profitability expectations, it is difficult to improve diversity by ignoring the benefits to services and users. Because user engagement is directly connected to user satisfaction and service profit, the latter and our study are essential to understand how diversity relates to them.

Our study is the first to attempt to analyze the relationship between diversity and user engagement in the news domain. In recent years, studies have analyzed the diversity dynamics of users as an active research topic [1, 5], but studies on the relationship between diversity and user engagement are rare [3].

We also introduce the concept of popularity diversity and propose a metric for it. In recent recommender systems studies, recommender system performance and its dependence on popular item bias have garnered attention, with some studies proposing evaluation metrics to reduce popularity bias [7, 23, 24]. Treviranus and Hockema [21] pointed out the influence of popularity on filter bubbles and echo chambers. However, previous studies on user behavior and diversity used only content diversity, which is the diversity of individual items [5, 15, 18]. In contrast, we define popularity diversity metrics and use them to analyze the relationship between diversity and user engagement.

First, we analyze content diversity and popularity diversity using user behavior logs. We compare both diversity metrics with user activity levels and determine which metrics are reflected by which characteristics. Next, we analyze diversity stability and confirm that most users maintain their trends over time. We confirm how the two diversity metrics relate to user activity levels and identify the meaning and characteristics of each metric. In other words, we find that the higher the click rate, the higher the diversity of both types.

We determine the relationship between user withdrawal and both diversity metrics. We assume that user withdrawal is correlated with user engagement. There are apparent differences in withdrawal rates between high and low diversity metrics. Users with low content diversity had a 224% greater withdrawal rate than those with high content diversity, and users with low popularity diversity had a 112% greater withdrawal rate than those with high popularity diversity. Although many studies have examined changes in diversity in recommender systems, we notice this trend is affected by user withdrawal. These diversity metrics are correlated with the number of clicks, and through detailed analysis, we find that popularity diversity better explains user withdrawal.
We develop a machine-learning model to predict user withdrawal using diversity as a feature, and we effectively predict such withdrawal with diversity metrics. These features have significance as variables that characterize user behavior.

Our contribution is threefold. First, we analyze the relationship between diversity and user engagement using user behavior logs from a news application and find that diversity can explain user engagement well. Second, to analyze the impact of diversity in the news domain, we introduce popularity diversity and propose a metric for it. Finally, we confirm that we can understand user behavior better when news analysis is separated by time rather than a fixed number of logs.

The rest of this paper is organized as follows. We introduce the related work in “Related works”. “Dataset” describes our dataset, and “Metrics” introduces our metrics to measure the diversity of users behavior. “Analyzing user behavior” and “Withdrawal analysis” present the details of our experiment and discuss its result. We conclude in “Conclusion”.

Related works

Our paper is related to two research areas: diversity analysis in recommender systems and popularity bias in recommender systems. We briefly review these two topics in this section.

Diversity in recommender systems

Bradley and Smith [6] proposed introducing diversity to recommender systems and referred to diversity by evaluating a new algorithm designed to diversify recommendations. Fleder and Hosanagar’s [11] experiments showed that many recommender systems reduce recommendation diversity by focusing on accuracy. Clarke et al. [8] proposed a new measure of the relevance of search documents by combining diversity and novelty. Adomavicius and Kwon [2] evaluated several item ranking-based methods and determined that many provided a variety of recommendations while maintaining comparable accuracy.

Since Pariser [19] pointed out that recommender systems effectively isolate people from a variety of viewpoints or content, many researchers have tested this hypothesis. Bakshy et al. [5] analyzed news articles on Facebook and argued that cross-cutting news (information originating from sources of opposing political perspectives) was not shared widely, revealing that recommender systems present news articles selectively on social media. Vargas et al. [22] focused on genre as a key attribute of diversity assessment and proposed a binomial framework for measuring the diversity of genres in each recommendation list. Ekstrand et al. [9] evaluated diversity as well as novelty, accuracy, satisfaction, and personalization and showed that satisfaction is negatively dependent on novelty and positively dependent on diversity and that satisfaction predicts users’ final choices. Javari and Jalili [14] introduced a probabilistic structure to solve the diversity and accuracy dilemma in recommender systems and proposed a hybrid model that can adjust for diversity and
accuracy. Nguyen et al. [18] proposed a method for measuring how recommender systems affect consumed content diversity over time and investigated the effect of collaborative filtering on MovieLens data.\(^1\) In their study, people who followed recommender systems consumed more diverse content than those who did not, although both groups experienced less diversity over time. Kakiuchi et al. [15] analyzed the environmental and behavioral characteristics of consuming various content. However, these studies did not discuss the influence of diversity on user engagement with the services. To the best of our knowledge, the only study describing the relationship of service engagement and diversity was published in April 2020 and was conducted using user behavior logs from Spotify, a popular music-streaming service [3]. Our work is the first attempt to analyze recommender systems in the news domain and identify the difference between the music and news domains.

**Popularity bias**

Many recommender systems suffer from popularity bias: popular items being recommended frequently, and less popular items being rarely recommended or not at all. Various researchers have discussed this bias in recommender systems, such as Treviranus and Hockema [21], who pointed out that recommender systems promote group polarization such as echo chambers [20]. Fleder and Hosanagar [10] discussed the impact of recommender systems on diversity and argued that some recommendations led to a net decrease in diversity and increase in personal diversity. Hosanagar et al. [13] found that iTunes sales diversity decreased with users purchasing more common items due to recommender systems.

Against this background, some studies have considered addressing popularity bias. Abdollahpouri et al. [1] introduced a personalized diversification re-ranking approach to represent less popular items in recommendations better while maintaining recommendation accuracy. Other researchers have advised avoiding the recommendation of too many similar items [7, 23, 24]. However, a quantitative assessment of how popularity bias affects user behavior has not yet been conducted. Our research adds to this body of work by analyzing the relationship between popularity bias and user engagement through the use of news-application activity logs.

**Dataset**

We analyzed user activity logs from a mobile news application provided by Gunosy Inc. covering an 8-month period from 2018 to 2019\(^2\). This provider offers news applications in Japan and has seen more than 40 million application downloads.

The application provides a list of news content for the user upon launch, presenting a title and thumbnail image for each news article. Upon clicking news of interest,

\(^1\) [http://www.movielen.org/](http://www.movielen.org/).

\(^2\) The exact date is not disclosed due to the existence of an agreement with Gunosy Inc, the data provider, to keep the data private.
the user can view the body of the news. The news list is personalized for each user based on the user’s click history. Users can also search for news by keywords, but most select news from the list. In this paper, we analyze user diversity with these click histories and determine user engagement with application launch histories to illustrate the relationship between diversity and user engagement.

Application launch histories were constructed of columns listing user identifications (IDs) and launch timestamps. News article click histories were constructed of user identifications IDs, article IDs, and timestamps for when users clicked on articles. We also employed master user data to connect user IDs and user registration timestamps and master article data, which included article titles. These data were separated from personal information and pseudonymized.

We extracted users who joined during a specific period of time in 2018–2019 for analysis. To exclude outlying users and align the conditions, we used the following rules to extract the users. First, we removed users with no clicks and users with clicks in the top 1% or more. Second, we removed users who had been using the service for less than 7 weeks.

**Metrics**

Diversity is an essential indicator of user experience, and good diversity metrics are expected to explain user behavior well. We assume that diversity has other aspects and propose the concept of popularity diversity, a state in which users only read popular articles as a state of reduced popularity diversity. We also propose a metric for measuring popularity diversity.

**Content diversity and average pairwise distance**

Content diversity is a criterion of how item content differs and is often used in recommender systems studies that involve algorithms and user experience [16, 18].

In this study, we analyze diversity based on the average pairwise distance (APD), a general method for measuring the diversity of a content list [25]. APD is defined as

$$D = \frac{\sum_{c_i \in L} \sum_{c_j \in L, c_j \neq c_i} d(c_i, c_j)}{|L|(|L| - 1)},$$

where $L$ is the item list and $c_i$ represents the vectors of each item, created from the titles of news articles using [term frequency]–[inverse document frequency] [4] weighted word2vec [17]. In this study, we use vectors of titles, because we evaluate user behavior with and without clicks, and only the title is available at the time of the click. To measure the similarity of two items, we compute the Euclidean distance between two item vectors.

A user with a high APD will consume a variety of items, whereas users with low APD will consume similar items. Within filter bubbles, a recommender system
isolates users from information that does not match their views, limiting information to a range of user interests. For example, the Wall Street Journal\(^3\) has shown the impact of obscuring political dissent around Brexit and the election of United States President Donald Trump [12]. The information that supporters of each party receive is dominated by content and opinions that are convenient for each party, which reduces their chances of being aware of the existence of opposing views. Lower user APD could indicate a more significant influence of a recommender system over that user. If filter bubbles limit the exposure of users to information that matches their interests, the items will be concentrated in a specific title, thus reducing APD.

**Popularity diversity and percentage of popular items**

When considering new diversity aspects, we focus on item popularity. For example, if users consume across various content while consuming only popular items from a news service, such behavior decreases diversity in terms of item popularity.

We introduce the concept of *popularity diversity* and define it concretely based on how much an item is consumed, clicked on, or purchased. To measure this popularity diversity, we propose a metric called the percentage of popular items (PPI). The idea of the PPI is simple, and it is a measure of how many popular items are included in the set of items consumed by a user. The PPI is calculated as

\[
PPI(u, r) = \frac{n_r}{N_u},
\]

where \(n_r\) stands for the number of popular items consumed by a user when the top \(r\%\) of items consumed are defined as popular items, and \(N_u\) stands for the number of all items consumed by a user \(u\).

A user with a high PPI is a user who consumes only popular items, which means that they mainly consume items that are presented at the top of the screen of the service. Users who behave in this way have low diversity of behavior, or diversity of popularity, in terms of whether the items they consume are biased toward only popular items.

**Analyzing user behavior**

We analyze how these two diversity metrics relate to user behavior and identify the meaning and relationship of each metric. We begin by clarifying the relationship between each diversity metric and users’ clicking activity. The number of clicks is an indicator of user satisfaction and suggests the relationships of user experience with our diversity metrics. We then check for diversity stability and determine correlative relationships between both diversity metrics.

\(^3\) [https://graphics.wsj.com/blue-feed-red-feed/](https://graphics.wsj.com/blue-feed-red-feed/)
We compute APD and PPI weekly. Because these metrics are computed for a set of items, it is necessary to cover items consumed over a certain period. We also analyze the time-series change, so we divide the calculation into 1-week segments.

We examine how to choose a period over which to measure diversity in Appendix A. Although existing studies have analyzed the changes in user diversities using blocks divided by certain numbers of clicks [15, 18], we confirm that we can better capture diversity changes if we analyze diversity on a weekly basis.

**Diversity and activity**

We begin by analyzing the relationship between content diversity and the number of clicks per user. We define the number of clicks as the level of activity. Figure 1 shows the distribution of average weekly APD over 7 weeks following the start of use for levels of activity higher than 20%, which are excluded for minimal clicks. This APD is the average of weekly APD for the first 7 weeks after the start of use. The distribution of APD increases as the amount of activity increases, confirming that increased activity produces greater content diversity. However, even for small amounts of activity, some users possess high APD, indicating that content diversity cannot be explained by activity alone. This result differs from the results of the musical domain [3]. We must consider the differences in services when considering diversity.

We also analyze the relationship between PPI and click numbers. Figures 2 and 3 show the distribution of PPI for every level of activity beyond 20% (using the same exclusion criterion used for the APD calculation), illustrating that PPI distribution decreases as activity increases. In other words, higher activity over a smaller percentage of popular articles read indicates greater popularity diversity. In contrast, content diversity shows little difference for each action, even with users possessing significant PPI, showing that popularity diversity also cannot be explained simply by activity. Although both PPI(0.5) and PPI(0.1) share this tendency, the distribution of PPI(0.5) is broader and more useful for indicating user behavior, so we apply it as the PPI in “Diversity stability” and “Withdrawal analysis”.

---

**Fig. 1** Distribution of APD for each level of activity, which increases as the amount of activity increases, confirming that increased activity produces greater content diversity. However, even for small amounts of activity, some users possess high APD, indicating that content diversity cannot be explained by activity alone.
We find that users with high activity levels have high content diversity and popularity diversity, confirming that user activity is related to both diversity metrics. PPI(0.5) and PPI(0.1) show the same trend, indicating that users with higher activity levels have higher popularity diversity.

**Diversity stability**

If APD and the PPI are stable over time, they can be considered inherent and used to analyze user behavior. As we saw in “Diversity and activity”, the PPI(0.5) and PPI(0.1) share the same tendency, indicating that users with high activity levels have high popularity diversity. The PPI(0.5) has a distribution with a large difference, so we treat the PPI(0.5) as the PPI in this section. We sample users who participated in the service for more than 14 weeks and then calculate the average of five-week metrics from 2–6 weeks after start of use and the average of five-week metrics from 10–14 weeks after start of use before comparing each period. Figures 4 and 5 display how the metrics of individual users change over the one-month period we consider. More specifically, they show user distribution for each period, drawn by random sampling to reduce computation times. Users taking the value of the metrics are
indicated as density approaches (yellow). Because there are many users with APD values between 1.3 and 1.8 in both periods and there are many users on the line $y = x$ for the PPI, many users have the same metrics values for the two periods, so the metrics values of users are stable over the period. The Pearson coefficient between the period of each metric exhibits moderate correlation, with a value of 0.551 for APD and 0.628 for the PPI, indicating broadly similar diversity for all user behavior.

Because it remains unclear how many users maintain the trend from a correlation analysis alone, we also divide users by the median value for each period and each indicator, aggregating whether user metrics maintain their sizes. For APD, 42.0% of users remain high, 30.6% remain low, 13.9% move to the low group, and 13.5% move to the high group. For the PPI, 37.4% of users remain high, 36.0% remain low, 13.4% move to the low group, and 13.3% move to the high group. Thus, most users maintain their trends over time.
We also find that the tendency to be high or low does not change significantly, confirming that content diversity and popularity diversity are stable. This result is consistent with the findings of Anderson et al. [3]. According to the above, these metrics can be used to represent the nature of a user in a long-term analysis.

**Withdrawal analysis**

This section compares diversity metrics with consecutive week and withdrawal rates. If diversity is an essential indicator for users, there must be some link between diversity metrics and continuity of services. The withdrawal rate is an indicator of user engagement and related to profitability, but understanding what behaviors encourage user withdrawal from a service is challenging. We analyze how click rates, APD, and the PPI are related to withdrawal to develop better recommender systems.

**Definition of withdrawal**

We define withdrawal as an absence of use lasting longer than four consecutive weeks. In “Diversity, activity, and withdrawal”, users who did not continue using the service after the 12th week were considered withdrawn, and other users were considered as continuing to use the service.

**Continuation week and diversity**

We examine each metric value for each week by dividing the number of users by the week of withdrawal to reveal the relationship between the continuation week and each metric. If metric values and change trends differ for each continuation week, that indicates user engagement and provides useful insights for discussing the decreasing diversity that allegedly causes social issues, such as filter bubbles and echo chambers.

Figures 6 and 7 show the values of APD and the PPI for each continuation week. A longer week duration indicates greater APD values and lower PPI values. Additionally, users who continue for more than 12 weeks show rising APD values and unchanging PPI values throughout the period; in other words, content diversity increases and popularity diversity remains unchanged over time from initial use for continuing users. However, users who leave after a short period show declining APD values and rising PPI values. We combine these metrics in the following analysis to show how APD and the PPI are related to withdrawal.
To clarify the relationships between service withdrawal, content diversity, popularity diversity, and activity level, we divide the users into two groups based on whether each metric value is high or low, regardless of the number of clicks at first. We then group users by their click histories in ascending order into 10-percentile groups (e.g., users in the 90th-100th percentiles are a group) and divide the users in each group into four quadrants based on APD and the PPI. The quadrants are denoted as “PPI: High, APD: High,” “PPI: High, APD: Low,” “PPI: Low, APD: High,” and “PPI: Low, APD: Low.” Each metric value is averaged over the metrics.

**Fig. 6** APD for each continuation week. A longer week duration indicates greater APD values and users who continue for more than 12 weeks show rising APD values throughout the period. The last week has a smaller APD value, which may be related to fewer clicks due to withdrawal.

**Fig. 7** PPI for each continuation week. A Longer week duration indicates lower PPI values and users who continue for more than 12 weeks show unchanging PPI values throughout the period. The last week has a higher PPI value, which may be related to fewer clicks due to withdrawal.

**Diversity, activity, and withdrawal**

To clarify the relationships between service withdrawal, content diversity, popularity diversity, and activity level, we divide the users into two groups based on whether each metric value is high or low, regardless of the number of clicks at first. We then group users by their click histories in ascending order into 10-percentile groups (e.g., users in the 90th-100th percentiles are a group) and divide the users in each group into four quadrants based on APD and the PPI. The quadrants are denoted as “PPI: High, APD: High,” “PPI: High, APD: Low,” “PPI: Low, APD: High,” and “PPI: Low, APD: Low.” Each metric value is averaged over the metrics.
by calculating the metrics for each week of the first 7 weeks after start of use and excluding weeks with no clicks.

Based on the number of clicks, we define the APD and PPI thresholds by the medians of both metrics for each user group. We do not use the medians for all users, as each indicator is related to the number of clicks, as mentioned in “Diversity and activity”. We confirm that the numbers of users in the four quadrants of both metrics are almost equal in each click-number group because they are divided based on the median value of each user group. By looking at the withdrawal rate of a group of users divided in this manner, we see how these indicators affect user engagement.

**Results**

We analyzed how the two diversity metrics proposed in “Metrics” relate to user behavior and identified the meaning and relationship of each metric. When we analyzed each metric, we found that users with low content diversity had a 224% greater withdrawal rate than those with high content diversity, and users with low popularity diversity had a 112% greater withdrawal rate than those with high popularity diversity.

We then compared diversity metrics with withdrawal rates and analyzed how the number of clicks, APD, and PPI affected the withdrawal rate to develop better recommender systems. Figure 8 shows the withdrawal rate of four quadrants. The overall average standardizes withdrawal rates. The overall trend was that the higher the number of clicks, the lower the withdrawal rate. For nearly all the clicks, if the PPI was low, including users who clicked articles that were not popular, the withdrawal rate was high. Furthermore, when the number of clicks was sufficiently high, the differences in withdrawal rates between the four quadrants were small.

The higher the PPI, the lower the withdrawal rate, regardless of the number of clicks and APD; when the number of clicks was high, however, the withdrawal rate was low, even though the PPI was low. In contrast, when the number of clicks was small, the APD and withdrawal rate were lower; when the number of clicks was sufficient, however, the APD was higher and the withdrawal rate was lower. When the number of clicks was sufficiently high, the content diversity and the satisfaction

---

**Fig. 8** Withdrawal rate by number of clicks. The overall average standardizes withdrawal rates. The overall trend is that the higher the number of clicks, the lower the withdrawal rate.
were higher. Because these users used the application more often and had more time, they were more likely to be satisfied with the variety of available information.

**Withdrawal prediction**

We predicted user withdrawal using these metrics as features to clarify the relationships among withdrawal, content diversity, popularity diversity, and number of clicks. If improving prediction accuracy is possible using these features, they have significance as variables that characterize user behavior.

As seen in “Dataset”, we removed users who had not been using the service for 7 weeks, and we used the log up to the 7th week as features for the classifier. In the model, we defined withdrawal as not using the service for more than four consecutive weeks, as in “Definition of withdrawal”. As a reference week, the number of weeks was varied from 8 to 12th. Three quarters of the data were randomly sampled as training data and one quarter as test data. We also downsampled the number of continuing users to match the number of users who withdraw when we trained the model, and we applied the model to the original percentage of users for inference.

First, we compared logistic regression, random forest, and multilayer perceptron as classifiers. We considered users who did not continue using the service after the 8th week to have left and all others to have continued, and we trained three classifiers using the number of clicks, PPI, and APD as features. We measured the prediction results using the F value and the area under the curve (AUC) of the precision-recall (PR) curve, because the data had disproportionately few users who withdrew and many who continued. Table 1 shows the prediction results for each classifier. It is more accurate to use multilayer perceptron than other classifiers. In the following, we use multilayer perceptron to compare features and to define withdrawal.

Table 2 shows the prediction results for each feature. As a reference week, the number of weeks was varied from 8 to 12th. Because the data were disproportional, F values varied from week to week, and AUCs were consistently less accurate as the weeks increased. This shows that it is difficult to predict the future with user logs

|                | F value | AUC  |
|----------------|---------|------|
| Logistic regression | 0.630   | 0.909|
| Random forest    | 0.609   | 0.907|
| Multilayer perceptron | 0.677   | 0.913|

---

4 We used these classifiers provided by scikit-learn (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html) with the default parameters, because parameter search made almost no difference.

5 We used a scikit-learn function (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html) to calculate the AUC-PR.
up to the 7th week. It is more accurate to use the PPI or APD than just the number of clicks. From the results, the PPI and APD seemed able to improve at capturing user behaviors. Figure 8 shows the withdrawal rate changes with each APD, PPI and number of clicks. Although more accuracy would be expected with these three features, there was no significant difference in prediction accuracy. As a future issue, it is necessary to create a model that takes into account the interaction between features.

Discussion

In “Continuation week and diversity”, we confirmed that continuing users have not only high APD and PPI values but also increasing diversity metrics as the week progresses. This shows that increasing diversity is beneficial not only in terms of solving social problems related to diversity and recommender systems but also in terms of increasing revenue.

We then divided the users by the metric sizes for each user activity level and found that users with high content diversity and popularity diversity were less likely to withdraw, but if users had the same activity level, those with lower popularity diversity were less likely to withdraw. This means that with the same activity level, reading popular articles made users more satisfied. This result could be an essential indicator for thinking about filter bubbles and other social issues related to diversity and is also useful for considering service management.

We also predicted withdrawal in “Withdrawal prediction”. When comparing withdrawal thresholds from the 8 to 12th week and features, in all conditions, using APD and the PPI was more accurate than using clicks alone. Thus, we confirmed that APD and the PPI are related to withdrawal. On the other hand, there was no difference between APD and the PPI, and using both APD and the PPI did not contribute to improving the accuracy. This might be due to predictions that ignore complex relationships between features, such as the presence of users with low activity levels.

| Table 2 | Evaluating withdrawal predictions for multilayer perceptron |
|---------|-----------------------------------------------|
| Features | 8th week | 9th week | 10th week | 11th week | 12th week |
| AUC-PR   |          |          |          |          |          |
| Clicks   | 0.902    | 0.870    | 0.837    | 0.837    | 0.786    |
| Clicks, APD | 0.914    | 0.899    | 0.857    | 0.848    | 0.817    |
| Clicks, PPI | 0.916    | 0.898    | 0.855    | 0.847    | 0.832    |
| Clicks, APD, PPI | 0.926    | 0.886    | 0.868    | 0.858    | 0.823    |
| F value  |          |          |          |          |          |
| Clicks   | 0.526    | 0.452    | 0.381    | 0.630    | 0.516    |
| Clicks, APD | 0.702    | 0.744    | 0.583    | 0.694    | 0.674    |
| Clicks, PPI | 0.719    | 0.738    | 0.572    | 0.691    | 0.738    |
| Clicks, APD, PPI | 0.826    | 0.635    | 0.673    | 0.749    | 0.702    |
but high diversity indicators, as identified in “Analyzing user behavior”. Using these factors for predictions is a task for future research.

**Conclusion**

Many studies have analyzed content diversity; however, we proposed adding popularity diversity as a diversity indicator. We analyzed content diversity and popularity diversity using user behavior logs, compared both diversity metrics with user activity levels, and found that they reflect different characteristics: users with high content diversity and high popularity diversity were more likely to click on a variety of articles than all users. We also analyzed the stability of both diversities and confirmed that most users maintain the trend over time. We confirmed that both metrics were inherent, so we can use them to analyze user behavior. We also analyzed the correlation of both diversity metrics, and their combination conveys the characteristics of user behavior.

Furthermore, we investigated how diversity related to user engagement, analyzing the relationship between diversity and withdrawal rate. We found that continuing users had greater content diversity and popularity diversity, and their diversities increased over time. However, users with the same activity level seemed satisfied and continued using the application when they clicked on popular articles. Although many studies have examined changes in diversity in recommender systems, we noticed this trend is affected by user withdrawal. This result could be an essential indicator for thinking about filter bubbles and other social issues related to diversity and is also useful for considering service management. We then developed a machine-learning model to predict user withdrawal using our two diversity metrics as features. These diversity metrics effectively predict user withdrawal. Although these results depend on specific service logs, we expect popularity diversity and content diversity to be useful metrics for various services using recommender systems.

For future work, we plan to develop a novel news recommender system using this knowledge about diversity and user engagement and consider other diversities that might be useful in explaining user behavior. For example, whether a news article is useful could be an essential factor in user behavior. If such diversities are analyzed sufficiently, we can better assess various issues related to recommender systems and deal with them appropriately to avoid fragmentation and polarization in society.

**A. Block analysis**

In this study, we show the relationship between diversity and user engagement by analyzing the weekly changes in diversity. We find that although the diversities depend on activity levels, the number of continuing users does not significantly decrease over time. This result is different from Nguyen’s findings [18]. Existing studies have analyzed the changes in user diversities using blocks divided by certain numbers of clicks [15, 18]. We call our method “weekly-based analysis” and this other method “block-based analysis.” In this section, we analyze the relationship
between diversity and user engagement using block-based analysis and discuss its differences from weekly based analysis.

A.1 Definition of block-based analysis

First, we describe block-based analysis in detail. We divide the click logs into blocks by a certain number of clicks and evaluate the diversity of each block based on Nguyen et al. [18]. This is expected to enable analyzing only the history of the click log independent of the frequency of application use. The number of clicks per week varies greatly depending on each user’s frequency of use, and even for the same user, the number of clicks might increase or decrease due to differences and changes in log-in days and times. However, it could be more susceptible to changes in daily news trends. We compare these features of different analysis methods to capture the characteristics of diversity in the news medium.

We first remove the first ten click logs for each user, as the first log after installation of the application. The period is presumably a period of learning how to use the application, not necessarily indicating interest in a particular article. After that, we divide the clicks into blocks and evaluate them. The basic analysis is based on the definition of 10 clicks as a block; when examining the impact of the number of clicks in a block, the analysis of 20 clicks as a block is performed simultaneously. To limit the number of users who have enough logs to analyze, we select users with more than three blocks for analysis, and logs that are divided into blocks and are a fraction of the last clicks are removed from the analysis target.

A.2 Block series analysis

We evaluate the change in the APD value on a block-by-block basis. As confirmed in “Withdrawal analysis”, the content diversity of continuing users increases while the diversity of withdrawal users decreases. We first divide users by the number of blocks included in the period of analysis, regardless of service withdrawal, to see how the APD value changes.

Figures 9 and 10 show the APD value per block when the number of clicks included in a block is defined as 10 and 20, respectively. In both definitions, these results show a decrease in APD, and users with many blocks, that is, many clicks, have a low initial APD and a small decrease, whereas users with few blocks have a high initial APD and a large decrease in APD. This shows a difference in initial usage between users who use the service more frequently and those who use it less frequently. However, as it stands, the impact of service withdrawal is not clear.

Therefore, we also analyze the relationship between APD change and service withdrawal in this block-based analysis. Figure 11 shows the APD for each block of continuing users. Unlike the continuation week analysis, we see a decrease in APD values. Figure 12 shows the APD for each block of withdrawal users. For withdrawal users, block-based analysis shows a large change in diversity. The reason for these results is the characteristics of the particular news medium.
A.3 Discussion

The block-based analysis indicates a decrease in diversity, but the magnitude of the decrease is in the order of 1/100. This decrease is much smaller than the decrease of withdrawing users in week-based analysis. Although the meaning is unclear, because there is no absolute measure of diversity, the reduction in diversity shown in block-based analysis is negligible.
This result might be influenced by the characteristics of the news medium. News content is affected by changes in society and changes in the content covered on a daily basis. When a significant news event occurs, the relevant content remains in the spotlight for several days. We assume that block-based analysis is affected by these news trends. Therefore, in analyzing news, it is better to analyze the news by dividing the range by the time of the week or other time units rather than blocks.

Of course, it is important to analyze the differences between week- and block-based analysis. Presumably, the content diversity of continuous users does not decrease week by week, whereas block by block, diversity decreases. In this application, news in the same category is grouped into tabs, and related articles are recommended for each news article. This facilitates selection of similar articles in the session, and the RS is suitable for the user, which makes it easier to continue the service. However, to prove this hypothesis, we must perform the analysis by overlapping the blocks like moving averages and varying the width of the definition of blocks more. The relationship between the fit of the recommender system and the withdrawal rate must also be examined in more detail. These issues must be addressed in the future.

Data availability The data that support the findings of this study are available from Gunosy Inc. but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of Gunosy Inc.

Declarations

Conflict of interest On behalf of all authors, Atom Sonoda states that there are no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long
as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abdollahpouri, H., Burke, R., & Mobasher, B. (2019). Managing popularity bias in recommender systems with personalized re-ranking. In: 32nd International Flairs Conference
2. Adomavicius, G., & Kwon, Y. (2011). Improving aggregate recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering, 24(5), 896–911.
3. Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic effects on the diversity of consumption on spotify. In: Proceedings of The Web Conference 2020, pp. 2155–2165
4. Baiza-Yates, R., Ribeiro-Neto, B., et al. (1999). Modern information retrieval (Vol. 463). ACM press.
5. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130–1132.
6. Bradley, K., & Smyth, B. (2001). Improving recommendation diversity. In: Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland, pp. 85–94. Citeseer
7. Castells, P., Vargas, S., & Wang, J. (2011). Novelty and diversity metrics for recommender systems: choice, discovery and relevance. In: International Workshop on Diversity in Document Retrieval at the ECIR 2011: the 33rd European Conference on Information Retrieval.
8. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 659–666
9. Ekstrand, M.D., Harper, F.M., Willemsen, M.C., & Konstan, J.A. (2014). User perception of differences in recommender algorithms. In: Proceedings of the 8th ACM Conference on Recommender systems, pp. 161–168
10. Fleder, D., & Hosanagar, K. (2009). Blockbuster culture’s next rise or fall: The impact of recommender systems on sales diversity. Management Science, 55(5), 697–712.
11. Fleder, D.M., & Hosanagar, K.: (2007). Recommender systems and their impact on sales diversity. In: Proceedings of the 8th ACM conference on Electronic commerce, pp. 192–199
12. Hooton, C. (2016). Social media echo chambers gifted Donald Trump the presidency. Independent
13. Hosanagar, K., Fleder, D., Lee, D., & Buja, A. (2014). Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. Management Science, 60(4), 805–823.
14. Javari, A., & Jalili, M. (2015). A probabilistic model to resolve diversity-accuracy challenge of recommendation systems. Knowledge and Information Systems, 44(3), 609–627.
15. Kakiuchi, K., Toriumi, F., Takano, M., Wada, K., & Fukuda, I. (2018). Influence of selective exposure to viewing contents diversity. The 10th International Conference on Social Informatics
16. Konstan, J. A., & Riedl, J. (2012). Recommender systems: From algorithms to user experience. User Modeling and User-Adapted Interaction, 22(1–2), 101–123.
17. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
18. Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., & Konstan, J.A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 677–686. ACM
19. Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin.
20. Sunstein, C.R. (1999). The law of group polarization. University of Chicago Law School, John M. Olin Law & Economics Working Paper 0(91)

21. Treviranus, J., & Hockema, S. (2009). The value of the unpopular: Counteracting the popularity echo-chamber on the web. In: 2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH), pp. 603–608. IEEE

22. Vargas, S., Baltrunas, L., Karatzoglou, A., & Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: Proceedings of the 8th ACM Conference on Recommender Systems, RecSys ’14, p. 209-216. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2645710.2645743.

23. Zhang, M., & Hurley, N. (2008). Avoiding monotony: improving the diversity of recommendation lists. In: Proceedings of the 2008 ACM conference on Recommender systems, pp. 123–130

24. Zhou, T., Kuscsik, Z., Liu, J. G., Medo, M., Wakeling, J. R., & Zhang, Y. C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences, 107(10), 4511–4515.

25. Ziegler, C.N., McNee, S.M., Konstan, J.A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web, pp. 22–32

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.