In Search of User Features for Identifying Different Inspection Behaviors on Recommended Items

Kibeom Lee
Graduate School of Convergence Science and Technology
Seoul National University
Seoul, Korea
kiblee@snu.ac.kr

Sangmin Lee
Graduate School of Convergence Science and Technology
Seoul National University
Seoul, Korea
pizzicato@snu.ac.kr

Kyogu Lee
Graduate School of Convergence Science and Technology
Seoul National University
Seoul, Korea
kglee@snu.ac.kr

ABSTRACT

In recent years, research in recommender systems have began focusing on other elements of recommender systems besides accuracy, such as novelty, diversity, and serendipity. Naturally, research in these areas concentrate on providing novel and relevant recommendations. However, when presented with such recommendations, it is important that users actually inspect the unknown, novel items. Encouraging users to inspect such items can be achieved through the system itself, such as building trust or using previews and explaining recommendations.

However, in this study, we analyze the users, rather than the system, to find features that are good indicators of the users' behaviors towards novel items. In order to achieve this, we carry out a user study and observe the user's interactions with the recommendations. These users are divided into different groups depending on their reactions to recommended items: Explorers and Indifferents. We then search for features in user profiles that can help distinguish the differences between the two groups.

Based on the results of independent samples t-tests, we propose that artist, genre, and tag diversity, in addition to widely-distributed listening behaviors across artists, are features that show significant differences in mean values between Explorers and Indifferents. We believe that these features can be utilized to add another layer of personalization to existing recommenders to adjust the novelty of recommendations according to the target user's group affiliation.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factors; H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Experimentation

KDD 2015 Workshop on Interactive Data Exploration and Analytics (IDEA'15) August 10th, 2015, Sydney, Australia.

Keywords

user features, novelty, recommender systems, recommendation interaction, recommendation inspection

1. INTRODUCTION

Recommender systems have been an extremely active field of research, with its importance growing in recent years due to the rapid advancements in technology and massive amounts of data available. Research on recommender systems first emerged in the 1990s, with the introduction of collaborative filtering [16, 18]. Throughout the years, other methods of recommender systems, such as content-based recommenders [9,14,15] and hybrid recommenders [2,3,6,17] were also proposed.

Until recently, the majority of research and development efforts on recommender systems were focused on accuracy: trying to predict the users' ratings on items. Algorithm rankings in competitions such as the Netflix Prize [4] and the KDD Cup were also based on accuracy metrics. In recent years, research in recommender systems that go beyond accuracy have emerged, stemming from findings that user satisfaction and recommender accuracy are not always correlated [12, 25]. This led to research in various aspects of recommender systems, such as increasing the diversity, novelty, and serendipity of the recommended items [7].

The various research on diversity and novelty have an inherent agreement that users will actually examine the provided novel recommendations. Thus, it is accepted that motivating users to inspect recommendations can be attained through various facets of recommender systems, such as trust, transparency, etc. However, in this paper, rather than focusing on the above facets of recommender systems from a technical viewpoint, we study the users themselves based on the interactions with recommended items. By doing so, our goal is to search for user features that can be used to differentiate two types of contrasting users: the type that proactively samples unknown recommended items, and the type that shows no interest in such recommendations.

2. RELATED WORK

Project Phoenix, a study carried out by media company Emap, surveyed 2,200 15-39 year olds about their music listening habits. The people were then divided into four tiers of interest in music: indifferents, casuals, enthusiasts, and savants [8]. Based on these groups, Celma suggested that

each of these groups would require different types of recommendations [5]. Besides Project Phoenix, there were no reported studies on differentiating groups of users according to their behavior or attitudes toward novel recommendation items to our knowledge.

Various recommenders were proposed that aimed to increase diversity and novelty of the recommended items [1,10, 11,21–23]. These studies identified the problems with previous recommender systems as only focusing on the accuracy recommendations. Each paper presented its own method to tackle this problem. However, while it is widely accepted that there is more than accuracy when providing recommendations, it is difficult to find research on what kinds of users inspect such recommendations. In this paper, we attempt to identify users who show interest in novel recommendations and those who are indifferent to such recommendations.

Besides studies on algorithms, there have also been research that found different aspects of recommender systems that were correlated to user satisfaction. Novelty, diversity, serendipity, trust, transparency, and social factors were some of the aspects that influenced user satisfaction of the recommender system [5, 7, 13, 19]. In particular, Swearingen and Sinha explored the design elements of recommender systems that enabled the system to introduce users to novel items and convince them to view them [19]. Based on the results of their user study, they suggested that different recommender systems would be needed to satisfy the needs of different users. Their proposed solution to this was to let the users decide what recommendations they wanted or to explicitly ask the kind of recommendations they desired at the beginning of each session. In this paper, we search for user features that can indicate their attitudes toward novel items, thus removing the need of requesting explicit feedback.

3. USER STUDY

We designed a user study in order to divide users into groups depending on their behavior toward novel items in their recommendations. With our criteria, users would be categorized largely into two extreme groups: (1) those who showed interest in the novel items in their recommendation lists, and (2) those who showed no interest to the recommendations at all.

3.1 Design of User Study

The most important part of the design of the user study was to capture the most natural behaviors of the participants towards recommendations. To achieve this, we refrained from any explicit instructions on the user study and avoided encouraging users in inspecting their recommendations. Instead, on the welcoming page, we declared that this was a user study on recommender systems and that the participants would be provided a list of artists based on their Last.fm profiles, which they were free to explore as they wish. They were also notified that a simple question would be asked in the end.

The user study was largely divided into three stages, as illustrated in Figure 1. Once the participant input their Last.fm ID, we generated 10 recommendations that included artists who had a high probability of being unknown (novel) to the user. Details of generating the recommendation list is provided in the next section. Next, the participants viewed their recommended artists and could click the links to access more information and listen to the artists on their respective Last.fm artist pages. During this exploration stage, we tracked the number of click-throughs of each artist. The participants could end their browsing session by clicking a button, which brought them to the last stage of the user study. Here, they were represented with identical recommendation list and were requested to select all the artists that they were already familiar with. By doing so, we were informed of which recommended artists were actually novel to the user.

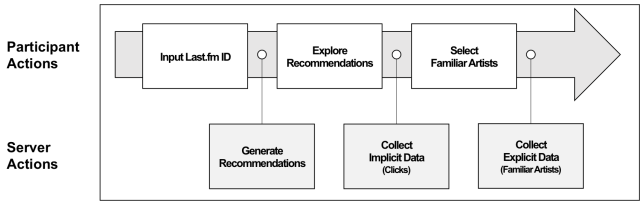


Figure 1: Process of the user study. Users input their Last.fm ID and receive 10 recommendations that include several novel items. While they explore the recommended items (if at all), we record the time spent on the user study webpage and any clicked artists. The only explicit information we require from them is to flag any familiar artists, which is collected after the exploration stage is complete.

3.2 Recommendations with Novel Items

3.2.1 Algorithm

The Myrrix[1] recommender system was used to generate the recommendations, which uses a variant of the ALS-WR algorithm [24]. The parameters used for the ALS-WR algorithm were the default $\lambda = 0.01$ and $\alpha = 40$.

The Myrrix recommender was trained with data gathered from Last.fm, which is discussed in detail below. From the recommendations provided by Myrrix, we took the top seven items as the extremely accurate items and took the 100th, 200th, and 300th items as the potentially novel items. By doing so, we aimed to bring novel items to the participants while keeping them moderately relevant instead of offering random, unknown artists. In total, a list of 10 recommendations were generated for each participant with seven items that were highly probable of being known to the participants and three items that had higher chances of being novel. The order in which the artists appeared were randomized.

We deliberately had the recommendations contain seven extremely accurate items in order to build trust on the recommender system, as indicated by several studies [19]. Thus, we aimed to provide trust in the recommender with the accurate items and observe the behavior towards the remaining novel items.

3.2.2 Data

The data used to train the Myrrix recommender was gathered with the Last.fm API. Listening data of 32,413 Last.fm users were gathered by querying their 50 most listened artists. The profiles of the 32,413 users covered 184,890 unique artists.

To train the Myrrix recommender with the collected Last.fm dataset, the playcount information was converted to ratings with a common rating scale of $[0, n]$ using the func-

[1]The Myrrix project has been discontinued as of December 31, 2013 and is currently part of the Oryx project

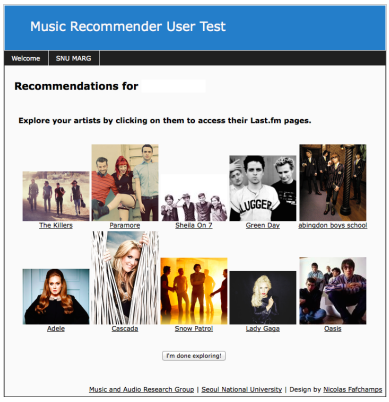


Figure 2: Screenshot of user study (username is erased) showing 10 recommendations for a user. The user is free to explore the artists by clicking on them to go to their respective Last.fm pages where they can read about the artists and listen to their songs.

tion $r(u,i) = n * F(\text{playcount}_{u,i})$, where $\text{playcount}_{u,i}$ is the playcount of user u on item i, and $F(\text{playcount}_{u,i})$ is the cumulative distribution function of $\text{playcount}_{u,i}$ defined by $|\{j \in \boldsymbol{u} | \text{playcount}_{u,j} \leq \text{playcount}_{u,i}\}|/|\boldsymbol{u}|$, using the items in u's profile - $\boldsymbol{u}$, as in [20].

To summarize, the training data for the recommender was based on 32,413 Last.fm user profiles. These profiles were converted to ratings, making up 2.9 million ratings.

4. GROUPING BASED ON SAMPLING BEHAVIOR

Participants for our user study were recruited through various outlets, such as Last.fm message boards, MIR mailing lists, and online communities. The only requirement was that they have a Last.fm ID. A total of 148 participants visited the user study, of which 110 participants actually took part in the user study. Among the 110 participants, we removed users who were not presented with any novel items in their recommendations using the feedback from the last stage of the user study, which left us with 92 participants. Lastly, we filtered out the bottom 10% of participants with the least cumulative playcounts, signifying developing profiles that were not yet 'mature'. In the end, our data was made up of 83 participants. A screenshot of the user study is shown in Figure 2

Based on the Last.fm personal profiles, we present demographics on the participants. Not all users had their profiles public, so the following statistics do not accurately represent all the participants. The average age was 24.94 with 11 women and 55 men (17 unknown) and while the nationalities were diverse, the most dominant nationalities were the U.S. (24) and the U.K. (13).

The goal of our experiment was to find features in user profiles that could help differentiate between users with opposing interests towards recommendations. Thus, we divided the participants into two groups - Explorers and Indifferents - based on their interactions with the recommended items. Explorers were users who showed interest in the novel (unknown) items and viewed additional information by clicking the artist links. In contrast, Indifferents were users who showed absolutely no interest in any of the recommended items.

Using the explicit feedback from the participants, we found that an average of 2.99 novel recommendations (out of 10 recommendations) were given to the 83 participants. The Explorers group was provided an average 3.12 novel recommendations and the Indifferents was given 2.97 recommendations.

Regarding Indifferents, we were aware that the lack of inspection, or interaction of items could be due to many factors. For apathetic behavior, in particular, it could be argued that users with such characteristics could wrongly belong in this group. However, we decided that these characteristics are natural attributes of the user. Thus, we believe that if such factors exist in the study, then they would also exist in the real-world, representing an accurate portrayal of real-world behavior.

5. USER FEATURES

In order to search for features that could work as good identifiers of these groups, we explored various facets of the data and defined several features that we predicted would differentiate Explorers and Indifferents. The search for features was done as a full exploration of features that could be acquired from the available data (music listening history, tags, personal profile, friends list, etc). Throughout the paper, we define the user's profile, denoted as $\boldsymbol{u}$, as the top 50 most listened artists from the user's library.

5.1 Features on Profile Diversity

Intuitively, we predict that users with diverse listening habits will likely be Explorers. Thus, we define features that measure diversity from four different perspectives, namely artists, genres, tags, and online friends.

5.1.1 Artist Diversity

We predict that artist diversity will be a good metric in distinguishing the two groups, as stated in Hypothesis 1.

HYPOTHESIS 1. *Explorers will have profiles with relatively higher artist diversity, while Indifferents will have profiles with relatively lower artist diversity.*

We measured the diversity of artists in a user profile by collecting the top 20 similar artists (obtained through the Last.fm API) for each artist in the profile and finding the ratio of unique artists to total artists [2], which we name Artist Diversity.

More formally, let $A = \{sim_{20}(x) | \forall x \in u\}$ be the multiset of top-20 similar artists of each artist in $\boldsymbol{u}$, and A_d be the set of distinct artists in A. Then, we define Artist Diversity (AD) as,

$$\text{AD} = \frac{|A_d|}{|A|} \quad (1)$$

5.1.2 Genre Diversity

Likewise, we predict that the two groups will have significant differences in genre diversity, as in Hypothesis 2.

HYPOTHESIS 2. *Explorers will have profiles with relatively higher genre diversity, while Indifferents will have profiles with relatively lower genre diversity.*

[2]Formula adapted from `http://anthony.liekens.net/pub/scripts/last.fm/supereclectic.php`.

Because Last.fm lacked genre metadata, we obtained the artists' genres using the Echonest API. The profiles of the participants spanned 922 genres, as Echonest assigns each artist with dozens of weighted genres. Using this data we formulated two varying methods of measuring genre diversity, which we labeled Genre Diversity and Genre-Space Uniformity.

To calculate Genre Diversity, we collect the associated genres with weights above a certain threshold for each artist in the user's profile $\boldsymbol{u}$. Genre Diversity is then given by the ration of unique genres to all genres.

Formally, let $G = \{\text{TopGenres}\,(x, w_{\text{genre}})|\forall x \in \boldsymbol{u}\}$ be the multiset of genres from each artist x in $\boldsymbol{u}$ with weights $\geq w_{\text{genre}}$, and G_d be the set of distinct genres in G. Then, we define Genre Diversity (GD) as

$$\text{GD} = \frac{|G_d|}{|G|} \tag{2}$$

As another way to measure genre diversity, we represent each artist in a user's profile as a vector in the genre space with the genre weight as entries. Thus, a user's profile creates an $M \times N$ matrix, where M is the number of artists in the user's profile and N is the number of genres.

Formally, let H be the set of all genres; $\text{GenreWeight}(\boldsymbol{u}_i, j)$ be the weight of genre j for artist i in $\boldsymbol{u}$; and S be an $M \times N$ matrix where $M = |\boldsymbol{u}|$, $N = |H|$, $S(i,j) = \text{GenreWeight}(\boldsymbol{u}_i, j)$, $Q = \{x|x \in 1 \cdot S, x > 0\}$ and $\mathbf{1}$ is a $1 \times M$ vector of all ones. Then, we define Genre-Space Uniformity (GSU) as,

$$\text{GSU} = \frac{\sum \mathbf{1} \cdot S}{|Q|} \tag{3}$$

5.1.3 Tag Diversity

Similarly, we anticipate that tag diversity will also be an effective method of differentiating Explorers and Indifferents, as we state in Hypothesis 3.

HYPOTHESIS 3. *Explorers will have profiles with relatively higher tag diversity, while Indifferents will have profiles with relatively lower tag diversity.*

Tag data for each artist was gathered using the Last.fm API, which returns tags weighted on a scale between (0...100]. Using these weighted tags, we calculate Tag Diversity the same way we did Genre Diversity.

Stated formally, let $T = \{\text{TopTags}\,(x, w_{\text{tag}})|\forall x \in \boldsymbol{u}\}$ be the multiset of tags from each artist x in $\boldsymbol{u}$ with weights $\geq w_{\text{tag}}$, and T_d be the set of distinct tags in T. Then, we get Tag Diversity (TD) with,

$$\text{TD} = \frac{|T_d|}{|T|} \tag{4}$$

5.1.4 Social Diversity

Regarding social diversity, we anticipate that Explorers will have friends with diverse listening habits while Indifferents will have friends with similar tastes, leading to less diversity overall, as stated in Hypothesis 4.

HYPOTHESIS 4. *Explorers will have social networks mainly comprised of friends with relatively differing musical tastes, while social networks of Indifferents will mainly be comprised of friends with relatively similar musical tastes.*

Social Diversity measures the average dissimilarity of musical taste between a user and his/her social network. This feature makes use of the Tasteometer metric in the Last.fm API, which measures the similarity between two users based on their profiles.

Let $\text{tasteometer}(u, v)$ be the similarity between user u and v, and let u_{friends} be the set of friends of u in Last.fm. Then, we calculate Social Diversity (SD) with,

$$\text{SD} = \frac{\sum \text{tasteometer}\,(u, v)}{min\,(|u_{\text{friends}}|\,, 50)}, \quad \forall v \in u_{\text{friends}} \tag{5}$$

5.2 Features on Listening Behavior

5.2.1 Profile Popularity

HYPOTHESIS 5. *Explorers will have profiles with relatively less popular artists, while Indifferents will have profiles comprised of relatively popular artists.*

We develop two methods of measuring the overall popularity of a user's profile. The first method uses the number of unique listeners of an artist on Last.fm as a quantitative measure of popularity, which we denote as $\text{pop}(x)$. Thus, we calculate Mean Profile Popularity (MPP),

$$\text{MPP} = \frac{\sum \text{pop}\,(x)}{|\boldsymbol{u}|}, \quad \forall x \in \boldsymbol{u} \tag{6}$$

The second method borrows the formula for calculating spectral centroids and applies it to playcounts, which we name Rank Centroid (RC). The formula is,

$$\text{RC} = \frac{\sum \text{pop}_{\text{rnk}}\,(x)\,|\text{plays}\,(u, x)|^2}{\sum |\text{plays}\,(x)|^2}, \quad \forall x \in \boldsymbol{u} \tag{7}$$

where $\text{pop}_{\text{rnk}}(x)$ is the global popularity rank of artist x, and $\text{plays}(u, x)$ is the playcount of artist x by user u.

5.2.2 Playcount Distribution

HYPOTHESIS 6. *Explorers will distribute their music listening to a relatively larger number of artists, while Indifferents will have skewed music listening towards a relatively smaller number of artists.*

We measure the distribution of playcounts in a user's playlist via two methods. The first method is done by sorting the artists in $\boldsymbol{u}$ in descending order by playcount. Using this distribution of playcounts across artists in a user's profile, we calculate Playcount Skewness (PS) by adapting the adjusted Fisher-Pearson standardized moment coefficient,

$$\text{PS} = \frac{n}{(n-1)\,(n-2)} \sum \left(\frac{\text{plays}(u, x) - \overline{\boldsymbol{u}}}{s} \right)^3, \quad \forall x \in \boldsymbol{u} \tag{8}$$

where $n = |\boldsymbol{u}|$, s is the sample standard deviation, and $\overline{\boldsymbol{u}}$ is the mean playcount of $\boldsymbol{u}$.

The second method, similar to RC, is based on calculating spectral spreads. We call this feature Rank Spread (RS) and define it as,

Table 1: Results of the independent samples t-tests performed on Explorers and Indifferents using the proposed features as variables. All tests were done at the 5% significance level. Tests that reject H0 are in bold.

	Mean (Std. Deviation)				
Feature	Explorers	Indifferents	*t*	*df*	*p*
Artist Diversity	0.74 (0.11)	0.69 (0.10)	2.03	61.84	**0.05**
Genre Diversity	0.18 (0.08)	0.20 (0.10)	-0.72	63.99	0.48
Genre-Space Uniformity	1.10 (0.20)	1.20 (0.22)	-2.09	63.70	**0.04**
Tag Diversity	0.46 (0.10)	0.40 (0.09)	2.71	62.47	**0.01**
Social Diversity	0.62 (0.22)	0.58 (0.19)	0.67	56.76	0.51
Mean Profile Popularity	8.73^{*} (3.94^{*})	7.33^{*} (4.78^{*})	1.31	64.33	0.19
Rank Centroid	$8.41^{\dagger}$ ($4.15^{\dagger}$)	$9.82^{\dagger}$ ($9.82^{\dagger}$)	-1.07	58.25	0.29
Playcount Skewness	2.69 (1.22)	3.01 (1.26)	-1.05	64.73	0.30
Rank Spread	$6.72^{\dagger}$ ($2.13^{\dagger}$)	$5.61^{\dagger}$ ($1.65^{\dagger}$)	2.37	58.24	**0.02**

$* : \times 10^5$, $\dagger : \times 10^2$

$$\text{RS} = \sqrt{\frac{\sum \left(\text{pop}_{\text{rnk}}(x) - \text{SC}_{\text{ArtistRank}}\right)^2 |\text{plays}\,(u,x)|^2}{\sum |\text{plays}\,(u,x)|^2}}, \quad \forall x \in \boldsymbol{u} \tag{9}$$

6. RESULTS & DISCUSSION

The distribution of the 83 participants were: 32 in Explorers and 35 in Indifferents. The remaining 16 participants did not fall into either group (i.e. they only inspected recommendations that they were familiar with). The act of inspecting only those artists that the participants were familiar with were not characteristics of Explorers nor Indifferents. Due to their vagueness, these participants were removed and the analysis was done on the two extreme groups.

In order to test the hypotheses, we performed an independent samples t-test on each feature comparing Explorers and Indifferents. The results of the tests are summarized in Table 1.

There were significant differences in mean values for Artist Diversity between Explorers and Indifferents, indicating that Explorers listen to a more diverse range of artists compared to Indifferents and supporting Hypothesis 1.

Genre Diversity was measured with $w_{\text{genre}} = 1.0$ and $w_{\text{genre}} \geq 0.9$ for TopGenres(x, w_{genre}). Results for $w_{\text{genre}} = 1.0$ are shown in Table 1. For $w_{\text{genre}} \geq 0.9$, the t-test also failed to reveal a statistically reliable difference between Explorers ($M = 0.19, SD = 0.07$) and Indifferents ($M = 0.19, SD = 0.08$); $t(64.91) = 0.1, p = 0.92$. On the other hand, mean values of Genre-Space Uniformity showed significant differences between Explorers and Indifferents. We believe that Genre Diversity fails to measure genre diversity accurately because of its misrepresentation of the real-world due to the lack of using genre weights. As can be seen in the formula for Genre Diversity, it does not take into account genre weights in the calculations but simply uses them as a threshold. Thus, all genres are treated equally regardless of weight, resulting in a limited method of expressing various user profiles via genres when artists are affiliated to different genres unequally. Therefore, by using Genre-Space Uniformity, we can support our assumptions in Hypothesis 2.

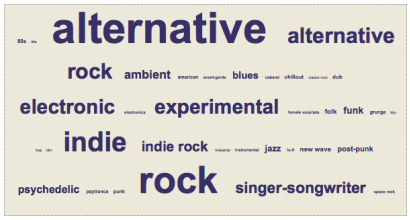
Results of Tag Diversity showed significant differences in mean values for the two groups. This feature was measured with $w = 100, 90, 80, 70, 60, 50$ for TopTags(x, w). Results of the t-tests for different w thresholds are shown in Table 2. In Last.fm, there is only one tag with maximum weight 100 assigned to each artist. Thus, for $w = 100$, each user is represented with tags that are equal in number with the number of artists in his/her profile, making it a conservative measure of diversity. As the threshold for w is lowered, tags that are less and less accurate begin to cloud the metric. The test fails for $w = 50$, where users are associated with an abundant amount of tags but are inaccurate. Such tags create too much noise in the data, making it difficult to extract meaningful interpretations. The t-test results support the idea that Explorers have higher tag diversity and Indifferents have lower tag diversity, as stated in Hypothesis 3. A real example of tag clouds of sample users from the two groups is shown in Figure 3.

Regarding Social Diversity, the two groups did not have any significant differences in mean values. In other words, friend relationships on the social network seem to be formed independent of similarities in musical tastes, contrary to what we predicted. This is interesting as Last.fm is also a social networking service centered on music and musical preferences. Here, we failed to find any supporting data for Hypothesis 4.

On profile popularity, results showed that both Mean Pro-

Table 2: Results of the independent samples t-tests using Tag Diversity as the variable with varying tag weight thresholds (w). All tests were done at the 5% significance level. Tests that reject H0 are in bold.

	Mean (Std. Deviation)				
w	Explorers	Indifferents	*t*	*df*	*p*
100	0.46 (0.10)	0.40 (0.09)	2.71	62.47	**0.01**
90	0.44 (0.09)	0.38 (0.08)	3.01	61.50	**0.00**
80	0.41 (0.09)	0.36 (0.07)	2.76	58.65	**0.01**
70	0.39 (0.09)	0.34 (0.06)	2.62	56.05	**0.01**
60	0.37 (0.09)	0.32 (0.06)	2.55	53.67	**0.01**
50	0.35 (0.09)	0.33 (0.07)	1.47	60.16	0.15



(a) Sample user from the Explorer group.

(b) Sample user from the Indifferents group.

Figure 3: Tag cloud of user's profiles. There is a perceivable difference in the variety of tags between a user from the Explorer group and a user from the Indifferents gruop. Tag cloud images generated from `http://anthony.liekens.net/pub/scripts/last.fm`

file Popularity and Rank Centroid failed to show significant differences in mean between Experts and Indifferents. We had anticipated that Explorers would be listening to long-tail artists and Indifferents would be concentrated towards popular artists. However, according to these results, looking at the popularity of artists is not an effective measure of classifying Explorers and Indifferents. Again, we were not able find supporting data for Hypothesis 5.

Lastly, t-test results on features measuring the listening distribution of users showed significant differences in mean for Rank Spread but failed for Playcount Skewness. We had predicted that Explorers would have relatively less skewed listening habits compared to Indifferents, resembling a balanced consumption of music. However, results do not support this assumption, which could be explained by the nature of how we consume music. Because songs are listened to multiple times, the formation of a power-law distribution in listening patterns may be inevitable when viewing user profiles that represent years of music consumption. Thus, it may be more meaningful to look at the skewness of listening patterns in time scales of a week or month, rather than overall.

Rank Spread, on the contrary, showed significant differences in mean between Explorers and Indifferents. Because the t-test failed for Rank Centroid, we assume that both groups have equal means in Rank Centroid but have significantly different means in Rank Spread. In other words, while both groups listened to similarly popular artists, the distribution of listening by Explorers were spread widely across other artists and the distribution of listening by Indifferents were less spread and more focused on a smaller range of artists, which is in agreement with Hypothesis 6.

7. CONCLUSION

There are numerous studies on increasing novelty and diversity in recommender systems. It is widely accepted that such research on recommenders are necessary as accuracy is simply one of many unknown factors that influence user satisfaction. Likewise, the act of inspecting recommendations, regardless of novelty, may depend on a range of factors, from various elements of the system such as trust, transparency, and user interface. In this paper, we suggest that besides the perspective of the system, there could be human factors that influence interactions with recommendations, which we believe would be embedded in user profiles. Thus, we ventured to find features in user profiles that could differentiate two extreme groups of users: Explorers, who were users that sampled unknown, novel items and Indifferents, who were users that refrained from inspecting any items.

Based on our experiments, the findings indicate that users who inspect unknown, novel items have certain characteristics in their user profiles that are indicators of their behavior.

When dividing the groups into Explorers and Indifferents, the features that distinguish those two groups seem to be Artist Diversity, Genre-Space Uniformity, Tag Diversity, and Rank Spread, which are in support of the Hypotheses 1, 2, 3, and 6.

By using the features proposed in this study, we believe that tailored recommender systems can emerge, in which the system generates different recommendations for users in Explorers and Indifferents groups. For instance, the system could generate more diverse and novel recommendations to users in the Explorers group at the cost of accuracy, while providing more conservative and accurate recommendations to users in the Indifferents group.

8. FUTURE WORK

The user study in this research was designed to be as unobtrusive as possible to the participants, because we wanted to capture their behavior that was the most representative of the real world. To do this, the participants were not explicitly instructed to click on recommendations, but were simply informed that they could through the hyperlinks. However, the implicit data that was collected through the user study may not be representing a user's true intentions. To overcome this problem, the user study could possibly have a post-study survey to record the participants' intentions.

In addition, as with all user studies, a larger sample size would have yielded a more reliable representation of the user population. Regarding features, we attempted to find a wide range of features that targeted different aspects of the user profiles. There were features that we anticipated would work but actually failed. With a larger sample size, there is a possibility that these features could be significant.

Although the presented study did have a rather limited sample size, we believe it does indicate potential features that can be used to predict user behavior towards novel recommended items. Nonetheless, it would be valuable to investigate whether it is indeed the case that user satisfaction can be improved by taking the user's propensity to explore into account. We believe that, with more robust studies regarding interactions on recommendations, this can lead to recommender systems that dynamically adjust its parameters to add another level of personalization for the user. This extra layer of personalization would decide, perhaps, the degree of novelty and diversity in the final recommendations. Such a system would result in a customized recommender for each user, which contrasts to existing recommender systems that use one-size-fits-all personalization algorithms to generate recommendations.

9. ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (490-20130014).

10. REFERENCES

[1] G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *Knowledge and Data Engineering, IEEE Transactions on*, 24(5):896–911, May 2012.

[2] M. Balabanović and Y. Shoham. Fab: Content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72, Mar. 1997.

[3] C. Basu, H. Hirsh, and W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 714–720. AAAI Press, 1998.

[4] J. Bennett, S. Lanning, and N. Netflix. The netflix prize. In *In KDD Cup and Workshop in conjunction with KDD*, 2007.

[5] O. Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[6] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR workshop on recommender systems*, volume 60. Citeseer, 1999.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, Jan. 2004.

[8] D. Jennings. *Net, blogs and rock'n'roll: how digital discovery works and what it means for consumers, creators and culture*. Nicholas Brealey Publishing, 2007.

[9] K. Lang. Newsweeder: Learning to filter netnews. In *in Proceedings of the 12th International Machine Learning Conference (ML95*, 1995.

[10] K. Lee and K. Lee. My head is your tail: Applying link analysis on long-tailed music listening behavior for music recommendation. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 213–220, New York, NY, USA, 2011. ACM.

[11] K. Lee and K. Lee. Using dynamically promoted experts for music recommendation. *Multimedia, IEEE Transactions on*, 16(5):1–10, 2014.

[12] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 116–125, New York, NY, USA, 2002. ACM.

[13] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1097–1101, New York, NY, USA, 2006. ACM.

[14] R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 195–204, New York, NY, USA, 2000. ACM.

[15] M. Pazzani and D. Billsus. Learning and revising user profiles: The identification ofinteresting web sites. *Mach. Learn.*, 27(3):313–331, June 1997.

[16] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[17] J. Salter and N. Antonopoulos. Cinemascreen recommender agent: Combining collaborative and content-based filtering. *IEEE Intelligent Systems*, 21(1):35–41, Jan. 2006.

[18] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating “word of mouth”. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 210–217, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[19] K. Swearingen and R. Sinha. Beyond algorithms: An hci perspective on recommender systems. In *ACM SIGIR. Workshop on Recommender Systems*, volume Vol. 13, Numbers 5-6, pages 393–408, 2001.

[20] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pages 109–116, New York, NY, USA, 2011. ACM.

[21] S. Vargas and P. Castells. Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 129–136, Paris, France, France, 2013. Le Centre De Hautes Etudes Internationales D'Informatique Documentaire.

[22] M. Zhang and N. Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 123–130, New York, NY, USA, 2008. ACM.

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

[24] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management*, AAIM '08, pages 337–348, Berlin, Heidelberg, 2008. Springer-Verlag.

[25] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.