Journal of Information Processing Vol.22 No.4 1–10 (Oct. 2014)

[DOI: 10.2197/ipsjjip.22.1]

Regular Paper

A Study of User Intervention and User Satisfaction in Recommender Systems

Yoshinori Hijikata[1,a)] Yuki Kai[1] Shogo Nishida[1]

Received: September 3, 2013, Accepted: June 18, 2014

Abstract: Recent recommender systems have achieved high precision in recommending favorite items to users. However, it has been reported that user satisfaction does not necessarily increase even when a recommender system recommends items high precision. User satisfaction is considered to be influenced by many factors. Among these factors, we focus in particular on user intervention. User intervention is a user control over a recommender system.
We provide three hypotheses: i) user intervention in the recommendation process itself improves the user satisfaction, ii) user intervention improves the user satisfaction when the intervention is reflected in the recommendation results, and iii) the degree of improvement in user satisfaction differs among the types of user interventions applied. In this study, we conducted a user experiment using several kinds of interventions, and clarify the relationship between user intervention and user satisfaction.

Keywords: recommender system, user intervention, user satisfaction, user profile, content-based filtering, user context

1. Introduction

Recommender systems have been introduced in many domains or services to solve the problem of information overload. Recommender systems select items (contents or goods) suited to a user's preference based on the user's preference data, buying history, browsing history, demographic information, and so on. A traditional approach in the study of recommender systems is to improve the recommendation accuracy [1]. This means that they consider selecting a user's favorite items correctly to be the most important evaluation index in this research area. However, investigations have revealed that user satisfaction with recommendations (hereinafter, *"user satisfaction"*) is not necessarily high even when the system achieves a high precision level [1], [2]. Therefore, along with the accuracy of the recommendation, many features and evaluation parameters are garnering attention as factors for improving user satisfaction. Examples of these are the diversity of the recommendation list and the user's understanding of the recommendation results [3], [4].

As described herein, we specifically examined user intervention in the system recommendation processes (hereinafter, *"user intervention"*) as a factor influencing user satisfaction. User intervention is a user activity in which the user not only receives recommendation results, but is also involved with the process of recommendation and controls the recommendation mechanism. Explicit feedback regarding the level of preference to an item, or an edit to a user profile by a user [5], [6], [7], [8], [9], is an example of a user intervention. It is generally considered that a user's load or burden in using a recommender system should be reduced to the greatest degree possible. Therefore, most recommender systems avoid explicit inputs from users. Nevertheless, this common belief might not be correct according to the time and circumstances. For example, MediaUnbound, a music recommender system, has received better evaluations from users than the Amazon.com recommender system, even though it asks users to give answers to no fewer than 35 questions [10]. The US matchmaking site, eHarmony, has attracted many users despite asking them to answer questionnaire that takes about an hour to complete [11]. Some successful factors might be gleaned from these two cases above. First, after answering many questions, user faith in the recommender system (an emotional factor such as *the system should produce good recommendation because I answered so many questions!*) might be established. Second, a user's understanding of the recommendation (to guess the reason why the recommendation result was provided) may have been improved.

Based on the facts described above, and upon our inference, we tested the following three hypotheses related to user intervention and satisfaction.

H1 User intervention itself improves user satisfaction.

H2 User intervention improves the user satisfaction when the intervention is reflected in the recommendation results.

H3 The types of user intervention affect the user satisfaction.

First, H1 indicates that user intervention itself has a psychological influence on the users, and that it improves the user satisfaction irrespective of the recommendation results. Even if a system allows users to intervene in the recommendation process and uses no intervention data to produce a recommendation, user satisfaction will be higher than a case in which the users cannot intervene. The reason for this is the psychological value of the user

[1] Guraduate School of Engineering Science, Osaka University
[a)] hijikata@sys.es.osaka-u.ac.jp

intervention itself. First, we examine whether user intervention itself influences user satisfaction.

When hypothesis H1 is not validated, it becomes evident that users will not be satisfied with the recommendation results unless they reflect the user interventions. We then tested the second hypothesis, H2. H2 indicates that a user intervention improves the user satisfaction when the intervention is reflected in the recommendation results. If users notice that their efforts in providing an additional intervention are rewarded, they will be more satisfied with the recommendation results.

H3 indicates that the type of user intervention influences the user satisfaction. When the types of user intervention differ, the types of information the users input also differ. This might influence the user satisfaction. Therefore, we examined the relationship between the intervention type and user satisfaction.

We tested these hypotheses for the music domain. Many people listen to music daily (e.g., on commuter trains, in the car, or during working hours). It was therefore easy to invite many people to participate in this experiment. We implemented a music recommender system through which the users can exert their influence on the recommendation process in several ways, and by which we can verify the three hypotheses through a user experiment.

The remainder of this paper is organized as follows. Section 2 describes the user interventions used in our user experiment. Section 3 explains the experimental method used. Section 4 presents the experiment results. Section 5 provides a further discussion based on these results. Section 6 introduces other works related to user satisfaction of a recommender system. Finally, Section 7 summarizes the results of the present study and provides a description of future work.

2. User Intervention

For this study, we prepared the following user interventions with different types of information to be input for assessing H3: (1) rating, (2) context input (CI), (3) content attribute selection (CAS), and (4) profile editing (PE), which are commonly used in existing recommender systems.

2.1 Rating

Rating is the simplest type of user intervention, and is a user's assigned evaluation of a provided item. Users can respond optionally with a rating to the recommender system based on their implicit input, such as purchasing or browsing history. Many recommender systems use ratings to create user profiles (a model representing a user's preference or interests) [12]. In this study, we also created user profiles based on rating data. We use a scale of 1 to 5 for inputting the rating values in our experimental system (1 = dislike very much, 2 = dislike, 3 = no preference, 4 = like, and 5 = like very much.). We define such rating as the baseline for when users do not intervene in any particular way (*no intervention*). Upon receiving a recommendation, the user does not input any additional information. In our study, users must assign ratings to items in advance for the cases of three other user interventions (described later).

2.2 Context input

It has been stated that user preferences vary depending on the contexts [13]. Many systems utilize the user contexts for making recommendations [13], [14], [15], [16], [17]. Ono [14] and Motomura [17] use a Bayesian network to recommend items depending on user contexts. They showed that using such contexts makes the recommendations more accurate.

The above systems ask users to input their current contexts. *Context input* (CI) is a user intervention by which users input their current contexts into a recommender system. In the systems described above, users input the current context when they want to receive a recommendation. They can input the context by choosing one of the provided alternatives in each context parameter. We use the same type of input method in this study. We adopt three parameters related to the contexts used in many other studies [13], [14], [15], [16]: when, where, and with whom. The provided options in each parameter are as follows: { *when*: *morning* / *daytime* / *night* }, { *where*: *home* / *office* / *driving* }, { *with whom*: *alone* / *friend* / *family* }. A screenshot of the context input page in our experimental system is presented in Figure 1-(a) (the upper-left screenshot in Figure 1).

Although, users choose only one of the provided options in each parameter, they must input their contexts each time they receive a recommendation. The system recommends songs considering the contexts input by the users.

2.3 Content Attribute Selection

We assume that users may occasionally want to select a category of items explicitly to receive a recommendation. For example, for a music recommendation, the user may want to hear only the latest music at certain times, but may prefer older music at other times. A user who generally likes rock music might occasionally choose to listen to pop music. However, recommender systems that recommend items based only on user preference data cannot cater to such temporary requests. *Content attribute selection* (CAS) is a user intervention by which users select the attributes (categories) of items that they want. Users can actively narrow down the items for a recommendation.

For this study, we use the following six content attributes (the details of attribute selection are explained in Section 3.1): { *country*: *international* / *domestic* }, { *genre*: *rock* / *pop* / *R&B and hip hop* / *anime* / *enka* }, { *sex*: *male* / *female* }, { *unit*: *solo* / *group* / *band* / *idol* }, { *year*: *before '70* / *'80* / *'90* / *'00* }, { *tune*: *ballade* / *medium* / *up* }. In our experimental system, the users can select music attributes by checking the checkboxes of the content attributes. A screenshot of the content attribute selection page of our experimental system is presented in Figure 1-(b) (the lower-left screenshot in Figure 1) .

2.4 Profile Editing

In common recommender systems, users cannot see their own user profiles or edit them directly. *Profile editing* (PE) is a function showing users their user profiles and allowing them to edit these profiles directly. Some researchers have asserted that it is important to be able to edit the user profiles directly [18]. In most systems, user profiles are created automatically using machine

(a) Context input page

(b) Content attribute selection page

(c) User profile editing page

Fig. 1 Screenshot of the experimental system

learning techniques. However, incorrect results are unavoidable in machine learning [19]. Profile editing can allow users to correct such errors manually [8]. Furthermore, presenting user profiles to users might enable them to understand more easily the reasons underlying the recommendations [9].

Many methods have been proposed for realizing profile editing capability. One method compels users to edit the parameters regarding their preferences [20], [21], [22], [23], [24]. Another visualizes a user's interest or comprehension of each item and compels the user to edit its value [5], [24], [25], [26]. Another method visualizes the models of user preferences learned using machine learning algorithms, and induces them to edit the models [6], [7], [8], [9], [27].

As described above, the format of a user profile depends on the recommendation algorithm used. We adopt a user profile that can be represented based on the user's preference levels of each value for the content attributes (the recommendation algorithm is explained in Section 3). This is the simplest presentation because it simply uses the content attributes. Figure 1-(c) (the right screenshot of Figure 1) portrays a screenshot of the profile editing page of our experimental system. On this page, a user profile is visualized as bar graphs. Each bar corresponds to a content attribute value. The value of a bar shows how much the user prefers an item with a content attribute value. Users can edit their profiles by manipulating these graphs.

3. Experimental Method

3.1 Experimental Music Recommender System

We implemented an experimental music recommender system equipped with the user intervention functions described in Section 2 in a server-client architecture. The server was built in Java Servlet, and the client was built in Flash and ActionScript. The method for utilizing the system is as follows. First, the user gives a rating to the presented songs to create a user profile. A song list, in which each song is equipped with a preview button, is presented to the user. The song list is composed of ten songs, and after finishing an evaluation of the ten songs, another song list is presented to the user. After finishing the evaluation, the user can receive a recommendation. The recommendation list includes five songs with a preview button. When the user wants to conduct an intervention, the user should invoke the intervention page for each intervention type (Figure 1). After completing an intervention, the user can receive a recommendation reflecting their intervention.

We applied a content-based filtering technique for realizing recommendations because such a technique can utilize the content feature for a user intervention. Among the many methods of content-based filtering, we applied a Bayesian estimation to our experimental system, which is the simplest and most popular model-based method [14], [28], [29], [30]. The system learns the frequency of each value of every content attribute in the favorite class and in the non-favorite (disliked) class. It calculates the probability of a new song for each class based on the above frequencies using Bayesian estimation. It recommends songs to the user in order of their probability of preference. The algorithm used in Bayesian estimation is explained in Subsection 3.2.

The user profile corresponds to the frequency of every content attribute value (a total of twenty attribute values). The profile editing page presents the graph of the frequencies in the favorite and non-favorite classes. To realize recommendations considering the user context, the system also learns the frequency of each content attribute value in each pair of context attribute value and favorite/non-favorite class.

Content-based filtering requires content attributes for a recommendation. Two types of attributes exist in music data: the category, such as genre and year, and features, representing various aspects of the music such as tempo and key. For this study, we adopted the category type, which is used in many music recommender systems. We selected six attributes and their categorical values, as presented in Section 2.3.

3.2 Recommendation Algorithm

This subsection describes our recommendation algorithm using Bayes' estimation. Bayes' estimation is a method of statistical inference in which evidence or observations are used to update the probability of a hypothesis being true. Given new evidence, Bayes' theorem adjusts the probabilities as

$$P(A \mid X) = \frac{P(X \mid A)\, P(A)}{P(X)} \tag{1}$$

where A and X are discrete random variables, A represents a specific hypothesis, $P(A)$ is called the prior probability of A inferred before new evidence X became available, and $P(A|X)$ is the posterior probability of A given X. Then, $P(X)$, which is called the marginal probability of X, is calculated as follows:

$$P(X) = \sum_{B} P(X \mid B)\, P(B) \tag{2}$$

In this equation, B is a complete set of mutually exclusive hypotheses. Furthermore, if X is a set of some mutually exclusive events $X = (x_1, ..., x_n)$, then the following equation is true:

$$P(X) = \prod_{i=1}^{n} P(x_i) \tag{3}$$

Consequently, equation (1) can be transformed as follows based on equation (2) and (3):

$$P(A \mid X) = \frac{\prod_{i=1}^{n} P(x_i \mid A)\, P(A)}{\sum_{B} \prod_{i=1}^{n} P(x_i \mid B) P(B)} \tag{4}$$

We define the algorithm for estimating the preference probability of each song in our dataset for each user using Bayes' estimation. Our system learns the features of a user's favorite and non-favorite (disliked) songs based on the user's ratings (the learned model becomes the user profile). Our system calculates the probability of the user liking each song (hereinafter, *"song score"*) using Bayes' estimation with the user profile and the song features. In equation (4), X corresponds to an event that represents the presence of a song, $x_i (i = i, ..., n)$ corresponds to values of the content attributes, A corresponds to an event showing that a user likes a song, and B corresponds to all events regarding the user's preferences (likes and dislikes).

Additionally, because we use the context input as a user intervention, our experimental system needs to calculate the song scores considering the user's context. Our system learns the features of the user's favorite and non-favorite songs under each value of the context attributes.

3.3 Experimental Method

We conducted an experiment to investigate the relationship between user intervention and user satisfaction using the system described above. Eighty-four participants (59 men and 25 women) joined in the experiment. They ranged in age from 19 to 25 (the average being 21.9). The participants were separated into two groups: one with intervention feedback (45 users) and one without intervention feedback (39 users).

Without intervention feedback means the following. Despite the users intervening, the system uses no intervention data to create the recommendation lists. The system presents recommendation lists just as it would if there had been no user intervention.

Compared to a recommendation list with no intervention, a recommendation list with some intervention becomes altered. If user satisfaction is improved after a user intervention, then it remains unexplained which factor influences the user satisfaction: the user intervention itself or the change in the recommendation list by the user intervention. We therefore verified the effects of user intervention alone by preparing a group without intervention feedback.

We originally created a music dataset of 1,000 songs for the experiment. Among these 1,000 songs, 100 were used for learning the user profiles, and 900 were used for testing the recommendations.

When the users participated in the experiment, they first answered the following questions about their original degree of interest in music: Q1, average time (hours) listening to music per day; Q2, number of CDs purchased; Q3, number of music files owned (MP3s etc.); Q4, frequency of concert attendance; Q5, availability of playing instruments; and Q6, general reason of listening to music. These questions were designed to determine how much the users enjoy listening to music in their daily life. User satisfaction was expected to vary concomitantly with the user's seriousness about the domain of the target items. We therefore quantified each user's degree of interest in music based on these questions, and used these degrees of interest for an analysis of the results.

The user tasks were as follows. Users first evaluated the songs in the training dataset. They then assigned ratings with no contexts, and three ratings with certain contexts. For the ratings with certain contexts, the users evaluated the songs according to whether or not they liked them under the context examples. The system chose three context examples for each song from the twenty defined examples (e.g., morning – home – alone), and presented these contexts to the users. The users could listen to all of the songs to assign a rating.

Second, the users conducted an intervention into the system and received recommendations. They conducted two out of four kinds of user intervention: rating, context input (CI), content attribute selection (CAS), and profile editing (PE). The type and order of the interventions varied among users. They received the recommendation lists ten times, each of which included five songs, for each kind of intervention. They then gave their degree of satisfaction of each recommendation list a level of 1 to 5 (1 = not satisfied, 5 = satisfied).

It was reported that user satisfaction for the same recommendation varies based on the conditions under which the users use the recommendation systems [35]. For instance, the evaluations may be harsher when the users are required to pay for acquiring the item compared to when the item is free of charge.

We asked the users to evaluate the following three types of satisfaction: (1) purchase satisfaction (sat-purchase), where users provide a satisfaction evaluation considering whether or not they will immediately purchase the recommended song; (2) listening satisfaction (sat-listening), where users provide a satisfaction evaluation for the current song at a free service site; and (3) satisfaction when finding a new interest (sat-interest), where users provide a satisfaction evaluation according to whether they are interested in a recommended song for a future purchase. The sat-purchase type considers cases in which users shop at a shopping site and should decide whether to buy their preferred items immediately. The sat-interest type considers cases in which users seek promising items for future purchases. In these cases, the users should not decide whether to buy the items on the spot. More unknown items might satisfy the users under this type of situation. Furthermore, users should answer the question "Why does the user conduct the intervention?"

Table 1 Reason for each type of user intervention

	(i)	(ii)	(iii)	(iv)
Rating	17.1 %	**51.4** %	20.0 %	11.4 %
CI	**40.5** %	8.1 %	18.9 %	32.4 %
CAS	23.7 %	13.2 %	**39.9** %	23.7 %
PE	5.3 %	15.8 %	18.4 %	**60.5** %

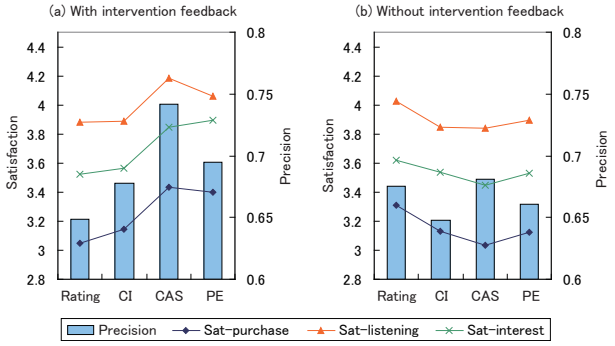


Fig. 2 Relationship between user intervention, precision and user satisfaction

3.4 Reason of Each User Intervention

Using the above question, we surveyed the users on why they conducted each user intervention. The users selected one or more of the following options: they were (i) looking for songs they would like to listen to now, (ii) reflecting their permanent preferences in their user profiles, (iii) looking for new songs, and/or (iv) testing the changes in the recommendation results according to their intervention. Table 1 shows the percentages of each reason based on the type of user intervention.

The reason evidently varies with the type of user intervention, as shown in Table 1. For rating, it is natural for reason (ii) to be dominant because applying rating is an intervention conveying the user's own preferences into the system. Context input is an intervention by which we presume that the users input their context for seeking songs that match the current context. For context input, it is natural that reason (i) be dominant. The rate of reason (iv) is also high for context input. Therefore, many users tested the differences in the recommendation results under different contexts. For content attribute selection, reason (iii) is popular. Under this reason, the user wants songs that are not usually recommended. For profile editing, reason (iv) is dominant, perhaps because it is difficult for users to understand how much changes in the parameters affect the recommendation results. However, when considering reasons other than (iv), users edit their user profiles to notify the system of their permanent preferences and for finding new songs (the percentages of which are almost same at 15.8% and 18.4%). This reflects the users' high-level requirements on the recommender system through this intervention.

4. Relationship between User Intervention and User Satisfaction

4.1 Overall Tendency

This section describes an analysis of the relationship between user intervention and user satisfaction. Figure 2 shows the precision and the three types of user satisfaction for each user intervention. The graph is separately depicted for the group with context feedback (Figure 2-(a)) and for the group without context feedback (Figure 2-(b)). Significant differences by *t*-test are shown in Table 2 for the group with context feedback, and in Table 3 for the group without context feedback. The total number of data points is 1,680 (combination of 84 users, two kinds of user intervention, and ten recommendation lists). The precision is the average precision of all recommendation lists provided, and the degree of satisfaction is the average degree of satisfaction for each recommendation list.

Table 2 Significant differences (*t*-test) in Figure 2 (a)

	Sat-purchase	Sat-listening	Sat-interest
Rating vs. CI			
Rating vs. CAS	***	***	***
Rating vs. PE	***	***	***
CI vs. CAS	***	***	***
CI vs. PE	***	***	***
CAS vs. PE		**	

*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

Table 3 Significant differences (*t*-test) in Figure 2 (b)

	Sat-purchase	Sat-listening	Sat-interest
Rating vs. CI	**	***	
Rating vs. CAS	***	***	**
Rating vs. PE	**	**	
CI vs. CAS			
CI vs. PE			
CAS vs. PE			

*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

First, we specifically examined the level of precision. In the group with intervention feedback (see Figure 2-(a)), the precision of users who performed an intervention is higher than that of users who did not (rating only). In contrast, in the group without intervention feedback (see Figure 2-(b)), the precisions of CI and PE are smaller than that of rating. There is only a slight difference in the precision between rating and CAS. The results therefore show that the changes in recommendation results by user interventions improve the precision.

Next, we specifically examined the level of satisfaction. In the group with intervention feedback (see Figure 2-(a)), the results show that user satisfaction increases when the users conduct CAS and PE. Significant differences are found using *t*-tests between rating and CAS, and rating and PE (see Tables 2). However, there is no difference between rating (no intervention) and CI. We cannot say that an intervention of any kind increases the user satisfaction. On the other hand, the satisfactions of CAS and PE are higher than that of CI. A possibility exists for the user satisfaction to differ among the types of user intervention. However, the correlation between satisfaction and precision is high (correlation coefficients r are 0.613 (sat-purchase), 0.720 (sat-listening), and 0.497 (sat-interest) ($p < 0.01$ for all)). It remains unclear whether the types of user intervention influence the user satisfaction. For the group without intervention feedback (see Figure 2-(b)), the user satisfaction after conducting an intervention is lower than when no intervention (rating only) takes place. The possibility exists that the user satisfaction will decrease when the intervention is not fed back into the recommendation results.

4.2 Grouping by User's Degree of Interest in Music

User satisfaction is expected to vary according to the user's

Table 4 Significant differences (t-test) in Figure 3 (a)

	Sat-purchase	Sat-listening	Sat-interest
Rating vs. CI			
Rating vs. CAS	**	***	
Rating vs. PE	***	***	***
CI vs. CAS	**	***	**
CI vs. PE	***	***	***
CAS vs. PE	*		***

*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

interest in the domain of the target items: in this experiment, music. Therefore, we quantified the user's degree of interest in music based on her answers to questions Q1–Q6 in Section 3.3. In fact, Q1–Q6 are the style of question in which users choose one from multiple options, and we assign scores to each option. We defined a user's degree of interest in music as their total score for Q1–Q6 (normalized in $[0, 1]$). Users with scores of less than 0.5 are categorized as a *low-interest group* (26 users in the group with intervention feedback, and 17 users in the group without intervention feedback), whereas users with a score of 0.5 or higher are categorized as a *high-interest group* (19 users in the group with intervention feedback, and 22 users in the group without intervention feedback).

Figure 3 shows graphs regarding the group with intervention feedback, and Figure 4 shows graphs for the group without intervention feedback. In the figure, panel (a) is a graph showing high-interest group data, and panel (b) is a graph showing low-interest group data. First, we examined the group with intervention feedback. In the high-interest group (see Figure 3-(a)), user satisfactions regarding CAS and PE are higher than those regarding rating and CI. In addition, user satisfaction for PE is higher than that for CAS. The correlation between satisfaction and precision is lower than that of all users (in Section 4.1) except for sat-interest (correlation coefficients r of 0.588 (sat-purchase), 0.707 (sat-listening), and 0.502 (sat-interest) ($p < 0.01$ for all)). Significant differences are found by t-test among some interventions (see Table 4). From this result, it remains possible that user satisfaction differs among the types of user intervention. In the low-interest group, user satisfaction depends on the precision (correlation coefficients r of 0.634 (sat-purchase), 0.736 (sat-listening), and 0.497 (sat-interest) ($p < 0.01$ for all)). We cannot determine the relations between user intervention and user satisfaction.

In the group without intervention feedback (see Figure 4), user intervention decreases the user satisfaction in the high-interest group. In the low-interest group, user satisfaction tends to depend on the precision (correlation coefficients r of 0.677 (sat-purchase), 0.776 (sat-listening), and 0.522 (sat-interest) ($p < 0.01$ for all)). From this result, users with a high interest in music may be disappointed with the results because they received worse recommendation lists in spite of their additional tasks.

In the following section, we specifically examine only those users who have a high interest in music.

4.3 Grouping by Precision

Although the correlation between satisfaction and precision is lower in the high-interest group, this fact does not indicate that no correlation exists. It remains unclear how much user intervention influences user satisfaction. Therefore, we classified each recommendation result into one of three groups based on the precision: (a) 0.0–0.2, (b) 0.4–0.6, and (c) 0.8–1.0. Because the precision is almost constant in each group, we can readily investigate the influence of user intervention on user satisfaction.

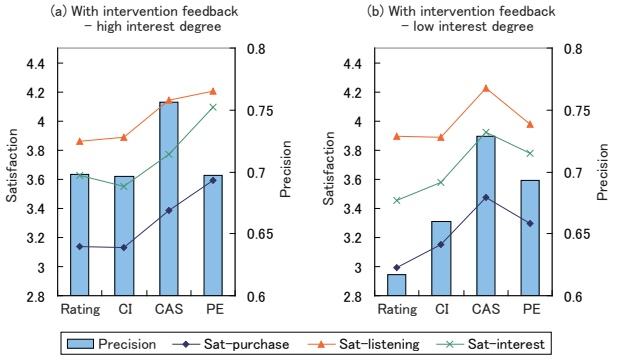


Fig. 3 Relationship between user intervention, precision, and user satisfaction in the group with intervention feedback considering user interests in the music domain

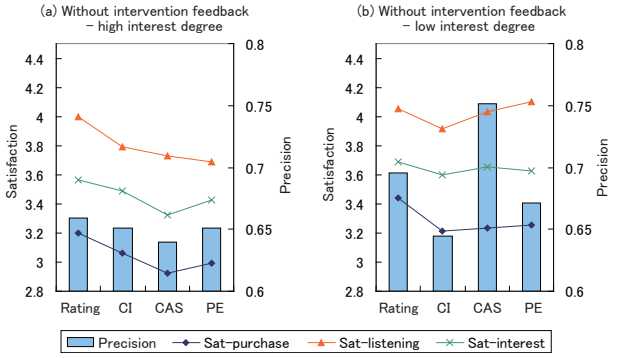


Fig. 4 Relationship between user intervention, precision, and user satisfaction in the group without intervention feedback considering user interests in the music domain

Figure 5 shows the results of users who are highly interested in music. Significant differences by t-test for precision (c), 0.8–1.0, are presented in Table 5 (for the group with intervention feedback) and in Table 6 (for the group without intervention feedback). We omitted those for precisions (a), 0.0–0.2, and (b), 0.4–0.6, because of space limitations).

First, we examined the low-precision group (a) and medium-precision group (b). In the group with intervention feedback (a-1) and (b-1), the probability exists that a user intervention improves the user satisfaction because only the satisfaction regarding PE is high. However, the satisfaction regarding the other two interventions is equivalent to that with no intervention (rating) (for (a-1), the satisfaction for interest regarding the three interventions is lower than that regarding rating.). It is therefore not clear whether user intervention influences the user satisfaction. In the group without intervention feedback, (a-2) and (b-2), user satisfactions decrease when a user conducts an intervention. We believe this is the result of betraying the users' expectations in spite of their additional workload (intervention).

We specifically examined the high-precision group (c). In the group with intervention feedback (c-1), user satisfactions regarding CAS and PE are higher than those regarding rating (no intervention) and CI. This shows that any type of user intervention does not necessarily improve the user satisfaction, which does

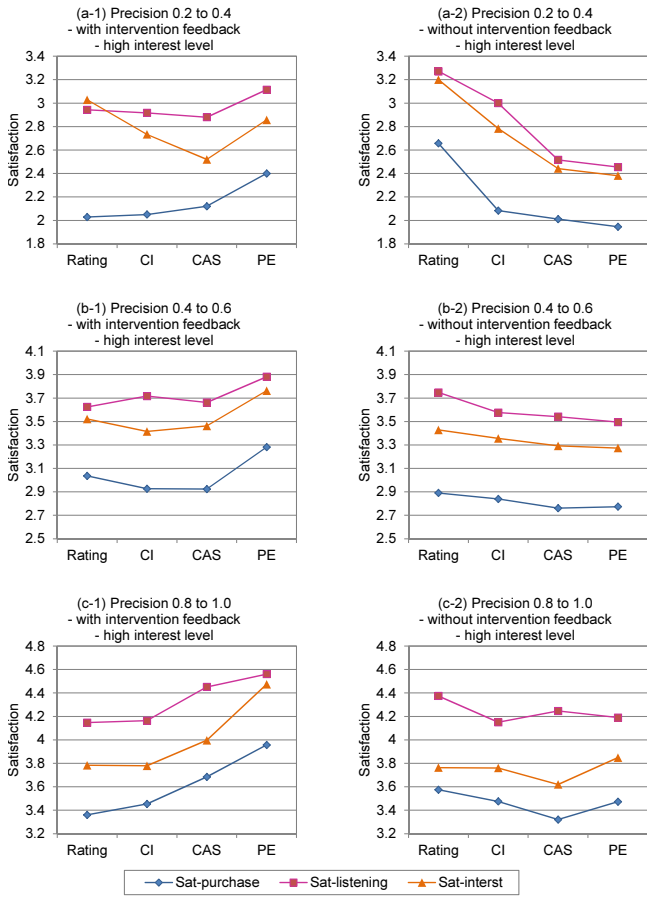


Fig. 5 Satisfactions grouped by precision: ((a) 0.2–0.8, (b) 0.4–0.6 and (c) 0.8–1.0)

Table 5 Significant differences (*t*-test) in Figure 5 (c-1)

	Sat-purchase	Sat-listening	Sat-interest
Rating vs. CI			
Rating vs. CAS	**	***	
Rating vs. PE	***	***	***
CI vs. CAS	*	***	**
CI vs. PE	***	***	***
CAS vs. PE	**		***

*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

Table 6 Significant differences (*t*-test) in Figure 5 (c-2)

	Sat-purchase	Sat-listening	Sat-interest
Rating vs. CI		**	
Rating vs. CAS			
Rating vs. PE		**	
CI vs. CAS			
CI vs. PE			
CAS vs. PE			

*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

not support H2. However, CAS and PE improve the user satisfaction, and this improvement is higher in PE than in CAS. It shows that user satisfaction differs among the types of user intervention, which supports H3.

In contrast, in the group without intervention feedback, (c-2), satisfaction regarding a user intervention is the same as that for no user intervention (rating). Comparing (c-2) and (c-1), in terms of CAS and PE, satisfaction in (c-2) is lower than in (c-1). No effect from a user intervention itself is evident: H1 is therefore not validated.

4.4 Summary of Results

According to the results of the experiment, when the system recommends items with high precision for users with a high interest in music, some types of user intervention (CAS and PE) improve the user satisfaction. However, another type of user intervention (CI) does not improve the user satisfaction. This means that whether a user intervention improves the user satisfaction is dominated by the intervention type. In addition, the satisfaction of PE is higher than that of CAS. This means that user satisfaction differs among the types of user intervention conducted. Finally, the effect of user intervention itself on user satisfaction cannot be proved because no improvement is found when the intervention is not fed back into the recommendation results. In addition, the precision strongly affects the user satisfaction.

From this result, we can say that recommending items with high precision is necessary for a recommender system. Presently, most recommender systems can produce accurate recommendations. Therefore, adding certain functions to existing systems by which users can intervene in the recommendation mechanism can improve the user satisfaction. Additionally, the results show that user satisfaction is influenced by the user's degree of interest in the domain of the target items. If recommender systems apply different interaction models for individual users, the user satisfaction will further increase.

5. Discussion

5.1 Intervention purposes and efforts

In the previous section, we did not find that user interaction itself influences user satisfaction. Initially, we thought that a user intervention improves the user satisfaction from the psychological effects of such action. However, the users confirmed the recommendation results in the experiment conducted. We asked the users their reasons for an intervention, as described in subsection 3.4. They selected different reasons when the types of intervention differ. Therefore, they might check whether the recommendation result meets their reason for an intervention.

The experimental results revealed that when a user intervention is fed back into the recommendation result, it does not necessarily improve the user satisfaction. We also found that the degree of satisfaction improvement differs among the types of intervention conducted. In detail, the improvement in satisfaction is larger in PE than in CAS. From this result, we believe that user satisfaction may be influenced by the user's effort regarding the intervention. We measured the time required for finishing an intervention (see Figure 6). The intervention time in PE is larger than in other intervention types ($p < 0.01$ by *t*-test). From this result, the user's effort regarding an intervention may be related to the user's satisfaction.

Furthermore, we believe that the time and effort required for an intervention might be longer and larger when the user intervenes in the recommendation process at a finer grade. We call such a grade the level of detail. The level of detail may be related to user satisfaction. When examining the level of detail, it is necessary to make the experimental conditions the same (e.g., the same type of intervention and the same intervention target) and test several grades of control for the intervention. For example, testing dif-

ferent rating scales (e.g., binary, five-level, ten-level scales), and testing different context grades (e.g., morning/daytime/night, 6 a.m. to 9 a.m./10 a.m. to 0 p.m./···) for a context input can be done during the experiment. We examined the type of intervention in this paper. The different levels of detail will be examined in the future.

Finally, when examining the level of intervention, a trade-off may exist between a user's efforts and the performance improvement of the recommendation results. If users have given a finer control in the recommendation process, the precision of the recommendation results may increase. However, this requires greater user efforts. We believe that user satisfaction is related to both factors. If the level of user intervention is higher, it sometimes requires additional learning data (a higher number of ratings). We combined the context and ratings in this experiment. The algorithm counted the frequency of music features occurring in each context. If the context is separated into more patterns, the algorithm will require additional learning data. Although the level of intervention includes complex problems, we need to clarify the optimal level of user intervention.

5.2 Practical use

This subsection describes how our findings can be used for a real commercial music site. User satisfaction was high in both CAS and PE in our experiment. However, this finding is true only for users with a high interest in the target domain (music). Therefore, commercial sites need to know the user's original interest in the domain. This information can be acquired by asking certain questions regarding the music domain. The questionnaires that were asked during our experiment, as described in Subsection 3.3, can be used here.

The main reason for conducting CAS is looking for new songs. It is better to propose that the user conducts CAS when looking for new songs. We therefore need to know the user's status while using the service. When the user browses through several different songs, or reads an item description of a song in a genre that they rarely listen to, it may be determined that the user is looking for new songs. The probability is higher that the user will accept the proposal for an intervention than during other situations.

When displaying the recommendation results, it is better to display a link or button to display the user's user profile. When the user doubts the validity of the recommendation or when they want to know why certain songs are recommended, the user might want to check their user profile. After checking their user profile, if some errors are discovered, the user may want to edit the profile. After editing their user profile and finding good songs, the user will be satisfied with the recommendation results.

6. Related Work

Many studies have been undertaken to examine or improve the user satisfaction of a recommender system. We next introduce studies of elements influencing user satisfaction, algorithms, and interaction models.

Liang et al. [31] presented three elements that might influence user satisfaction: (1) accuracy and the amount of information, (2) the user's motivation for a recommendation, and (3) user involvement. The authors investigated the respective effects of these elements through experimentation. As a result, they concluded that (1) and (2) strongly influence user satisfaction, but that (3) has no influence. Nevertheless, they investigated only explicit feedback as user involvement. In contrast to his study, we investigated user intervention more precisely, considering the several types of intervention available. Through a user experiment, Bollen et al. [32] showed that the size of the recommendation list influences the attractiveness of the recommendation and the difficulty in making a choice. Cremonesi et al. examined the relationship between the length of the user profiles (the number of ratings the user has input) and the user satisfaction [33]. They reported that the user satisfaction never decreases when the user perceives the relevance of a recommendation even when the system insists that the user input many ratings. Knijnenburg et al. examined the relationship between the inspectability and control (user intervention) based on structural equation modeling [34]. They reported that the user satisfaction increases when the user perceives that they can control the recommender system. However, they did not examine the intervention types in their experiment.

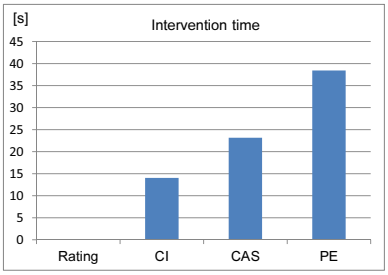


Fig. 6 Users' intervention time

Some researchers studying recommendation algorithms have considered user discoveries in recommendation lists [4], [35]. The topic diversification algorithm presented by Ziegler et al. [4] diversifies recommendation lists by combining the similarity list of items and the recommendation list output through collaborate filtering. This method elicited higher a user satisfaction than normal collaborative filtering. Hijikata et al. [35] sought to improve the novelty by combining collaborative filtering that uses preference ratings with collaborative filtering that uses acquaintance ratings. Novelty is an evaluation parameter representing how many new and favorite items are recommended by the system [1]. Their method elicits a higher user satisfaction than normal collaborative filtering. The novelty and diversity evaluation parameters (presented above) for a recommendation are important for satisfying users. However, these parameters are used independently for the evaluation. Vargas and Castells proposed a model presenting both novelty and diversity under a common framework [36].

Some studies have aimed at improving a user's understanding of a recommendation by explaining the rationale or basis of the recommendation. Herlocker et al. [37] compared several recommendation reasons for a movie recommender system using collaborative filtering. Schafer [38] also implemented a movie recommender system that explains how much the recommended items match the users' demands regarding the screening time, distance to the theater, and so on. In terms of profile editing, some

studies have aimed not only at improving the precision but also at obtaining the short-term preferences of users. Terveen et al. [6] and Bostandjiev et al. [24] represented user preferences regarding music through bar graphs that users can edit themselves. Ahn et al. [9] examined user trust in a recommender system when users edit their user profiles. Our study also applied profile editing and considered short-term preferences. Nevertheless, our study differs from the studies presented above in that it investigated the relationship between the types of user intervention and user satisfaction.

7. Conclusion

As described herein, we specifically examined user intervention as one factor influencing user satisfaction of a recommender system. We produced and tested three hypotheses: the action of user intervention itself improves the user satisfaction; a user intervention improves the user satisfaction when the intervention is reflected into the recommendation results; and the types of user intervention affect the user satisfaction. We conducted an experiment to compare certain types of user interventions to verify the hypotheses presented above. We also tested two cases: in one case, the given intervention is not fed back into the recommendation results, and in the other case, the given intervention is fed back into the recommendation results.

We analyzed the results of the experiment from two viewpoints: the precision and the users' degree of interest in the music domain. The results demonstrate that user satisfaction increases when certain types of user intervention are conducted; however, it decreases when other types of user intervention are conducted. When examining only the types of intervention that increase the user satisfaction, the degree of improvement differs according to the intervention type. We therefore proved that the types of user intervention affect the user satisfaction. However, a user intervention itself does not contribute to an improvement in user satisfaction.

We expect that the findings presented in this paper will be helpful for recommender service providers. For recommendation services, it is ideal to create a rich user experience and improve the user satisfaction. This paper described the necessity of maintaining good accuracy and altering the form or style of the user intervention according to the user's interest in the domain of the target item. These findings will aid service providers in designing recommender system functions and services.

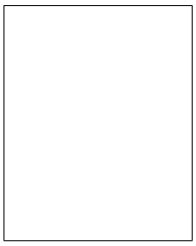
Finally, some research problems have become apparent based on the results of this experiment. The time and effort for an intervention may be related to user satisfaction. They are also related to the levels of detail of the intervention. We need to study the relationships among user satisfaction and the levels of detail of an intervention, the intervention time, and a user's effort. Furthermore, user satisfaction may be related to a user's understanding of the recommendation mechanism. When users are familiar with the recommendation mechanism, their satisfaction may increase. An examination of the users understanding of such mechanisms should be conducted in the future. We will study the above problems and provide further insight regarding the psychological aspects.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 20700085 and 24700091.

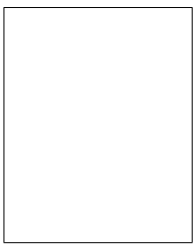
References

[1] Herlocker, J.L., et al.: Evaluating Collaborative Filtering Recommender Systems, ACM Trans. on Information Systems, Vol. 22, Issue 1, pp. 5–53 (2004)
[2] McNee, S.M., Riedl, J. and Konstan J.A.: Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems, Proc. of CHI'06, pp. 1097–1101 (2006)
[3] Swearingen, K. and Sinha, R.: Beyond algorithms: An HCI perspective on recommender systems, Proc. of SIGIR Workshop on Recommender Systems (2001)
[4] Ziegler, C.N., et al.: Improving Recommendation Lists Through Topic Diversification, Proc. of WWW'05, pp. 22–32 (2005)
[5] Weber, G. and Brusilovsky, P.: ELM-ART: An Adaptive Versatile System for Web-based Instruction, International Journal of Artificial Intelligence in Education, Vol. 12, pp. 351–384 (2001)
[6] Terveen L., et al.: Specifying Preferences Based on User History, Proc. of CHI'02, pp. 315–322 (2002)
[7] Waern, A.: User Involvement in Automatic Filtering, User Modeling and User-Adapted Interaction, Vol. 14, pp. 201–237 (2004)
[8] Hijikata, Y., et al.: Content-based Music Filtering System with Editable User Profile, Proc. of ACM SAC'06, pp.1050–1057 (2006)
[9] Ahn, J., et al.: Open User Profiles for Adaptive News Systems: Help or Harm?, Proc. of WWW'07, pp. 11–20 (2007)
[10] Swearingen, K. and Sinha, R.: Beyond algorithms: A Human-Centered Evaluation of Recommender Systems, SIMS 213, UC Berkeley (2002)
[11] Financial Times "Win the Hearts of Bachelors: Marriage Meeting Sites Battle in the USA" http://news.goo.ne.jp/article/ft/life/science/ft-20061226-01.html (2006)
[12] Hijikata, Y.: User Profiling Technique for Information Recommendation and Information Filtering, Journal of JSAI, Vol. 19, No. 3, pp. 365–372 (2004)
[13] Adomavicius, G., et al.: Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach, ACM Trans. on Information Systems, Vol. 23, No. 1, pp. 103–145 (2005)
[14] Ono, C., et al.: A Context-Aware Movie Preference Model Using a Bayesian Network for Recommendation and Promotion, Proc. of User Modeling 2007, LNCS Vol. 4511, pp. 257–266 (2007)
[15] Oku, K., et al.: Context-Aware SVM for Context-Dependent Information Recommendation, Proc. of FMUIT'06, IEEE, pp.119–122, (2006)
[16] Oku, K., et al.: A Recommendation System Considering Past / Current / Future Contexts, Proc. of Workshop on Context-Aware Recommender Systems (2010)
[17] Motomura, Y. and Iwasaki, H.: Bayesian Network Technology: Modeling of Users and Customers, and Uncertain Reasoning, Tokyo Genki University Press (2006)
[18] Vassileva, J.: A Practical Architecture for User Modeling in a Hypermedia-Based Information Systems, Proc. of the 4th International Conference on User Modeling (UM), pp. 115–120 (1994)
[19] de Rosis, F., De Carolis, B. and Pizzutilo, S. : User Tailored Hypermedia Explanation, Proc. of INTERCHI'93, pp. 169–170 (1993)
[20] Mathe, N. and Chen, J.: User-centered Indexing for Adaptive Information Access, User models and User Adapted Interaction, Vol. 6, No. 2–3, pp. 225–261 (1996)
[21] Hohl, H., Bocker, H.D. and Gunzenhauser, R.: Hypadapter: An Adaptive Hypertext System for Exploratory Learning and Programming, User Models and User Adapted Interaction, Vol. 6, No. 2–3, pp. 131–156 (1996)
[22] Beaumont, I.: User Modeling in the Interactive Anatomy Tutoring System ANATOM-TUTOR, User Models and User Adapted Interaction, Vol. 4, No. 1, pp. 21–45 (1994)
[23] Kay, J. and Kummerfeld, R.: An Individualised Course for the C Programming Language, Proc. of WWW'94, pp. 17–20 (1994)
[24] Bostandjiev, S., O'Donovan, J. and Hollerer, T.: Taste Weights: a visual interactive hybrid recommender system, Proc. of ACM RecSys'12, pp. 35–42 (2012)
[25] Cook, R. and Kay, J.: The Justified User Model: a Viewable, Explained User Model, Proc. of User Modeling (UM), pp. 145–150 (1994)
[26] Kay, J.: The UM toolkit for Cooperative User Models, User Models and User Adapted Interaction, Vol. 4, No. 3, pp. 149–196 (1994)
[27] Fujimori, H., Hijikata, Y. and Nishida, S.: Visualization of the Neighborhood in Collaborative Filtering, Journal of Human Interface Society, Vol. 7, No. 1, pp. 69–81 (2005)
[28] Breese, J.S., Heckerman, D., and Kadie, C.: Empirical Analysis of

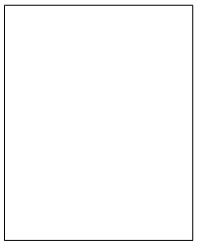
Predictive algorithms for Collaborative Filtering, Proc. of UAI'98, pp. 43–52 (1998)
[29] Condliff, M.K., et al.: Bayesian Mixed-Effects Models for Recommender Systems, Proc. of ACM SIGIR Workshop on Recommender Systems (1999)
[30] Zhang, Y. and Koren, J.: Efficient Bayesian Hierarchical User Modeling for Recommendation Systems, Proc. ACM SIGIR'07, pp. 47–54 (2007)
[31] Liang, T.P., Lai, H.-J. and Ku, T.-C.: Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings, Journal of Management Information Systems, Vol. 23, No. 3, pp. 45–70 (2007)
[32] Bollen, D., et al.: Understanding Choice Overload in Recommender Systems, Proc. of ACM RecSys'10, pp. 63–70, (2010)
[33] Cremonesi, P., Garrzotto, F. and Turrin, R.: User Effort vs. Accuracy in Rating-based Elicitation: Proc. of ACM RecSys'12, pp. 27–34, (2012)
[34] Knijnenburg, B.P., Bostandjiev, S., O'Donovan, J. and Kobsa, A.: Inspectability and control in social recommenders, Proc. of ACM RecSys'12, pp. 43–50, (2012)
[35] Hijikata, Y., Shimizu, T. and Nishida, S.: Discovery-oriented collaborative filtering for improving user satisfaction, Proc. ACM IUI'09, pp. 67–76 (2009)
[36] Vargas, S. and Castells, P.: Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems, Proc. of ACM RecSys'11, pp. 109–116 (2011)
[37] Herlocker, J.L., Konstan, J.A. and Riedl, J.T.: Explaining Collaborative Filtering Recommendations, Proc. ACM CSCW'00, pp. 241–250 (2000)
[38] Schafer, J.B, Konstan, J.A. and Riedl, J.: Meta-recommendation System: User-controlled Integration of Diverse Recommendation, Proc. ACM CIKM, pp. 43–51 (2002)



Yoshinori Hijikata Yoshinori Hijikata was born in Kobe, Japan. He received the B.E. and M.E. degrees from Osaka University in 1996 and 1998, respectively. In 1998, he joined IBM Research, Tokyo Research Laboratory. After working on Web technologies there, he received Ph.D. degree from Osaka University in 2002. He visited University of Minnesota, GroupLens Research as a visiting researcher in 2014. Currently, he is associate professor in Osaka University. His research interests are on Web intelligence, recommender systems and text mining. He received the best paper awards from IPSJ Interaction'05, Interaction'13, WebDB'11, WebDB'12, WebDB'13, IEICE DEWS'06, ACM IUI'06. He also received IPSJ Yamashita SIG Research Award in 2013. He is a member of the IPSJ, IEICE, JSAI, HIS and DBSJ.



Yuki Kai Yuki Kai was born in Hyogo Prefecture, Japan. He received the B.E. and M.E. degrees from Osaka University in 2008 and 2010, respectively. Currently, he is in NS Solutions. His research interests are on recommender systems and personalization.



Shogo Nishida Shogo Nishida was born in Hyogo Prefecture in 1952. He received the B.S. M.S. and Ph.D. degrees in Electrical Engineering from the University of Tokyo, in 1974, 1976 and 1984, respectively. From 1976 to 1995, he worked for Mitsubishi Electric Corporation, Central Research Laboratory. From 1984 to 1985, he visited MIT Media Laboratory, Boston, Massachusetts, as a visiting researcher. In 1995, He moved to Osaka University as a Professor of Graduate School of Engineering Science. He was the dean of Graduate School of Engineering Science from 2004 to 2007 and was the Trustee and Vice President of Osaka University from 2007 to 2011. His research interests include CSCW, Media Technology, Human Interfaces and Human Communication. He is a Fellow of IEEE (1998) and a Fellow of IEE in Japan (2008) and Honorary Member of Human Interface Society of Japan (2007).