Evaluating an Adaptive Music-Clip Recommender System

Tingshao Zhu and Russ Greiner

University of Alberta, Edmonton, Alberta T6G 2E8, Canada,
{tszhu, greiner}@cs.ualberta.ca

Abstract. In this paper, we propose an experiment design to address three evaluation goals of "First Adaptive System Evaluation Challenge" [1], and demonstrate how to achieve each of these goals.

1 Introduction

The system to be evaluated is a recommender system, specifically suggest music clip, either for individual users, or groups of users. Recently, the system has also been extended to model how happy each individual is as a consequence of having seen the clips so far, and the challenge is how to evaluate the user models (i.e., modelling the preferences of each individual) that have been used to recommend music clips.

Primary input for the models is provided in the form of ratings from 1 to 10 for each music clip for each individual. There are three proposals for modelling the happiness (i.e., M^1, M^2, and M^3). The system uses a selection model to determine which item to show next (i.e., what is the next most suitable item to place in the clip sequence), based on the individual's preferences of the viewed clips. The evaluation aims to provide answers to the following questions:

Relatively Valid Prediction
Which of the three proposals for modelling happiness succeed in making relatively valid predictions (i.e., when they predict that a clip will make an individual unhappy this is indeed the case, and vice versa, independently from the level of un-/happiness)?

Inter-individual Difference
Which of the three proposals is best at predicting inter-individual differences in happiness (i.e., managing to determine that clip C would make user U_1 happier than user U_2)?

Precision
Which of the three proposals achieves the highest modelling precision (i.e., manages to more precisely predict the user's happiness after having watched a clip/series of clips)?

[1] http://www.easy-hub.org/hub/workshops/um2005/challenge.html

2 Experiment Design

To evaluate the performance of the three happiness models, we can conduct an user study. The participants are given access to the recommender system, browsing the music clips by their own choices and ask for recommendations any time they want.

Everytime a subject is presented with a recommended clip, s/he is required to give an evaluation based on her/his happiness, from 1 to 10 (i.e., $ScoreEvaluated$). Here, since *the input is provided in the form of ratings from 1 to 10 for each music clip for each individual*, we thus use the primary input of the user's rating for the suggested clip as its $ScoreEvaluated$. The higher score, the happier the subject.

Whenever a subject asks for recommendation, the system will first randomly pick out one model from M_1, M_2, and M_3; calculate each clip how happy the user would be (i.e., $ScoreComputed$) if that clip were presented next (given the visited clips so far); then normalize $ScoreComputed$ to $[1, 10]$; and finally rank all the remaining clips based on their $ScoreComputed$.

The system will randomly choose one of the following two policies to select a clip as recommendation:

1. If there is only one clip with the highest score (i.e., $ScoreComputed = 10$), output it, or randomly output one clip if there are multiple clips that result in maximum happiness; otherwise follow the second option.
2. Choose one clip randomly, and output it.

For each evaluation, we will record the following information:

$$< UserID, ClipSeq, ModelID, ScoreComputed, ClipSuggested, ScoreEvaluated >$$

UserID
The identification of the subject.

ClipSeq
The music clip that the subject has visited so far.

ModelID
Which model has been chosen to generate the recommendation.

ScoreComputed
A continuous value from 1 to 10, denotes the happiness score of the suggested clip according to the selected model.

ClipSuggested
The music clip that presented to the subject as recommendation.

ScoreEvaluated
The primary input to show the subject's happiness for the suggested clip.

3 Evaluation Goals

Here, we will demonstrate how the proposed experiment design is able to address each of these evaluation goals.

3.1 Relatively Valid Prediction

Alternatively, we want to find which model will generate the least difference (statistically significant) between *ScoreComputed* and *ScoreEvaluated*. It is expected that the model which generates less differences will be more promising for making relatively valid predictions. Note that the primary goal focuses on un-/happy, we will test these suggested clips with extreme *ScoreComputed*, either happy ($ScoreComputed = 10$) or unhappy ($ScoreComputed = 1$). Since *ScoreComputed* can approach *ScoreEvaluated* in either way, we only care about the absolute value of the difference, that is, $|ScoreComputed - ScoreEvaluated|$.

For each subject s_i, we collect the suggested clips with $ScoreComputed = 10$ or $ScoreComputed = 1$, then compute the mean of the differences between *ScoreComputed* and *ScoreEvaluated* for each of three models (i.e., M^1, M^2, and M^3).

Table 1. Subject s_i's ($|ScoreComputed - ScoreEvaluated|$) for all three models

	M^1	M^2	M^3
	5	0.6	2.4
	4	0.55	1.5
		0.45	0.9
Mean	4.5	0.53	1.6

For example, in Table 1, s_i has asked 8 times for recommendations, in which the suggested clips had *ScoreComputed* either 10 or 1. Among these 8 recommendations, the system has selected M^1 2 times, 3 times for both M^2 and M^3. For each suggested clip, we compute the absolute difference between its *ScoreComputed* and *ScoreEvaluated*. Then we calculate the average difference for each model which is shown in Table 1. After we have computed the average difference for each model of each subject, we can build a happiness difference matrix as shown in Table 2.

Table 2. Happiness Differences

Subject	M^1	M^2	M^3
⋮			
s_i	4.5	0.53	1.6
⋮			

At first, we run Friedman test [2] on Table 2 using $k = 3$. The null hypothesis states that there is no significant difference among the three models.

If no significant difference can be detected (i.e., $p > 0.05$), we can conclude that there is no significant difference among these three model for making relatively valid predictions.

If there does exist significant difference (i.e., $p \leq 0.05$), we can then run Wilcoxon test [3] on any pair of models to identify the best model(s) for making relatively valid predictions. For example, after we have concluded that there exists significant difference among the three models on Table 2, we then run Wilcoxon test to verify two hypotheses: $M^2 \leq M^1$, $M^2 \leq M^3$. If both result in $p \leq 0.05$, then we can make conclusion that M^2 is the best model for making relatively valid predictions.

3.2 Inter-individual Difference

The intuition here is that the model that can predict the most inter-individual difference in happiness is the model that can produce the maximum number of significant differences among the subjects.

At first, for each model ($M \in \{M^1, M^2, M^3\}$), we compare each pair of subjects (*e.g.*, s_i and s_j), to detect whether M can identify significant difference between them. To do so, we just collect all the suggested clips for each subject, and compute the difference :

$$ScoreComputed - ScoreEvaluated$$

For example, Table 3 summarizes the difference produced by M for subject s_i and s_j.

Table 3. Inter-Individual Difference

M	s_i	s_j
	$\vdots$	$\vdots$
	0.15	-4.4
	-0.08	5
		6.2

We run Mann-Whitney test on Table 3 to detect whether there exists significant difference between s_i and s_j (i.e., Yes if $p \leq 0.05$, otherwise No), then construct a matrix to present the difference between any pair of the subjects for model M.

[2] Friedman is a statistical measure of two-way analysis of variance by ranks, with k repeated (or correlated) measures.

[3] Wilcoxon test is a nonparametric test that can be used for 2 repeated (or correlated) measures.

$$\left(\begin{array}{cccccc} & s_1 & s_2 & \dots & s_i & \dots \\ s_1 & & Y & \dots & N & \dots \\ s_2 & & & \dots & Y & \dots \\ \vdots & & & & & \\ s_i & & & \dots & & \dots \\ \vdots & & & & & \end{array}\right)$$

Note that the Mann-Whitney test that we run is non-directional, so the matrix is symmetric. We define the inter-difference score of M as :

$$\text{Inter-Difference}(M) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} sgn(s_i, s_j)$$

where

$$sgn(s_i, s_j) = \begin{cases} 1 \text{ there eixsts significant difference between } s_i \text{ and } s_j; \\ 0 \; otherwise. \end{cases}$$

The model that has the highest Inter-Difference Score will be the best model to predict inter-individual differences.

3.3 Precision

We follow the process that described in Section 3.1, the only difference here is that we use all the evaluations, not only these $ScoreComputed = 10/1$. The model that achieves the highest precision will be the model that produce the least difference between $ScoreComputed$ and $ScoreEvaluated$.