

Evaluating an Adaptive Music-Clip Recommender System

Tingshao Zhu Russ Greiner
University of Alberta



A Music Clip Recommender System





Recommender System

- For individual users, or groups of users
- Suggest music clip as a consequence of having seen the clips so far
- using the user models (modeling the preferences of each individual)

Evaluating User Models

Input:

- ratings (1-10) for each music clip for each individual

				...	
	2	1	7		10
	3	5	1		6
	8	4	9		2
	5	2	1		7



Experiment Design

Conduct a user study

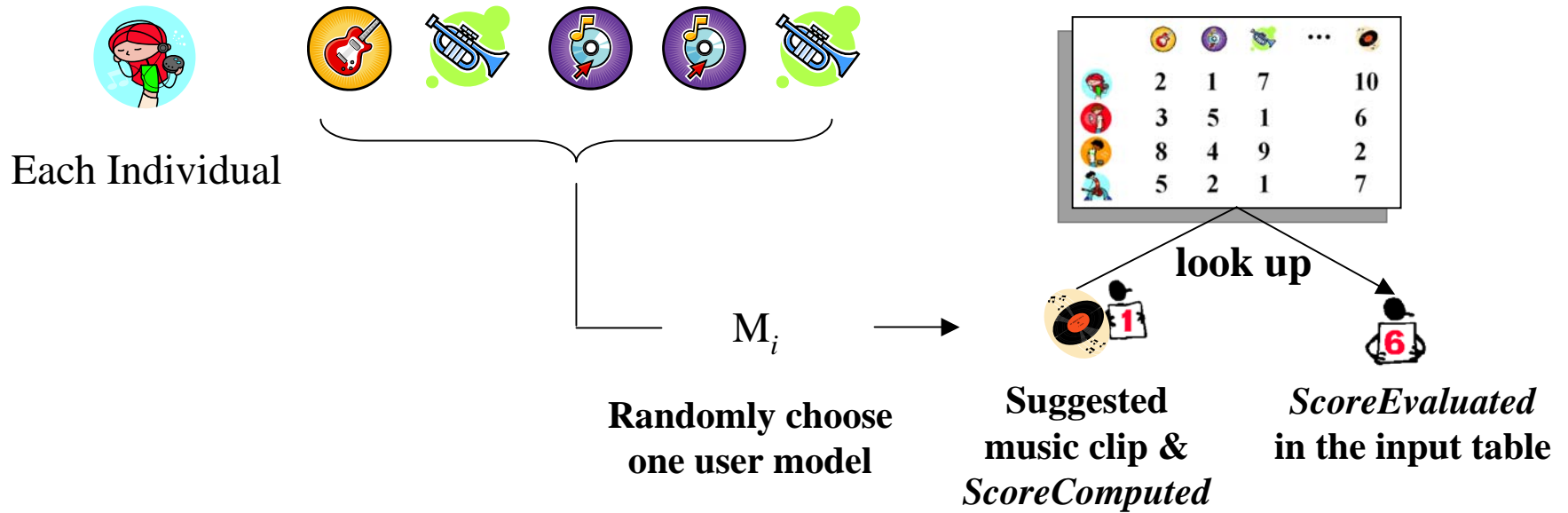
- randomly pick out one model when the subject requests
- compute the happiness (*ScoreComputed*) of all clips based on the observed clip sequence
- return the clip with the highest *ScoreComputed* as the suggested clip

The subjects are given access to the recommender system

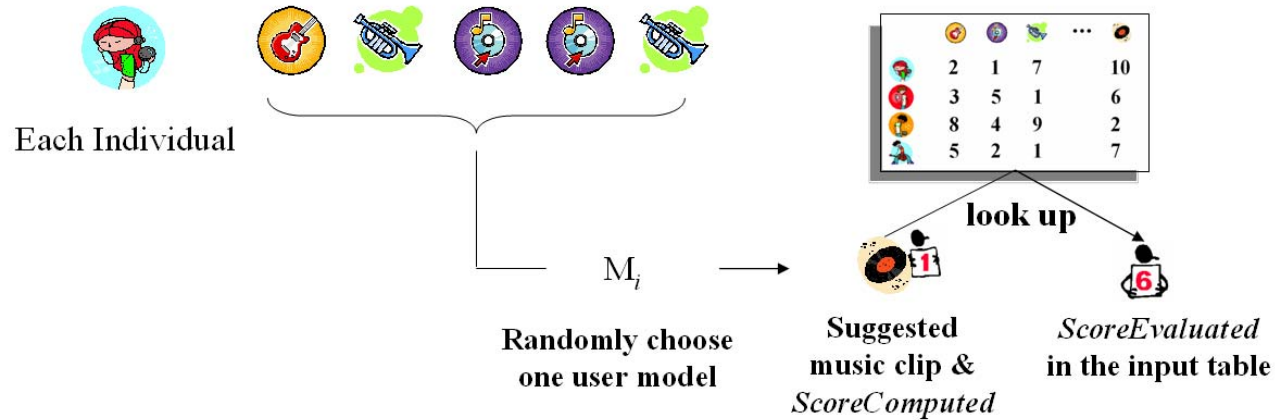
- ask for recommendation any time they want
- evaluate each recommended music clip (i.e., *ScoreEvaluated*).




Here, *ScoreEvaluated* is define as the primary input of the user's rating for the recommended Clip

Experiment Design



Experiment Design



User ID	# ClipSession	# Model	Suggested Clip	<i>ScoreComputed</i>	<i>ScoreEvaluated</i>
		1		1	6
	⋮				



Relatively Valid Prediction

Which of the three proposals for modeling happiness succeed in making **relatively valid predictions** ?

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.







Relatively Valid Prediction


Assumption:

The model which generates less differences (statistically significant) between *ScoreComputed* and *ScoreEvaluated* will be more promising for making relatively valid predictions.

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).

Relatively Valid Prediction


User ID	# ClipSession	# Model	Suggested Clip	$Score_{Computed}$	$Score_{Evaluated}$	$ Score_{Computed} - Score_{Evaluated} $
	 	1		1	6	5

Collect the values of $|Score_{Computed} - Score_{Evaluated}|$ of  , where $Score_{Computed}=10/1$

	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

Subject s_i 's ($|Score_{Computed} - Score_{Evaluated}|$) for all three models

Relatively Valid Prediction

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

Average Happiness for all subjects



Relatively Valid Prediction

1. **Friedman** Test ($k=3$), detect whether there exists significant difference among three models.
2. **Wilcoxon** test (pair-test) to decide which model(s) is the best model for making relatively valid predictions.

For example, Wilcoxon test on two hypotheses: $M^2 < M^1$, $M^2 < M^3$. If both result in *p-value* < 0.05 , then M^2 is the best model.



Inter-individual Difference

Which of the three proposals is best at predicting **inter-individual differences** in happiness?

managing to determine that clip C would make user U_1 happier than user U_2 .



Inter-individual Difference

Assumption:

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.



Inter-individual Difference

For each model M , detect whether there exists significant difference between any pair of subjects based on (*ScoreComputed* - *ScoreEvaluated*).

For each subject, collect all the values of (*ScoreComputed* - *ScoreEvaluated*) on each recommendation

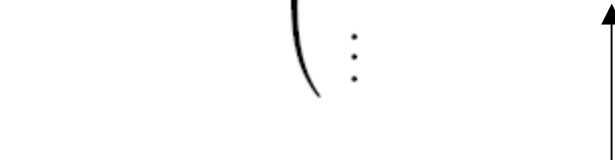
M	s_i	s_j
	\vdots	\vdots
	0.15	-4.4
	-0.08	5
		6.2

Inter-individual Difference

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

Mann-Whitney Test

To detect whether there exists significant difference between the two subjects.

$$\begin{pmatrix} & s_1 & s_2 & \dots & s_i & \dots \\ s_1 & & Y & \dots & N & \dots \\ s_2 & & & \dots & Y & \dots \\ \vdots & & & & & \\ s_i & & & \dots & Y & \dots \\ \vdots & & & & & \end{pmatrix}$$


Inter-individual Difference

We define the inter-difference score of M as :

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

$$\begin{pmatrix} s_1 & s_2 & \dots & s_i & \dots \\ s_1 & Y & \dots & N & \dots \\ s_2 & & \dots & Y & \dots \\ \vdots & & & & \\ s_i & & \dots & Y & \dots \\ \vdots & & & & \end{pmatrix}$$

where

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

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



Precision

Which of the three proposals achieves the highest modeling **precision**?

manages to more precisely predict the user's happiness after having watched a clip/series of clips.



Precision

- Same process as “Relatively Valid Prediction”
- Use all the values of $|ScoreComputed - ScoreEvaluated|$, not only these $ScoreComputed=1/10$.
- The model that achieves the highest precision will be the model that produce the least difference ($|ScoreComputed - ScoreEvaluated|$).



Thank you!