Lenovo Research 联想研究院 | 香港浸會大學 HONG KONG BAPTIST UNIVERSITY | KU LEUVEN

MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction

Yucheng Jin[1], Wanling Cai[2], Li Chen[2], Nyi Nyi Htun[3], Katrien Verbert[3]

[1]Lenovo Research, [2]Hong Kong Baptist University, [3]KU Leuven

jinyc2@lenovo.com, {cswlcai, lichen}@comp.hkbu.edu.hk, {nyinyi.htun, katrien.verbert}@cs.kuleuven.be

Abstract

Critiquing-based recommender systems aim to elicit more accurate user preferences from users' feedback toward recommendations. However, systems using a graphical user interface (GUI) limit the way that users can critique the recommendation. With the rise of chatbots in many application domains, they have been regarded as an ideal platform to build critiquing-based recommender systems. Therefore, we present *MusicBot*, a chatbot for music recommendations, featured with two typical critiquing techniques, user-initiated critiquing (UC) and system-suggested critiquing (SC). By conducting a within-subjects (N=45) study with two typical scenarios of music listening, we compared a system of only having UC with a hybrid critiquing system that combines SC with UC. Furthermore, we analyzed the effects of four personal characteristics, musical sophistication (MS), desire for control (DFC), chatbot experience (CE), and tech savviness (TS), on the user's perception and interaction of the recommendation in *MusicBot*. In general, compared with UC, SC yields higher perceived diversity and efficiency in looking for songs; combining UC and SC tends to increase user engagement. Both MS and DFC positively influence several key user experience (UX) metrics of *MusicBot* such as interest matching, perceived controllability, and intent to provide feedback.

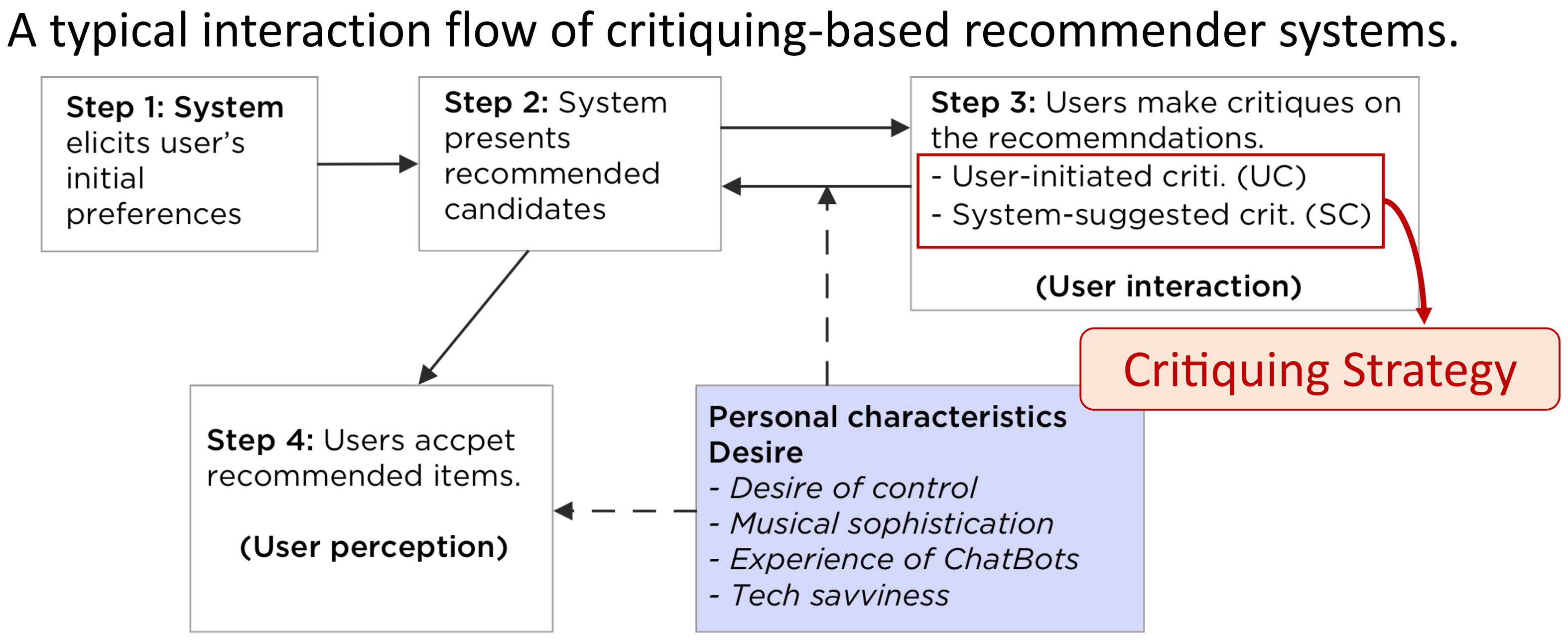
Background

Critiquing-based Recommender Systems

Users could make critiques on the recommended items to allow the system to iteratively update user preference model and provide users with desired recommendations.

Conversational Interaction

A typical interaction flow of critiquing-based recommender systems.



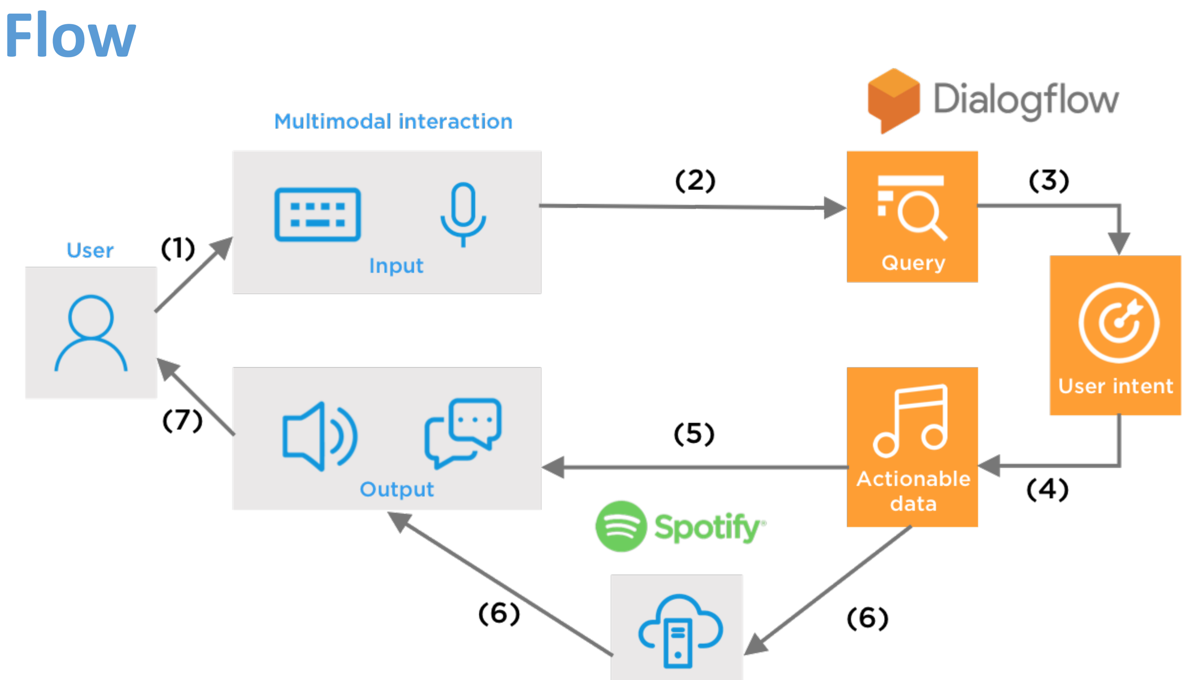
Research Questions

RQ1: Which critiquing setting, UC versus HC, is better suited for controlling music recommendations?

RQ2: Which personal characteristics (e.g. musical sophistication, desire for control, chatbot experience, and tech savviness) might influence user's perception and interaction of recommendations?

System Design

Work Flow



Algorithm

Recommendation Algorithm

The Spotify API generates recommendations based on three types of seeds, i.e., songs, artists, and music genres.

Critiquing-based Algorithm [1]

1. Critique pattern vector (e.g., {(energy, higher), (danceability, similar)})
2. Association rule mining algorithm (i.e., Apriori algorithm)
3. Multi-attribute utility theory (MAUT)
4. A set of personalized and diversified critiques

MusicBot UI

User Experiments

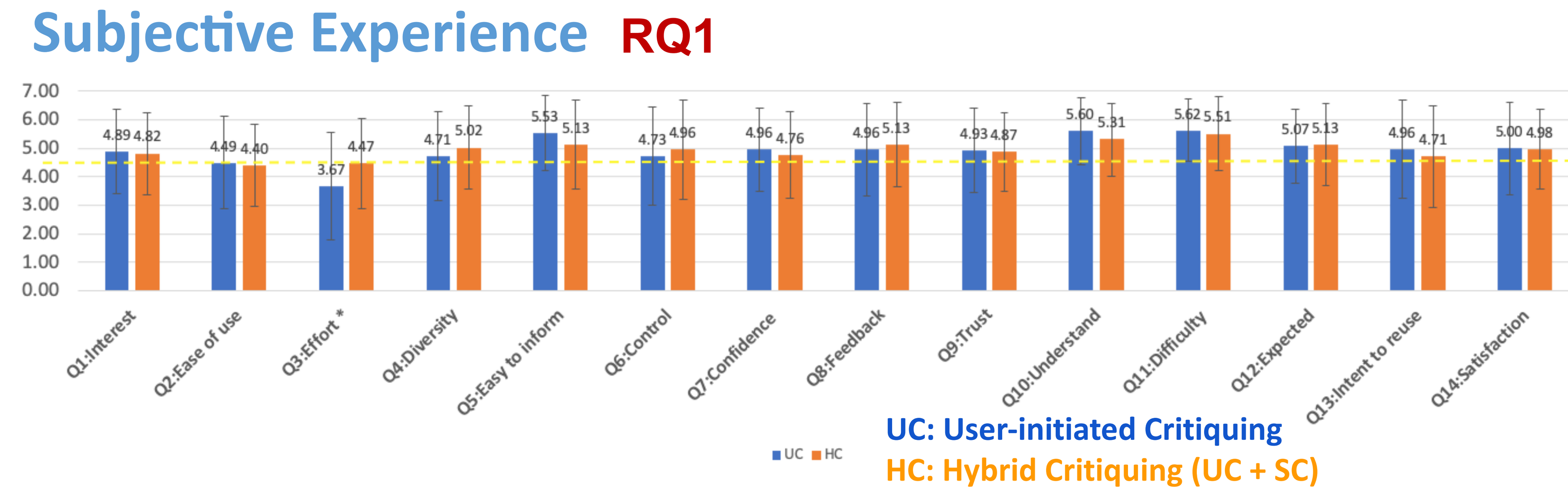
Online user study (45 participants, Age: 20-30(36), 30-40(6),41-50(1), > 50(2); Gender: Female = 19, Male = 26). A prize draw (each voucher: 10 USD)

Task: Find 5 songs in two scenarios and give ratings

Procedure: ①Watch Video Tutorial → ②Build User Profile → ③Pre-Study Questionnaire → ④Warm Up → ⑤Interact with MusicBot → ⑥Post-Study Questionnaire

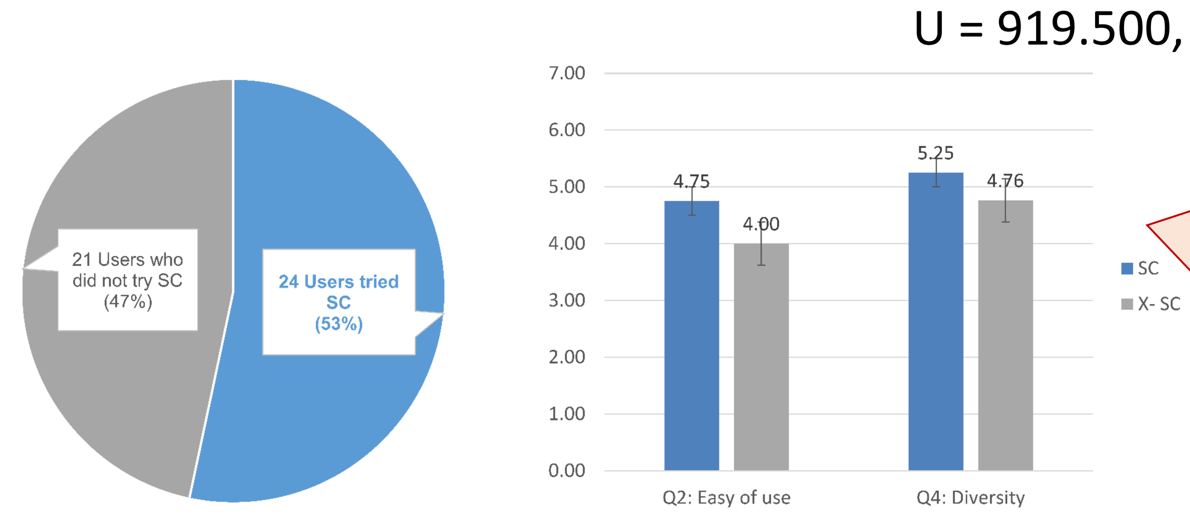
Results & Discussion

Subjective Experience RQ1



UC: User-initiated Critiquing
HC: Hybrid Critiquing (UC + SC)

- Users *positively* rated UC and HC in most of the UX metrics.
- Significant difference on effort of looking for songs (Mann-Whitney test, U = 919.500, p = .02)



Users who tried SC tend to *perceive higher ease of use and diversity than* users who did not try SC in HC.

Interaction Behavior

Interaction metrics	UC (mean,sd)	HC (mean,sd)
#Listened songs**	(10.67, 4.99)	(13.13, 6.09)
Rating (stars)	(4.05, 0.47)	(4.08, 0.44)
Completion time* (minutes)	(5.40, 4.19)	(6.98, 4.16)
#Turns(times)**	(12.29, 8.21)	(16.11, 9.35)
#Btn(times)***	(9.18, 3.38)	(12.64, 7.07)
#Typing(times)	(3.09, 4.78)	(3.07, 4.21)
#Voice(times)	(1.24, 7.90)	(0.71, 2.97)
#Words	(2.13, 1.92)	(2.28, 1.84)
#Unknown utterances	(1.78, 6.46)	(0.78, 1.80)

- HC leads to more dialogue turns, more completion time, more listened songs.

Personal Characteristics RQ2

PC	Q1:Interest	Q2:Ease of use	Q3:Effort	Q4:Diversity	Q5:Easy to inform	Q6:Control	Q7:Confidence
CE	0.15 (0.33)	0.14 (0.37)	0.07 (0.66)	0.03 (0.84)	-0.03 (0.86)	0.11 (0.46)	0.05 (0.73)
TS	-0.01 (0.98)	-0.13 (0.40)	**0.36 (0.02)***	0.10 (0.51)	-0.08 (0.59)	-0.19 (0.21)	-0.12 (0.43)
MS	**0.40 (0.01)***	0.25 (0.10)	-0.22 (0.14)	0.17 (0.26)	0.10 (0.53)	**0.31 (0.04)***	0.29 (0.05)
DFC	0.23 (0.14)	0.03 (0.84)	0.13 (0.41)	0.24 (0.11)	0.22 (0.15)	**0.35 (0.02)***	0.25 (0.10)

PC	Q8:Feedback	Q9:Trust	Q10:Understand	Q11:Difficulty	Q12:Expected	Q13:Intent to reuse	Q14:Satisfaction
CE	0.06 (0.70)	-0.01 (1.00)	-0.07 (0.65)	0.02 (0.88)	0.06 (0.69)	0.21 (0.17)	0.10 (0.52)
TS	0.16 (0.29)	0.07 (0.66)	-0.12 (0.42)	-0.04 (0.77)	0.04 (0.78)	-0.12 (0.42)	-0.19 (0.10)
MS	**0.55 (<0.001)*****	**0.37 (0.01)***	0.09 (0.57)	0.13 (0.38)	0.23 (0.14)	**0.31 (0.04)***	0.22 (0.15)
DFC	0.06 (0.68)	0.16 (0.29)	**0.30 (0.04)***	**0.38 (0.01)***	0.22 (0.14)	0.28 (0.06)	0.20 (0.19)

Correlation analysis between personal characteristics and user perception

MS(+): Interest matching, control, trust, intention to give feedback and reuse.

DFC(+): Control, easy to understand and use.

References

[1] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: Survey and emerging trends. UMUAI 22, 1-2 (2012), 125–150.