Rating-based Preference Elicitation for Recommendation of Stress Intervention

Helma Torkamaan
University of Duisburg-Essen
Duisburg, Germany
h.torkamaan@acm.org

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

In recent years, recommender systems have emerged as a key component for personalization in health applications. Central in the development of recommender systems is rating-based preference elicitation, based both on single-criterion and multi-criteria rating. Though its use has already been studied in various domains of recommender systems, far too little attention has been paid to preference elicitation in health recommender systems (HRS). The purpose of this paper is to develop a better understanding of this preference elicitation by studying the criteria that users consider when they rate a health promotion recommendation from HRS, and accordingly, to offer a design solution as a functional feedback model for mobile health applications. This paper investigates the user-perceived importance of various criteria, as well as latent factors for eliciting user feedback on the recommendations. It also reports the relationship of explanation and trust to the overall rating. By aggregating a list of all possible criteria, we further discover that not all criteria are equally important to users, and that the effectiveness of a recommendation plays a dominant role.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing → Human computer interaction (HCI)**; • **Applied computing → Health care information systems**.

KEYWORDS

Health recommender systems, Multi-criteria rating, Mobile health, Preference elicitation, Behavioral intervention

ACM Reference Format:
Helma Torkamaan and Jürgen Ziegler. 2019. Rating-based Preference Elicitation for Recommendation of Stress Intervention. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3320435.3324990

1 INTRODUCTION

Recommender systems are widely used tools in various domains such as e-commerce, tourism, and entertainment, but also have potentials in the health domain [4, 12] in the form of Health Recommender Systems (HRS)[1]. This could be particularly interesting for mental health promotion and behavior change, in which automatic personalization of health services and communication, including behavioral interventions [2], is still a challenge. Using HRS, interventional messages could be sent as personalized recommendations to users. However, in order to personalize the recommendations, one needs feedback from users on the perceived quality or relevance of the provided recommendations. This is a typical problem of preference elicitation in the recommender systems area.

Usually, explicit user feedback in recommender systems is captured using a bounded set of natural numbers. In case of a single-criterion rating, the feedback is captured as a single number that represents an overall rating of a recommended item, whereas in the case of a multi-criteria rating, the feedback could contain several important dimensions for the rating of an item. For example, in a hotel recommendation, these important dimensions or criteria could be the hotel's location, cleanliness, staff friendliness, etc. This extra information on the dimensions in multi-criteria rating can reveal more of the users' preferences and, as a result, may increase the quality of the recommendations. In the past 15 years, researchers [2, 6, 7, 9, 11, 13] have looked into the application of multi-criteria rating in order to capture a wider range of an individual's subjective opinion and to overcome the shortcomings of single-criterion rating for user preference elicitation. Adomavicius and Kwon [1] have mentioned several potentials for multi-criteria rating, such as enhancing the quality of the recommendations, providing additional information, presenting the complexity of user's preferences, and addressing multi-objective recommendations strategies or multiple performance criteria of a recommender system. Considering these potentials, multi-criteria rating seems to be suitable for HRS – a domain with possibly multiple objectives and various stakeholders.

Though a multi-criteria rating approach seems suitable, it is still an open issue which kind of criteria could be used to elicit users' preferences in HRS, and in particular, in the health promotion domain. In this paper, we consider a health promotion application that recommends behavioral interventions to users in order to improve their health and general mental well-being. This application focuses on stress and the reduction of its negative effect in daily life. Such an application in health promotion heavily relies on the user's subjective opinion, and therefore it is vital that explicit preference elicitation is well-explored.

UMAP '19, June 9–12, 2019, Larnaca, Cyprus

[1] HRS refers to the user of recommender systems in both medical and health promotion domains. It could also be a recommender system that uses health information and provides recommendations in other domains.
[2] Behavioral intervention is a combination of services, advice, and support for an individual in order to change existing behavior or shape a new one.

The aim of our research is to broaden the current knowledge of HRS by focusing on the design of a rating-based preference elicitation solution in a stress intervention application. Imagine the recommendation in Figure 1. How should the user preference elicitation be designed and implemented? What would an overall rating for such a recommendation capture? What does a high rating indicate? Which criteria do users consider in the first place? These questions are common challenges for all such applications that intend to use HRS. The rest of the paper is organized as follows: Section 2 discusses the background and our expectations, and explains the research questions; Section 3 explains the design of our user study; and Section 4 and Section 5, respectively, report the results of our analysis and our conclusion.

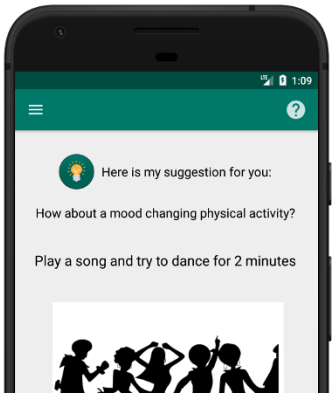

Figure 1: An example mockup (cropped) that shows a recommendation in the physical activity group. This recommendation is for reducing the negative effect of stress in daily life.

2 BACKGROUND AND OBJECTIVES

In order to study a multi-criteria rating-based preference elicitation, one needs to first know which criteria exist for such a rating and then determine the relative importance of these criteria to the users and in comparison with each other. In a preliminary study [14], we investigated which criteria exist in the first place. However, to the best of our knowledge, there existed no list of criteria that one could use to run the study; therefore, we created an initial list of 19 criteria. Inspired by the following theories: Fogg's behavior model for persuasive design [5]; Health Belief Model [10]; self-efficacy theory [3]; and goal-setting theory [8], we listed nine criteria: *cost, time-consumption, effort, difficulty, social deviance, emotional gain, effectiveness, easiness, and meeting the goals*. We added several other criteria, such as *interestingness*, *enjoyment*, and *meeting user preferences*, and we extended some of the items. In total, we considered 19 criteria [14].

As part of this study [14], independent open-ended questions captured users' opinions about health promotion recommendation in general as well as for various scenarios separately. The analysis of 296 open-ended answers resulted in 14 additional criteria beyond our initial considerations, which could not be coded into the initial 19 derived criteria. The mentioned study [14] was only based on the preliminary list of criteria and did not take into account any of the 14 obtained criteria which could be essential in explaining the overall recommendation quality. These results alone were insufficient for designing HRS applications, and therefore we conducted the following study, which addresses the shortcomings of the previous study and answers further questions. Based on the results, we obtained a list of 33 criteria that users may consider for the rating of a health promotion recommendation. These criteria are possibly able to capture a wider range of various user preferences and could also represent the factors one would consider while making a decision.

Using the list of criteria, we examined the following research questions. This study attempts to *(Q1)* investigate if any latent factor or structure can be extracted for a criteria-based rating of a health promotion recommendation. It also examines *(Q2)* if the overall satisfaction with a recommendation can be explained by its criteria-based rating. Since explanation and trust are two influential criteria for health domain, *(Q3)* this study also looks into the explanation and its effect on the acceptability, overall rating, and trust to a recommendation (implicitly). In this study, we also find out what the participants' behavior of choice would be when they receive a recommendation in various contexts and with different required effort *(Q4)*. This helps up to both learn when users most likely read or engage with a recommendation on their mobile devices and suggest a practical design solution in health applications.

3 METHOD

To answer our research questions, we conducted an online study to capture user preferences and opinions about the rating of the recommendation from a stress intervention application. The study started with the description of PAX – a mental health promotion application for stress intervention on smartphones – and its functions, i.e., giving recommendations. To investigate Q1, we asked participants to rate how important each of the given criteria is when they give an overall rating to a recommendation. In order to have a manageable list for the participants, of 33 criteria we only kept 21 and removed the redundant and extended criteria that were seldom mentioned[3]. For example, we removed *giving a good feeling* in favor of *emotional gain*. Both of these criteria represent the same thing and therefore could easily bore participants.

For Q2, we gave participants four randomly ordered recommendations using mockups, (see Figure 1). The recommendations offered four distinct suggestions (physical activity, positive thinking, social engagement, and mindfulness)[4] to capture a broad spectrum of possible interventions and also to prevent a potential bias associated with some specific domain-related recommendations. We asked the participants to first try following the recommendations

[3] 33 criteria are: effectiveness, emotional gain, giving a good feeling*, possibility to follow, personal preferences, likability, time, location, easiness, fulfillment of user goal, difficulty, social acceptance, effort required, time consumption, interestingness, enjoyment, interruption, cost, suitability for gender, novelty, explanation, trust, intrusiveness*, suitability for my mood*, relevance for me*, understandability*, need*, encouragement*, funniness*, easiness of reading*, empowerment*, transparency*, human-like*. Criteria with a star in front were omitted since they were either redundant or less frequently mentioned by the users in the open-ended answers [14].

[4] Sample recommendations for reduction of the negative effect of stress in daily life: *Positive thinking* asks participants to remember the most significant achievements they had and recall how happy they were at that time; *Mindfulness* is a four-step instruction to try relaxing and focusing on visual stimuli and being present at the moment; *Social engagement* encourages the participants to contact a friend and say something supportive or positive to them and explains the importance of seeking social support as one of the coping mechanisms.

to both engage them and help them focus on the questions and then give an overall rating to each recommendation. Participants then revisited the recommendation, and this time the survey asked for the rating of the following criteria for each recommendation: *effectiveness, emotional gain, difficulty, required effort, location, and time consumption*[5].

Q3 deals with trust in the source and explanation of a recommendation. In our study [14], participants mentioned *trust* and *explanation*; for example, subject 18 stated: *"Rating the recommendations exactly depends on the kind of recommendations. I do not trust the unknown new application until it is certified by the psychologists or the doctors."* Subject 45 mentioned that he would look for the scientific reason behind the recommendations when rating them. Several other subjects stated that they like to know the reason for receiving a specific recommendation and also why or how following such a recommendation could be effective. Observing such sentiments and also considering the importance of trust in the source and explanation for a health recommendation, we considered Q3. With a within-subject design, we added a textual explanation to two out of four recommendations (positive thinking and social engagement), and we asked participants' opinions about acceptability and validity of the recommendations (understanding the source of recommendation and their judgment on the recommendation being founded) as factors for determining user trust in the recommendations.

Learning more about user behavior, help us design better functional solutions for mobile health applications. To design user-centered systems that can personalize recommendations for each user, we should also consider that a user may receive a recommendation in various contexts. For example, they may be on the move, and therefore it is important to consider possible user behavior while receiving a recommendation in various scenarios or conditions. To address Q4, we gave users a task with five different scenarios that asked participants to imagine a stress reduction application on their phone that gives them recommendations. The task asked users to express their most likely behavior for each scenario by either choosing from five predefined answers or writing their answers. Each scenario depicted various everyday life situations or context that a user might be in as well as different required efforts, such as time consumption and difficulty of following the recommendation. These scenarios[6] help us to learn about possible user behavior that can be reflected in a practical solution.

4 RESULTS AND DISCUSSION

For the experiment, we recruited 95 participants (61 F, 34 M) with an average of ($M = 27.36, SD = 9.78$) years. The survey was designed in three languages, and the participants were from 25 different countries of origin with the majority (50/95) being from Germany.

Descriptive analysis of the participants' responses for the importance of the criteria shows a higher mean for *effectiveness* ($M = 8.87, SD = 1.40$), *emotional gain* ($M = 8.52, SD = 1.78$), *enjoyment* ($M = 8.39, SD = 1.61$), *liking* ($M = 8.38, SD = 1.58$), and *interestingness* ($M = 8.29, SD = 1.57$) in comparison with other criteria. These results indicate that for the rating of a recommendation, the *effectiveness, emotional gain, enjoyment, liking*, and *interestingness* of the recommendation are more important than other criteria such as those representing effort or user context aspects. One should consider that criteria such as *emotional gain, enjoyment, liking* and *interestingness* could also possibly (directly or indirectly) contribute to the perception of effectiveness (existing moderate correlations).

Table 1: Factor loadings of the importance of a criterion. Effectiveness <- $r = 0.314$ - >Effort/context

Criteria	Effect	Effort	Communality
Effectiveness	0.665		0.563
Emotion Gain	0.605		0.616
Fit Preferences	0.531		0.620
Enjoyment	0.767		0.463
Difficulty		0.743	0.434
Location	0.554		0.662
Time Consumption		0.652	0.573
Trust	0.565		0.699
Fulfill Goals	0.515		0.654
Liking	0.578		0.584
Social Acceptance		0.668	0.580
Effort Required		0.667	0.522
Cronbach's alpha	0.819	0.777	

To investigate Q1, we conducted an exploratory factor analysis using principal axis factoring using oblimin rotation. Results of the Kaiser-Meyer-Olkin measure of sampling adequacy was .74, well above the recommended value, and Bartlett's test of sphericity was significant ($\chi^2(210) = 617.54, p < .01$). Based on both the Scree plot from parallel analysis and MAP results, we extracted two factors. Following several steps, a total of nine items were eliminated because they had either lower factor loadings (< .5) or lower communalities (< .25). The two extracted factors count for 59.2% of the variance overall. The loadings of each of the criteria on these factors are presented in Table 1. The result of this factor analysis shows that a two-dimensional structure can explain the importance of the criteria in users' judgment of a recommendation. The first dimension explains the effect of a recommendation and the second one addresses the personalization of a recommendation based on user abilities. Looking back to the results of the descriptive analysis, we observe that none of the criteria in the effort dimension are among the top five criteria for the users. These results together highlight the importance of the criteria in the *effect* dimension for eliciting user preferences.

In Q2, because of a positive correlation between the overall rating of a recommendation and its criteria-based rating ($.25 < \tau b < .59, P < .01$), it was possible to carry out a hierarchical multiple regression analysis to observe if the overall rating of a recommendation can be predicted from its criteria-based rating. Since there is a possibility that criteria-based rating for different recommendation leads to varying predictions of the overall score, we considered the moderating effect by adding dummy variables representing each

[5]We selected these criteria based on the result of the previous study [14] considering at least two criteria that the users could judge as representatives of each possible group of criteria: effort, effect, and context. Additionally, other criteria were included for Q3.
[6]For example, we asked for the most likely behavior of the participants when they are *very busy*, i.e., the location or time is not suitable, namely when they are in a meeting, classroom or in the gym. Another example is asking for the most likely behavior when the participants are busy; namely, they are sitting in public transport but the recommendation fits their location and asks for a mindfulness technique.

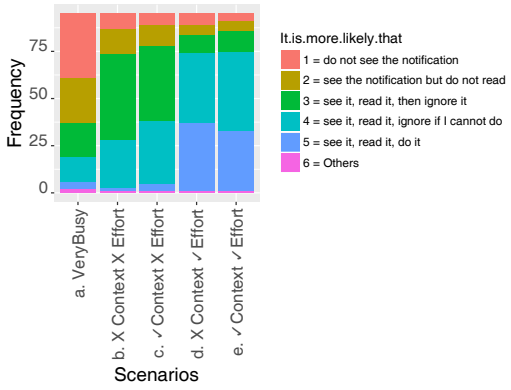

Figure 2: Users' self-reported behavior of choice when they receive a recommendation in various scenarios. Others show open-ended answers. In scenarios, suitable: ✓; wrong: x.

type of recommendation. We found statistically insignificant b-coefficient by including moderating effect. The results suggested a multicollinearity (Variance Inflation factor < 2.5) and, subsequently, we used the step-wise method. The regression model gives a significant regression equation ($F(4, 374) = 105.692, p < .000$), with an R^2 of .531. *Effectiveness* was selected for entry into the analysis first and explained 47.4% of the variance in overall rating. Following four steps at the end, the four predictors of *effectiveness, emotional gain, location*, and *required effort* account for 53.1% of the variance in overall rating. The slope of the regression in the final model on *effectiveness* b is .412, on *emotional gain* b is .231, on *location* b is .089, on *required effort* b is .089 and the intercept α is .609. Besides, we repeated the regression analysis for each recommendation separately with a step-wise multiple regression for predicting the overall rating. For all recommendation, a significant regression equation was found, and similar to previous analysis, *effectiveness* was found with the highest b value for the slope of regression. *Effectiveness* regardless of the type of recommendation seems to be the best predictor of the overall user satisfaction compared to other criteria.

To assess Q3, we captured participants' self-rated answers to acceptability (A) of the recommendations. As discussed previously in the method section, we also asked participants' opinions about the recommendation being founded (F) and if they know the source of the recommendation (S). In all of the recommendations, there were significant Kendall's tau-b correlation between the overall rating of a recommendation and A ($.35 < \tau b < .64, P < 0.01$). However, such a correlation was not consistently observed for S or F variables. We conducted a one-way, repeated measures ANOVA to compare the effect of recommendation type on A, F, and S. Repeated measures ANOVA did not reveal any statistically significant results related to A nor F. However, there was a significant effect of the recommendation type on S. The correlation Kendall's tau-b correlation between overall rating and S variable was insignificant. Therefore, although there could possibly be an effect between the level of explanation and the users' perceptions of the validity of the source, one would need further research to investigate if such an effect influences the user satisfaction of a recommendation. As a result, considering the limitations of our study, *explanation*, and *trust* seems to have no statistically significant influence on the users' overall ratings.

Figure 2 shows the results of Q4. Five different scenarios depict various everyday life conditions in which a user might be. In scenario (a), a user receives a recommendation when she is very busy. In scenario (b), the recommendation comes in the wrong context (time or location), and the required effort for following it is too high. Scenario (c) is a situation where the user is free and the context is right, but the effort of following the recommendation is not appropriate, such as being difficult or time-consuming. Scenario (d) describes a condition where the context is not right, but the effort required is adapted to fit the current context of the user. Finally, scenario (e) is a situation where both the required effort and context are suitable. As Figure 2 shows, when the required effort for following a recommendation is appropriate, it is more likely that the majority of our participants will read and follow the recommendation. However, when the effort is either not appropriate, too high, or does not fit the current context, users would probably not engage with the recommendation. Therefore, the required effort seems to be influential in the user's decision-making for following a recommendation. User context appears to influence whether a user sees, reads, or also depending on the effort, ignores a recommendation.

The results of Q1-4 provide a pathway for designing a health promotion application. While there could be various criteria characterizing the subjective user preferences, not all these criteria are equally important to the users. *Effectiveness, emotional gain*, and *enjoyment* are the most important criteria to the users and represent the effect that a user may perceive from a recommendation and could also determine if a user is satisfied with a recommendation. It is, therefore, crucial to consider such an effect for the personalization and design of HRS. Effect and effort are two latent factors representing an underlying two-dimensional structure of the criteria. Criteria in the effort dimension, however, seem to be necessary for the user's decision-making to engage with a recommendation possibly rather than their satisfaction with it. It is very likely that the users ignore a recommendation that is too difficult to follow. This could also be the case for a recommendation that may come in the wrong context. Which criteria to consider at the end depends on the objectives of HRS, and one can benefit from a combination of both explicit and implicit approaches for the preference elicitation.

5 CONCLUSIONS AND OUTLOOK

Our work started with an attempt to discover criteria for the multi-criteria rating of a health promotion recommendation. Following the research questions, we summarized a list of possible criteria and their relative importance. We also derived an underlying structure for the rating of a health promotion recommendation and found simple models explaining the overall rating of a recommendation. We found that users' evaluations of a recommendation most likely reflect its effect either directly (i.e., effectiveness), or indirectly (e.g., through emotional gain). We recommend that if the design decision is to use a rating-based preference elicitation, one should consider that while user feedback most likely shows the effectiveness of the recommendation rather than other criteria, it does not mean that the context/effort factors are not important. However, context/effort factors would be of particular importance for user engagement. Therefore, instead of an explicit user feedback elicitation, it seems suitable that such criteria would be considered implicitly in design.

REFERENCES

[1] Gediminas Adomavicius and YoungOk Kwon. 2015. *Multi-Criteria Recommender Systems.* Springer US, Boston, MA, 847–880. https://doi.org/10.1007/978-1-4899-7637-6_25

[2] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Trans. Inf. Syst.* 23, 1 (Jan. 2005), 103–145. https://doi.org/10.1145/1055709.1055714

[3] Albert Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84, 2 (1977), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

[4] André Calero Valdez, Martina Ziefle, Katrien Verbert, Alexander Felfernig, and Andreas Holzinger. 2016. *Recommender Systems for Health Informatics: State-of-the-Art and Future Perspectives.* Springer International Publishing, Cham, 391–414. https://doi.org/10.1007/978-3-319-50478-0_20

[5] BJ Fogg. 2009. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09).* ACM, New York, NY, USA, Article 40, 7 pages. https://doi.org/10.1145/1541948.1541999

[6] Dietmar Jannach, Fatih Gedikli, Zeynep Karakaya, and Oliver Juwig. 2012. Recommending Hotels based on Multi-Dimensional Customer Ratings. In *Information and Communication Technologies in Tourism 2012.* Springer, Vienna, 320–331. https://doi.org/10.1007/978-3-7091-1142-0_28

[7] Dietmar Jannach, Zeynep Karakaya, and Fatih Gedikli. 2012. Accuracy Improvements for Multi-criteria Recommender Systems. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12).* ACM, New York, NY, USA, 674–689. https://doi.org/10.1145/2229012.2229065

[8] Edwin A. Locke and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist* 57, 9 (2002), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

[9] Nikos Manouselis and Constantina Costopoulou. 2007. Analysis and Classification of Multi-Criteria Recommender Systems. *World Wide Web* 4, 10 (2007), 415–441. https://doi.org/10.1007/s11280-007-0019-8

[10] Irwin M. Rosenstock. 1974. Historical Origins of the Health Belief Model. *Health Education Monographs* 2, 4 (1974), 328–335. https://doi.org/10.1177/109019817400200403

[11] Fernando Sanchez-Vilas, Jasur Ismoilov, Fabín P. Lousame, Eduardo Sanchez, and Manuel Lama. 2011. Applying Multicriteria Algorithms to Restaurant Recommendation. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '11).* IEEE Computer Society, Washington, DC, USA, 87–91. https://doi.org/10.1109/WI-IAT.2011.124

[12] Hanna Schäfer, Santiago Hors-Fraile, Raghav Pavan Karumur, André Calero Valdez, Alan Said, Helma Torkamaan, Tom Ulmer, and Christoph Trattner. 2017. Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health.* ACM, 157–161.

[13] Dharahas Tallapally, Rama Syamala Sreepada, Bidyut Kr. Patra, and Korra Sathya Babu. 2018. User Preference Learning in Multi-criteria Recommendations Using Stacked Auto Encoders. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18).* ACM, New York, NY, USA, 475–479. https://doi.org/10.1145/3240323.3240412

[14] Helma Torkamaan and Jürgen Ziegler. 2018. Multi-Criteria Rating-Based Preference Elicitation in Health Recommender Systems. In *Proceedings of the 3rd International Workshop on Health Recommender Systems (HealthRecSys'18) co-located with the 12th ACM Conference on Recommender Systems (ACM RecSys 2018) (CEUR Workshop Proceedings).* 18–23.