

Introduction to the First Adaptive System Evaluation Challenge

Abstract. The Fourth Workshop on Evaluation of Adaptive Systems launched the first “evaluation challenge”. The challenge concerns an adaptive system, which recommends sequences of music clips to groups of users. Focusing on a real world problem, the challenge aimed to foster the development of innovative evaluation designs as well as encourage controversial discussion during the workshop. Participation in the challenge entailed proposing an empirical evaluation design purposely created to answer specific design questions regarding the system’s modeling component. This chapter describes the task and participation requirements given to the participants. The following two chapters contain the two submissions received.

1. Introduction

The Evaluation Challenge is about an adaptive system with a need for evaluation. Focusing on a real world problem, the challenge aimed to foster the development of innovative evaluation designs as well as encourage controversial discussion during the workshop. Potential participants were required to submit a short proposal on how this system can be evaluated. The proposals received were presented during the workshop, and are included in these proceedings. Workshop participants discussed the presented proposals in the context of a dedicated workshop track. The subsequent sections contain the detailed system description and express evaluation requirements that proposals were asked to address.

2. The System

2.1. Description

The system to be evaluated is a recommender system. The recommendation domain is music clips. Specifically, the system recommends sequences of music clips, either for individual users, or groups of users (the later being the most typical use case of the system, and the one addressed by this challenge). Recommendations are based on models of individual user preferences in relation to individual clips. The primary goal of the adaptive component of the system is to recommend a sequence of clips that will achieve reasonable levels of satisfaction for all members of the group, throughout the sequence. The rest of this section presents certain aspects of the system in more detail.

2.1.1. The domain

The system's application domain is music clips. The following information is available to the system for each clip: (a) performing artist(s); (b) compilation(s) in which the clip has appeared; (c) music genre(s); (d) recording company. This information is available for all clips in the system's database. The system has no way of further analyzing clips or locating / deriving additional information about them.

2.1.2. User modeling

The system is capable of modeling the preferences of each individual for any music clip. Primary input for the models is provided in the form of ratings from 1 to 10 for each music clip for each individual. The system uses that input to perform a hybrid of content-based and collaborative filtering-based modeling. The content-based part identifies generalizable "patterns" in the user's preferences (e.g., preference for a particular music genre). The collaborative filtering-based part performs dynamic user clustering based on explicit clip preferences and inferred general preferences, and follows traditional approaches in enriching individual user models and maintaining aggregate virtual models for clusters.

Recently, the system has been extended to also model how happy each individual is as a consequence of having seen the clips so far. This is the main aspect of the system that the challenge addresses. It is the intention that the happiness of individuals be used as input for an improved selection algorithm (see next section for an overview of the system's recommendation algorithms).

Based on literature, the following assumptions have been made to underlie the happiness modeling

- A1. Mood impacts evaluative judgment: when people are in a good mood, they evaluate more positively.
- A2. People's affective forecasting can change their actual emotional experience: if you expect to like something, then you might end up liking it more than if you did not have any expectations (this is called assimilation).
- A3. Emotional reactions become less intense with time: happiness wears off.
- A4. The difference between a rating of 9 and 10 might feel higher than the difference between a 5 and a 6.
- A5. Mood cannot be of unbounded intensity.
- A6. Actual feelings experienced differ from those reported retrospectively.

Initially, when no clips have been viewed yet, the happiness of each individual is modeled as zero: $Happiness(<>) = 0$. The happiness of an individual who has viewed item i after already having viewed a sequence $items$ is modeled as a function of their happiness with sequence $items$, and the impact on their happiness of new item i . There are three proposals for this modeling (with $0 \leq \delta \leq 1$):

1. $Happiness(items + <i>) = \delta * Happiness(items) + Impact(i)$
2. $Happiness(items + <i>) = (\delta * Happiness(items) + Impact(i)) / (1 + \delta)$
3. $Happiness(items + <i>) = \delta * Happiness(items) + Impact(i, \delta * Happiness(items))$
with $Impact(i, s) = Impact(i) + (s - Impact(i)) * \epsilon$, for all s and $0 \leq \epsilon \leq 1$

Where for item i : $Impact(i) = \begin{cases} (Rating(i) - 5.5)^2, & \text{if } Rating(i) \geq 5.5 \\ -(Rating(i) - 5.5)^2, & \text{if } Rating(i) < 5.5 \end{cases}$

and $\text{Rating}(i)$ is the inferred rating for a given user for item i .

Multiplying $\text{Happiness}(\text{items})$ with δ is done because of assumption A3. The quadratic definition of $\text{Impact}(i)$ is because of assumption A4. Dividing by $(1 + \delta)$ in proposal 2 is to use some kind of average, rather than summation, and is partly done because of assumption A5. The use of $\text{Impact}(i, \delta * \text{Happiness}(\text{items}))$ in proposal 3 is because of assumptions A1 and A2. Epsilon (ϵ) in proposal 3 models the extent to which the mood a user is already in influences that user's evaluative judgment (assumption A1): with $\epsilon = 0$ there is no such influence, with $\epsilon = 1$ the influence is so immense that the new clip cannot have any effect on the user's mood.

2.1.3. Recommendation algorithm

The system uses a selection algorithm to determine what item to show next (i.e., what is the next most suitable item to place in the clip sequence), based on the preferences of the individuals in the group. At the moment, it uses a Multiplicative Utilitarian Selection Algorithm (MUSA), which basically multiplies the (inferred) preference ratings of individuals, to arrive at the preference of the group as a whole.

The aim of the system is to keep the group as whole happy, ensuring that everybody in the group remains reasonably happy throughout the sequence. Based on the recent additions for happiness modeling, as described in the previous section, a new selection algorithm is being developed. In broad terms, this algorithm is based on MUSA, but excludes items that might bring an individual's happiness below a certain threshold from being added to the list.

2.2. Evaluation Goals

The evaluation aims to provide answers to at least one of the following questions:

- Which of the three proposals for modeling happiness (see previous section) 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)?
- 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 U1 happier than user U2)?
- Which of the three proposals achieves the highest modeling precision (i.e., manages to more precisely predict the user's happiness after having watched a clip / series of clips)?

There is only one restriction that the evaluation has to observe, namely it has to be of an empirical nature, involving end users. It is also perfectly acceptable if your evaluation design comprises more than one components, with only one of them being of empirical nature.

3. The Challenge

Challenge entries are, in essence, proposals of how the described system should be evaluated. Specifically, entries were required to explicitly:

- Propose how the described system, and in particular the happiness modeling and its underlying assumptions, can be evaluated (see the specific questions in the previous section). This should include a full description of one or more empirical designs: sampling, setting and material used, treatments (if applicable), dependent and independent variables, and analysis.
- Discuss what the main difficulties for this evaluation are, and how you propose to overcome them.

For the purpose of this challenge it can be assumed that the preference modeling has been shown by a previous, dedicated study, to be quite accurate.

4. Summary

The challenge has been widely advertised and care has been taken to keep the barrier for participation as low as possible. Nevertheless, only two challenge entries were received. Both entries were reviewed by the workshop's program committee and accepted for presentation at the workshop. Revised versions of the entries are included in these proceedings.

The workshop participants discussed the merits and shortcomings of the entries. An amalgamation of the proposals received, coupled with the comments made by workshop participants during the respective discussions, will form the basis for the evaluation of a real-world system that is very close to the one described for this challenge (the evaluation will be overseen by members of the workshop's organizing committee). The proposers will be included as co-authors on any publications that are based on the evaluation proposal.