



Fourth Workshop on the Evaluation of Adaptive Systems

Held in conjunction with the
10th International Conference on User Modeling (UM'2005)
Edinburgh, UK, July 24-30 2005

<http://www.easy-hub.org/hub/workshops/um2005/index.html>

First Adaptive System Evaluation Challenge

1. Introduction

In this Evaluation Challenge, we present you with an adaptive system, with a need for evaluation. Participation is easy: write a short proposal on how this system can be evaluated. This can be as short as one page. The best proposals will be presented during the workshop, and published in the proceedings. Workshop participants will discuss the presented proposals and vote a winning entry. This is an ideal way to contribute to the workshop, if you do not have your own evaluation to report on. Also, an ideal way to share your evaluation expertise, and perhaps be recognized for it by winning! We hope that this challenge will lead to animated interactions during the workshop, focussing the discussion on a real world problem.

News and updates on the challenge can be found at the [“Evaluation Challenge” section of the workshop’s web site](#). If you have questions regarding the challenge, or simply want to participate in related discussions, please visit the dedicated discussion forum [Fourth EAS Workshop \(UM2005\) - Evaluation Challenge](#) accessible from the [Easy-Hub fora page](#).

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.

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 analysing clips or locating / deriving additional information about them.

User modelling

The system is capable of modelling 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 modelling. The content-based part identifies generalisable “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 modelling

- A1. Mood impacts evaluative judgement: 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 modelled as zero: $Happiness(<>) = 0$. The happiness of an individual who has viewed item i after already having viewed a sequence $items$ is modelled 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 modelling (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 $Rating(i)$ is the inferred rating for a given user for item i .

Multiplying $Happiness(items)$ with δ is done because of assumption A3. The quadratic definition of $Impact(i)$ 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 $Impact(i, \delta * Happiness(items))$ 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 judgement (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.

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 modelling, 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 modelling 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 modelling 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

3.1 Requirements

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

- Propose how the described system, and in particular the happiness modelling 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.

In designing the evaluation activities, you can assume that the preference modelling has been shown by a previous, dedicated study, to be quite accurate.

3.2 Process

Before the workshop:

- Prepare and submit an entry following the guidelines and using the material in this document (deadline: March 7th).
- Proposals to be presented at the workshop will be selected by the workshop's programme committee on the basis of the extent to which they address the questions put forward in section 2.2 above (at least one question needs to be addressed), completeness of the proposal (e.g., experimental design, materials to be used in experiment), discussion of interesting evaluation difficulties, and originality of the approach. (notification of authors: April 7th).
- The authors of selected proposals will have to deliver a camera-ready version of their proposal for inclusion into the workshop's proceedings (deadline: May 9th).

During the workshop

- The selected proposals will be presented during a dedicated workshop track.
- The workshop participants will discuss the merits and shortcomings of different entries, with the assistance of an "evaluation challenge panel" made up by members of the workshop's programme committee.
- Following the discussion session, workshop participants will be invited to participate in a voting session that will identify the "winning" entry. The winner(s) will be declared at the end of that session.
- The winner(s) will also be announced as part of the report on the workshop's outcomes to be delivered during the main UM conference.

After the workshop

- The winning proposal 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 organising committee).
- The winner(s) will be included as co-author(s) on any publications that are based on the evaluation proposal.

3.3 Submitting an Entry

The submission process for challenge entries is the same as for papers intended for the main track of the workshop. This applies to deadlines (included below for easy reference), formatting requirements and the actual submission of the file containing the entry. Note that, unlike papers intended for the main track of the workshop, challenge entries are limited to a maximum of 3 pages (note that submissions with fewer than three pages are also welcome). Refer to the ["Submission" section of the workshop website](#) for additional details.

Important dates

March 7th, 2005: Submission of entries

April 7th, 2005: Notification of authors

May 9th, 2005: Delivery of camera-ready copy

July 24th or 25th, 2005: Workshop day; the conference lasts from the 23rd to the 29th