User Model User-Adap Inter
DOI 10.1007/s11257-010-9079-z

ORIGINAL PAPER

Using affective parameters in a content-based recommender system for images

Marko Tkalčič · Urban Burnik · Andrej Košir

This is an author-created version of the paper originally published in User Modeling and User-Adapted Interaction (2010), Volume 20, Number 4, Pages 1-33,
DOI:10.1007/s11257-010-9079-z
The final publication is available at http://www.springerlink.com/content/0924-1868/

Received: 15 January 2010 / Accepted in revised form: 10 September 2010

Abstract There is an increasing amount of multimedia content available to end users. Recommender systems help these end users by selecting a small but relevant subset of items for each user based on her/his preferences. This paper investigates the influence of affective metadata (metadata that describe the user's emotions) on the performance of a content-based recommender (CBR) system for images. The underlying assumption is that affective parameters are more closely related to the user's experience than generic metadata (e.g. genre) and are thus more suitable for separating the relevant items from the non-relevant. We propose a novel affective modeling approach based on users' emotive responses. We performed a user-interaction session and compared the performance of the recommender system with affective versus generic metadata. The results of the statistical analysis showed that the proposed affective parameters yield a significant improvement in the performance of the recommender system.

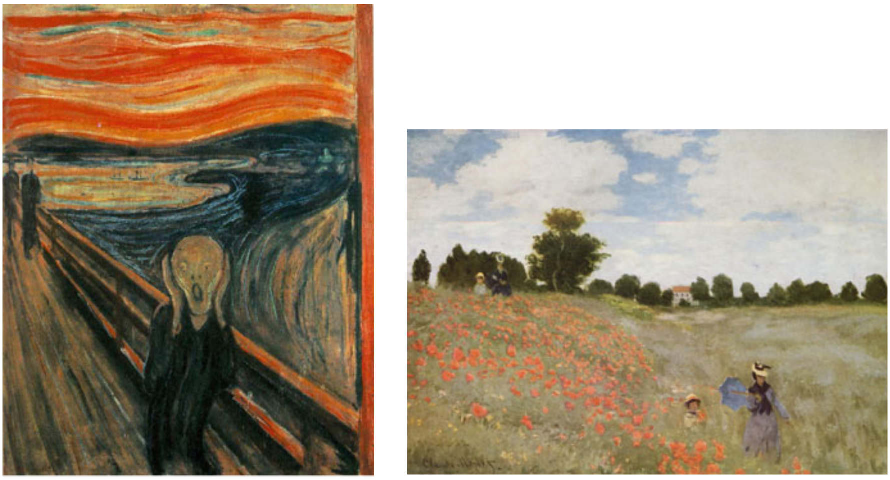
Keywords Affective modeling · Content-based recommender system · Emotion induction · IAPS · Item profile · Machine learning · Metadata · User profile · Valence-arousal-dominance

1 Introduction

The growing amount of multimedia content is making it hard for end users to find relevant content. The goal of recommender systems is to assist users by finding a small subset of relevant multimedia items for each user. There are several implementations of recommender systems, for example the TiVo system (Ali and Van Stam 2004) or the Netflix system (Koren et al. 2009), and current research (see surveys by Burke

M. Tkalčič (✉) · U. Burnik · A. Košir
Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: marko.tkalcic@fe.uni-lj.si

2002; Adomavicius and Tuzhilin 2005; Pazzani and Billsus 2007). Generally, there are two types of recommender systems: content-based recommender (CBR) systems and collaborative-filtering (CF) recommender systems. This paper deals with CBR systems. In such systems the items are annotated with metadata (e.g. actors, genre, subject matter, etc.) that are stored in a data structure called the *item profile*. A CBR system makes an estimation of the relevancy of an observed item based on the inclination of the user toward the item's metadata values. The user preferences are stored in the *user profile*. A crucial point in the design of CBR systems is the choice of fields (also referred to as *features*) in the item and user profiles. The chosen metadata fields must carry enough information to allow the CBR system to efficiently separate relevant items from non-relevant items for any observed user. In this paper we propose the usage of metadata fields containing emotional parameters in order to increase the precision rate of a CBR system. The underlying assumption is that emotional parameters contain information that accounts for more variance than generic metadata. We compare the performance of a CBR with the proposed metadata and a generic metadata CBR. The reason for the inclusion of emotive metadata fields lies in the assumption that end users differ in the target emotive state they are seeking when choosing multimedia content to view. For example, the famous paintings *Scream* by Edvard Munch and *Poppies Blooming* by Claude Monet (see Fig. 1) elicit different emotive states in viewers (anxiety and calmness, respectively, according to the authors of this paper). However both paintings have their respective admirers, which reflects the assumption that some people like paintings that cause anxiety and some people like paintings that induce calmness (there are probably people that like both as well as people who do not like either of them). Our hypothesis is that these individual differences can be exploited to yield better



(a) Scream (E. Munch) **(b)** Poppies Blooming (C. Monet)

Fig. 1 Two famous paintings that elicit different emotive states in viewers: anxiety and calmness, respectively, according to the authors of this paper (*source*: wikipedia.org) recommendations. To test this general hypothesis we built a CBR system for static color images and performed an experiment with users. We then evaluated how the inclusion of emotional parameters influences the performance of the recommender system.

1.1 Related work

The work related to affective recommender systems covers one or more areas of the affective recommender chain, which can be divided into four key steps: (i) emotion detection, which is a prerequisite for any affective modeling, (ii) item modeling, which deals with the description of items, (iii) user modeling, which models users' preferences and (iv) the recommender system, which uses the item and user models to compile personalised sets of relevant items for each user. Each of these steps can be tackled in a number of different ways, which are summarized in Table 1.

The majority of early recommender systems, especially movie recommenders, used metadata fields provided by content producers via databases like imdb.com for the description of items and users (see the work carried out by Basu et al. 1998; Pogačnik et al. 2005 and surveys by Adomavicius and Tuzhilin 2005; Burke 2002). Typical examples of such content producers' metadata are *genre*, *actors*, *subject matter*, etc. We will refer to this kind of metadata as *generic metadata* (GM).

However, in recent years the research work on recommender systems has started to follow *the affective computing* (Picard 2000) and *social signal processing* (Vinciarelli et al. 2009) paradigms. The work by González et al. (2004), Nunes et al. (2008), Arapakis et al. (2009), Joho et al. (2009), Shan et al. (2009) are examples of such systems. Affective computing is a broad area that deals with the detection and interpretation of human emotions and the generation of machine emotions in human–computer interaction. Social signal processing is a novel area that is more focused on the ability of computer systems to recognize human social signals. Various methods for the unobtrusive acquisition of human feedback, developed following these paradigms, are now making it possible to build applications and services that are based on affective and social information.

These paradigms caused a shift from GM to more human-oriented factors for the description of items and users. We will refer to metadata that are related to the users' personality and emotive responses as *affective metadata* (AM).

González et al. (2004) carried out one of the first investigations of affective modeling in recommender systems. They built the *smart user model*, a data structure based on users' emotional intelligence. However, they did not provide sufficient information to assess the success rate of their approach or to reproduce the experiment.

Nunes et al. (2008) modeled the users of a recommender system with two metadata sets: (i) *identity*, which was a set of self-reported personality metadata and (ii) *reputation*, which was calculated from other users' opinions of the observed user. They used this approach to find the nearest neighbors in a collaborative recommender system. They claimed to have achieved an accuracy of between 80 and 100% on a dataset where the users voted for one of three possible presidential candidates.

Table 1 Related work comparison based on the approaches taken in covering various steps of the affective recommender chain

Related work	Emotion detection	Item modeling	User modeling	Recommender system
Basu et al. (1998)	NA	IMDB data on movies	Affinity with items' metadata	Similarity map between user and item model
Pogačnik et al. (2005)	NA	TV Anytime hierarchical genre structure	Affinity with items' metadata	Similarity map between user and item model
González et al. (2004)	Emotional intelligence questionnaire	Non-affective, domain specific	A collection of key-value pairs describing demographic, subjective data and personality	Similarity measure as a weighted sum of item and user profile
Nunes et al. (2008)	Personality detection through the IPIP questionnaire	NA—collaborative filtering approach	Combination of personal and social opinions in the form of key-value pairs	NP
Arapakis et al. (2009)	Real-time detection of universal emotions	Keywords metadata	Ratings were assessed with click-through feedback and binary valence of the detected emotion	Support Vector Machine (SVM) ML technique
Joho et al. (2009)	Motion Units features mapped to three-level valence values	Low-level video and audio features and manual annotations	NP	NP
Shan et al. (2009)	Movie's induced emotion detected through the movie's music score low-level features	Induced emotion labels	NA	NA—the system describes an affective query approach rather than a personalized approach
Ioannou et al. (2005)	Detection of universal emotion classes and valence-arousal values with a neural network algorithm	NA	NA	NA
Proposed solution	IAPS dataset	Statistical moments of the induced emotion in the VAD space	ML techniques using supervised learning	ML

NA not applicable, *NP* not performed

Arapakis et al. (2009) developed a system that performs a binary classification (relevant/non relevant) of the consumed video items based on the video stream of the face of the user. Their approach employs a support vector machines (SVM) classifier that uses implicit feedback in the form of emotions detected through a camera and click-through data. Their results showed that the inclusion of affective feedback improves the performance of a recommender system. Similar work has been carried out by Joho et al. (2009); they developed a system that extracts affective labels of video clips from the facial expressions of viewers.

Shan et al. (2009) built a query ranking system for movies that extracts low-level features from the movie's audio track to detect the induced emotive state of the movie segment. The search query is composed of a list of desired emotive states and the system builds a list of ranked items. Their system does not model individual users according to their affective preferences, so no personalized recommendations are made.

The usage of AM for *affective modeling*, as described by Carberry and de Rosis (2008), is not limited to recommender or information retrieval systems. There are several sub-areas of affective computing that deal with the detection of emotions (Ioannou et al. 2005; D'Mello et al. 2008; McQuiggan et al. 2008; Yannakakis et al. 2008; Zeng et al. 2009; Caridakis et al. 2010) and affective modeling for adapting the user interfaces to the emotive state of the users (Porayska-Pomsta et al. 2008; Batliner et al. 2008; Conati and Maclaren 2009).

A direct comparison of the related work is not possible because this related work covers different aspects of the whole affective recommender chain. Furthermore, there are no publicly available datasets with affective metadata that could be used for a direct comparison of affective recommender systems. However, we provide a soft comparison in Table 1, where selected related work is compared based on the approaches taken in covering various parts of the affective recommender chain.

1.2 Problem formulation

Existing CBR systems based on GM do provide good results; however, they can be improved. The above-mentioned recommender systems that use AM do provide good results as well, but they cannot provide a direct comparison of both kinds of metadata on the same CBR system and dataset. Such a comparative study is needed in order to prove the hypothesis that AM carry more of the information needed to distinguish the relevant items from the non-relevant items than GM. Furthermore, there are several issues to explore regarding the usage of AM in CBR systems, like the influence of specific metadata fields and different algorithms on the performance of the CBR system.

The question that has not been answered so far is whether the inclusion of AM improves the performance of a CBR recommender system based on GM. We believe that it does, which leads us to formulate our hypothesis as follows.

H: A CBR system for images based on AM performs better than the same CBR system with GM.

The reasoning behind the hypothesis is that AM contain more information needed for the separation of relevant items from non-relevant items than GM, as illustrated

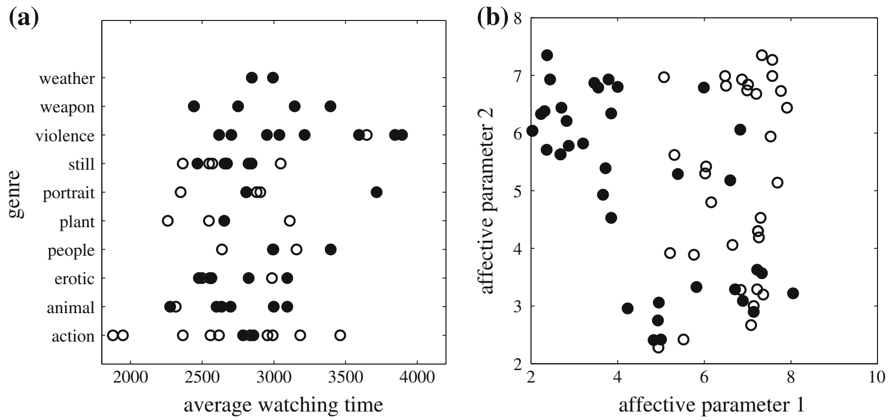


Fig. 2 Distribution of relevant (*white circles*) and non-relevant (*black circles*) image items for a single user in (**a**) a GM feature space (g and $\bar{t_w}$ as defined in Sec. 2.1.1) and (**b**) an AM features space ($\bar{v}$ and $\bar{a}$ as defined in Sec. 2.1.3). In ML terminology, AM carry more information for discriminating relevant items from non-relevant than GM because the AM feature space has a better ratio of between-class and within-class variance than the GM feature space

in Fig. 2. The calculation of item relevancy in CBR systems is done with machine learning (ML) algorithms, which require features with a high ratio of between-class and within-class variance in order to efficiently discriminate the classes (Hastie et al. 2001).

In order to validate the hypothesis we performed separate offline CBR experiments using AM and GM. We chose a subset of images from the IAPS database (Lang et al. 2005) as items for the users' consumption. We acquired the users' ratings in a dataset-acquisition session with real users. We designed a novel set of AM that consists of the first two statistical moments of the induced emotive responses in viewers in the *valence-arousal-dominance* (VAD) emotive space. We annotated all the images used with AM. We further annotated the images with a set of GM composed of the genre and the watching time. We performed an offline CBR experiment that yielded predictions of the binary ratings for all the items and all the users. The experiment was repeated for AM, GM and different ML techniques used in the CBR recommendation procedure. We compared the predicted binary ratings with the ground truth data acquired in the dataset-acquisition phase. We performed statistical tests on the confusion matrices given by the comparisons in order to see whether the differences in the means of the performances of the CBR were significant.

1.3 Organisation of the paper

The remainder of the paper is organized as follows. In Sect. 2 we describe our approach to modeling items and users with AM. We give a brief overview of how CBR works and describe the GM used in our comparative study. Then we provide a taxonomy of the emotive notations from which we derive the proposed affective modeling approach for items and users. In Sect. 3 we describe the experimental procedure employed; we provide arguments for the choice of the experimental approach and we describe the dataset acquired and argue the quality of the sample of users. Details about the evaluation methodology used are also given. In Sect. 4 we provide the results of the experiments according to the evaluation methodology. In Sect. 5 we discuss the results, identify the pending issues and provide the guidelines for future work. Section 6 provides the conclusions based on the work carried out.

2 Affective modeling of items and users

A generic CBR scenario is composed of a database of multimedia items (images in our case) and a set of users. We denoted the users that use the recommender system as U and the items available in the recommender system as H. For each user $u \in U$ the system is designed to separate the relevant items $H_R \subset H$ from the non-relevant items $H_{\mathrm{NR}} \subset H$, where $H_R \cup H_{\mathrm{NR}} = H$. Figure 3 shows a simplified example of how this is done in CBR systems. Each item h is described with the item profile $md(h)$, which is a set of metadata key-value pairs. The example item profiles from Fig. 3 have, besides the *id* (which is not a metadata field), the *title* and *genre* metadata fields: $md(h_1) = \{\text{Girl, Erotic}\}$, $md(h_2) = \{\text{Basketball, Sport}\}$ and $md(h_3) = \{\text{Kitchen, Still life}\}$.

The description of the user preferences is stored in the user profile $up(u)$ which is a data structure based on the user's past behavior (referred to also as *the usage history*).

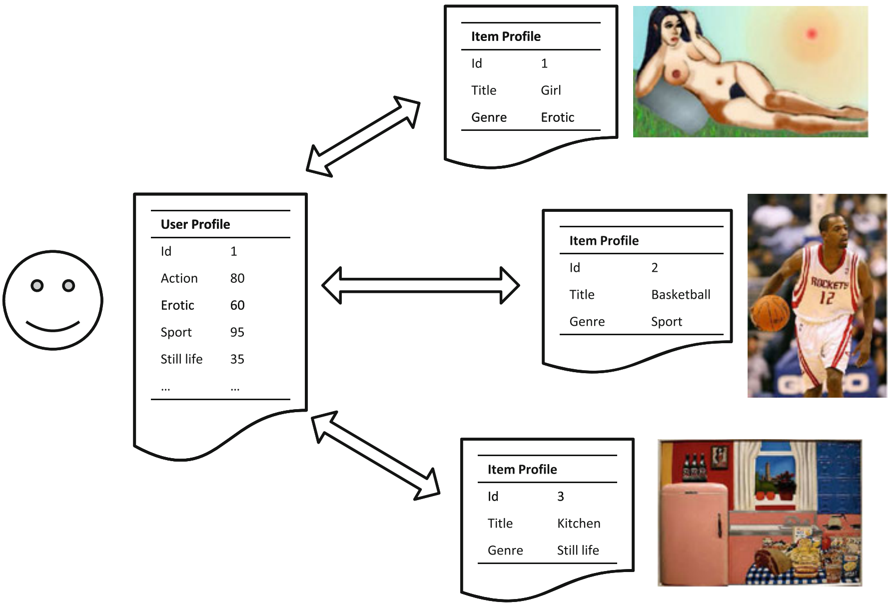


Fig. 3 Matching of the item profiles with the user profiles in CBR systems. In a generic CBR approach values in the item profile and related values in the user profile are matched with a similarity function. An aggregation algorithm classifies each item into one of the two classes: relevant or irrelevant for the observed user. In this example the items with ids 1 and 2 are classified as relevant, while item 3 is classified as non-relevant for the user with id 1

The user profile can have different forms (the form of the user profiles proposed in this paper depends on the ML algorithm used, as we will show in Sect. 2.2), but it generally reflects the metadata used in the item profile. The example in Fig. 3 shows the user profile as a list of affinities (scalar values between 0 and 100) of the observed user towards specific genre values that are to be found in the item profiles.

The recommender algorithm matches all the item profiles with the observed user profile. The output of the algorithm is a binary classification of all the items, as either *relevant* or *non-relevant* for the observed user. We denote these two classes as C_1 and C_0, respectively, and they form the set of possible binary ratings $\Omega = \{C_0, C_1\}$. The calculated classes are estimations of the binary ratings that the observed user u would give to any item h if she/he were to view them. We denote these estimations as $\hat{e}(u, h) \in \Omega$. The example profiles $md(h_1)$ and $md(h_2)$ in Fig. 3 yield high ratings $\hat{e}(u_1, h_1) = C_1$ and $\hat{e}(u_1, h_2) = C_1$, while the profile $md(h_3)$ yields a low estimated rating $\hat{e}(u_1, h_3) = C_0$.

In order to get to know the user preferences the CBR system needs to collect feedback information. The most common form of feedback collection is explicit feedback, where users explicitly express their opinions about items in the form of a rating. Ratings in recommender systems can be binary ratings (like it/don't like it) or Likert ratings (e.g. on a scale from 1 to 5, which are usually thresholded to binary ratings later on) (Adomavicius and Tuzhilin 2005). We denote the explicit binary rating given by user u to the item h with $e(u, h) \in \Omega$. When validating the performance of the recommender system the explicit ratings $e(u, h)$ represent the ground truth.

2.1 Item profile

In the example shown in Fig. 3 we used a simple item profile for illustration purposes. In this section we first describe the GM used for the comparative study. Then we discuss the models for describing emotions and we argue for the choice of model for our needs. Finally, we describe the proposed emotive parameters used as AM in the proposed modeling scheme for the items.

2.1.1 Generic metadata

The GM set used in our comparative study is composed of the genre g and the average watching time $\bar{t_w}$ of the item h. Both attributes are widely used in recommender systems (Adomavicius and Tuzhilin 2005; Pogačnik et al. 2005; Kim et al. 2005) and thus suitable for our comparison. The genre was set manually and was chosen from a set of ten available genres. The average watching time was calculated by averaging the watching times of all the users who have watched the item U_h. We denote the GM set as the double

$$\mathcal{A} = (g, \bar{t_w}) \tag{1}$$

where g stands for the genre of the item and $\bar{t_w}$ is the average watching time for that item.

2.1.2 A brief taxonomy of emotions

In order to make a sound choice of the AM we compared various models for describing emotions. The definition and description of emotions is a problem that has been known for a long time (Scherer 2005; Cowie et al. 2001). There are two main approaches for describing the emotive state of a user: (i) the *universal emotions model* and the (ii) *dimensional model*.

The universal emotions model is the consolidation of the work started by Darwin (1872) and is based on the observable features of the face. It describes each emotive state as a distinct state or a combination of distinct universal emotions. There is no unanimity as to which are the universal emotions. Cowie et al. (2001) use the Plutchik's (Plutchik 2001) wheel of eight emotions (joy, acceptance, fear, surprise, sadness, disgust, anger and anticipation) while Ekman (1999) defined a list of seven universal emotions, which have different observable facial features (neutral, anger, disgust, fear, happiness, sadness and surprise) and 11 additional universal emotions that do not exhibit facial-muscle changes. Several other sets of universal emotions have been defined or used (Scherer 2005; Schröeder et al. 2010; Shan et al. 2009). The wheel model proposed by Plutchik (2001) consists of four pairs of opposite basic emotions. He suggests that emotions are managed by a brain circuitry similar to the one that drives color perception.

On the other hand, the dimensional model, which was introduced by Mehrabian (1996) as the *pleasure-arousal-dominance* (PAD) space, describes each emotive state as a point in a three-dimensional space. The *pleasure* dimension has been referred to as *valence* by many authors (Posner et al. 2005; Villon and Lisetti 2006; Bradley and Lang 2007). Some authors, like Ioannou et al. (2005), refer to the *arousal* dimension as *activation*. In this paper we will refer to the dimensional model as the *valence-arousal-dominance* (VAD) space because our work relies on the dataset provided by Lang et al. (2005), which uses that terminology. The dimensions of the space are valence v (accounts for the pleasantness of the emotion), arousal a (accounts for the strength of the emotion) and dominance d (describes whether we are in control of our emotions or not). Posner et al. (2005) connected both models by introducing the *circumplex model*, which maps the universal emotions to the VA plane of the VAD space (see Fig. 4). In order to standardize the notation of emotions for use in computers, the W3C consortium is formalizing a markup language called EmotionML (Schröeder et al. 2010).

According to Scherer (2005) one of the divisions of emotions when describing emotive responses to stimuli is *aesthetic* or *utilitarian* and *intrinsic* emotions. Aesthetic emotions are produced by the appreciation of intrinsic emotions contained in the observed item. In the presented work we are modeling the emotions of end users while they are viewing digital items. The AM thus refer to the emotions that are induced in end users when they view digital items. It is important to note that we model the aesthetic emotions of viewers (e.g. *the picture made me happy*) and not the content's intrinsic emotions contained in the pictures' characters (e.g. *the picture shows a happy person*).

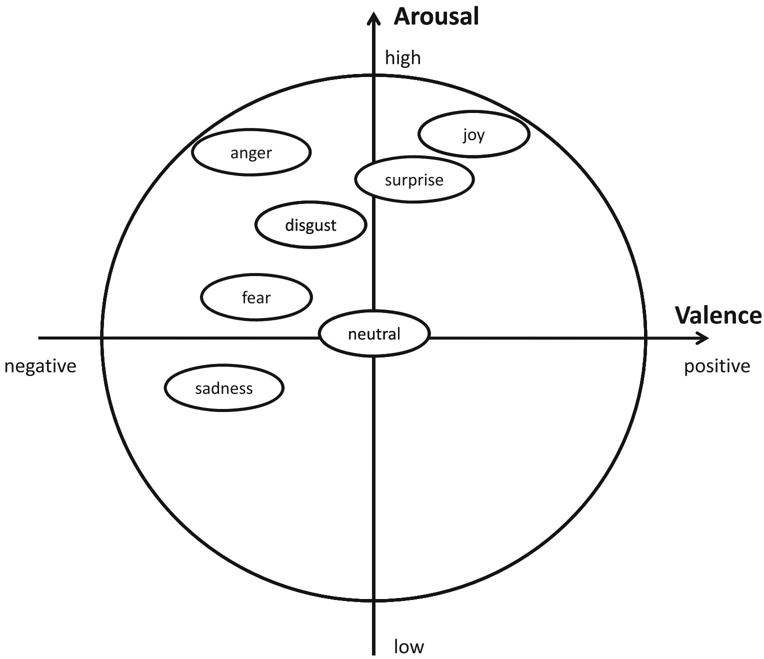


Fig. 4 The circumplex model of affect, which maps the universal emotions in the valence-arousal plane. The universal emotions depicted as distinct classes are in fact bell-shaped curves. Adapted from Pantic and Vinciarelli (2009)

2.1.3 The proposed affective metadata

Among the described models we chose to use the VAD model for the description of the emotive responses of users because it has a finer resolution than the coarse classes (labels) of the universal emotions model. The emotive quantum in our scenario is a single emotive state that has been induced in the user u by the image h. We denote this emotive response as $er(u, h)$. The emotive response is a triple of the scalar values valence, arousal and dominance $er(u, h) = (v, a, d)$. We denote the set of users who have viewed item h with U_h. The emotive responses of the users U_h form the set $ER_h = \{er(u, h) : u \in U_h\}$. We propose to use AM that include the emotive responses of many users. We thus use the first two statistical moments of the known emotive responses ER_h to an item. This yields the proposed set of AM in the form of the six tuple

$$\mathcal{V} = (\bar{v}, \sigma_v, \bar{a}, \sigma_a, \bar{d}, \sigma_d) \tag{2}$$

The underlying assumptions needed for the calculation of the statistical moments are (i) that each item h has been viewed by several users and (ii) that their emotive responses have been recorded, either directly through a questionnaire (like the Self Assessment Manikin developed by Lang et al. 2005) or in an unobtrusive manner, like the ones overviewed by Zeng et al. (2009).

Table 2 Example of a combined item profile with generic metadata $\mathcal{A}$ (genre g and average watching time $\bar{t_w}$) and affective metadata $\mathcal{V}$ (first two statistical moments of the induced emotion values v, a and d)

	Metadata field	Value
	Image id	1234
Generic metadata (GM)	g	Action
	$\bar{t_w}$	3198
Affective metadata (AM)	$\bar{v}$	3.12
	σ_v	1.13
	$\bar{a}$	4.76
	σ_a	0.34
	$\bar{d}$	6.28
	σ_d	1.31

Both metadata sets can be combined into a larger one, denoted with $\mathcal{A} \times \mathcal{V}$. Table 2 shows an example of an item profile composed of both metadata sets.

2.2 CBR and user profiles

In contrast to the most used user modeling approach, as employed by Pogačnik et al. (2005) (we used a simplified version of it for illustration purposes in Fig. 3) or surveyed by Adomavicius and Tuzhilin (2005), we propose to model the users' preferences toward emotive states with a user profile—a data structure that is the result of training the ML algorithm based on past ratings (e.g. Fig. 5 shows the data structure of a trained decision tree ML algorithm). In supervised learning a ML algorithm takes as its input a training set of data that consists of several records containing vectors of feature values with their respective class values. The ML algorithm learns the relations between the features and the classes and stores these relations in a classifier-dependent data structure that represents the learned knowledge. The algorithm then uses the classifier data structure to classify the new feature vectors into classes.

We use ML techniques with supervised learning to build the user model. The ML algorithm takes the past binary ratings $e(u, h)$ of an observed user u as the training set to learn the patterns of the user's preferences. The parameters of the ML algorithm corresponding to a specific user are the user model. Such a user model is not necessarily human readable. If a tree classifier is used, like in Fig. 3, the profile is human readable. If some other classifier is used, like the SVM, then the user profile is an unreadable set of support vector parameter values. After the training phase the ML algorithm takes the user model and applies it to non-rated items to classify them into binary rating estimates $\hat{e}(u, h)$ for the observed user.

We assume that any observed user u has been using the CBR system for some time so a sufficient quantity of explicit ratings $e(u, h)$ are available for inferring the preferences of the user. By making this assumption we avoid any discussion of the *new user problem*, which is beyond the scope of this paper and has been dealt with elsewhere (Rashid et al. 2002; McNee et al. 2003; Adomavicius and Tuzhilin 2005;

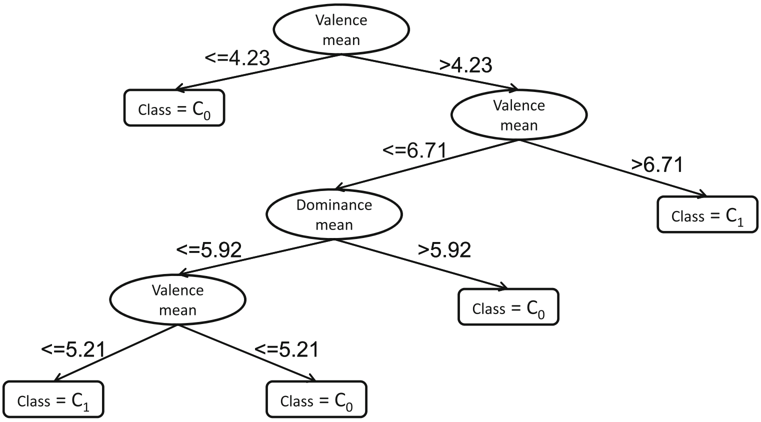


Fig. 5 Example of a user profile when the C4.5 tree classifier is used for inferring the user's preferences. The user preferences are learned by the ML algorithm and stored in a data structure called the user profile. Depending on the machine learning technique used the structure of the user profile can take various forms

Table 3 The usage history of an observed user contains the GM (g and average watching time $\bar{w_t}$) and AM (first two statistical moments of the induced emotion v, a and d) of an item along with the explicit ratings e (C_1 represents a relevant item, C_0 an irrelevant) for the viewed images. The empty spaces in the last column indicate that the observed user has not viewed and rated the item

Image id	g	$\bar{t_w}$	$\bar{v}$	σ_v	$\bar{a}$	σ_a	$\bar{d}$	σ_d	e
10	Action	2435	6.2	1.8	6.2	2.6	5.7	1.8	C_1
11	People	3487	6.2	0.5	3.7	0.8	3.1	2.0	C_0
12	Still	1667	6.4	1.1	6.5	0.7	5.1	1.6	
13	Violence	4871	4.1	1.6	4.9	0.5	5.3	0.8	C_1
14	Still	3500	7.9	0.8	7.7	0.9	4.8	0.1	

Berger et al. 2007). Based on the user's past ratings the ML algorithm learns the user model and is able to calculate estimates of the unrated items $\hat{e}(u, h)$.

When learning the user model of the observed user u the ML algorithm takes as the training set the items' metadata values and the user's ratings $e(u, h)$. Table 3 shows an excerpt from the usage history dataset we used in our experiment (see Sect. 3), where items with ids 10, 11 and 13 form the training set of the ML algorithm because they contain the explicit ratings e. In the experimental part of this paper we evaluated four ML techniques: AdaBoost, C4.5, NaiveBayes and SVM (Witten and Frank 2005). The structure of the user model thus depends on the ML technique used. Figure 5 shows an example of a user model when the tree classifier C4.5 is used for inferring the user's preferences. Based on the user model the ML algorithm classifies the remaining items (those that have not been rated by the observed user) into relevancy classes, which we denote with the mapping $\delta : H \rightarrow \Omega$. It generates the rating estimations $\hat{e}(u, h)$. Based on these, the items are arranged into the sets of recommended items H_R and non-recommended items H_{NR}

$$H_R = \{h : \hat{e}(u, h) = C_1\} \tag{3}$$
$$H_{\mathrm{NR}} = \{h : \hat{e}(u, h) = C_0\} \tag{4}$$

3 Materials and methods

3.1 Experimental overview

The hypothesis under question is whether the proposed AM bring a significant performance improvement over the GM in an image CBR system. We also evaluated four different ML algorithms for the calculation of predicted ratings $\hat{e}(u, h)$. Furthermore, we evaluated the suitability of each of the proposed AM parameters as a metadata field for CBR systems.

First, we acquired the dataset, then we performed an offline simulation of the CBR and finally we tested the hypothesis. Figure 6 shows our experimental setup. Each user u was shown a set of image stimuli $\{h\}$ that induced a set of emotive responses

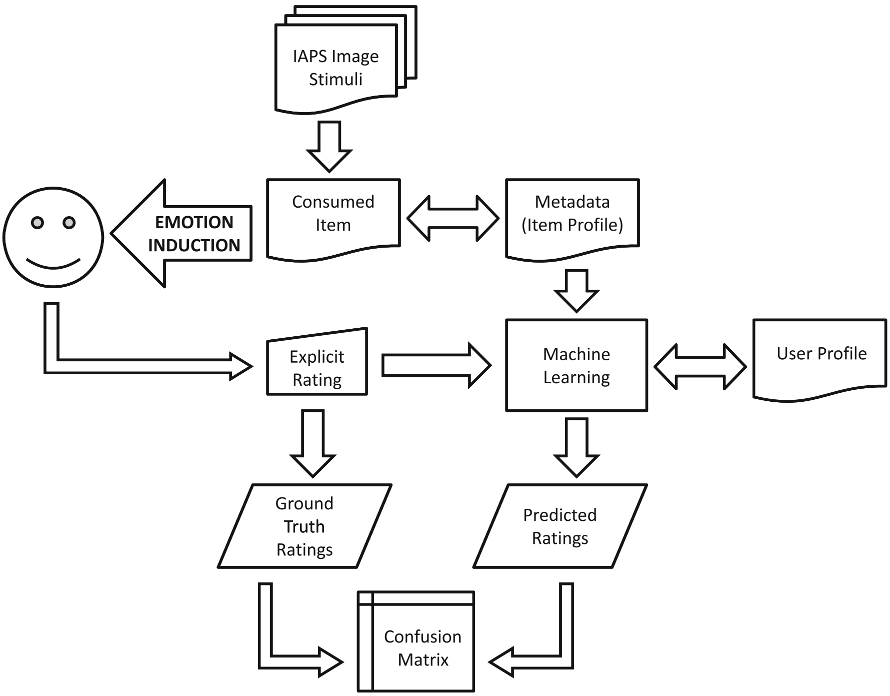


Fig. 6 Experimental setup. A sequence of images from the IAPS database is shown to the subject, which provides an explicit rating for each image according to the user task. An ML algorithm builds the user profile based on the explicit ratings and metadata contained in the item profile (generic and affective metadata). The user profile is used to calculate the rating predictions for non-rated images. Non-rated images are generated through a folding scheme and compared with ground truth data to yield the performance of the CBR in the form of confusion matrices for each user separately

$\{er(u, h)\}$. She/he gave explicit ratings $\{e(u, h)\}$ that represented the ground truth for the evaluation of the CBR. Based on the explicit ratings $\{e\}$ and item profiles $\{md(h)\}$ a ML algorithm built the user profile $up(u)$. The validation of the CBR for each individual user was made by calculating the predicted ratings $\{\hat{e}(u, h)\}$ with the ML algorithm. These ratings were compared with the ground truth ratings $\{e(u, h)\}$ and yielded the confusion matrices. We performed a statistical test to determine whether the confusion matrices yielded by the sets of metadata under observation (GM and AM) were significantly different. Furthermore, we calculated the scalar measures—precision, recall and F-measure to explain whether the significant differences achieved meant an improvement or not. Details are given in the following subsections.

3.2 Choice of the experimental technique

Our experimental design builds on the *emotion induction* technique (also referred to as *emotion elicitation*). The emotion induction technique is a well-known scientific approach in psychology, neuroscience and psychiatry (Lang et al. 2005; Bradley and Lang 2007; Coan and Allen 2007). According to Bradley and Lang (2007) it is being used to study disturbances in emotional development, to assess the physiological impact of stress, to determine the level of emotional impairment in brain-damaged patients and to construct more efficient treatments for fear, anxiety and depression. It has also been used in affective computing, especially in the development of emotion detection techniques (see Scheirer et al. 2002; Rottenberg et al. 2007; Lichtenstein et al. 2008; Zeng et al. 2009). These are implemented as classification problems where we need to have ground truth data for training and evaluating the system. As already stressed by Scheirer et al. (2002), the ground truth in an emotive human–computer interaction is a non-trivial problem. In emotion induction experiments the emotion has been induced but not confirmed. For example, the visual stimulus of a snake may cause fear in some people but not in others. The presence of noise in ground truth data implies uncertainty due to the uncontrolled nature of the experiment. The alternative would be to ask the user each time about her/his emotive state. This would bring more control but would cause a shift of focus that would diminish the credibility of the results. If we want to keep the focus of the users on the chosen user goal, we need to use a set of standardized emotional stimuli whose quality has also been validated in terms of cross-user and cross-cultural studies. We chose a subset of the IAPS database (Lang et al. 2005) as the content items and emotional stimuli. Because the IAPS dataset contains images annotated with their respective induced emotions on viewers we did not have to break our users' flow with questionnaires on their emotive response after viewing each image. And because the IAPS dataset has been validated in cross-cultural studies (Irun and Moltó Brotons 1997; Ribeiro et al. 2005; Verschuere et al. 2007) the uncertainty in the induced emotions for our users was low. Thus the choice of the IAPS dataset allowed us to (i) keep the focus of the experiment intact and (ii) keep the uncertainty caused by the uncontrolled emotion induction low.

3.3 Dataset acquisition: the emotion induction experiment

In the dataset-acquisition phase we induced emotive responses in end users through visual stimuli and acquired their explicit feedback regarding their preferences toward the presented visual stimuli.

3.3.1 Users

We had $N_U = 52$ users taking part in our experiment. They all gave their written consent to participate in the study. All the users were students of the 4th grade in a secondary school and they were all aged between 17 and 20 with an average age of 18.3 years and a standard deviation of 0.56. The sample consisted of 15 males and 37 females.

The quality of the users sample was assessed in terms of their individual differences in the cognitive task of giving ratings to items that induce different emotions. Because we model the individual user's preferences with emotive parameters we would like our sample to represent all the users in terms of their heterogeneity of attitude toward emotions. According to Westen (1999) and Yik et al. (2002) personality accounts for the individual differences of emotions in motivation and decision making. We thus chose personality as the criterion of the sample's heterogeneity. We used the IPIP questionnaire for assessing the users' personality through the five-factor model (FFM) (Goldberg et al. 2006). The FFM is composed of five scalar components (openness, conscientiousness, extraversion, agreeableness, and neuroticism) forming a five tuple for each user. It is difficult to compare the distribution of the five factors from our sample and other users due to the non-existing norms for the IPIP five factors. The authors of the IPIP instrument argue that such norms are misleading and therefore they do not provide these norms (Goldberg et al. 2006). From a visual inspection of the distributions of users in Fig. 7, which shows the scatter plots of pairs of the five factors along with the distribution histograms for each personality factor (the diagonal elements in Fig. 7), we conclude that the users from our experiment cover a sufficiently wide range of personalities.

3.3.2 Content items

The IAPS set of images that we chose as the source for our content items represents a set of stimuli which was compiled in a controlled experiment (Lang et al. 2005). Several cross-cultural studies have been carried out that showed consistency and thus confirmed its suitability for eliciting specific emotive responses (Irun and Moltó Brotons 1997; Ribeiro et al. 2005; Verschuere et al. 2007). Because of the results of the replicated cross studies mentioned above we were confident that the uncertainty of the induced emotive responses was kept as low as possible.

We had additional reasons for using images from the IAPS dataset. The first reason was that the IAPS dataset is made up of generic content and it was thus easy to define the task for the users involved in our experiment. The second reason was that we could simplify the modeling of the emotive response to a single emotive state. If we

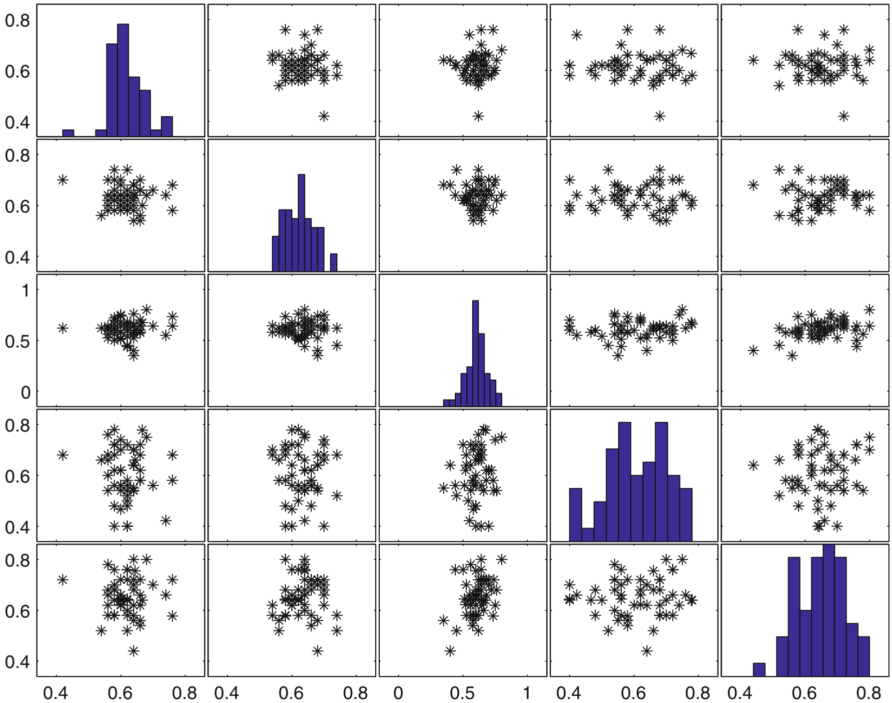


Fig. 7 Scatter plots of pairs of the five factor model personality factors of the users sample. The histograms on the diagonal show the distribution of each personality factor. The distribution of the personalities of the users that participated in the data acquisition shows the heterogeneity of the users sample

had used audio or video sequences the emotive response would be a more complex time-dependent data structure.

We chose a subset of 70 generic color images from the IAPS database. The number of images used in the experiment was a compromise between having as many items as possible and having a limited time for each user. The images were chosen randomly with the constraint that they cover the whole value-arousal plane equally. We divided the valence-arousal plane into three by three equal quadrants and chose the images in such a way that each quadrant had seven or eight images. The size of the images was 1,024 per 768 pixels and they were annotated with AM and GM. The AM values were provided by the IAPS dataset. Each image was annotated with the first two statistical moments of the induced emotion in users in the VAD space. The acquisition of induced emotions was carried out by Lang et al. (2005) with the Self Assessment Manikin (SAM) questionnaire. For the images used in our experiment the statistical moments were calculated for each image based on samples of 10–14 users. As already stated at the beginning of this section, cross-cultural studies validated the statistical moments of the VAD values in the IAPS dataset.

The genres in the GM were annotated manually. We annotated each item with a single genre value g out of a pool of ten possible genres. The choice of the genres was made independently by three persons from our group. Each person annotated all the images with custom tags representing genres. In order to narrow the number of distinct genres we used hypernyms to map the tags into semantically higher-level genre classes. For example, if one annotator tagged the image with *tiger* and another with *cat* both were mapped into the hypernym *animal*. For each item we applied a majority vote to obtain a single genre per item. In cases where there were three dissonant votes (genres) we randomly chose one. This procedure yielded ten distinct genres for the items in our dataset: *action, animal, erotic, people, plant, portrait, still, violence, weapon* and *weather*. For the given ten genres, three annotators and 70 items the Fleiss' kappa inter rater agreement measure (Fleiss 1971) was $\kappa = 0.69$, which meant substantial agreement among the annotators.

In a dataset of 70 images the number of distinct genres available can play a crucial role in the classification of items in binary classes. If we had too few distinct genres available (two or three) the genre feature could not carry enough information on the whole variance, which would make it a useless feature even at the experiment-design stage. This would be an unfair comparison where the affective features would be in advantage. On the other hand, if the number of genres was close to the number of all items the genre feature would account for an unrealistically large amount of variance, since in real systems the number of genres is much smaller than the number of items. We further annotated automatically each item with the average watching time $\bar{t_w}$ (as part of the GM) calculated from the acquired dataset.

3.3.3 Users' task

The participants were asked to select images for their computer's desktop. They were instructed to rate the images on a Likert scale from 1 to 5 (see Table 4). They were informed that images with ratings of 4 and 5 would be chosen for the computer wallpaper. They consumed the content items that induced an emotive response and gave explicit ratings to the images. According to the taxonomy of users' tasks given by Herlocker et al. (2004) the task in our experiment falls into the category of *find all good items*. The acquisition procedure was performed using a Matlab-based GUI application, which is depicted in Fig. 8. The participants were shown the first image from the image subset and watched it as long as they pleased. After giving an explicit rating to the image by clicking on the appropriate button the next image from the sequence appeared on the GUI. The presentation sequence was the same for all users. The explicit rating $e(u, h)$, which was first given on a Likert scale from 1 to 5, was later thresholded to a binary rating by assigning $e(u, h) = C_0$ to Likert ratings lower than 4 and $e(u, h) = C_1$ to the Likert ratings higher than or equal to 4. This procedure

Table 4 Likert scale for explicit ratings: during the dataset acquisition procedure the subjects gave explicit ratings to each image

Rating	Description
5	I like it very much
4	I somewhat like it
3	I neither like it nor dislike it
2	I somewhat dislike it
1	I dislike it very much

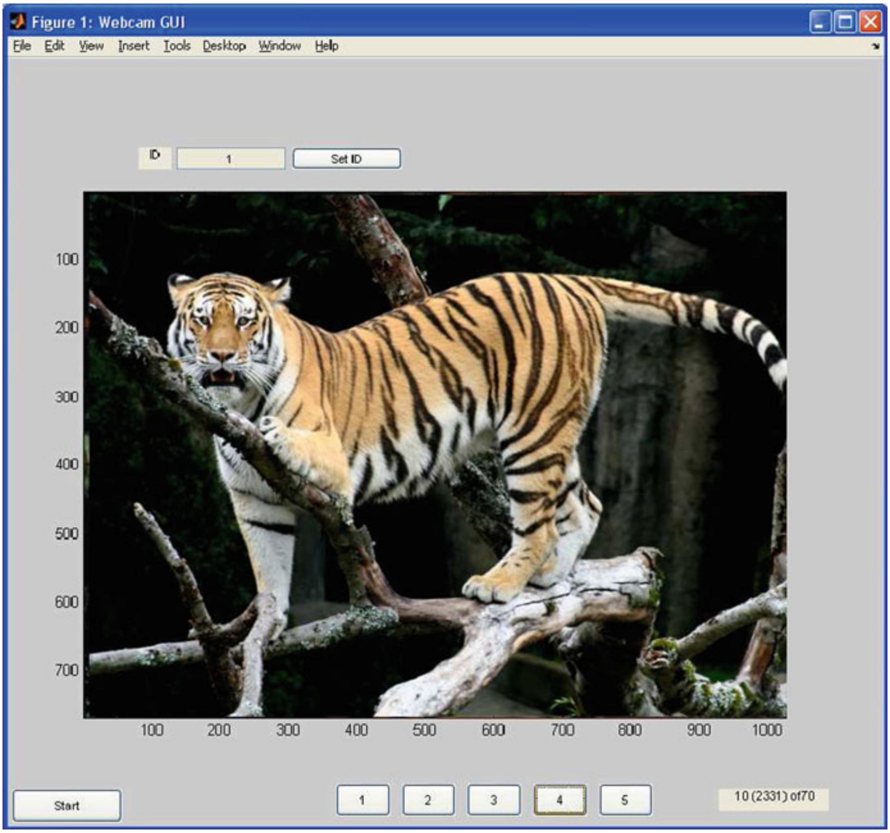


Fig. 8 Snapshot of the graphical user interface: the application for gathering users' responses in the induced-emotion experiment

continued until all 70 images were shown and rated. The GUI application recorded the watching times and explicit ratings given to each image and provided a usage history file for each user. The order of the item presentation during the data-acquisition stage was randomly chosen before the dataset acquisition and was equal for all the users.

3.3.4 Properties of the dataset

After the dataset-acquisition procedure we compiled the dataset of the usage history of the participants (see Tkalčič et al. 2010 for more information on the dataset). An overview of the dataset statistics is given in Table 5. The dataset contains data on item ratings by the users involved in the experiment. Each item is annotated with the metadata sets $\mathcal{A}$ and $\mathcal{V}$. The acquired dataset has full density as all the users have rated all the items. In total we had 52 users rating all 70 images, which yields a total of 3,640 ratings. Among these there were 1,460 items rated as relevant by the users and 2,180 items rated as non-relevant. If we used a random binary classifier for

Table 5 Overview of the dataset statistics: the number of ratings given by the users are reported in the rightmost two columns for relevant and non-relevant

		Number of ratings given by users	
		Non-relevant $e(u, h) \leq 3$	Relevant $e(u, h) > 3$
	Total	2180	1460
Number of users	Gender		
15	Males	681	369
37	Females	1499	1091
Number of items	g (item genre)		
12	Action	260	364
7	Animal	222	142
8	Erotic	242	174
5	People	150	110
4	Plant	105	103
6	Portrait	191	121
11	Still	317	255
9	Violence	386	82
6	Weapon	264	48
2	Weather	43	61
Mean = 2859 ms	$\bar{t_w}$ (watching time)		
	Less than mean	1333	984
	More than or equal to mean	847	476
Mean = 5.41	$\bar{v}$		
	Less than mean	1280	384
	More than or equal to mean	900	1076
Mean = 1.64	σ_v		
	Less than mean	1129	743
	More than or equal to mean	1051	717
Mean = 5.12	$\bar{a}$		
	Less than mean	908	600
	More than or equal to mean	1272	860
Mean = 2.16	σ_a		
	Less than mean	1016	648
	More than or equal to mean	1164	812
Mean = 4.42	$\bar{d}$		
	Less than mean	917	435
	More than or equal to mean	1263	1025
Mean = 1.79	σ_d		
	Less than mean	507	325
	More than or equal to mean	1673	1135

The information about the distribution of the given ratings is shown per user gender and per metadata attributes (g, $\bar{t_w}$, $\bar{v}$, σ_v, $\bar{a}$, σ_a, $\bar{d}$ and σ_d). For each attribute we give the number of occurrences of each rating for two intervals: less than the attribute mean value and more than or equal to the attribute mean value estimating the ratings $\hat{e}(u, h)$ we would obtain an overall scalar measure of approximately $P \approx \frac{0.5\times1460}{0.5\times3640} = 0.40$, which represents the worst-case performance.

3.4 The CBR procedure

We performed an offline CBR experiment for the validation of the different metadata sets employed. We also evaluated the metadata sets with four different ML classifiers.

We evaluated three decision maps: (i) $\delta^A : H \rightarrow \Omega$ where the item profile was composed of the standard metadata set $\mathcal{A}$, (ii) $\delta^V : H \rightarrow \Omega$ where the item profile was composed of the affective metadata set $\mathcal{V}$ and (iii) $\delta^{AV} : H \rightarrow \Omega$ where the item profile was composed of both metadata sets $\mathcal{A} \times \mathcal{V}$.

We denoted the evaluated ML algorithms with Γ. The set of evaluated classifiers was $\Gamma = \{AdaBoost, C4.5, NaiveBayes, SVM\}$. We chose the classifiers based on the following criteria: (i) related work (Adomavicius and Tuzhilin 2005; Pogačnik et al. 2005; Lew et al. 2006), (ii) to cover different families of ML algorithms and (iii) based on the visual inspection of the dataset. We performed the visual inspection of the data on plots like the one shown in Fig. 2 (right plot), which reveals that the points of different classes are mixed at the class borders. We did not expect good results from a nearest-neighbors classifier because the mixed border would yield lots of misclassifications. We included the Bayes classifier because in practice it gives good performance for a wide range of applications. The SVM was a natural choice because we could easily build the borders between the classes with a sequence of straight lines. Finally, boosting algorithms are known to perform well in a wide range of applications, so we included the AdaBoost. We performed a preliminary test of several classifiers and rejected the ones that did not perform well.

We used Matlab as the scripting language for the experiment and for the visualisation of the results. We used the weka package for folding the data and evaluating the ML algorithms.

The experiment consisted of splitting the dataset into the training and test sets and employing ML algorithms as the implementations of the mappings δ^A, δ^V and δ^{AV}.

We carried out the performance evaluation separately for each user $u \in U$ (52 users in total), each metadata set $\mathcal{A}$, $\mathcal{V}$ and $\mathcal{A} \times \mathcal{V}$ and each classifier $\gamma \in \Gamma$, which represented the three independent variables of the process under evaluation (see Fig. 9). In each iteration through the set of users, set of classifiers and the metadata sets we calculated the estimation of the ratings $\hat{e}(u, h)$ for all the items $h \in H$ (a total of 70 items). We split the items into training and test sets following the ten-fold cross-validation scheme, as proposed by Kohavi (1995). The outcomes of the classifications, the predicted ratings $\hat{e}(u, h)$, were compared to the ground truth ratings $e(u, h)$. From this comparison we calculated the number of correct and incorrect classifications, which yielded the confusion matrices $M(\delta^A)$, $M(\delta^V)$ and $M(\delta^{AV})$ (see Table 6).

3.5 Evaluation measures and statistical testing

In order to validate the hypothesis and provide answers regarding the suitability of the employed ML algorithms and AM fields we performed the following evaluation

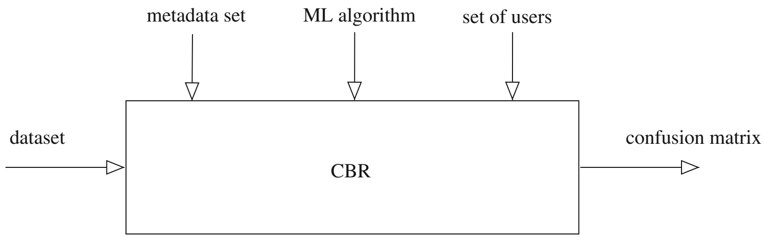


Fig. 9 The model of the process under evaluation with three independent variables: the metadata set ($\mathcal{A}$,$\mathcal{V}$), the ML algorithm (AdaBoost, C4.5, NaiveBayes, SVM) and the set of observed user

Table 6 A confusion matrix is the result of the CBR procedure, which performs a binary classification of items into relevant (C_1) or non relevant (C_0) for the observed user. Items can be correctly classified (True Positive (TP) or True Negative (TN)) or incorrectly classified (False Positive (FP) or False Negative (FN)). The numbers of correctly and incorrectly classified instances are reported in the confusion matrix

Predicted class $\hat{e}(u, h)$	Actual class $e(u, h)$ C_1 relevant	C_0 non relevant
C_1 relevant	TP	FP
C_0 non-relevant	FN	TN

methodology. We assessed the performance of the CBR with scalar measures derived from the confusion matrices. This was done separately for each user, ML algorithm and combination of metadata (see Fig. 9). As the nature of the scalar evaluation measures is probabilistic due to the nature of the dataset employed, we performed further statistical tests in order to determine whether the observed differences were real or coincidental. In order to determine the quality of the single AM fields for modeling in the CBR we also performed the analysis of variance (ANOVA). Before applying significance tests for small sample sizes we verified that the test conditions were met.

3.5.1 Scalar measures

We evaluated the performance of the CBR system through measures derived from the classifier's confusion matrices $M(\delta^u)$. We used the measures precision P, recall R and F-measure F_β (where $\beta = 1$) as defined by Herlocker et al. (2004) and commonly used in the evaluation of recommender systems (Herlocker et al. 2004; Pogačnik et al. 2005; Adomavicius and Tuzhilin 2005; Witten and Frank 2005). The precision tells us the rate of relevant items among all the selected items. The recall describes the rate of the relevant items that were actually selected. The F measure aggregates the precision and recall into a single scalar value.

According to the taxonomy of the recommender system's tasks proposed by Herlocker et al. (2004) our CBR falls into the category *find all good items*. The scalar measure that is best suited to the assessment of such systems is the precision P. In fact, what the users really want is that the list of recommended items H_R has as few false positives (non-relevant items classified as relevant) as possible. From the user's perspective it is more annoying to have false positives than false negatives (relevant items classified as non-relevant). Because the false negatives are never disclosed the user is not aware of the relevant items that have not been included in the recommended list.

We calculated all three scalar measures for the class of relevant items C_1. The basis for the calculation was the sum of the confusion matrices over all the users and all the folds.

3.5.2 Statistical tests

We transferred the statistical testing of the confusion matrices into the testing for the equivalence of two estimated discrete probability distributions (Lehman and Romano 2005). We compared the confusion matrices yielded by the CBR that used the GM set $\mathcal{A}$ with the confusion matrices of the CBR that used the combined GM and AM set $\mathcal{A} \times \mathcal{V}$ and all four ML algorithms.

The zero hypothesis to be tested was $H_0 = [M(\delta^A) \simeq M(\delta^{AV})]$ where $\simeq$ stands for the equivalence of the underlying discrete distributions. The natural choice here was the Pearson χ^2 test (Lehman and Romano 2005). It tests whether a sample $(n_1, \ldots, n_N)$ is drawn from a multinomial distribution $B(n, \mathbf{p})$ with parameters $n = n_1 + \cdots + n_N$ and $\mathbf{p} = (p_1, \ldots, p_N)$. Assuming $p_i > 0$ for all i, the test statistics is $Q = \sum_{i=1}^{N} \frac{(n_i - np_i)^2}{np_i}$ distributed as $\chi^2(N-1)$ if $np_i \gg 1$ for all $1 \leq i \leq N$ and n is large enough (Lehman and Romano 2005). Experimental studies showed that $np_i \gg 1$ in practice means $np_i \geq 5$. In our case where we have only two classification classes $N = 2$ the distribution is $\chi^2(1)$.

3.5.3 Evaluation of metadata fields

In ML feature selection we prefer features that maximize the ratio of the between-class variance and the within-class variance (Hastie et al. 2001). In order to assess the quality of the AM as features for ML we performed the ANOVA for each scalar metadata field f from the set $f \in \{\bar{t_w}, \bar{v}, \sigma_v, \bar{a}, \sigma_a, \bar{d}, \sigma_d\}$. The analysis was performed separately for each user to determine whether the difference in the mean values of the observed metadata field for each of the classes C_0 and C_1 was significant (as in Fig. 11b) or not (as in Fig. 11a). If the difference was significant we assumed that the observed metadata field is good for separating the classes C_0 and C_1. In other words, when the difference in the means was significant we assumed that the ratio of the between-class variance and the within-class variance was large enough. For each metadata field we repeated this procedure on a per user basis. We defined the measure of quality of a metadata field for usage in a CBR system q_f as the ratio between the number of users where the difference in the means was significant, which we denoted with N_S, and the number of all the users N_U

$$q_f = \frac{N_S}{N_U} \tag{5}$$

Once we have the scalar ratio q_f of a feature $f \in \{\bar{t_w}, \bar{v}, \sigma_v, \bar{a}, \sigma_a, \bar{d}, \sigma_d\}$ we want to find out whether q_f is significantly different than 0. By running the ANOVA we compare q_f with a hypothetical useless feature that has $q_f = 0$.

4 Results

This section is structured as follows. First we present the numerical results of the tests with different AM fields and different ML algorithms. Then we describe the results of the statistical analysis. Finally we provide the results on the quality of the metadata fields.

4.1 Results of scalar measures

Here we present the scalar performance measures of the CBR experiment as described in Sect. 3.4 and Fig. 9. The validation procedure yielded ten confusion matrices per each fold for each of the 52 users, four classifiers and three metadata sets $M(\delta_i^{A,u})$, $M(\delta_i^{AV,u})$ and $M(\delta_i^{V,u})$, which accounts for a total of 6,240 confusion matrices. As we were comparing the performance of the CBR in terms of the different metadata sets and classifiers used we first summed the confusion matrices over all the folds of the different users with the same metadata set and classifier. Then we summed these confusion matrices over all the users with the same metadata set and classifier, which yielded three summary confusion matrices $M(\delta^A)$, $M(\delta^{AV})$ and $M(\delta^V)$.

We calculated the scalar measures P, R and F from the confusion matrices. These values are given in Table 7. The results are grouped by metadata sets. Within each metadata set the results are presented separately for each of the four classifiers.

Figure 10 shows the dependence of the precision, the P column from Table 7, on the different metadata sets and classifiers. Each of the four lines in Fig. 10 represents a single classifier's performance with respect to the metadata-set used.

4.2 Statistical results

As the scalar results are calculated from the particular dataset they are not suitable for determining whether the differences between the different metadata sets and classifiers

Table 7 Precision, recall and F measures for the three metadata sets and four classifiers: the results are grouped by metadata sets. Within each metadata-set row the results are further split by classifiers

Metadata set	Classifier γ	P	R	F
$\mathcal{A}$	AdaBoost	0.57	0.42	0.48
	C4.5	0.60	0.46	0.52
	NaiveBayes	0.58	0.58	0.58
	SVM	0.61	0.55	0.58
$\mathcal{A} \times \mathcal{V}$	AdaBoost	0.63	0.56	0.59
	C4.5	0.64	0.57	0.60
	NaiveBayes	0.57	0.64	0.61
	SVM	0.65	0.61	0.63
$\mathcal{V}$	AdaBoost	0.64	0.56	0.60
	C4.5	0.62	0.54	0.58
	NaiveBayes	0.57	0.60	0.58
	SVM	0.68	0.55	0.61

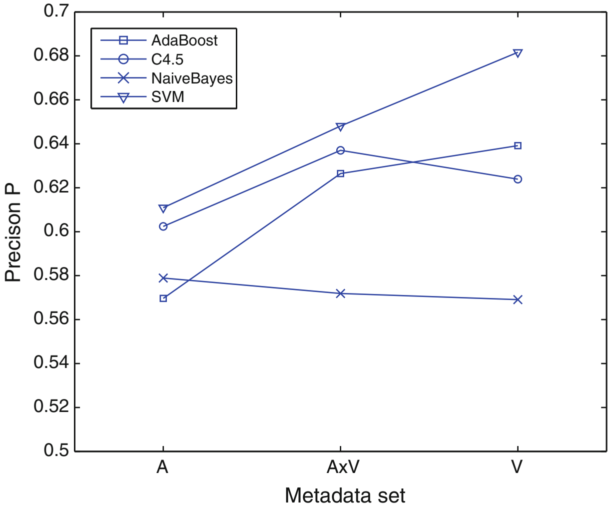


Fig. 10 Dependency of precision P on metadata sets using different classifiers: the values of precision P of each classifier over different metadata sets are connected with a *broken line*

are significant or not. To determine this we applied the Pearson χ^2 statistical significance test (Lehman and Romano 2005) to compare the confusion matrices $M(\delta^A)$ and $M(\delta^{AV})$.

The distribution of the test statistics Q is $\chi^2(1)$ and the critical value at risk level $\alpha = 0.05$ is 3.84. Since all the listed values are larger than the critical value, all the p values are way below the risk level and we conclude that the performances of all the classifiers are significantly different when applied using the $\mathcal{A}$ or the $\mathcal{A} \times \mathcal{V}$ metadata set.

4.3 Metadata fields results

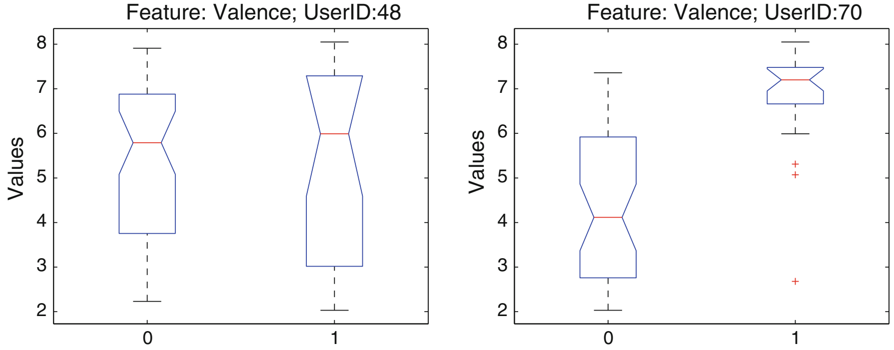
For each of the scalar metadata fields f we performed a one-way ANOVA, as described in Sect. 3.5, and calculated the ratios q_f, as defined in Eq. 5. The ratios q_f are summarized in Table 8. For example, the metadata field $\bar{v}$ had the highest ratio $q_{\bar{v}} = 0.71$, which means that for 71% of the users the differences in the mean values of the metadata field $\bar{v}$ for the classes C_0 and C_1 were significantly different (similar to Fig. 11b) and for the other 29% the difference in the means was not significant (similar to Fig. 11a). The ratios q_f are all significantly higher than 0 at the risk level $\alpha = 0.05$, except for the feature σ_a.

5 Discussion

In this section we discuss the outcome of the main hypothesis testing. We also discuss the effect of the choice of different ML algorithms and the quality of each AM field. The limitations of the proposed solution are discussed and indications of open issues for the future are given.

Table 8 Results of the ANOVA for the scalar features: the ratio q_f for each analyzed feature f tells us the portion of all users where the difference in the mean values of the feature f (between the groups of items marked as relevant and non-relevant) was significant. The rightmost column of the table reports the p value of the comparison of the observed q_f with a hypothetical $q_f = 0$

Feature f	Ratio q_f	p value
$\bar{t_w}$	0.17	0.002
$\bar{v}$	0.71	0.000
σ_v	0.10	0.022
$\bar{a}$	0.21	0.000
σ_a	0.06	0.080
$\bar{d}$	0.31	0.000
σ_d	0.12	0.011



(a) user u_{48} having $p = 0.98$ **(b)** user u_{70} having $p \approx 10^{-9}$

Fig. 11 Box plots representing the distribution of the feature $\bar{v}$ in relevant (C_1) and non-relevant (C_0) items for two different users: the user on the **a** exhibits similar distributions of the feature $\bar{v}$ in both classes, while the user on the **b** exhibits a clear preference toward items with high values of the $\bar{v}$ feature. This means that for the user on the left the parameter $\bar{v}$ does not account for any between-class variance, while for the user on the right most of the between-class variance is contained in the parameter $\bar{v}$

5.1 The effect of AM on CBR performance

The experimental results showed that our main hypothesis is true. The inclusion of the proposed AM brings a significant increase in performance over GM in the presented CBR system.

However, this increase in the performance requires further discussion. The statistical test showed that there is a significant difference between the confusion matrices when we used the $\mathcal{A}$ over the $\mathcal{A} \times \mathcal{V}$ metadata sets for all the ML algorithms. To find out whether this difference is an improvement or deterioration in the CBR performance for end users we must observe the scalar measures. We already argued in Sect. 3.5.1 why the precision P is the most indicative scalar measure for the CBR task our system performed (*find all relevant items* from the taxonomy proposed by Herlocker et al. 2004).

When we use AM the CBR always performs better than with GM only, with the only exception being the NaiveBayes classifier, which is discussed later in Sect. 5.2. It is also interesting to note that in some cases the usage of AM only, yields a better precision than the usage of both AM and GM (this is the case for the AdaBoost and the SVM classifier). The improvements in the precision range between 3% (the C4.5 classifier with AM only) and 12% (the AdaBoost classifier with AM only). The absolute values of the precision when AM are used are not so impressive, somewhere between 0.62 and 0.68, which means that approximately two out of three recommended items are truly relevant for the end user. The reference point of a random classifier is approximately $P \approx 0.40$. Although the absolute values of the precision are not extremely high we observe a significant improvement when we use AM together with GM.

A similar pattern of improvement, but with higher rates, can be observed for the recall. This means that the inclusion of AM has a positive effect on the reduction of false negatives as well.

5.2 The effect of ML algorithms on the CBR performance

The results showed that the choice of ML algorithms has an effect on the CBR performance. There are differences between the ML algorithms regarding the metadata set used and the scalar measure we choose for the observation. Overall, the SVM turned out to yield the best performance, while the NaiveBayes showed ambiguous results.

We did not perform any test to see whether the differences among the ML algorithms were significant, but we found the scalar results to be indicative. If we look at the results in Table 7 we observe that the employed ML algorithms have different behaviour in terms of scalar measures. If we observe the precision only we can see that the NaiveBayes classifier yields better results when only GM are used, which is an unexpected outcome. In the case of the other three classifiers the precision was higher when AM were included, although it is not clear whether it is better to use AM only or AM combined with GM. In the case of the C4.5 classifier the combination of AM and GM was better than AM only. In the case of the AdaBoost and SVM classifiers the employment of AM only outperformed the combination AM + GM.

Among the ML algorithms evaluated the NaiveBayes stands out for its ambiguous results. It is interesting to note that, although the differences in the confusion matrices were significant, the scalar measures precision and recall showed different trends in the case of the NaiveBayes classifier. There was a slight deterioration of precision ($P = 0.58$ and $P = 0.57$ for the $\mathcal{A}$ and $\mathcal{A} \times \mathcal{V}$ sets, respectively) and a substantial increase of recall ($R = 0.58$ and $R = 0.64$ for the $\mathcal{A}$ and $\mathcal{A} \times \mathcal{V}$ sets, respectively). Such ambiguous results call for a deeper analysis of the costs of false-positive and false-negative rates for the selected user scenario in order to give a more sound interpretation of the dissonance of the scalar measures.

The natural question regarding the choice of ML algorithms is which one is the most suitable. Our results show that the SVM performs better than the other three in terms of precision and F-measure. We must stress that the focus of the research presented here was not an evaluation of ML algorithms but the comparison of the AM and GM sets for CBR systems. So we believe there is still room for an improvement in performance by evaluating other ML approaches and spending more resources in fine tuning.

5.3 Quality of AM fields for the CBR system

During feature selection for machine learning we usually want to reduce a large number of features to a manageable, smaller number. We do this by using methods like the Principal Component Analysis (PCA) to extract a limited set of features that account for most of the variance. In our case we already have a limited set of proposed AM features, six only. So we do not need to reduce their number; we just want to assess whether each of them, as novelties in affective modeling, is suitable or not for CBR systems. We simplify the usual approach of feature-quality assessment (i.e. the ratio of between-class variance and within-class variance) to testing whether the differences in the mean values of the observed feature for the two classes (C_0 and C_1) are significant or not for a user. We defined the ratio q_f as a measure of the share of users where there is a significant difference in the means (see Eq. 5). The results showed that the AM feature $\bar{v}$ had significant mean differences in 71% of the users. This was expected since $\bar{v}$ describes the pleasantness of the emotion and the user task implied the search for pleasant content. The other two first statistical moments, $\bar{a}$ and $\bar{d}$, had significant mean differences in 21 and 31% of the users. Although these values are lower than for $\bar{v}$, they still carry enough information to be used in CBR systems. The second statistical moments, σ_v, σ_a and σ_d, had substantially lower ratios q_f, 10, 6 and 12%, respectively.

In a further analysis we compared the obtained ratios q_f with the ratio of a hypothetical useless feature that has $q_f = 0$. The ANOVA showed that all features have significantly higher ratios q_f except for the feature σ_a at the level $\alpha = 0.05$. This means that, at the given confidence interval, the feature σ_a can be considered as useless for affective modeling in CBR systems.

We conclude that the first statistical moments, $\bar{v}$, $\bar{a}$ and $\bar{d}$, are very suitable for CBR systems, while the second statistical moments σ_v, σ_a and σ_d carry less information and can be treated as optional. Furthermore, we can observe that among the three emotive parameters, v, a and d, it is a that carries the least information for separating the relevant items from the non-relevant. This holds for both the first and the second statistical moments $\bar{a}$ and σ_a. This can be seen as a guideline for the design of the acquisition phase, where we could leave out the acquisition of the affective parameter a.

5.4 Sample size implications

The quality of the sample used in such experiments is very important. Here we discuss the size of the sample used in the presented research. We observe this issue from two points of view: (i) the number of users/the number of items and (ii) the ML training set size.

The number of users (52) and the number of items (70) in our experiment is not huge, but it is on the same level as related work that reports the following figures: Pogačnik et al. (2005): 41 users, Nunes et al. (2008): ten users, Arapakis et al. (2009): 23 users and Joho et al. (2009): six users. From the strictly statistical point of view, when validating hypotheses, it is of paramount importance to verify the conditions for applying the strongest test possible. A key parameter here is the sample size. The sample size has an influence on the type II error, which is difficult to estimate. In our case we applied the statistical tests and reported the results correctly. A larger sample would, however, improve the power of the test and thus give greater credibility to the results (i.e. affective recommender outperforms generic recommender) by reducing the type II error.

In the presented system, as in other recommender systems, ML techniques are applied on a per-user basis. This means that for each observed user we took the already rated items to learn the user model. More concretely, in the ten-fold cross-validation scheme, we used 63 ratings to train the user model and seven for testing in each fold. This might appear small compared to other ML applications (for instance it is quite easy and inexpensive to build a dataset of several thousand items for face recognition). But in the case of recommender systems this number is actually quite realistic. In fact, real recommender application datasets contain lots of users, lots of items, but very few ratings per user. This problem is usually referred to as the matrix-sparsity problem (Adomavicius and Tuzhilin 2005). In theory, we would achieve better accuracy for the recommended items if we had more ratings per user, as our model would be trained on a larger training set. But in practice (in real recommender systems) the number of ratings per user is relatively low, so the user models are not as good as they could be if the users had rated more items. The widely used datasets each movie and jester have 38 and 54 ratings per user, respectively (Grouplens Data Sets).

In order to identify the implications caused by the (relatively) small dataset this should be observed through the perspective of parameter estimations of the concrete classifier. For example, when training the SVM classifier we observe our process as an estimation of parameters of the support vectors. Generally the smaller the training set, the smaller the confidence interval for the estimated classifier parameters. However it is hard to estimate what is the concrete influence of the confidence interval of the parameters of a concrete ML algorithm.

In summary, the relatively low number of ratings per user in our dataset reflects real-life recommender systems. This means that the results reported in this paper reflect the results that such a system would yield if employed in a real-life application. When comparing both metadata sets we applied the correct statistical tests so the results given are valid.

5.5 Shortcomings and open issues

Based on our results we can conclude that the inclusion of AM does improve the performance of the CBR system. There are, however, several open issues that we are aware of and should be addressed in the future.

One of the biggest issues we wish to explore is context dependency. The baseline hypothesis of our work is that people differ in the emotion they are seeking. Some are thrill seekers, while some prefer calm experiences. But the same person might seek different emotions in different situations. The nature of the tasks given to the user in information seeking (e.g. a specific search task vs. a broader search task) might also influence the amount of variance contained in affective metadata. We expect that the context would explain a substantial part of the variance and could thus contribute to exploiting the affective metadata in a better way.

Our experimental design was based on a priori known, or perhaps it would be better to say assumed, emotive responses of the users. We already discussed the implications and shortcomings of the experimental design in Sect. 3.2. Thus the second issue we want to address in the future is the ongoing and implicit affective tagging of multimedia items directly from the users' responses. A lot of work is going on in the field of automatic emotion detection (Ioannou et al. 2005; Zeng et al. 2009; Joho et al. 2009). By employing such methods we can easily compute the proposed AM on the fly, as new information arrives. The inclusion of these methods would lead us to a standalone recommender system that can be used in real applications.

The transition from laboratory experiments, such as the one presented here, to real applications, as suggested in the previous paragraph, leads inevitably to the cold-start problem, which occurs when a new user or a new item is added and the system does not have enough information to build a useful profile. A possible approach to overcoming the new item problem is to use a method for extracting affective parameters from the item's low-level features. Such work has already been carried out (Hanjalic 2006; Shan et al. 2009) and could be integrated into a recommender system.

In the presented work we used affective metadata related to the aesthetic emotions of users and not the intrinsic emotions contained in the items. An interesting area would be to compare both affective properties and analyze whether they account for a significant part of the variance.

In order to strengthen the significance of the reported work, repetitions of the experiment with a larger number of users and different sets of content items would be welcomed.

6 Conclusion

The work presented in this paper aims at establishing the importance of the inclusion of affective metadata in CBR systems. The study contributes to the knowledge on affective computing in recommender systems.

Our results showed that the usage of the proposed affective features in a CBR system for images brings a significant improvement over generic features. We presented a simple yet efficient method for modeling items and users through the first two statistical moments of the users' emotive responses in the valence-arousal-dominance space. Our experimental results also indicate that the Support Vector Machine algorithm is a good candidate for the calculation of items' rating estimates. Finally, we showed that among the proposed affective features, the first statistical moments carry more information for the separation of the relevant items from the non-relevant items than the second statistical moments.

This work sets a reference for the affective modeling of items and users in CBR systems. It is important though, that the work is continued. Further investigations are needed in the fields of context-aware affective modeling and the inclusion of the automatic affective tagging of items.

Despite the drawbacks mentioned in the previous section, our study clearly showed that affective metadata improve the performance of a CBR system and that the first two statistical moments represent a sound affective modeling approach, especially when used in conjunction with the Support Vector Machine algorithm. Furthermore, we offered a simple methodology for the assessment of the quality of individual affective features. Finally, we indicated the major issues to be addressed in the future for an improvement to the proposed approach.

Acknowledgments We would like to thank the members of the LDOS group (http://www.ldos.si) for their assistance in the dataset-acquisition procedure. We would also like to express our gratitude to the students who participated in the experiment and the responsible people from the Gimnazija Poljane school for their support and understanding. We are very grateful to Jernej Trnkoczy for his careful reading and valuable feedback. We would also like to thank the anonymous reviewers who helped us to substantially improve the quality of the paper. This work was partially funded by the European Commission within the 6th framework of the IST under grant number FP6-27312. All statements in this work reflect the personal ideas and opinions of the authors and not necessarily the opinions of the European Commission. The work has been also supported by ARRS, the Slovenian Research Agency.

References

Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE. Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)

Ali, K., Van Stam, W.: TiVo: making show recommendations using a distributed collaborative filtering architecture. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 394–401. ACM, New York, NY, USA (2004)

Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., Jose, J., Gardens, L.: Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1440–1443 (2009)

Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: using social and content-based information in recommendation. In: Proceedings of the National Conference on Artificial Intelligence, pp. 714–720. Wiley, New York (1998)

Batliner, A., Steidl, S., Hacker, C., Noth, E.: Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. User Model. User-Adapt. Interact. J. Pers. Res. **18**(1), 175–206 (2008). doi:10.1007/s11257-007-9039-4

Berger, H., Denk, M., Dittenbach, M., Pesenhofer, A., Merkl, D.: Photo-based user profiling for tourism recommender systems. In: Psaila, G., Wagner, R. (eds.) E-Commerce and Web Technologies, vol. 4655, pp. 46–55. Springer, Berlin (2007)

Bradley, M.M., Lang, P.J.: The International Affective Picture System (IAPS) in the Study of Emotion and Attention, chap. 2. Series in Affective Science. Oxford University Press, Oxford (2007)

Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adapt. Interact. **12**(4), 331–370 (2002). doi:10.1023/A:1021240730564

Carberry, S., de Rosis, F.: Introduction to special issue on affective modeling and adaptation. User Model. User-Adapt. Interact. **18**(1), 1–9 (2008)

Caridakis, G., Karpouzis, K., Wallace, M., Kessous, L., Amir, N.: Multimodal users affective state analysis in naturalistic interaction. J. Multimodal User Interfaces **3**(1), 49–66 (2010)

Coan, J., Allen, J.: Handbook of Emotion Elicitation and Assessment. Oxford University Press, New York (2007)

Conati, C., Maclaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Model. User-Adapt. Interact. **19**(3), 267–303 (2009)

Cowie, R., Cowie, E.D., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)

Darwin, C.: The Expression of the Emotions in Man and Animals. Oxford University Press, New York (1872)

D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., Graesser, A.: Automatic detection of learner's affect from conversational cues. User Model. User-Adapt. Interact. J. Pers. Res. **18**(1), 45–80 (2008). doi:10.1007/s11257-007-9037-6

Ekman, P.: Basic emotions. In: Dalgleish, T., Power, T. (eds.) Handbook of Cognition and Emotion. Wiley, New York (1999)

Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The international personality item pool and the future of public-domain personality measures. J. Res. Pers. **40**, 84–96 (2006)

González, G., López, B., de la Rosa, J.L.L.: Managing emotions in smart user models for recommender systems. In: Proceedings of 6th International Conference on Enterprise Information Systems ICEIS 2004, vol. 5, pp. 187–194 (2004)

Grouplens Data Sets.: http://www.grouplensorg/node/12. Accessed Sept 2010

Hanjalic, A.: Extracting moods from pictures and sounds. IEEE Signal Process. Mag. **23**(2), 90 (2006)

Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2001)

Herlocker, J., Konstan, J., Terveen, L., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 53 (2004)

Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K., Kollias, S.: Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Netw. **18**(4), 423–435 (2005)

Irun, M., Moltó Brotons, F.: Looking at pictures in North America and Europe: a cross-cultural study on the IAPS. In: Poster presented at the 1997 FEPS Meeting in Konstanz (1997)

Joho, H., Jose, J., Valenti, R., Sebe, N.: Exploiting facial expressions for affective video summarisation. In: Proceeding of the ACM International Conference on Image and Video Retrieval, pp. 1–8. ACM (2009)

Kim, Y., Yum, B., Song, J., Kim, S.: Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites. Expert Syst. Appl. **28**(2), 381–393 (2005)

Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, no. 12, pp. 1137–1143. Morgan Kaufmann, San Mateo (1995)

Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009). doi:ieeecomputersociety.org/10.1109/MC.2009.263

Lang, P., Bradley, M., Cuthbert, B.: International affective picture system (IAPS): affective ratings of pictures and instruction manual. Technical Report a-6. Technical Report, University of Florida, Gainesville, FL (2005)

Lehman, E.L., Romano, J.: Testing Statistical Hypotheses. Springer, New York (2005)

Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimed. Comput. **2**(1), 1–19 (2006)

Lichtenstein, A., Oehme, A., Kupschick, S., Jürgensohn, T.: Comparing Two Emotion Models for Deriving Affective States from Physiological Data, pp. 35–50. Springer-Verlag, Berlin (2008). doi:10.1007/978-3-540-85099-1_4

McNee, S., Lam, S., Konstan, J., Riedl, J.: Interfaces for eliciting new user preferences in recommender systems. In: User Modeling 2003: 9th International Conference, UM 2003, Johnstown, PA, USA, June 22–26, 2003: Proceedings, pp. 178–187. Springer-Verlag, Berlin (2003)

McQuiggan, S., Mott, B., Lester, J.: Modeling self-efficacy in intelligent tutoring systems: an inductive approach. User Model. User-Adapt. Interact. J. Pers. Res. **18**(1), 81–123 (2008). doi:10.1007/s11257-007-9040-y

Mehrabian, A.: Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr. Psychol. **14**(4), 261–292 (1996)

Nunes, M.A., Cerri, S., Blanc, N.: Improving recommendations by using personality traits in user profiles. In: Proceedings of I-KNOW '08 8th International Conference on Knowledge Management and Knowledge Technologies, pp. 92–100. Graz, Austria (2008)

Pantic, M., Vinciarelli, A.: Implicit human-centered tagging. IEEE Signal Process. Mag. **26**(6), 173–180 (2009)

Pazzani, M., Billsus, D.: Content-Based Recommendation Systems. The Adaptive Web, pp. 325–341. doi:10.1007/978-3-540-72079-9_10 (2007)
Picard, R.W.: Affective Computing. MIT Press, Cambridge (2000)
Plutchik, R.: The nature of emotions. Am. Sci. **89**(4), 344–350 (2001)
Pogačnik, M., Tasič, J., Meža, M., Košir, A.: Personal content recommender based on a hierarchical user model for the selection of TV programmes. User Model. User Adapt. Interact. **15**, 425–457 (2005)
Porayska-Pomsta, K., Mavrikis, M., Pain, H.: Diagnosing and acting on student affect: the tutor's perspective. User Model. User-Adapt. Interact. J. Pers. Res. **18**(1), 125–173 (2008). doi:10.1007/s11257-007-9041-x
Posner, J., Russell, J.A., Peterson, B.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. **17**, 715–734 (2005)
Rashid, A., Albert, I., Cosley, D., Lam, S., McNee, S., Konstan, J., Riedl, J.: Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the 7th International Conference on Intelligent User Interfaces, January, ACM, pp. 13–16 (2002)
Ribeiro, R., Pompéia, S., Bueno, O.: Comparison of Brazilian and American norms for the International Affective Picture System (IAPS). Revista Brasileira de Psiquiatria **27**, 208–215 (2005)
Rottenberg, J., Ray, R.D., Gross, J.J.: Emotion Elicitation Using Films, chap. 2. Oxford University Press, London (2007)
Scheirer, J., Fernandez, R., Klein, J., Picard, R.W.: Frustrating the user on purpose: a step toward building an affective computer. Interact. Comput. **14**(2), 93–118 (2002). doi:10.1016/S0953-5438(01)00059-5
Scherer, K.: What are emotions? And how can they be measured?. Soc. Sci. Inf. **44**(4), 695 (2005)
Schröeder, M., Baggia, P., Burkhardt, F., Oltramari, A., Pelachaud, C., Peter, C., Zovato, E.: Emotion markup language (emotionml) 1.0. W3C Working Draft 29 July 2010. http://www.w3.org/TR/2010/WD-emotionml-20100729/ (2010)
Shan, M.K., Kuo, F.F., Chiang, M.F., Lee, S.Y.: Emotion-based music recommendation by affinity discovery from film music. Expert. Syst. Appl. **36**(4), 7666–7674 (2009). doi:10.1016/j.eswa.2008.09.042
Tkalčič, M., Tasič, J, Košir, A.: The LDOS-PerAff-1 corpus of face video clips with affective and personality metadata. In: Kipp M (ed.) Proceedings of the LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (2010)
Verschuere, B., Crombez, G., Koster, E.: Cross Cultural Validation of the IAPS. Ghent University, Ghent, Belgium. http://users.ugent.be/~bvschuer/Iaps.pdf (2007)
Villon, O., Lisetti, C.: A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. In: The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006, ROMAN 2006, pp 269–276 (2006)
Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
Westen, D.: Psychology: Mind, Brain and Culture. 2nd edn. Wiley, New York (1999)
Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco (2005)
Yannakakis, G., Hallam, J., Lund, H.: Entertainment capture through heart rate activity in physical interactive playgrounds. User Model. User-Adapt. Interact. J. Pers. Res. **18**(1), 207–243 (2008). doi:10.1007/s11257-007-9036-7
Yik, M., Russell, J.A., Ahn, C.k., Fernandez Dols, J.M., Suzuki, N.: Relating the five-factor model of personality to a circumplex model of affect: a five-language study. In: McCrae, R.R., Allik, J. (eds.) The Five-Factor Model of Personality Across Cultures, pp. 79–104. Kluwer Academic Publishers, New York (2002)
Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009). doi:10.1109/TPAMI.2008.52

Author Biographies

Marko Tkalčič received his B.Sc. and M.Sc. degrees in September 1999 and July 2003, respectively. He was awarded with the Prešeren student award in February 1999. Since October 1999 he is working as a research scientist in the Digital Signal, Image and Video Processing Laboratory at the University of Ljubljana, Faculty of Electrical Engineering. His areas of research include user modelling, human perception, device color calibration, digital image processing, database-driven dynamic web applications and remote education systems. He is currently doing research on the usage of affective and personality parameters for modeling users and content in various applications.

Urban Burnik Ph.D. works at the Faculty of Electrical Engineering in Ljubljana as a Senior Lecturer and as a Teaching Assistant in the field of digital signal processing and mobile communications. He was visiting researcher at the University of Westminster, London, where he specialised in signal processing. His current activities are dedicated to multidimensional signal processing and to telecommunication services, especially in the field of interactive TV and universal multimedia access. He actively participates in European research projects, is vice-secretary of the COST 276 Action and a member of the IEEE.

Andrej Košir Ph.D. is an Associate Professor at the Faculty of Electrical Engineering, University of Ljubljana. He was awarded the Vidmar prize for his educational prowess. He is active in several research fields, including signal, image and video processing, optimization (numerical optimization, genetic algorithm), and user interfaces. He was a guest researcher at the University of Westminster, London, UK, at the University of Waterloo, Canada, and at the North Carolina State University, USA. He is currently leading projects from the field of multimedia, optimization methods and digital signal processing—especially in object recognition on digital images, intelligent networks, user interfaces and user modeling.