DRAFT, to be published in IJHCS
https://doi.org/10.1016/j.ijhcs.2018.04.004

Moodplay: Interactive Music Recommendation based on Artists' Mood Similarity

Ivana Andjelkovic[a,*], Denis Parra[b], John O'Donovan[a]

[a] *University of California - Santa Barbara, Santa Barbara, CA 93106, USA*
[b] *Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Santiago, Chile*

Abstract

A large amount of research in recommender systems focuses on algorithmic accuracy and optimization of ranking metrics. However, recent work has unveiled the importance of other aspects of the recommendation process, including explanation, transparency, control and user experience in general. Building on these aspects, this paper introduces *MoodPlay*, an interactive music-artists recommender system which integrates content and mood-based filtering in a novel interface. We show how *MoodPlay* allows the user to explore a music collection by musical mood dimensions, building upon GEMS, a music-specific model of affect, rather than the traditional *Circumplex* model. We describe system architecture, algorithms, interface and interactions followed by use-case and offline evaluations of the system, providing evidence of the benefits of our model based on similarities between the typical moods found in an artist's music, for contextual music recommendation.

Finally, we present results of a user study (N=279) in which four versions of the interface are evaluated with varying degrees of visualization and interaction. Results show that our proposed visualization of items and mood information improves user acceptance and understanding of both the underlying data and the recommendations. Furthermore, our analysis reveals the role of mood in music recommendation, considering both artists' mood and users' self-reported mood in the user study. Our results and discussion highlight the impact of visual and interactive features in music recom-

*Corresponding author
Email addresses: iva.andel@gmail.com (Ivana Andjelkovic), dparras@uc.cl (Denis Parra), jod@cs.ucsb.edu (John O'Donovan)

mendation, as well as associated human-cognitive limitations. This research also aims to inform the design of future interactive recommendation systems.

Keywords: Recommender Systems, Music Recommendation, Mood Context, Context-aware Recommendation, Affective Computing, Recommendation Interface

1. Introduction

Recommender systems are increasingly relied on in many domains for identifying relevant, personalized content from very large information spaces. Well established algorithms, such as Collaborative Filtering [1], Content-Based Filtering [2] and Matrix Factorization [3], are used across a variety of domains to recommend digital content or merchandise. Due to its unique consumption characteristics, music falls into a domain where alternative approaches to the traditional recommendation problem can help. These characteristics can be demonstrated by comparing music to two types of content widely offered to users via recommender systems: online movies and merchandise. For example, movies typically require undivided attention for 1-3 hours and most often one movie is watched per sitting. On the other hand, we can listen to music throughout the day in almost any situation – while working, exercising, commuting, cooking, socializing and so forth. Similarly, online shopping is usually a focused action that most people engage in for a shorter period of time compared to music listening. While both can depend on a user's taste, music listening is more often guided by feelings rather than practical reasoning. Overall, compared to other domains, music listening is more context dependent and closely tied to our emotional state. There are several music recommender systems that employ different types of context (daily activity [4], time of the day [5], music genre [6], etc.). However, previous work that integrates affective context for music discovery into a visual and interactive recommendation system is scarce (e.g. [7]). In this paper, we focus on prototyping and evaluating an interactive recommender system that suggests music bands based on artists' mood similarity and user input as an indicator of current preference.

Experimental evidence shows a strong relation between emotion and music [8] and previous research in affect-based recommender systems produced improvements over their non-contextual alternatives [9, 10]. Previous studies, e.g. [11, 12, 13, 14, 15, 16, 17] demostrate the importance of building interactive recommender interfaces, that go beyond the static-ranked list paradigm to improve user satisfaction. This trend is further supported by results showing that user satisfaction does not depend on recommendation accuracy only, but on factors such as serendipity, novelty, control and transparency as well [18, 19]. Our goal is to build a prototype recommender system with an interactive interface (Figure 1) that supports users in discovering unknown, interesting items via interaction in an affective-aware visualization. As a proof of concept, we have designed and implemented MoodPlay, a system that: (a) visually represents affective metadata for a music recommender system, and (b) supports interaction, explanations and control over such visualization. We frame our work around the following research questions (RQ):

- **(RQ1)** What are the effects of interactive visualizations on the user experience with a recommender system, and what is the right amount of interaction for a music recommender?
- **(RQ2)** Does affective information improve recommendation accuracy and user experience versus when it is not included?

In our effort to answer the research questions, we have produced the following key contributions:

- **A novel visual interface for recommendation.** A visualization that maps moods and music artists in the same two-dimensional space, supporting item exploration and user control. The space is built using mood tags associated with artists, collected from an established, public database. We extend and visualize the music-specific emotion model - GEMS (defined in section 2), to better fit a mood-aware music recommendation system.
- **Affect-aware recommendation method.** A novel hybrid recommendation algorithm for mood-aware and audio content-based music recommendation. The algorithm uses both mood tags of artists and audio content of their most popular songs.

DRAFT, to be published in IJHCS
https://doi.org/10.1016/j.ijhcs.2018.04.004

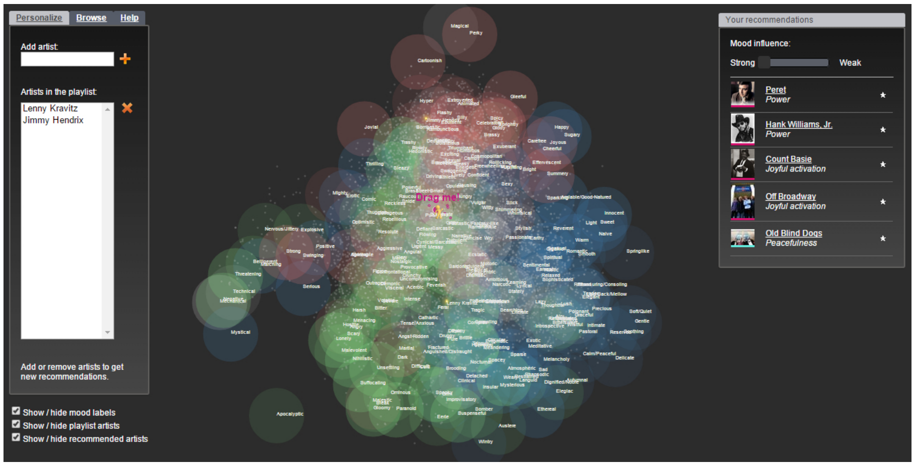


Figure 1: Screenshot of the MoodPlay interface, divided into three sections: (left) pane for generating user profile by entering artist names, (center) snapshot of the mood space visualization, (right) recommendation list, along with slider for adjusting mood influence. Demonstration video is available at `https://youtu.be/eEdo32oOmcE`

- **Enhanced interaction techniques.** We introduce several new interaction mechanisms for hybrid recommendation on a visual mood space. For instance, trail-based and radius-based techniques.

- **Empirical evidence for avoiding high cognitive load.** We present an evaluation of the system through an online experiment (N=279). We discover interesting relations between user interaction, trust, and user perception. We also provide some lessons for interface design in the context of exploratory tasks on recommender systems.

In contrast to our previous work on this system presented in [20], graphical design and several interface features have been improved based on the user feedback from the first experiment. The mood space visualization has been updated to show smoother transition between mood categories. We enabled live streaming within the application to provide a more real-world context to the experiment and we display artist information upon clicking on artist nodes in the visualization. Using this improved system, a new experiment was designed, conducted and presented in this paper. Detailed comparison to the previous experiment can be found in 6.

To evaluate our system, we conducted a user study over four different conditions:

(1) static recommendations in the form of ranked lists, generated based on a user's selection of seed artists (2) static recommendations, highlighted in a mood space visualization, (3) dynamic recommendation lists generated via user interaction in the mood space visualization, using current user's preference and (4) dynamic, interaction driven, trail-based recommendations.

The rest of this article is structured as follows. In the first two sections we provide important definitions and present related work. Then we describe MoodPlay interface, visual affective model and interactions, followed by a section detailing system architecture and recommendation methods. Next, we describe the user study setup and discuss the key results:

- In general, the system was rated as highly novel and fun to use.
- Visualization of mood information in a visual space significantly improved users' understanding of recommendations.
- Trail-based interaction (example shown in Figure 5) was considered too confusing.
- Visual conditions (2), (3) and (4) tend to improve system trust, after trust propensity is controlled.

Finally, we share our ideas for future work and the resulting implication for design of future interactive recommender systems.

2. Definitions

Research presented in this article focuses on mood based recommendation. However, throughout the article we use related terms, *emotion* and *affect*, to explain different concepts. Here we provide definitions for each of the terms and other relevant constructs.

Affect: Colloquial term that covers a broad range of feelings. It encompasses both emotions and moods [21, 22].

Emotions: Intense, short lived feelings, speculated by most researchers to be directed at someone or something [23, 24].

Moods: General, low intensity feeling states that often lack a contextual stimulus [25]. While the duration of emotions is typically measured in minutes, moods may last several hours or days and cause us to think or brood for a while [25, 26]. Emotions become mood states when grouped into positive and negative categories because such grouping allows us to look at emotions more generally instead of in isolation [25]. Therefore, emotion models such as Circumplex model of affect [27] are often used to represent moods as well.

Visual Mood Space: We use this term to refer to the 2-D space used to plot artists and moods in *Moodplay*'s visualization.

GEMS: This acronym stands for Geneva Emotional Music Scales. It is a music-specific emotion model proposed and validated by Zentner et al. [28, 29], which we used to categorize large number of music related moods. As stated by its authors, "GEMS is the first model and rating instrument specifically designed to capture the richness of musically evoked emotions" [1]. GEMS is a hierarchical model, with three root emotions (Sublimity, Vitality, Unease), 9 corresponding sub-levels, and, in the third level, 45 emotion labels (details in Section 5, Table 1)

3. Related Work

The following aspects are the most relevant to our research: affective-aware recommendations, recommendation of music bands, visual approaches to present recommendations beyond a rating list and affective-aware visualizations of music collections.

3.1. Affective Computing and Recommendations

Research in affective computing has been gaining extensive attention in recent years. Proliferation of mobile and wearable computer devices makes it both necessary and possible to achieve natural and harmonious human-computer interaction. Such devices enable us to track a variety of sources that carry emotional content. For example,

[1] http://www.zentnerlab.com/content/musically-evoked-emotions

https://doi.org/10.1016/j.ijhcs.2018.04.004

different aspects of bodily movement and gestures have been used to recognize emotions: head and hands motion [30], gout patterns [31], body posture [32], to name a few. In the speech domain, vocal parameters such as pitch, speaking rate, formants and modulation of spectral content have also been successfully used to classify emotions in [33, 34, 35]. Furthermore, currently the largest data repository of face videos (2 million) owned by Affectiva[2] is efficiently used to train computers in detecting emotions from facial expressions in real time.

For recommendation purposes, Masthoff *et al.* [36] integrated affective state in a group recommender system by modelling satisfaction as mood, while González *et al.* [37] incorporated the emotional context in a recommender system for a large e-commerce learning guide. More related to our work, Park *et al.* [38] developed probably the first context-aware music recommender that exploited mood inferred from context information. And more recently, Tkalcic *et al.* [39, 40] discussed the role of emotions in recommender systems and introduced a framework to identify the stages where emotion can be used for recommendation.

In the music recommendation domain, several works infer the users' mood for music recommendation based on movements, temperature and weather [41] or from the music content [42]. For example, Griffiths *et al.* [43] measure a variety of contextual and physiological indicators (temperature, light, heart activity) in order to detect mood and recommend music by mapping both user's mood and music on the same emotion map. Zwaag *et al.* [44] take target mood as an input from user and then select songs that direct the user towards the desired mood, while measuring skin conductance to verify the change. Skin temperature [45] and arm gestures [46] have also been used for inferring mood and querying music collections. Compared to these studies, in our up to date work we use mood information associated with a set of seed artists provided by user to suggest new artists in similar moods. In addition, we propose a rich interface to help users explore mood space and choose music in a desired mood. In the future, this proposed system would be greatly enhanced by incorporating a method for detection of user's mood, using sensors available on wearable devices, social media activity or

[2] http://www.affectiva.com

other contextual information.

3.2. Recommendation of Music Bands

Recommendations in the music domain is a well-established field within recommender systems, which have shown, among many others, approaches to recommend tracks [47, 48], albums [49], playlists [6, 50, 51], music targeted at specific venues [52], music targeted at daily activities [4], and artists and music bands [13, 53]. Since our proposed interface aims at recommending music bands, we focus on presenting related work in this sub-field. Hijikata *et al.* [53] used a Naive Bayes recommender to present recommendations of music bands, while Bostandjev *et al.* [13] used a hybrid controllable recommender system with a visual interactive interface, TasteWeights. Compared to these previous approaches, we innovate by using professionally curated mood tags associated with bands to compute similarity, by introducing a user-controllable recommendation interface and by allowing users to explore the music band dataset interactively.

3.3. Visual Approaches to Recommendation

MacNee *et al.* [18] highlights the importance of user-centric approaches to evaluating recommender systems, and of developing interfaces and interaction designs, instead of focusing solely on improving recommendation algorithms. Konstan *et al.* [19], who shows that small improvements in recommender accuracy do not necessarily improve users' satisfaction with a system. However, the development of interfaces that present recommended items in a visual model different from a static ranked list is rather scarce. For example, SFVis [54] and Pharos [55] employ visualizations of social, latent communities to recommend new friends and social websites respectively. Other examples include collaborative filtering recommenders with rich user interactions such as PeerChooser [56] and SmallWorlds [12], and interactive visualizations for recommending conference talks – TalkExplorer [15] and SetFusion [16]. There is also a range of systems that support dynamic critiquing of an algorithm, such as Pu *et al.* [57] and Chen *et al.* [58]. Finally, Nagulendra and Vassileva [17] created an interactive visualization which provides users of social networking sites with awareness of the personalization mechanism. For a detailed review of visual and interactive recommender systems, read Chen *et al.* [59]. Although not focused on personalized recommendation, but rather on navigation of musical datasets, Knees et al. [60] introduced a virtual 3D landscape which allows the user to freely navigate a collection. To the best of our knowledge, *Moodplay* is the first interactive music recommender system that maps the artists in a latent, navigable, affective visual space based on the recently developed music-specific mood model GEMS [28], further explained in section 4.2.

3.4. Affective-aware Visualizations of Music Collections

Although affective-based music selection and recommendation are gaining popularity in both research and commercial settings, the development of visual aids for affective information is still scarce. Nearly all existing visualizations are built upon Russell's circumplex model of affect [27]. This model is today commonly used to represent emotions and moods as a mixture of two dimensions, valence and arousal, positioning them in the coordinate system. Yang *et al.* [61] incorporated it into their music retrieval method, and a commercial application Habu[3] uses it as a platform for music selection based on mood.

However, many emotions cannot be uniquely characterized by valence and arousal values [62]. For example, fear and anger, two distinctive emotions, both have high arousal and negative valence, and are commonly placed close to each other in the circumplex model [63]. It is also important to note that models derived from general research in psychology, such as Russell's, may not be suitable for musical emotions. One reason being that music, unlike other life events, possibly induces more contemplative range of emotions [29]. To address this problem, we propose a novel visual representation of music specific affective dimensions, built upon the GEMS model derived from an extensive psychological study by Zentner *et al.* [28] (see section 4.2 for details).

[3] http://habumusic.com

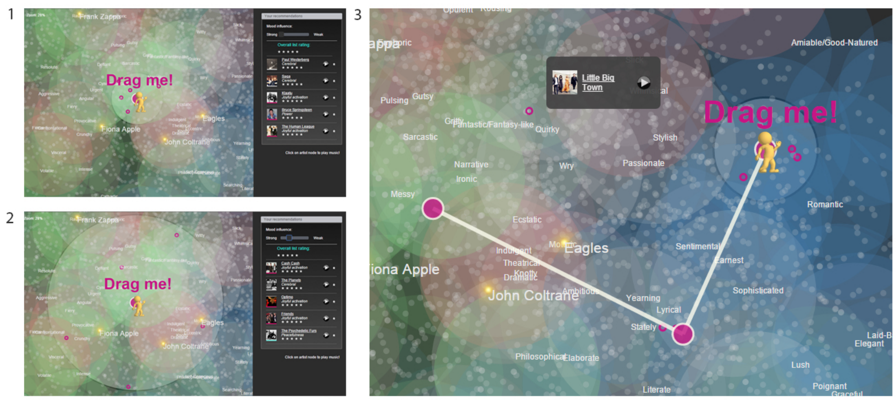


Figure 2: Screenshots of interactive features in *MoodPlay*: (1) and (2) - the varying recommendation catchment area around user avatar, controlled by a hybridization slider, (3) - trail based interaction, along with a display of artist information box upon clicking on artist node in the visualization.

4. System Overview

The MoodPlay system is accessible via web browser and consists of three sections: input, visualization and recommendation panel. Users construct profiles by entering names of artists via an interactive drop-down list (Figure 1-left). Based on the mood information associated with profile artists (see section 4.2 for explanation), the system positions a user avatar in a precomputed visual mood space (Figure 1-center) and recommends new artists (Figure 1-right). In this section we provide an overview of the user interface and explain the method for constructing the mood space.

4.1. Interface Design

Visualization. Visualization of the mood space along with the artists within it is central to solving the problem of navigation through the music collection and explanations of recommendations. The space consists of 266 moods - similar ones being positioned closer to each other than dissimilar ones (the construction method is detailed in section 4.2). Furthermore, moods form a hierarchy with three primary categories at the top - vital, sublime and uneasy (see Table 1), portrayed on canvas in different colors. Red, generally associated with passion and high energy [64], is used for vital mood category. Blue, more serene color [64], denotes sublime category, which includes tender and peaceful moods among others. Lastly, uneasy mood nodes are green

- a color that is occasionally associated with sickness [64], but here chosen mainly for aesthetic reasons to complement the other two category colors. Mood nodes are semi-transparent, their size is equal and purposefully large enough to cause overlap. This produces an interplay of colors, thus forming the space with gradual transitions between mood categories. Artists from our database are placed within the mood space based on moods associated with their music. Users can stream their music in real-time and see additional artist information by clicking on the nodes.

Our interface design follows Shneirdermann's visual information seeking mantra "*Overview first, zoom and filter, then details on-demand*" [65]. As indicated in Table 1 and explained in 4.2, moods form a hierarchy with 3 top categories displayed in distinct colors, and 15 subcategories. User can choose to display individual subcategories via *Browse* tab on the left interface pane (Figure 1). Such a hierarchical view of the large number of moods allows the user to explore the mood space by starting from broad terms and then filtering down selection to specific subset of moods in the visualization, while zooming and panning to more closely inspect areas of interest. Furthermore, we implement a dynamic mood labeling algorithm in order to reduce cognitive load in a dense mood space. We rank the moods based on the frequency of their usage to describe different artists, and show only limited number of the most popular moods at a given zoom level.

Recommendations. An ordered list of recommended artists is displayed in the right panel (Figure 1) and the corresponding artist nodes are highlighted in the mood space. In this way, we aim to provide transparency, trust, efficiency and satisfaction to the user, which are four out of the seven criteria identified by Tintarev and Masthoff [66] to design explanations in recommender systems (the other three are scrutability, effectiveness and persuasiveness). Items in the recommendation list are linked to audio streams via Rdio[4] and to Last.fm[5] profiles of artists. For each recommended artist we also display artist's picture, color of the top mood category (red, blue or green) and name of the sub-category the artist belongs to, with the goal to help users gain some un-

[4]Rdio streaming service was discontinued in December 2015.
[5]http://www.last.fm/api/intro

derstanding of the music upon visual inspection. Furthermore, because recommended items change as a result of the user's interaction with the system, we display up or down arrows next to artist name if its position changed, a horizontal line if it remained the same or a star if the recommendation is new. Rating of recommended items is enabled only for the purpose of the user study, and is achieved by clicking on one of the five stars below artist names.

Interaction. Adaptivity of music recommenders is particularly important due to the dynamic nature of the listening context [67]. Keeping this in mind, we model the gradual change of a user's preference by enabling the movement of the avatar (Figure 2.1) in the mood space and maintaining the array of trail marks, weighted by distance from the current position (Figure 2.3). As the user navigates away from the initial position, we incorporate the mood information associated with each trail mark into the recommendation algorithm. Removing any of the trail marks is possible by simply clicking on it, and deleting the whole trail is achieved by clicking on the initial position of the avatar.

Finally, fine-tuning of recommendations is further supported by controlling the hybridization of recommendation process. Our recommendation approach accounts for the fact that mood-based similarity between artists does not necessarily match audio based similarity (e.g. techno and punk artists are both energetic, but they do not sound similar). Therefore, we allow users to adjust the mood influence via a slider control which dynamically re-sizes a catchment area around the current avatar position (Figure 2.1 and 2.2). The weaker the mood influence, the more we rely on audio similarity to calculate recommendations, and vice-versa.

4.2. Music Specific Visual Model of Affect

A key challenge of this research was showing and explaining inter-relationships between artists and moods in a two-dimensional space. To that end, we analyzed mood metadata for 4,927 artists collected from Rovi[6], which is to our knowledge the most comprehensive collection of professionally curated mood-artist tags. Each artist in

[6]http://developer.rovicorp.com/io-docs

https://doi.org/10.1016/j.ijhcs.2018.04.004

our dataset is characterized by approximately 5 to 20 weighted mood words out of 289 available ones, and represented with a vector $X \in \mathbb{R}^{289}$. Finally, correspondence analysis [68] was used to reduce data dimensionality, which resulted in a latent variable space containing moods and artists.

Numerous emotion models, both continuous and categorical have been proposed in the psychology field [69, 70, 71]. For the purpose of identifying potential clusters in our mood space, we explore whether our visual map fits into a hierarchical, and therefore categorical, music-specific emotion model proposed by Zentner *et al.* [28]. This model, from now on referred to as *Geneva Emotional Music Scales* or GEMS, consists of 3 main categories (vitality, uneasiness, sublimity), 9 sub-categories and 45 music relevant emotion words distributed across different sub-categories. Our hypothesis was that such hierarchy should emerge in the visual mood space built upon professionally curated artist-mood associations. It is important to note that psychology researchers focus on deriving emotion models rather than mood models, and for recommendation purposes, music is generally tagged with mood descriptors. In this paper we use either of the terms depending on the field we address, and an overarching term, affect, in the context of our proposed system.

To perform our hierarchical classification of moods, we employed a WordNet[7] similarity tool[8] and calculated similarity scores between 289 Rovi and 45 GEMS mood words. Furthermore, since similarity between terms in WordNet is based on semantic relatedness and not strictly on synonimity, we evaluated mood classification by subjective observation. For example, the word *volatile* was found to be related more closely to *tender* than *tense* and was placed into sub-category *Tenderness*, rather than *Tension*. Hence, the following steps were taken to reduce observed classification error rate: (1) mood hierarchy is expanded to accommodate moods that do not belong to any of the GEMS categories, (2) 23 of the least frequently used mood words to describe artists in Rovi were discarded, (3) the set of remaining misclassified words are placed into categories that they are more likely to belong to. Table 1 shows the final list of categories

[7]https://wordnet.princeton.edu/
[8]http://maraca.d.umn.edu/cgi-bin/similarity/similarity.cgi

https://doi.org/10.1016/j.ijhcs.2018.04.004

Category	Sub-category	No. of moods	Example moods
Sublimity	Tenderness	24	Delicate, romantic, sweet
	Peacefulness	22	Pastoral, relaxed, soothing
	Wonder	24	Happy, light, springlike
	Nostalgic	9	Dreamy, rustic, yearning
	Transcendence	10	Atmospheric, spiritual, uplifting
Vitality	Power	29	Ambitious, fierce, pulsing, intense
	Joyful activation	32	Animated, fun, playful, exciting
Unease	Tension	32	Nervous, harsh, rowdy, rebellious
	Sadness	18	Austere, bittersweet, gloomy, tragic
	Fear *	10	Spooky, nihilistic, ominous
	Lethargy *	8	Languid, druggy, hypnotic
	Repulsiveness *	10	Greasy, sleazy, trashy, irreverent
Other *	Stylistic *	19	Graceful, slick, elegant, elaborate
	Cerebral *	12	Detached, street-smart, ironic
	Mechanical *	7	Crunchy, complex, knotty

Table 1: Structure and description of *MoodPlay* mood hierarchy. Categories and sub-categories marked with * are the expansions from the original GEMS model.

and distribution of associated moods.

5. Technical Design and Implementation

MoodPlay uses diverse data collected from different sources, mostly through public Web APIs. Recommendations have to be computed very quickly, since they are immediately presented in the interface as a result of user interaction. Therefore, the system requires an appropriate architectural design. As depicted in Figure 3, it has two main components: one for building the library of items with their metadata (*Dataset Construction*) and a second component that generates user recommendations (*Recommendation Framework*). The following subsections describe the architecture design and implementation in detail.

5.1. Dataset

MoodPlay relies on a static dataset of 4,927 artists, obtained in several iterations. First, 3,275 artists were randomly selected from a subset of the Million Songs Dataset[9].

[9]http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset#subset

https://doi.org/10.1016/j.ijhcs.2018.04.004

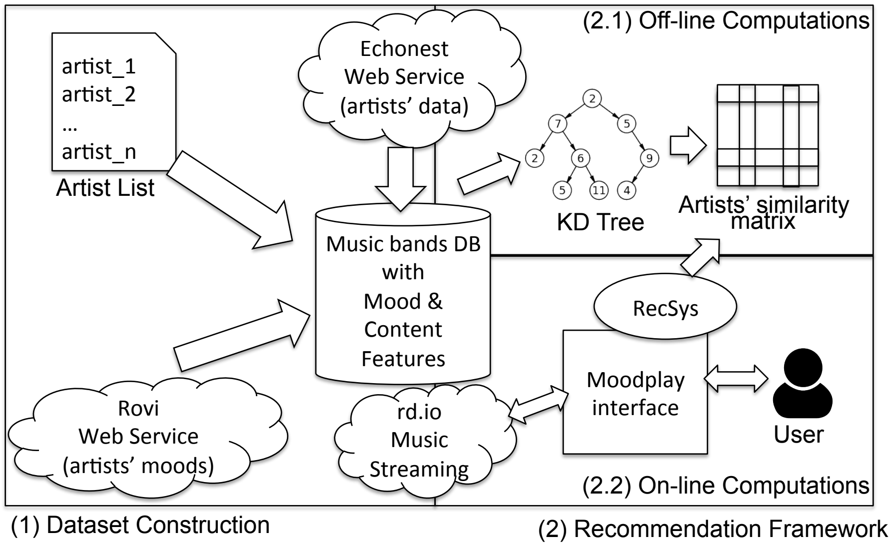


Figure 3: MoodPlay system architecture indicating the modules for: (1) dataset construction and (2) recommendation framework.

Artists ranged from very popular to less known, and played music in a variety of genres and over different decades. The pool was then expanded by 2,000 most popular artists from the public EchoNest[10] database, as measured by proprietary metrics *familiarity* and *hottness*. We complemented the initial set in order to better facilitate an online user study with participants of different ages from different parts of the world. Artists for which we were not able to obtain mood or song data were discarded. Next, mood data for each artist was obtained via Rovi API and the top ten most popular songs for each artist along with corresponding audio analysis data were obtained from EchoNest. Different versions of the same song, having the same title in EchoNest database were discarded. Rdio API was used for music streaming in MoodPlay. Finally, artist pictures and links to Last.fm profiles were obtained via Last.fm API.

[10]http://developer.echonest.com

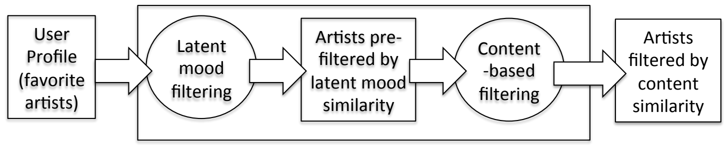


Figure 4: Schematic representing our hybrid cascading recommender which pre-filters based on mood similarity and then post-filters based on content similarity.

5.2. Recommendation Approaches

Our hybrid cascading recommender [72] operates in two stages as shown in Figure 4: (1) using the user profile as an input, our system produces a first candidate set of recommendations based on mood similarity, and (2) the output of the first recommender becomes the input to an audio content-based recommender, which re-ranks the artists and produces the final recommendation list. Such layered approach supports our goal to help the user control and understand how recommendations are generated while navigating mood space. The following paragraphs describe the recommendation process in detail.

Offline computation of artist similarity. Artists' pairwise similarity, based on mood and audio content, is calculated offline and stored in two separate data structures. Mood-based similarity between any two artists is a function of their Euclidean distance in the affective space produced by correspondence analysis. To calculate audio-based similarity, we first identify the 10 most popular songs for each artist in our database via the EchoNest API and obtain audio analysis data for the total of 49,270 songs from the same source. We used timbre, tempo, loudness and key confidence attributes, which amounted to approximately 10,000 numerical values per song. In order to make the similarity calculations efficient, we represent each song with a vector $v_i \in \mathbb{R}^{515}$ [73] and build artist data into a KD-tree [74]. Finally, an accelerated approach for nearest-neighbor retrieval that uses maximum-variance KD-tree data structure was used to compute similarity between songs, since it is has a good balance of accuracy, scale and efficiency [73]. In this way, time complexity of constructing a similarity matrix was reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$, while the search for the K nearest neighbors of a given artist is reduced from $\mathcal{O}(K \cdot n)$ to $\mathcal{O}(K \cdot \log n)$. To compute artist similarity, first, for each song we rank all other songs from the dataset from most to least similar.

DRAFT, to be published in IJHCS
https://doi.org/10.1016/j.ijhcs.2018.04.004

We then calculate average similarity rank of songs per artist [47], thus obtaining the artist similarity matrix (Algorithm 1).

Algorithm 1 Algorithm for computation of audio similarity

```
Input:
    Set of artists: A = {a_1, a_2, ..., a_n}
    Set of songs for all artists: S = {Sa_1 ∪ Sa_2 ∪ ... ∪ Sa_n}
Output: Audio similarity ranks: ARanks = {a_i → {a_j → rank_ij}}
function COMPUTEAUDIOSIMILARITYRANKS
    ARanks = {}                                  ▷ dictionary of artist similarity ranks
    for each artist a_i in A do
        SRanks = {}                              ▷ dictionary of song similarity ranks
        for each song s_k in Sa_i do
            SRanks[s_k] = COMPUTESIMILARITYMAPOFSONGRANKS(s_k, S)
        end for
        for each artist a_j in A do
            ARanks[a_i][a_j] = COMPUTEAVERAGESONGSIMILARITY(SRanks, Sa_j)
        end for
    end for
    return ARanks
end function

function COMPUTESIMILARITYMAPOFSONGRANKS(s, S)
    Rank all songs from S based on audio similarity to song s
    for each s_j in S do
        similarityMapOfSongRanks[s_j] = rank_j
    end for
    return similarityMapOfSongRanks
end function

function COMPUTEAVERAGESONGSIMILARITY(SRanks, Sa)
    average = 0
    for each song s_i in Sa do
        for each song s_j in SRanks.keys do
            average += SRanks[s_j][s_i]
        end for
    end for
    average = average / (Sa.size * SRanks.size)
    return average
end function
```

Online recommendation. During a user session, MoodPlay recommends new artists similar to the artists the user enters into her profile. First, we determine the overall mood by calculating the centroid $C(u) = (c_x, c_y)$ of profile artist positions, where we then place the user avatar. The coordinates c_x and c_y are calculated as in

https://doi.org/10.1016/j.ijhcs.2018.04.004

Equation 1, where P is the set of artists in the user profile, a_x is the $x - axis$ and a_y is the $y - axis$ coordinate of artist a in profile P.

$$c_x = \sum_{a \in P} \frac{a_x}{|P|} \qquad c_y = \sum_{a \in P} \frac{a_y}{|P|} \tag{1}$$

Artists found within the adjustable radius around the centroid are all potential candidates for recommendation because they are considered to reflect the latent moods derived from the user's input. Among the candidate artists, we select the ten most similar to the user profile based on pre-computed audio similarity data, rank them by distance from the user position and display first five as recommended artists (Algorithm 2).

Algorithm 2 Basic algorithm for online music recommendation

Input:
Artists in user profile: $P = \{a_1, ..., a_n\}$
User position: $u = Centroid(a_1, ..., a_n), a_i \in P$ or u is a position from user's trail $T = \{u_1, ..., u_n\}$
Recommendation radius: r
Audio similarity ranks: $ARanks = \{a_i \to \{a_j \to rank_{ij}\}\}$
Number of recommendations: n_{rec}
Output:
Recommended artists: $R = \{a_1, ..., a_n\}$

```
function RecommendMusic(u)
    M = []                                  ▷ artists within mood radius
    for a_i in A − P do
        if distance(a_i, u) < r then M[i] = a_i
        end if
    end for
    H = {}                                  ▷ dict. of artists & similarity with P
    for a_i in M do
        H[a_i] = AverageSimRanking(a_i, P)
    end for
    sort(H)                                 ▷ sort artists by audio similarity
    R = H[1..n_rec]
    return R
end function

function AverageSimRanking(a, P)
    average = 0
    for each a_i in P do
        average += ARanks[a][a_i]
    end for
    return average / P.size
end function
```

Trail-based recommendation. Furthermore, we propose a novel, adaptive recommendation approach that accounts for the preference change in terms of mood, reflected by the repositioning of user's avatar in the affective space. We keep track of each new position and apply a decay function to the preference trail when recommending new artists. Recommendations from the last position in the trail are assigned the greatest weight, because we presume that the most recent mood area of interest is the most relevant to user. The weights further decrease as a function of hop distance from the end of the trail. Pseudocode for the trail based recommendation algorithm is given in Algorithm 3 and here we outline the steps.

At each trail mark, we apply the recommendation algorithm described in the previous sub-sections, which produces an initial set of recommendation candidates. We then calculate adjusted distances d_a between trail marks and surrounding recommendation candidates in two steps. First, we normalize distances between the trail mark and artists because the radius can vary among trail marks. If the distances were not normalized, many relevant artists would be falsely considered irrelevant and would not appear in the final recommendation list. Next, we adjust the normalized distances for each trail mark based on the corresponding weights using the formula $d_a = d_n + \Delta \times (|T| - 1 - i)$, where d_n is a normalized distance, Δ is a decay constant, $|T|$ is a total number of trail marks and i is an iterator over the trail marks. After several tests, we found that weight constant Δ performs the best when calculated as: $\Delta = r_{min}/4$, where r_{min} is the minimal recommendation radius. The larger the value of Δ, the steeper the decay function. Finally, the recommendation candidates are sorted based on adjusted distances, and top five are recommended to the user.

6. Evaluation

Preliminary evaluation of an early version of MoodPlay has been described in [20]. Compared to the previously published study, here we present modified experiment design and more comprehensive analysis of results. We performed a crowd-sourced study with entirely new set of participants. Number of participants was 378, but 279 remained after filtering out those that we did not deem as valid, i.e. those who incorrectly answered attention check questions or ended the study prematurely.

The focus of the evaluation was to understand the effects of mood-based interac-

Algorithm 3 Hybrid recommendation with provenance trails.

Input:
Trail of user positions: $T = \{u_1, u_2, ..., u_n\}$, where u_1 is the profile based position and consecutive u_i are positions that user navigated to
Current recommendation radius: r
Minimum recommendation radius: r_{min}
Number of recommendations: n_{rec}
Output:
Recommended artists: $R = \{a_1, ..., a_n\}$

```
function RecommendMusicBasedOnTrail
    R = {}                                          ▷ dict. of recommended artists
    Δ = r_min/4
    for u_i in T do
        for a_j in RecommendMusic(u_i) do
            d_s = Scale(distance(u_i, a_j), r, r_min)
            d_a = d_s + Δ × (T.size − 1 − i)
            R[a_j] = d_a
        end for
    end for
    sort(R)                                         ▷ sort artists in R by d_a
    return R[1..n_rec]
end function

function Scale(d, r, r_min)
    d_c = Convert d from range [0, r] to [0, r_min]
    return d_c
end function
```

tions with a recommendation algorithm and to independently evaluate the influence of the MoodPlay visualization from an explanatory perspective. To improve the previous experiment design, in this study we gave users more freedom to naturally interact with the system and we tracked additional interaction metrics. Furthermore, in the previous study we focused the evaluation on user characteristics, interaction and experience, and placed less attention on ratings-based analysis. Here we report the results of both quantitative and qualitative analysis and address impact of mood based interactions on user experience.

6.1. Mood data in automated recommendation

Before we proceed with our main experiments on interaction with mood data, we first set out to understand the impact that mood information can have on *automated*, non-interactive algorithms. This is an important step that can provide insight into the utility of different inputs about mood during the interactive process that we describe later. In particular, we describe the results of an automated experiment to show quan-

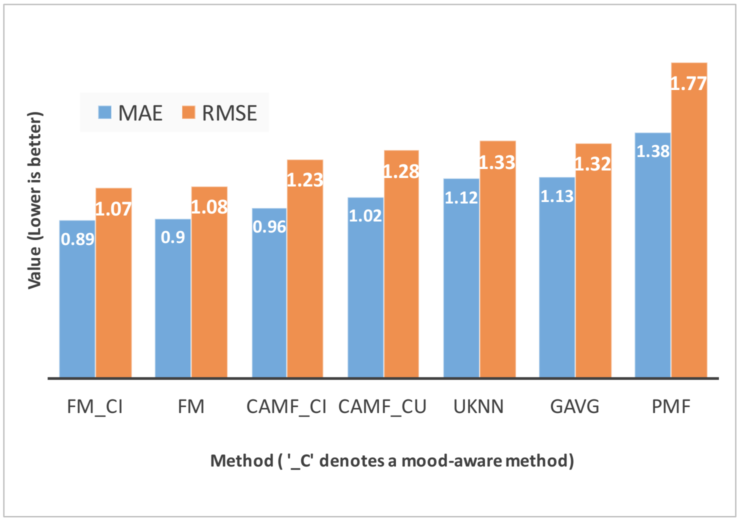


Figure 5: Predictive error (MAE and RMSE) for five recommendation algorithms:. PMF (Probabilistic Matrix Factorization), GAVG (Global average rating), UKNN (User-based K Nearest Neighbors), CAMF (Context-aware Matrix Factorization), FM (Factorization Machines). CAMF_CI, CAMF_CU and FM_CI are tuned with mood data at the item (CI) and user levels (CU), respectively, while the other three are benchmarks and are only tuned on a traditional user-item ratings matrix.

titatively that using mood as context during recommendation can improve recommendation quality. To do this, we use two versions of context-aware matrix factorization method from Zhang et al. [75], trained using a traditional item-rating matrix, along with mood information from our GEMS model, for each item. We compute prediction error to compare against three traditional recommendation algorithms, which are only trained on a matrix of user and item ratings. One limitation of this experiment is that we are predicting over ratings that were gathered through the Moodplay system, which may have influenced the rating in different ways. We plan an additional experiment with ratings gathered from LastFM to verify our results on a separate data set. We run a 5-fold cross validation on a collection of 2,548 ratings of 593 items (musical artists) with 5 associated mood tags per item. Data density was 0.29% and the mean item rating was 2.75 on a 1-5 Likert scale, with a standard deviation of 1.31. Figure 5 shows the results of this experiment for the five methods on the x-axis. The x-axis groupings are MAE and RMSE scores for each algorithm, two popular error metrics used to measure the predictive accuracy of recommendation algorithms [76]. Y-axis shows the value for those metrics. The *CAMF* or context-aware matrix factorization methods

Feature	(1)	(2)	(3)	(4)
Profile generation	x	x	x	x
Ordered list of recommendations	x	x	x	x
Display of mood space		x	x	x
Navigation in mood space			x	x
Hybridization control			x	x
Trail based recommendations				x
Number of subjects	70	69	70	70

Table 2: Table of experimental conditions and associated features. Conditions increase in complexity, (1) having only two and (4) having all available features.

were trained with additional mood information. The *CAMF_CI* algorithm computes mood information over items, while the *CAMF_CU* computes it over individual users. Figure 5 is sorted from left to right by best performance. It is clear that the three mood based methods outperformed the benchmarks, on both metrics. For the benchmarks, we chose a classic user-KNN collaborative filtering algorithm, along with a simple global average predictor and a pure matrix factorization approach. We note that the poor performance of the latter may be a result of sparsity in the ratings matrix – which enables us to make the point that perhaps mood-aware algorithms can be a good bootstrapping mechanism that helps to combat the sparsity problem for traditional matrix factorization algorithms when such data is available.

6.2. Setup

As in the previous study, we set up four conditions having different features, shown in Table 2. The conditions have increasing visual and interaction complexity (see subsection *Interaction* in 4.1 for description of more complex features). In each of the conditions users create a profile by entering artist names. The system uses this information to generate recommendations and display them as a list. In condition 1, users see the list of recommendations but do not see the visualization and mood information. In condition 2, mood space and the user's avatar within it are visible, but interaction is not enabled. Condition 3 allows users to navigate the mood space, move the avatar without keeping track of previous positions, and modify the size of catchment area around the avatar. Finally, condition 4 (Figure 2) tests the full system, including trail based recommendations.

https://doi.org/10.1016/j.ijhcs.2018.04.004

6.3. Study Procedure

Participants accepted the study on Mechanical Turk[11] and were redirected to a Qualtrics[12] pre-study survey with demographic and propensity related questions. Among the questions in this survey, we collected users' current mood by asking them to choose one of the following options: (a) Sublime (e.g.: joyful, warm and tender moods), (b) Vital (e.g.: stimulating moods such as "lively", "energetic" or "fierce") or (c) Uneasy (e.g.: negative moods such as "sad", "tense" or "fearful").

Following this, they were assigned to a random condition and performed the main task. Finally, participants gave qualitative feedback in a post-study survey, also administered through the Qualtrics platform.

During the main task, participants were given step by step instructions in the form of interactive MoodPlay system tutorial. They were asked to enter at least three profile items (music bands) from a drop-down list, shown on the left in Figure 1. In all conditions, this profile was used to generate a list of 5 recommendations, that were shown on the right side of the screen. Ratings were collected for the recommendation list as a whole and 5 individual items in the list. Participants were then allowed to interact freely with the system and generate as many intermediate recommendation lists as they wished. Once satisfied, they again rated the full list of items prior to finishing the MoodPlay interaction task. In our study, ratings were an indicator of users' perceived recommendation quality, or simply how well they liked suggested artists. Although music is subjective, and users may have different criteria for rating (e.g. expectations at a given moment, taste, current mood, similarity of suggestions to profile items), by comparing ratings across different conditions we can evaluate the impact of MoodPlay's features on the perceived quality of recommendations. To ensure that users spent sufficient time in the experiment, we displayed a non-numerical timer and gave users the opportunity to proceed to the post-study after at least 1.5 minutes of interaction.

[11] https://www.mturk.com
[12] http://www.qualtrics.com

6.4. Participants

The 279 valid participants were equally distributed across all 4 conditions: 70, 69, 70 and 70. Studies lasted an average of 20 minutes and participants were paid an amount of $1.00 per study. Age ranges of participants were reported from 18 to over 65, with an average range of 25-30. 52% were female. 13% did not finish college, 40% had a four-year college degree and 47% had a graduate degree. 74% were familiar with data visualization; 66% used a mouse for the interactive study and 34% had a trackpad. When asked about music tastes, 89% said they listen to music frequently. Reported use of streaming services such as Pandora was normally distributed. 71% of participants reported that they preferred a mix of popular and esoteric music. Participants were asked to rate the statement *I am a trusting person* on a scale of 1 to 5, in order to evaluate whether their trust in the recommendation system stems from their trust propensity or interaction with the system. The results were approximately evenly distributed across low, medium and high trust bins. During the design stage of this experiment, approximately 10 informal lab-based studies were also conducted and participants were interviewed to gauge their experiences with the system. Among the questions in the pre-study survey, we collected users' current mood by asking them "*Which of the following best describes your current mood?*". They choose one of the following options: (a) Sublime (e.g.: joyful, warm and tender moods), (b) Vital (e.g.: stimulating moods such as 'lively', 'energetic' or 'fierce') or (c) Unease (e.g.: negative moods such as 'sad', 'tense' or 'fearful')." The results indicated that most people felt to be in a sublime mood (56%), followed by vital (29%), and the fewest, unease(15%).

7. Results

We present our results in five subsections. We first provide details on how subjects interacted with the interface in the different conditions, in order to understand how the design decisions affected the user behavior. Then, in the second subsection, we present results in terms of the diversity and accuracy of the system, comparing artist streaming activity, rating and nDCG[13] differences among conditions. In the third subsection we analyze the effect of mood in the results – both self-reported user mood prior to the study and artists' mood category. Next, in the fourth section we present qualitative results to understand subjective aspects of user behavior. Finally, in section five, we summarize the results by combining both quantitative and qualitative data into an integral analysis. This allows us to explain how visual and interactive aspects of each condition affect the results of objective and subjective metrics.

7.1. User Behavior from Log Analysis

We recorded the amount of time users spent on the interface, but we found no significant differences among conditions. We also logged several user interactions with the system, most of which were clicks on different interface components as shown in Figure 6. While some of these actions were available across all conditions since they were recorded on the user profile and recommendation panels (adding and removing artists in the profile list, playing music by clicking on artists in the recommendation list), other actions were available only in the *visually-enhanced* conditions (clicking on artist nodes and playing music through artist nodes in the visualization). Finally, two interactions were available only in the most advanced condition, where users could actually draw a trail when moving the avatar (creating and removing trail marks). We highlight two results from analyzing these actions and detailed statistics can be found in Table 3.

Preference Elicitation. In MoodPlay, music artists were recommended based on their mood similarity to the artists in the user profile. We found that users added significantly more artists in conditions 1 (M=4.47, SE=0.23), $p < .002$, and 2 (M=3.78, SE=0.13), $p < .003$, than in condition 3 (M=3.03, SE=0.02) and 4 (M=3.09, SE=0.09). Furthermore, while 16 and 11 users removed artists from their profile in conditions 1 and 2 respectively, only one user removed an artist in condition 4 and no user did it in condition 3. In conditions 1 and 2 the only way that users could update their list of recommendations was by adding or removing artists in their profile. In conditions

[13] Normalized Discounted Cumulative Gain - a measure of ranking quality in terms of usefulness of recommended item based on its position in the recommendation list.

DRAFT, to be published in IJHCS
https://doi.org/10.1016/j.ijhcs.2018.04.004

	Condition			
Statistic	(1)	(2)	(3)	(4)
# users	70	69	70	70
avg. artists added/user	$4.47^{2,3,4}$	$3.78^{3,4}$	3.03	3.09
avg. artists removed/user	$0.37^{3,4}$	0.21	0	0
# users who removed artists	$16^{3,4}$	$11^{3,4}$	0	1
avg. total interactions/user	18.11	18.91	$24.27^{1,2}$	$23.25^{1,2}$
avg. recommended lists/user	1.87^{2}	1.53	$2.3^{1,2}$	$2.24^{1,2}$

Table 3: Statistics describing user interactions with the interface in different conditions. Superscript numbers indicate conditions over which the significant difference was found. Significance is obtained via multiple t-test with Bonferroni correction, except for *# users who removed artists*, where it was obtained via multiple proportion test.

3 and 4, users could update the recommendation list simply by moving the avatar in the interactive visualization. On average they generated more recommendation lists than users in conditions 1 and 2. Despite the ease to get a new set of recommendations in conditions 3 and 4 compared to 1 and 2, the results still show that users in all conditions were interested in exploring recommendations beyond the first list - either because they were curious, they enjoyed the system or were not content with the initial recommendations list.

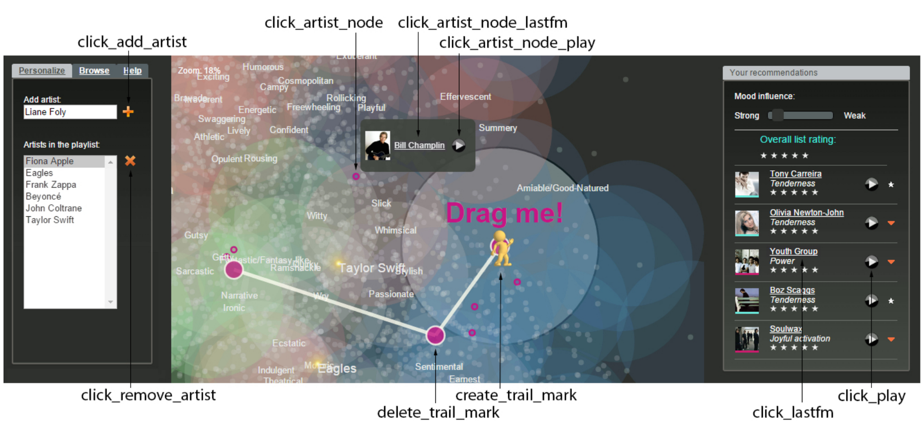


Figure 6: User actions logged during the user study, contextualized on the interface of condition 4, which includes all system's features.

Diversity. One of the most interesting results of our study is that the right amount of interaction functionality in a visual interface can promote diversity among the con-

https://doi.org/10.1016/j.ijhcs.2018.04.004

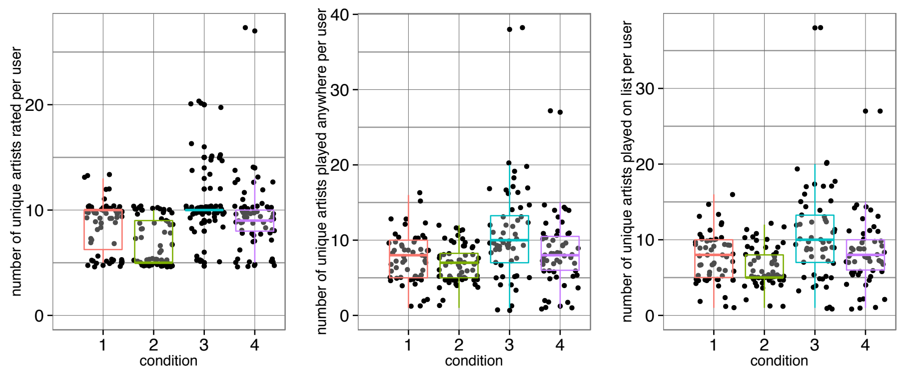


Figure 7: Consumption of unique items per user: rating-based (left) , and played-based interactions all over the interface (center) and on the recommendation list only (right).

sumed items. Promoting diversity in recommender systems is one of the most important topics in the area [77], particularly helpful in preventing the creation of filter bubbles [78]. We measured this effect by comparing the number of unique artists rated and played per user in each condition. With respect to artists played, we considered "playing activity" in any component of the interface (visualization and recommendation panel) and in the recommendation list only, to make a fair comparison against condition 1. Plots in Figure 7 show these distributions. Significant differences were assessed with Wilcoxon signed-rank tests since data departs from normality. We accounted for multiple comparisons with Bonferroni correction. The most important result is that condition 3 significantly outperforms all the other conditions in the three aforementioned metrics: amount of rated items (M=10.59 , S.E.=0.41), $p < .001$, number of artists played anywhere (M=10.71, S.E.=0.81), $p = .002$, and artists played on the recommendation panel only (M=10.61, S.E.=0.82), $p < .003$. Also notable, condition 1 shows significantly more diversity than condition 2 in terms of unique artists rated (M=8.56, S.E.=0.28), $p < .001$, and played in the recommendation list (M=7.63, S.E.=0.43), $p < .02$.

Ranking. In addition to differences in diversity of explored artists, we analyzed differences in ranking among the different conditions. During the study, users had to rate an initial and a final list of recommendation, and they were free to rate more lists in between. We used the metric nDCG [79], since it is a common metric used in rec-

	NDCG and standard error per condition			
	1	2	3	4
First recommended list	0.58 ± 0.02	0.54 ± 0.02	0.58 ± 0.02	0.55 ± 0.02
Last recommended list	0.54 ± 0.03	0.47 ± 0.03	0.58 ± 0.02	0.53 ± 0.03

Table 4: Normalized Discounted Cumulative Gain (NDCG) and standard error for the first and last rated recommendation list, per condition.

ommender systems [80]. nDCG measures the gain of a recommendation discounted by the logarithm of its position in the list. This accumulated gain is high when relevant items (rated 4 or 5) appear at the top of the list and the non-relevant elements (rated 1,2 or 3) are placed at the bottom. Table 4 shows the average nDCG of the first and last lists at each condition. To analyze the differences in nDCG ranking between conditions, we conducted multiple pairwise t-tests with Bonferroni correction. We found no differences in nDCGs at the initial lists, but by comparing the nDCGs at the end of the study, we found that condition 3 had a significantly larger nDCG (M=0.58,SE=0.02) than condition 2 (M=0.47, SE=0.02), p=0.048. Since the recommendation algorithm was the same in all four conditions, only the visualization and interaction could explain the observed differences among conditions. Condition 2 provides a visualization which allows users to explore the dataset (artists) in terms of mood, but unlike condition 3, it does not allow them to update the recommendation list through the visual mood space. This might explain the better ranking quality that was observed in condition 3.

7.2. Mood Analysis

We designed *Moodplay* to investigate the effect of music exploration based on mood categories upon user satisfaction. In this investigation, we also explored whether there is a connection between users' self-reported and the affective profile of the music that they listened to. During the pre-study, we collected users' self-reported mood as described in Section 6.3: Sublime, Vital or Uneasy.

For the sake of understanding the context, the distribution of artists' primary mood categories and users' self-reported mood are shown in Figure 8. The axes of users' self-reported mood distribution are flipped to facilitate comparison with Table 5, which

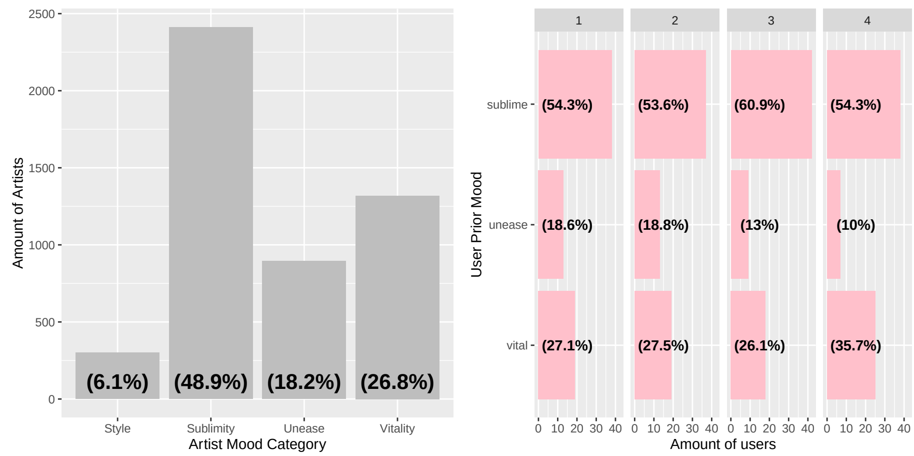


Figure 8: Distributions of primary artists' mood (left) and users' self-reported moods in each condition, prior to the beginning of user study (right).

		Average Artists' Moods Weights			
	User-Reported Mood	Art. Sublimity	Art. Uneasy	Art. Vitality	Art. Style
	(overall dataset)	0.38	0.21	0.27	0.14
Artists in User Profile	Sublime	0.37	0.22	0.27	0.14
	Unease	0.36	0.24	0.27	0.13
	Vital	0.37	0.20	0.29	0.14
Artists Rated	Sublime	0.40	0.16	**0.29**^{U}	0.15
	Unease	0.38	**0.20**^{S,V}	0.26	0.16
	Vital	0.38	0.16	**0.30**^{U}	0.15

Table 5: Average weight of each artist mood (columns) in the whole dataset (first row), among artists added to user profile (rows 2-4), and among artists rated by users (rows 5-6). Statistical tests were performed column-wise. Within each column, we found statistically significant weights only among artists rated, showing that usage of the MoodPlay system actually changed the consumption depending on user's self-reported mood.

shows the average weights of each primary mood category for artists (Sublimity, Uneasy, Vitality, Style) versus users' self reported mood (Sublime, Unease, Vital). In addition, Table 5 compares two groups of data: artists that users added to their profile, and artists that were rated. We observe the following main trends in this analysis:

- In Figure 8, we observe that sublimity is the most frequent mood in both: (a) artists' primary mood category, and (b) users' self-reported mood. This is followed by vitality and uneasy. In the case of artists, the least frequent primary mood is style, the one we created in our research to expand the GEMS model.

- Artists have weights in several mood categories. In the first row of Table 5, we show the average weight of each category over the whole artist dataset. The mood category with largest average weight is Sublimity (0.38), followed by Vitality (0.27), Uneasy (0.21), and finally Style (0.14).
- In rows 2-4 of Table 5 we split the data based on users' self-reported mood and consider the artists added to user profiles (965 total). When comparing each artist primary mood category (columns) across the three potential users' reported moods (rows), we found no statistically significant differences. This means that, on average, users added artists with similar primary mood distribution to their profiles, independent of their self-reported mood.
- The last three rows in Table 5 show the average weights in each primary mood, for each users' self reported mood, but this time considering the artists rated by users (2,704 ratings in total). We found a couple of statistically significant differences. For example, users who had Unease as their reported mood were more likely to listen to/rate music with high uneasy mood (0.2) than users in Sublime (0.16) or Vital (0.16) self-reported mood. Additionally, users in either Sublime (0.29) or Vital (0.3) reported mood, rated/listened to artists with significantly higher Vitality than users who reported feeling unease (0.26).

7.3. Analysis of Post-Study Survey

In the post-study survey, we analyzed user perception of the system. As expected, perceived ease of use drops-off with higher interface complexity and confusion increases, with condition 1 being significantly less confusing than all the rest, and also easier to use than conditions 2 and 4. Interestingly, the difference is not significant when compared to condition 3, which is visually as complex as condition 2, but offers more controllable functionality (e.g. draggable user avatar).

https://doi.org/10.1016/j.ijhcs.2018.04.004

	Mean agreement and standard error per condition			
Statement	1	2	3	4
I trusted recommendations from the system	37.1 ± 3.6	44.6 ± 3.5	48.8 ± 3.5	38.4 ± 3.6
Interaction with the interface increased my trust in the recommendations	43.4 ± 3.6	47.1 ± 3.8	49.4 ± 3.8	39.2 ± 3.7
The recommendations were diverse	60.9 ± 3.3	65.3 ± 3.3	68.6 ± 3.3	59.9 ± 3.3
The interface helped me understand and compare moods of different artists	49.4 ± 3.5	55.7 ± 3.5	55.7 ± 3.3	46.3 ± 3.4
The interface helped me understand how recommendations were generated	42.8 ± 3.6	54.4 ± 3.9	$\mathbf{58.6^{1} \pm 3.8}$	50.3 ± 3.7
The interface allowed me to control the recommendations	42.7 ± 3.6	53.8 ± 3.4	$\mathbf{63.8^{1} \pm 3.9}$	52.7 ± 3.5
The interface was confusing	23.1 ± 3.1	$\mathbf{45.3^{1} \pm 4.1}$	$\mathbf{46^{1} \pm 3.9}$	$\mathbf{52.6^{1} \pm 3.9}$
Overall, the recommendations were accurate	36.2 ± 3.6	40.7 ± 3.6	$\mathbf{49.8^{1} \pm 3.7}$	38.7 ± 3.5
The system was easy to use	$\mathbf{73.9^{2,4} \pm 3.6}$	58.3 ± 3.8	63.8 ± 3.6	53.2 ± 4
The interface was slow	22.2 ± 3.6	32.2 ± 3.84	31.9 ± 3.6	$\mathbf{37.8^{1} \pm 3.2}$
The tutorial explained the system reasonably well	$\mathbf{72.6^{4} \pm 3.3}$	62.9 ± 3.4	65.8 ± 3.2	57.5 ± 3.6
By the end of the session I was satisfied with the recommendations	42.2 ± 4.1	44.2 ± 4	49.3 ± 3.9	38 ± 3.6

Table 6: Summary of the most relevant variables in the post-study survey. Numbers indicate average user agreement (on a scale from 1-100) with mean $\pm$ S.E. Values in bold indicate significant difference over a condition indicated by the superscript number. Multiple comparisons were adjusted using Bonferroni correction.

Furthermore, we did not see clear differences in average ratings per condition, but the perception of accuracy in condition 3 is significantly higher than in condition 1. This result is very interesting, because we are using exactly the same recommendation algorithm in both conditions, but the perception of accuracy changes with the addition of visualization and draggable avatar. As expected, people also felt that condition 3 allowed them significantly more control than condition 1.

Several aspects were not perceived significantly different among conditions, such as trust (rows 1 and 2), diversity of recommendations and helpfulness of interface to compare moods of different artists. Nonetheless, users perceived condition 3 more helpful in understanding how recommendations were generated than condition 1.

All of these results indicate that the visual layout of moods and artists, with the addition of ability to re-position the avatar in the mood space and control the hybrid recommendation algorithm, gradually improve user experience. However, introduction of trails in condition 4 has a negative effect, most likely because of the cognitive overload. In addition, we suspect that trails may be perceived as limiting for the exploration. It is possible that users expected to receive recommendations based on the most

Cond.	Positive comments	Negative comments
1	All good.	Add more bands/artists to the search- for example, neither Silversun Pickups nor Smashing Pumpkins were found to add to my list.
	It was really fun.	The recommendations didn't seem to match the artists I chose.
	I enjoyed using this!	Show more information on how the mood of a song/artist is determined.
2	I think this could be a great tool. Good luck with the progress I am anxious to give it a try when it is finished	I put in 3 rappers and it gave me like oldies and pop songs. Genre plays roles in certain moods.
	I really liked this, it is a new concept that I've never seen. It helped introduce me to artists in different genres that I had never heard before and were very good.	It runs a little slow, should improve optimization for older computers.
	The mood cloud is awesome, and I didn't know there could be so many different music moods, that was great, but not being able to explore the artists within each specific mood circle causes some frustration. Making the cloud more dynamic to dragging and clicking would enhance the tool.	I really didn't understand it.
3	Really good player, i would change nothing it actually made me listen to a couple of artists i did not know about and liked their music.	Make the interface simpler and more concise. Speed up loading times
	An interesting concept. I use Pandora a lot, and my stations are usually based off of my mood that day. This tool would be useful for randomization of choices of music.	It was slow and laggy and some of the recommendations didn't have a play button. I'd like the option to buy a track if I heard one I really liked, or to save a playlist if I really enjoyed it.
	This is really cool, I do not listen to much music and I think this would help me find some new artists or even be used as a therapy tool.	Larger music selection, possibly change the strong week slider, to broad or specific to the particular mood you are feeling.
4	its a cool design	Some of recommended artists didn't relate to my mood close enough.
	Neat program! If I could practice with it more I think I would really enjoy it.	There is a lot of text on the page and it's a little overwhelming. Instead of starting off with so many "moods," maybe just have 20 initially listed.
	It was excellent! Thanks to the developers for developing wonderful tool.	Make the interface faster and smoother. There was too much choppiness when I was using the visualization tool.

Table 7: Selected positive and negative user feedback grouped by experimental condition.

recent trail point, whereas the system accounts for all previous trail points.

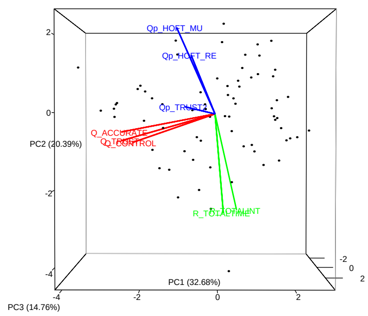
640 *7.4. Qualitative Analysis*

In each condition, participants were asked in the post-study survey to leave feedback on their experience and give suggestions for improving the system. Table 7 lists representative comments, grouped by condition and sentiment. On the positive side, many users had fun using *MoodPlay* and enjoyed discovering new artists in different moods in conditions 3 and 4. Participants reported in all four conditions that the artist database was small, compared to commercial systems, and also had mixing of genres in the recommendation lists. In addition, visualization rendering was sluggish for some users. These problems can be addressed in the future by considering genre in the recommendation algorithm and by optimizing the visual design for even larger artist database.

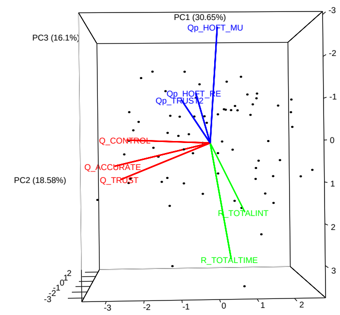
7.5. Connecting Behavioral and Perception Measures

In order to explore the relationships between quantitative and qualitative experimental results collected during the user study, we performed Principal Component Analysis (PCA) [81], a technique for dimensionality reduction, over the variables that have shown significant effects in previous studies [13, 14, 82, 83]. Our analysis focused on the following variables:

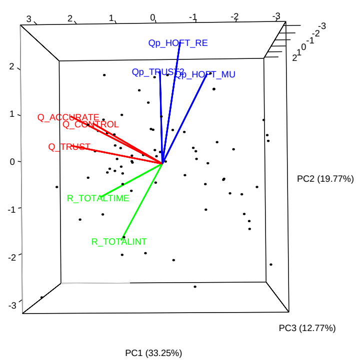
- Qp_HOFT_MU (pre-study question): *How often do you listen to music online?*
- Qp_HOFT_RE (pre-study question): *How often do you use recommender systems?*
- Qp_TRUST2 (pre-study question): *Are you a trusting person?*
- Q_ACCURATE (post-study question): *How accurate do you think the recommendation were?*
- Q_CONTROL (post-study question): *Did you feel in control of the interface?*
- Q_TRUST (post-study question): *How much do you trust the recommendations suggested during the study?*
- R_TOTALINT: Number of user interactions with the system (clicks, music plays, ratings, etc.)
- R_TOTALTIME: Duration of the user study.



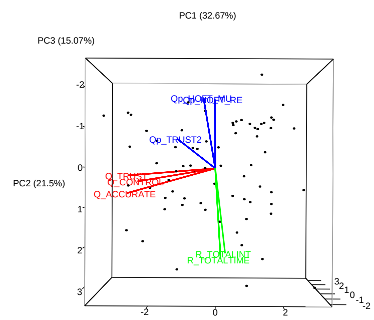
(a) Condition 1: no visualization, static list of recommendations.



(b) Condition 2: visualization without interactive update of recommendations.



(c) Condition 3: visualization with avatar and interactive recommendations.



(d) Condition 4: visualization with avatar, trails and interactive recommendations.

Figure 9: 3D Biplots for Principal Component Analysis of the experiment variables: (i) pre-study survey (blue), (ii) post-study survey (red) and (iii) user interaction (green).

Figure 9 shows biplots drawn from the output of PCA for each condition separately, with arrows denoting each of the variables in the above list. For interpreting PCA plots we used the guidelines described in [81]: (a) the length of a vector represents the variance of that variable (within the principal components used in the biplot), (b) the cosine of the angle between a vector and an axis indicates the importance of the contribution of the variable to the corresponding principal component, (c) the cosine of the angle between pairs of variables indicate how correlated they are, and (d) uncorrelated variables are at right angles with respect to each other.

Based on these guidelines, we plan to compare the influence of user interaction data such as time and clicks, on the results (accuracy, control, trust) versus the influence of user characteristics (trust propensity) and prior experience (familiarity with recommenders and music).

Condition 1. We observe that the red post-study variables (Q_CONTROL, Q_TRUST, Q_ACCURACY) are strongly correlated with each other. On the other side, they are almost orthogonal to variables R_TOTALINT and R_TOTALTIME, indicating that the amount of time spent and interaction with the interface had small or no effect on variables which have been shown to influence the final user satisfaction with the system [14]. Furthermore, blue-colored post-study variables are loading in the same direction as users' pre-existing level of trust (Qp_TRUST2), but the short length of this vector tells us that its total variability is not well explained by principal components PC1, PC2 and PC3. Finally, the familiarity of users with music (Qp_HOFT_MU) and how often they listen to music online (Qp_HOFT_RE) are strongly correlated between each other, but they do not explain the variability of red post-study variables.

Condition 2. Similar to condition 1, the prior levels of trust (Qp_TRUST2) and familiarity with recommendation systems (Qp_HOFT_RE) load in the same direction as perceived control, trust and accuracy (Q_CONTROL, Q_TRUST, Q_ACCURACY) in PC1. However, Q_CONTROL departs from Q_ACCURATE and Q_TRUST on the projection over PC2, which implies that the user perception of accuracy and trust diverted from the perception of control, compared to condition 1. This observation may explain some previous negative results in condition 2, such as the low number of unique artists rated and played compared to other interfaces. Since users were able to explore the mood space visualization, but they could not update the list of recommendations by interacting with it, their perception of control diverted from the perception of trust and accuracy.

Condition 3. The preference of users for this condition, shown in the previous analyses, can be explained holistically with the PCA plot. This is the only condition where Q_TOTALTIME and Q_TOTALINT load in the same left direction as Q_CONTROL, Q_ACCURATE and Q_TRUST in PC1 – the PC which explains most of the data variance in this condition. Notably, the acute angle between the red post-study variables and R_TOTALTIME shows that the amount of time that users spent on the interface explains the post-study variables, and especially Q_TRUST, more than in any other condition. This is an important result since it might indicate that both visualization and interaction combined help to increase the trust that users have in the recommendation system. By looking at the angles between variables in PC1, we can also see that Q_TRUST (final perceived trust) is more correlated to R_TOTALTIME than to Qp_TRUST2 (the initial level of user trust).

Condition 4. Similar to conditions 1 and 2, and unlike 3, this condition shows a disconnection between the amount of time and interaction on one hand, and the user perception in terms of Q_ACCURATE, Q_CONTROL and Q_TRUST on the other. The acute angle between Qp_TRUST2 and Q_TRUST shows that user's trust in the system is more likely to be determined by the inherent user trust than by the time spent interacting with the system. The plot also shows an opposite relation between the prior user experience with recommender and the amount of user interaction (since vectors form roughly a 180° angle), which may suggest that the user was confused and not fully taking advantage of advanced interface features. This could explain the drop in user satisfaction of this interface compared to condition 3.

8. Discussion

There is a fertile ground for expansions and branching of this research in several directions. The overarching idea is to build a system that recommends music according to user's musical taste, and guides the user from her current mood to the desired (target) mood. One caveat about the discussion, and our results, is that our experiment is based on single-session interactions with the system. Ideally, a longitudinal study in a real-world music listening context should be performed, and the authors are exploring possible ways to to achieve this. We now discuss the research questions in light of the study results and summarize the development process and features of MoodPlay system. In the next section, we follow with limitations and avenues for future work.

(RQ1) What are the effects of interactive visualizations on the user experience with a recommender system, and what is the right amount of interaction for a music recommender?

The user study results clearly showed that the interface design and a certain combination of interactive features improve objective and perceived recommendation accuracy, as well as self-reported user satisfaction. We have shown that introduction of hybridization control for recommendation algorithm and the ability to move a user avatar, yielded positive effects across a variety of examined metrics. However, tracking of user mood states in the form of locations on the provenance trail introduced undesirable effects. First, we suspect that this increased system complexity beyond a comfortable threshold and caused cognitive overload, although another potential reason is that the trail did not match the users' mental model, preventing them from navigating the collection of artists effectively for the purpose of identifying relevant artists based on mood. Second, users who are unfamiliar with the system and participate in short listening sessions may be more inclined to rapidly investigate the mood space than those who are familiar with it and use the system in a more natural setting. Thus, the trail may have been perceived as a limitation during relatively short experiment sessions. Nevertheless, modeling of changing mood preference is a fruitful research endeavor and our future work can address trails that follow smoother mood transitions, that are optional and are used during longer listening sessions.

(RQ2) Does affective information improve recommendation accuracy and user experience versus when it is not included?

Our analyses in section 7.2 show evidence that both user mood prior to the study and the primary mood associated with the artists have an effect on the distribution of ratings. This difference in distributions is more pronounced between the interface with no visualization (condition 1) versus the other interfaces which have visualization (2, 3, 4). This result shows that making people aware of the mood of the artists combined with appropriate interactivity in a music recommender, can change the way they perceive the accuracy of the same music algorithm. In particular, when users' self-reported mood was *Unease* (associated with anger, sadness, depression) their overall rating decreased compared to users with different self-reported mood (Sublime and Vital) only at the conditions with visualization of moods. Moreover, when users conducted the study while in *Sublime* mood, they were more likely to provide higher ratings to artists with mood *Style*, while the interfaces with visualization received higher ratings for artists with primary mood *Unease*. Overall, there is a need for further research to establish a causal link between users' self-reported mood, artists' mood, and interface, but the results of our study hint towards a direction where all these variables play an important role in building recommender system interfaces.

Visualization of affective metadata for a music recommender system

Mood information has been visually represented in several preceding works, with the goal of enabling user selection of artists in desired moods. Typically, users choose a mood point in the visual space and the system plays music associated with the selected mood. To our knowledge, all up to date visualizations of moods for this purpose are based on a circumplex model of affect, which represents moods along valence and arousal dimensions (see section 3.4). We argue that there exists a need to use a music specific mood model for the purpose of music recommendation, and propose an approach to fulfill it. Specifically, a dimensionality reduction method was applied to high-dimensional data containing mood-artist associations. It was then shown that a mood model, previously developed in music psychology research, emerges in the obtained two-dimensional, visual mood space.

To use this space during recommendation process, and help users to get a better understanding of it, several design aspects were addressed when incorporating it into an interactive system. Though not all explicitly tested in the user study: choice of colors, item sizes and transparency, dynamic labeling of mood nodes and node filtering based on mood categories, they all aim to explain the mood space and support the recommendation.

Supporting interaction, explanations and control over such visualization

The interaction with the system ranges from zooming and panning the visualization to explore the moods and artists, to controlling the hybridization level of the recommendation algorithm. Both user profile items and recommended artists are highlighted in the visualization, which helps users understand how those two sets are related based on moods. The explanation and exploration are further supported by providing links to external artist profiles, music streaming on demand, and displaying mood categories for recommended items. Moreover, a user avatar is positioned within the mood space at the centroid of user profiles items. The ability to move the avatar and form a trail of mood markers serves as a mechanism for modeling the change in user preference.

This way users can control the recommendations, which are regenerated whenever the position is changed. A second way to influence the recommendation algorithm is by setting the ratio between mood and audio based filtering. This is achieved via simple slider control and visually explained to the user by re-sizing the catchment area around the avatar when the slider is moved. As the area increases, the recommendation results depend more on the audio similarity to the profile items and less on the mood metadata.

9. Limitations and Future Work

Visual mood space. In this research, a key goal was to explore the use of a (visual) hierarchical mood model, capable of handling different granularity in the way moods are represented. Our model encompasses moods that are difficult to represent on the traditional valence arousal scale (Russell's Circumplex model). Our aim was also to enable location and exploration of moods via hierarchy, and therefore bring more diversity to research in mood scales. In the future work, however, we would like to compare our results with the traditional model as a benchmark.

Recommendation algorithm and interface. In terms of accuracy, users perceived the system as lacking prediction power, since the final survey resulted in an average evaluation among 36.2 - 49.8 out of a maximum of 100. We acknowledge this weakness of our implementation, and highlight the following aspects for improvement. First, building the mood space using a larger artist database could improve the mood-based component of the recommendation algorithm. Next, the MoodPlay system accounts for audio similarity when recommending music, but audio content analysis does not always accurately distinguish between music genres. Therefore, the recommendation algorithm can be improved by incorporating genre information. In addition, the system uses an audio similarity method that previously yielded satisfactory results but further investigation and comparison of algorithms could produce a better outcome. Finally, although the off-line algorithm evaluation found strong results when using state-of-the-art methods such as factorization machines and context-aware matrix factorization, for the user study in Mechanical Turk we used a simpler approach. Since we faced a cold-start problem (we had no previous preference information of the users), we relied

https://doi.org/10.1016/j.ijhcs.2018.04.004

on a version of our hybrid content and mood-based algorithm which led to less accurate predictions in some cases, but could be easily tuned with increased training data. Importantly, the comparison between interfaces and the effect of mood, which are the main aspects of this research are not affected by this.

Identifying user mood and musical preference. In the current implementation of MoodPlay, users build their profile by manually selecting several artists and we make recommendations based on the overall mood derived from that profile. We argue that it is acceptable to use mood data at the artist level, rather than on song level, because multiple moods associated with each artist in our database describe that artist's repertoire of songs. Nevertheless, using individual songs as an input and recommending tracks accordingly could perhaps yield greater precision. Another important consideration is that MoodPlay was introduced to users as a platform where they can create a list of favorite artists and be recommended new artists in similar moods. Hence, user's profile in MoodPlay reflects musical taste, but possibly it reflects the combination of both user's taste and mood. Although the core of our study was to explore how different interactive features affect the user experience with a mood aware recommendation system, and not to build a taste profile or auto-detect user's mood, it is important to note that taste and current preference based on mood can be treated as separate, but related parameters. Both can be determined by explicitly asking a user to provide the information, or implicitly, based on relevant data that has been collected automatically. In the following paragraphs, we focus on ideas for determining current preference as reflected by current mood, which could also capture the dynamics of a user's musical taste if tracked over longer period of time.

An approach used by some commercial recommendation systems (e.g. Spotify[14], GooglePlay[15]) is to let users type in a mood or select it from a predefined list. This is not always an efficient method nor an easy task for users given the large number of available mood tags. In particular, mood data in our system is very detailed and attempts to capture nuances that characterize different artists (e.g. rowdy, playful, grace-

[14] http://www.spotify.com
[15] http://play.google.com/music

ful, elegant), whereas typical users may employ a smaller vocabulary and less specific words to describe mood of the music (e.g. happy, sad, energetic, calm). An alternative approach is to use the mood hierarchy embedded in MoodPlay to pick a top mood category from a list, followed by a subcategory and finally select specific mood. In this manner we can mitigate matching issues that might arise from granularity or choice of language for mood.

Implicitly determining user mood in an automated fashion on a granular level is even more challenging. However, an implicit approach can be effective if used with less specificity because it can entirely free the user from interaction. If greater granularity is desired, it can be improved by asking for some minimal input. Extensive research in affective computing, and discussed in section 3.1 considers multiple mechanisms for improving data collection in MoodPlay. For example, a user's mood and current preference could be determined from contextual data such as: social media statuses, time of the day, weather, activity automatically inferred from GPS location or proximity of friends in the network, facial expression captured by mobile device or bodily functions measured by wearable devices. Another key benefit that arises from rich passive profiling data, is that mood can be inferred through behavior, and can serve to inform the system in a better way than a direct self-report from the user. In addition, there are indirect ways to measure users' mood, by asking them to choose certain colors, images or sound clips, which reflect how they are feeling at a certain time. For instance, the system AMARA (Affective Museum of Art Resource Agent) allows the users to explore art collections by asking them simple questions about their current feelings and interests in artwork [84]. This is especially useful when the user is in some state of denial about a current mood, or has any other metacognitive issue with reporting current mood.

Identifying target mood. It is not always desirable to play music that directly matches the listener's current mood. Instead, listeners may be interested in hearing music that changes how they feel. For example, happy music can uplift a listener who is feeling sad. Conversely, some people enjoy bitter-sweet music when sad, while at other times they might prefer springlike or playful songs. Target mood largely depends on a personal preference and current conditions, and therefore the recommender requires complex input or a highly advanced sensing algorithm to determine it.

Depending on the listening context and preference, the recommender can either suggest music in the target mood or find and follow a path from current to target mood. Commercial recommendation systems already offer playlists for different moods and activities (e.g. mellow, music for work or gym), which are effective for short term, action-based listening. However, to the best of our knowledge, there are no recommenders that allow transitions from one state to another or adapt to changes in how user feels or changes in listening context.

Adaptive recommendation systems have been an active area of research in recent years. Looking beyond their applications in entertainment, adaptive music recommenders can be of particular value in music therapy. Recent studies show positive effects of music on recovery of movement (e.g. in patients with stroke or Parkinson's disease) and speech [85]. Music therapy with the goal to modulate emotions has been studied less extensively, but its benefits to pain and mood management have been documented [86, 87]. The current version of MoodPlay has attracted interest from music therapists because its engaging interface can aid choosing music during therapy sessions for hospitalized children and elderly people with dementia. However, in a broad sense, adaptive recommendation systems can help to create a profound impact on a listener's well being, outside of formal therapeutic settings. By being able to continuously monitor feedback about a user's state and context, and adapt to changes, the therapeutic benefits of music can be improved. Our future work will look at these monitoring mechanisms with a view to tuning MoodPlay to adapt readily to observed changes in patients' physical and emotional contexts.

Path from one mood state to another. The trail algorithm in MoodPlay can be viewed as a crude way to create a trail (path) from one mood state to another and generate recommendations accordingly. Through the evaluations we have performed with the system, we observed via numerous metrics that users preferred recommendations obtained by navigating the music collection freely, over the recommendations given by a trail based algorithm. We do not assume that this means that path-based computation of music recommendations are a bad thing, but rather that we need to improve our visual and interaction design for this aspect. User feedback comments and interviews

https://doi.org/10.1016/j.ijhcs.2018.04.004

lead us to believe that level of control of the path-based algorithm is key factor in user satisfaction. For example, users could be given a choice whether to use the system in an exploratory mode and freely navigate, or in preference modeling mode where they build the trail. Depending on a user's activity, available time and listening context, she could choose to engage more or less with the system. In cases when the user chooses to build a trail, recommending items along the trail (in between the trail marks) could provide more gradual change in the recommendations and possibly offer a more enjoyable listening experience during long sessions. Such a recommendation method would require evaluation in a more natural setting, and over a longer period of time.

Scalability. MoodPlay was developed on a database of 5,000 artists. In comparison, online streaming services offer access to tens of millions of artists. In order to maximally scale the system, extensive work is needed in several areas. Even though there are efficient ways to perform dimensionality reduction of millions of data points, visualization design has to be adapted to accommodate such a large number. One simple way to achieve this is to show only limited number of artists on different zoom levels, according to some criteria such as popularity or relevance to a user based on preference data. A challenge in such a filtering method is to determine what artists the user is interested in seeing, and to show popular artists but also encourage discovery by introducing less known artists.

In a survey of dimensionality reduction methods, Fodor [88] argues for the usefulness of dimensionality reduction for high dimensional data. The argument is that not all of the variables are "important" for understanding the phenomenon of interest in a high-dimensional data set. Our multidimensional data, where many moods are associated with each artist, poses a similar challenge. We apply a correspondence analysis, but note that other methods (for example genetic approaches, factor analysis or multi-dimensional scaling) may yield different layout of moods, and therefore different recommendations. However, an exhaustive comparison of these methods is beyond the scope this work. Furthermore, dimensionality reduction can introduce noise, creating clusters which did not exist in the original high-dimensional data. In future work, we believe it will be revealing to inspect the results of several dimensionality reduction techniques, with a tool such as the one introduced in Stahnke et al. [89], and make advances in dynamic interactive labeling of the reduced space to help inform users of the underlying semantics of the space.

10. Conclusion

This paper presented and evaluated *MoodPlay* –a hybrid recommender system for musical artists which introduces a novel interactive visualization of moods and artists. The system supports explanation and control of a recommender system via manipulation of an avatar within the visualization. Design and implementation of an online experiment (N=279) was presented to evaluate the system through four conditions with varying degrees of visualization, interaction and control. Our key results have shown that interface design and a certain combination of interactive features improve objective and perceived recommendation accuracy, as well as self-reported user satisfaction with the recommender system (RQ1), and that making people aware of the typical mood of an artist's music, combined with appropriate interactivity in a music recommender, can change the way users perceive the accuracy of the recommendation algorithm (RQ2).

Funding Sources. This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S.Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. The author Denis Parra was funded by the Chilean research agency CONICYT, FONDECYT grant number 11150783.

References

[1] M. D. Ekstrand, J. T. Riedl, J. A. Konstan, Collaborative filtering recommender systems, Found. Trends Hum.-Comput. Interact. 4 (2) (2011) 81–173. doi:10.1561/1100000009.
URL http://dx.doi.org/10.1561/1100000009

[2] M. J. Pazzani, D. Billsus, The adaptive web, Springer-Verlag, Berlin, Heidelberg, 2007, Ch. Content-based Recommendation Systems, pp. 325–341.
URL http://dl.acm.org/citation.cfm?id=1768197.1768209

DRAFT, to be published in IJHCS
https://doi.org/10.1016/j.ijhcs.2018.04.004

[3] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (8) (2009) 30–37. `doi:10.1109/MC.2009.263`.
URL `http://dx.doi.org/10.1109/MC.2009.263`

[4] X. Wang, D. Rosenblum, Y. Wang, Context-aware mobile music recommendation for daily activities, in: Proceedings of the 20th ACM international conference on Multimedia, ACM, 2012, pp. 99–108.

[5] L. Baltrunas, X. Amatriain, Towards time-dependant recommendation based on implicit feedback, in: Workshop on context-aware recommender systems (CARS'09), 2009.

[6] F. Maillet, D. Eck, G. Desjardins, P. Lamere, et al., Steerable playlist generation by learning song similarity from radio station playlists., in: ISMIR, 2009, pp. 345–350.

[7] V. Castaignet, F. Vavrille, Musicovery, `http://musicovery.com/`.

[8] S. Koelsch, A neuroscientific perspective on music therapy, Annals of the New York Academy of Sciences 1169 (1) (2009) 374–384.

[9] I. Fernández-Tobías, I. Cantador, L. Plaza, An emotion dimensional model based on social tags: Crossing folksonomies and enhancing recommendations, in: E-Commerce and Web Technologies, Springer, 2013, pp. 88–100.

[10] M. Tkalčič, U. Burnik, A. Košir, Using affective parameters in a content-based recommender system for images, User Modeling and User-Adapted Interaction 20 (4) (2010) 279–311.

[11] B. Faltings, P. Pu, M. Torrens, P. Viappiani, Designing example-critiquing interaction, in: Proceedings of the 9th international conference on Intelligent user interfaces, ACM, 2004, pp. 22–29.

[12] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, T. Höllerer, Smallworlds: Visualizing social recommendations, in: Computer Graphics Forum, Vol. 29, Wiley Online Library, 2010, pp. 833–842.

[13] S. Bostandjiev, J. O'Donovan, T. Höllerer, Tasteweights: a visual interactive hybrid recommender system, in: Proceedings of the sixth ACM conference on Recommender systems, ACM, 2012, pp. 35–42.

[14] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, A. Kobsa, Inspectability and control in social recommenders, in: Proceedings of the sixth ACM conference on Recommender systems, ACM, 2012, pp. 43–50.

[15] K. Verbert, D. Parra, P. Brusilovsky, E. Duval, Visualizing recommendations to support exploration, transparency and controllability, in: Proceedings of the 2013 international conference on Intelligent user interfaces, IUI '13, ACM, New York, NY, USA, 2013, pp. 351–362. doi:10.1145/2449396.2449442.
URL http://doi.acm.org/10.1145/2449396.2449442

[16] D. Parra, P. Brusilovsky, C. Trattner, See what you want to see: Visual user-driven approach for hybrid recommendation, in: Proceedings of the 19th International Conference on Intelligent User Interfaces, IUI '14, ACM, New York, NY, USA, 2014, pp. 235–240. doi:10.1145/2557500.2557542.
URL http://doi.acm.org/10.1145/2557500.2557542

[17] S. Nagulendra, J. Vassileva, Understanding and controlling the filter bubble through interactive visualization: A user study, in: Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT '14, ACM, New York, NY, USA, 2014, pp. 107–115. doi:10.1145/2631775.2631811.
URL http://doi.acm.org/10.1145/2631775.2631811

[18] S. M. McNee, J. Riedl, J. A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: CHI'06 extended abstracts on Human factors in computing systems, ACM, 2006, pp. 1097–1101.

[19] J. A. Konstan, J. Riedl, Recommender systems: from algorithms to user experience, User Modeling and User-Adapted Interaction 22 (1-2) (2012) 101–123.

[20] I. Andjelkovic, D. Parra, J. O'Donovan, Moodplay: Interactive mood-based music discovery and recommendation, in: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP '16, ACM, New York, NY, USA, 2016, pp. 275–279. doi:10.1145/2930238.2930280.
URL http://doi.acm.org/10.1145/2930238.2930280

[21] J. M. George, Individual differences and behavior in organizations, San Francisco: Jossey-Bass, 1996, Ch. Trait and State Affect, p. 145.

[22] D. Västfjäll, Emotion induction through music: A review of the musical mood induction procedure, Musicae Scientiae 5 (1 suppl) (2002) 173–211.

[23] N. H. Frijda, Moods, emotion episodes and emotions, in: M. Lewis, J. M. Haviland (Eds.), Handbook of Emotions, New York: Guilford Press, 1993, pp. 381–403.

[24] H. M. Weiss, R. Cropanzano, Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work.

[25] D. Hume, Emotions and moods, Organizational behavior (2012) 258–297.

[26] R. J. D. Paul Ekman, The Nature of Emotion: Fundamental Questions, Oxford University Press, 1994.

[27] J. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (6) (1980) 1161–1178.

[28] M. Zentner, D. Grandjean, K. R. Scherer, Emotions evoked by the sound of music: Characterization, classification, and measurement., Emotion 8 (4) (2008) 494–521.

[29] M. Zentner, T. EEROLA, Self-report measures and models, Handbook of Music and Emotion: Theory, Research, Applications.

[30] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, K. Scherer, Toward a minimal representation of affective gestures, Affective Computing, IEEE Transactions on 2 (2) (2011) 106–118.

[31] M. Karg, K. Kühnlenz, M. Buss, Recognition of affect based on gait patterns., IEEE Transactions on Systems, Man, and Cybernetics, Part B 40 (4) (2010) 1050–1061.

[32] A. Kleinsmith, N. Bianchi-Berthouze, Recognizing affective dimensions from body posture, in: Proc. of the 2Nd International Conference on Affective Computing and Intelligent Interaction, ACII '07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 48–58.

[33] V. A. Petrushin, Emotion recognition in speech signal: experimental study, development, and application, in: In: Proc. ICSLP 2000, 2000, pp. 222–225.

[34] F. Yu, E. Chang, Y. qing Xu, H. yeung Shum, Emotion detection from speech to enrich multimedia content, in: Second IEEE Pacific-Rim Conference on Multimedia, 2001, pp. 550–557.

DRAFT, to be published in IJHCS
https://doi.org/10.1016/j.ijhcs.2018.04.004

[35] S. Wu, T. H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features., Speech Communication 53 (5) (2011) 768–785.

[36] J. Masthoff, The pursuit of satisfaction: affective state in group recommender systems, in: User Modeling 2005, Springer, 2005, pp. 297–306.

[37] G. Gonzalez, J. L. De La Rosa, M. Montaner, S. Delfin, Embedding emotional context in recommender systems, in: Data Engineering Workshop, 2007 IEEE 23rd International Conference on, IEEE, 2007, pp. 845–852.

[38] H.-S. Park, J.-O. Yoo, S.-B. Cho, A context-aware music recommendation system using fuzzy bayesian networks with utility theory, in: Fuzzy systems and knowledge discovery, Springer, 2006, pp. 970–979.

[39] M. Tkalcic, A. Kosir, J. Tasic, Affective recommender systems: the role of emotions in recommender systems, in: Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems, Citeseer, 2011, pp. 9–13.

[40] B.-j. Han, S. Rho, S. Jun, E. Hwang, Music emotion classification and context-based music recommendation, Multimedia Tools and Applications 47 (3) (2010) 433–460.

[41] S. Cunningham, S. Caulder, V. Grout, Saturday night or fever? context-aware music playlists, Proc. Audio Mostly.

[42] S. Rho, B.-j. Han, E. Hwang, Svr-based music mood classification and context-based music recommendation, in: Proceedings of the 17th ACM international conference on Multimedia, ACM, 2009, pp. 713–716.

[43] D. Griffiths, S. Cunningham, J. Weinel, A discussion of musical features for automatic music playlist generation using affective technologies., in: K. Delsing, M. Liljedahl (Eds.), Audio Mostly Conference, ACM, 2013, pp. 13:1–13:4.

[44] M. D. van der Zwaag, J. H. Janssen, J. H. D. M. Westerink, Directing physiology and mood through music: Validation of an affective music player., T. Affective Computing 4 (1) (2013) 57–68.

[45] J. H. Janssen, E. L. van den Broek, J. H. D. M. Westerink, Tune in to your emotions: a robust personalized affective music player., User Model. User-Adapt. Interact. 22 (3) (2012) 255–279.

[46] D. Amelynck, M. Grachten, L. van Noorden, M. Leman, Toward e-motion-based music retrieval a study of affective gesture recognition., T. Affective Computing 3 (2) (2012) 250–259.

[47] O. Celma, P. Herrera, A new approach to evaluating novel recommendations, in: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys '08, ACM, New York, NY, USA, 2008, pp. 179–186. doi:10.1145/1454008.1454038.
URL http://doi.acm.org.ezproxy.puc.cl/10.1145/1454008.1454038

[48] B. Logan, Music recommendation from song sets., in: ISMIR, 2004.

[49] D. Parra, X. Amatriain, Walk the talk: Analyzing the relation between implicit and explicit feedback for preference elicitation, in: Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization, UMAP'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 255–268.
URL http://dl.acm.org/citation.cfm?id=2021855.2021878

[50] C. Baccigalupo, E. Plaza, Case-based sequential ordering of songs for playlist recommendation, in: Advances in Case-Based Reasoning, Springer, 2006, pp. 286–300.

[51] N. Hariri, B. Mobasher, R. Burke, Context-aware music recommendation based on latenttopic sequential patterns, in: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, ACM, New York, NY, USA, 2012, pp. 131–138. doi:10.1145/2365952.2365979.
URL http://doi.acm.org/10.1145/2365952.2365979

[52] M. Kaminskas, F. Ricci, Location-adapted music recommendation using tags, in: User Modeling, Adaption and Personalization, Springer, 2011, pp. 183–194.

[53] Y. Hijikata, Y. Kai, S. Nishida, The relation between user intervention and user satisfaction for information recommendation, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM, 2012, pp. 2002–2007.

[54] L. Gou, F. You, J. Guo, L. Wu, X. L. Zhang, Sfviz: interest-based friends exploration and recommendation in social networks, in: Proceedings of the 2011 Visual Information Communication-International Symposium, ACM, 2011, p. 15.

[55] S. Zhao, M. X. Zhou, X. Zhang, Q. Yuan, W. Zheng, R. Fu, Who is doing what and when: Social map-based recommendation for content-centric social web sites, ACM Transactions on Intelligent Systems and Technology (TIST) 3 (1) (2011) 5.

[56] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, T. Höllerer, Peerchooser: visual interactive recommendation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2008, pp. 1085–1088.

[57] P. Pu, B. Faltings, L. Chen, J. Zhang, P. Viappiani, Usability guidelines for product recommenders based on example critiquing research, in: F. Ricci, L. Rokach, B. Shapira, P. B. Kantor (Eds.), Recommender Systems Handbook, Springer US, 2011, pp. 511–545. doi:10.1007/978-0-387-85820-3_16.
URL http://dx.doi.org/10.1007/978-0-387-85820-3_16

[58] L. Chen, P. Pu, Interaction design guidelines on critiquing-based recommender systems, User Modeling and User-Adapted Interaction 19 (3) (2009) 167–206. doi:10.1007/s11257-008-9057-x.
URL http://dx.doi.org/10.1007/s11257-008-9057-x

[59] C. He, D. Parra, K. Verbert, Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities, Expert Systems with Applicationsdoi:10.1016/j.eswa.2016.02.013.

[60] P. Knees, M. Schedl, T. Pohle, G. Widmer, An innovative three-dimensional user interface for exploring music collections enriched, in: Proceedings of the 14th ACM International Conference on Multimedia, MM '06, ACM, New York, NY, USA, 2006, pp. 17–24. doi:10.1145/1180639.1180652.
URL http://doi.acm.org/10.1145/1180639.1180652

[61] Y.-H. Yang, Y.-C. Lin, H. T. Cheng, H. H. Chen, Mr. emo: music retrieval in the emotion plane., in: A. El-Saddik, S. Vuong, C. Griwodz, A. D. Bimbo, K. S. Candan, A. Jaimes (Eds.), ACM Multimedia, ACM, 2008, pp. 1003–1004.
URL http://dblp.uni-trier.de/db/conf/mm/mm2008.html#YangLCC08

[62] G. L. Collier, Beyond valence and activity in the emotional connotations of music, Psychology of Music 35 (1) (2007) 110–131. arXiv:http://pom.sagepub.com/

content/35/1/110.full.pdf+html, doi:10.1177/0305735607068890.
URL http://pom.sagepub.com/content/35/1/110.abstract

[63] K. R. Scherer, T. Johnstone, G. Klasmeyer, Vocal expression of emotion, Handbook of affective sciences (2003) 433–456.

[64] N. A. Nijdam, Mapping emotion to color, 2005.

[65] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, in: Visual Languages, 1996. Proceedings., IEEE Symposium on, IEEE, 1996, pp. 336–343.

[66] N. Tintarev, J. Masthoff, Designing and evaluating explanations for recommender systems, in: Recommender Systems Handbook, Springer, 2011, pp. 479–510.

[67] S. Stober, A. Nürnberger, Adaptive music retrieval—a state of the art, Multimedia Tools Appl. 65 (3) (2013) 467–494. doi:10.1007/s11042-012-1042-z.
URL http://dx.doi.org/10.1007/s11042-012-1042-z

[68] N. J. Salkind (Ed.), Encyclopedia of Research Design, 0th Edition, SAGE Publications, Inc., 2010. doi:http://dx.doi.org/10.4135/9781412961288.
URL http://dx.doi.org/10.4135/9781412961288

[69] C. Izard, Human Emotions, Emotions, Personality, and Psychotherapy, Springer, 1977.
URL https://books.google.com/books?id=A8K1G9FXMsEC

[70] K. R. Scherer, On the nature and function of emotion: A component process approach., Lawrence Erlbaum, Hillsdale, NJ, 1984, Ch. 14, pp. 293–317.

[71] U. Schimmack, A. Grob, Dimensional models of core affect: A quantitative comparison by means of structural equation modeling, European Journal of Personality 14 (4) (2000) 325–345.

[72] R. Burke, Hybrid recommender systems: Survey and experiments, User Modeling and User-Adapted Interaction 12 (4) (2002) 331–370. doi:10.1023/A:1021240730564.
URL http://dx.doi.org/10.1023/A:1021240730564

[73] B. McFee, G. R. G. Lanckriet, Large-scale music similarity search with spatial trees., in: A. Klapuri, C. Leider (Eds.), ISMIR, University of Miami, 2011, pp. 55–60.
URL http://dblp.uni-trier.de/db/conf/ismir/ismir2011.html#McFeeL11

[74] J. L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18 (9) (1975) 509–517. doi:10.1145/361002.361007.
URL http://doi.acm.org/10.1145/361002.361007

[75] Y. Zheng, B. Mobasher, R. Burke, Carskit: A java-based context-aware recommendation engine, in: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1668–1671. doi:10.1109/ICDMW.2015.222.

[76] D. Parra, S. Sahebi, Recommender systems: Sources of knowledge and evaluation metrics, in: Advanced Techniques in Web Intelligence-2, Springer Berlin/Heidelberg, 2012, pp. 149–175.

[77] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on World Wide Web, ACM, 2005, pp. 22–32.

[78] E. Pariser, The filter bubble: What the Internet is hiding from you, Penguin UK, 2011.

[79] C. D. Manning, P. Raghavan, H. Schtze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.

[80] D. Parra, S. Sahebi, Recommender systems: Sources of knowledge and evaluation metrics, in: J. V. et al. (Eds.) (Ed.), Advanced Techniques in Web Intelligence-2: Web User Browsing Behaviour and Preference Analysis, Springer-Verlag, Berlin Heidelberg, 2013, Ch. 7, pp. 149–175.

[81] P. M. Kroonenberg, Applied multiway data analysis, Vol. 702, John Wiley & Sons, 2008.

[82] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction 22 (4-5) (2012) 441–504.

[83] D. Parra, P. Brusilovsky, User-controllable personalization, Int. J. Hum.-Comput. Stud. 78 (C) (2015) 43–67. doi:10.1016/j.ijhcs.2015.01.007.
URL http://dx.doi.org/10.1016/j.ijhcs.2015.01.007

[84] S. J. Park, G. Chae, C. MacDonald, R. Stein, S. Wiedenbeck, J. Kim, Amara: The affective museum of art resource agent, in: CHI '12 Extended Abstracts on Human Factors in Computing Systems, CHI EA '12, ACM, 2012, pp. 1035–1038. doi:10.1145/2212776.2212379.
URL http://doi.acm.org.ezproxy.puc.cl/10.1145/2212776.2212379

[85] M. H. Thaut, G. C. McIntosh, Neurologic music therapy in stroke rehabilitation, Current Physical Medicine and Rehabilitation Reports 2 (2) (2014) 106–113. doi:10.1007/s40141-014-0049-y.
URL http://dx.doi.org/10.1007/s40141-014-0049-y

[86] S. L. Siedliecki, M. Good, Effect of music on power, pain, depression and disability, Journal of Advanced Nursing 54 (5) (2006) 553–562. doi:10.1111/j.1365-2648.2006.03860.x.
URL http://dx.doi.org/10.1111/j.1365-2648.2006.03860.x

[87] P. N. Juslin, P. Laukka, Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening, Journal of New Music Research 33 (3) (2004) 217–238. doi:10.1080/0929821042000317813.

[88] I. Fodor, A survey of dimension reduction techniques, Tech. rep., Lawrence Livermore National Lab., CA (US) (2002).

[89] J. Stahnke, M. Dörk, B. Müller, A. Thom, Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions, IEEE transactions on visualization and computer graphics 22 (1) (2016) 629–638.

11. Vitae

Ivana Andjelkovic works in music and technology, approaching the field through visualization of music related data, music recommendation and audio signal processing. She holds BS and MS degrees in Computer Science and a PhD in Media Arts and Technology from University of California, Santa Barbara. She currently works as a Senior Audio Software Engineer, developing algorithms for audio manipulation.

https://doi.org/10.1016/j.ijhcs.2018.04.004

Denis Parra is Assistant Professor at the Department of Computer Science, School of Engineering at PUC Chile. Professor Parra holds a B.Eng. from Universidad Austral de Chile and a Ph.D. from the University of Pittsburgh, PA, USA. His main research interests are Recommender Systems, applications of Data Mining and Machine Learning, Information Visualization and Social Networks. Professor Parra has published in important journals in the area such as IJHCS, ComCom, ESWA and ACM TiiS, as well as in IUI, UMAP, Hypertext, RecSys, CSCW, and WWW conferences. He leads the SocVis Laboratory at PUC Chile.

John O'Donovan is an Associate Researcher at the Department of Computer Science, University of California Santa Barbara, where he co-directs the FourEyes laboratory. His research background is in AI and HCI, with a focus on recommender systems. He has published more than 60 research papers in peer reviewed conferences and journals in this area. His research on human computer interaction and intelligent interfaces has won multiple best paper awards. He received a highest impact paper award from ACM IUI in 2017. He served as general co-chair of ACM IUI conference for 2016, and program co-chair of ACM RecSys for 2018.