4th Workshop on Emotions and Personality in Personalized Systems (EMPIRE) 2016

Boston, MA, USA, September 16th, 2016

Proceedings

edited by

Marko Tkalčič

Berardina de Carolis

Marco de Gemmis

Andrej Košir

in conjunction with

10th ACM Conference on Recommender Systems (RecSys 2016)

Preface

This volume contains the papers presented at the 4th Workshop on Emotions and Personality in Personalized Systems (EMPIRE), hold as part of the 10th ACM Conference on Recommender System (RecSys), in Boston, MA, USA.

RecSys is the premier international forum for the presentation of new research results, systems and techniques in the broad field of recommender systems. Recommendation is a particular form of information filtering, that exploits past behaviors and user similarities to generate a list of information items that is personally tailored to an end-user's preferences.

The EMPIRE workshop focuses on the usage of psychologically-based constructs, such as emotions and personality, in delivering personalized content. The 9 (5 long, 4 short) technical papers included in the proceedings were selected through a rigorous reviewing process, where each paper was reviewed by three PC members.

The EMPIRE chairs would like to thank the RecSys workshop chairs, Elizabeth Daly and Dietmar Jannach, for their guidance during the workshop organization. We also wish to thank all authors and all workshops participants for fruitful discussions, the members of the program committee and the external reviewers. All of them secured the workshop's high quality standards.

August 2016

Marko Tkalčič
Berardina de Carolis
Marco de Gemmis
Andrej Košir

EMPIRE 2016 Workshop Organization

Chairs: Marko Tkalčič, *Free University of Bozen-Bolzano, Italy*
Berardina De Carolis, *University of Bari Aldo Moro, Italy*
Marco de Gemmis, *University of Bari Aldo Moro, Italy*
Andrej Košir, *University of Ljubljana, Slovenia*

Proceedings Chair: Marco de Gemmis, *University of Bari Aldo Moro, Italy*

Web Chair: Marko Tkalčič, *Free University of Bozen-Bolzano, Italy*

Program Committee: Ioannis Arapakis, *Eurecat*
Matthias Braunhofer, *Free University of Bozen-Bolzano*
Iván Cantador, *Universidad Autónoma de Madrid*
Fabio Celli, *University of Trento*
Li Chen, *Hong Kong Baptist University*
Matt Dennis, *University of Aberdeen*
Mehdi Elahi, *Polytechnic University of Milan*
Bruce Ferwerda, *Johannes Kepler University*
Sabine Graf, *Athabasca University*
Peter Knees, *Johannes Kepler University*
Fang-Fei Kuo, *University of Washington*
Neal Lathia, *University of Cambridge*
Matija Marolt, *University of Ljubljana*
Fedelucio Narducci, *University of Bari Aldo Moro*
Giuseppe Palestra, *University of Bari Aldo Moro*
Viviana Patti, *University of Turin*
Matevž Pesek, *University of Ljubljana, Slovenia*
Marco Polignano, *University of Bari Aldo Moro*
Olga C. Santos, *aDeNu Research Group (UNED)*
Björn Schuller, *University of Passau / Imperial College London*
Giovanni Semeraro, *University of Bari Aldo Moro*
Man-Kwan Shan, *National Chengchi University*
Yi-Hsuan Yang, *Academia Sinica*
Martijn Willemsen, *Eindhoven University of Technology*
Yong Zheng, *DePaul University*

External Reviewer: Muhammad Anwar

Table of Contents

Long Papers

Short Papers

Adapt to Emotional Reactions In Context-aware Personalization

Yong Zheng
Department of Information Technology and Management
School of Applied Technology
Illinois Institute of Technology
Chicago, Illinois, USA
yzheng66@iit.edu

ABSTRACT

Context-aware recommender systems (CARS) have been developed to adapt to users' preferences in different contextual situations. Users' emotions have been demonstrated as one of effective context information in recommender systems. However, there are no work exploring the effect of emotional reactions (or expressions) in the recommendation process. In this paper, we assume that users may give similar ratings even if they present different emotional reactions or expressions on the movies. We further model the traits of emotional reactions and incorporate them into context-aware matrix factorization as regularization terms. Our experimental results based on the LDOS-CoMoDa movie data set validate our assumptions and prove that it is useful to take emotional reactions into consideration in context-aware recommendations.

Keywords

context, recommendation, emotion, emotional reactions

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Retrieval models; H.1.2 [**Models and Principles**]: User/Machine Systems - *human information processing*

1. INTRODUCTION AND BACKGROUND

Recommender systems (RS) are an effective way in alleviating information overload by tailoring recommendations to users' personal preferences. Context-aware recommender systems (CARS) take contextual factors (such as time, location, companion, occasion, etc) into account in modeling user profiles and in generating recommendations. For example, users' choice on movies may be very different if the user is going to watch the movie with *children* rather than with his or her *partner*.

Context, is usually defined as, "*any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves* [12]". In CARS, we view the dynamic attributes as the observed contexts which may change when the user performs the same activity repeatedly [38]. For example, the *time* and *location* may change every time when a user tries to watch a movie. The *season* or *trip type* may change when a user is going to reserve a hotel. In addition to these factors, users' emotional states are one of these dynamic variables. And emotions may change anytime in the process of user interactions with the items or the applications. These emotional information have been demonstrated as effective and influential context in previous research [45, 44].

Emotional reactions or expressions are highly correlated with the traits of user personalities. Personality accounts for the most important ways in which individuals differ in their enduring emotional, interpersonal, experimental, attitudinal and motivational styles [24]. In the domain of recommender systems, personality can be viewed as a user profile, which may be context-independent and domain-independent. Both emotional information [18, 11, 34, 45] and user personality [31, 19, 35] have been successfully incorporated into recommender systems by existing research.

Our previous research [45] has successfully utilized emotional variables as contexts in recommender systems to improve recommendation performance. Unfortunately, as far as we know, there are no research on exploring the effect of emotional reactions or expressions. We believe that users' emotional reactions or expressions are also useful to model users' preferences or rating behaviors in real practice. For example, two different users may give high ratings on a same tragedy drama movie. One of them indicated his or her emotional state as "*happy*" when finishing the movie, because this user thought it was a really good movie. By contrast, another user may express his or her feeling as "*sad*" since the user was impressed or moved by the tragedy movie. As a result, the two users have same rating behaviors on the movie but with different emotional reactions or expressions. One of the potential reasons is that different user personalities may result in different ways or habits for users to express their emotions.

Therefore, the users' rating profiles associated with different (or even opposed) emotional reactions therefore could be useful to assist recommendations. In this paper, we propose to incorporate emotional reactions (or expressions) as regularization terms in the context-aware matrix factorization approach, and further explore its effect on the performance of context-aware recommendations.

The following sections are organized as follows: Section 2 introduces related work, including the background of context-aware recommendation, and the role of emotions and personality in recommender systems. Section 3 gives the preliminary description of essential information, such as the LDOS-CoMoDa movie data which contains emotional variables, and the introduction about the CAMF technique. Section 4 discusses our methodology that incorporates the emotional reactions as regularization terms

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

into the CAMF approach. Section 5 describes our experimental results and discussions, followed by Section 6 which concludes our findings and discusses our future work.

2. RELATED WORK

One of the goals in the recommender systems (RS) is to assist users' decision making by providing a list of recommendations. Due to the fact that users' choice usually varies from time to time and from context to context, context-aware recommender systems (CARS) [2, 1] are promoted and developed to adapt to users' preferences in different contextual situations.

In rating-based RS applications, such as movie or book ratings, the standard formulation of the recommendation problem begins with a two dimensional matrix of ratings, organized by user and item: *Users* $\times$ *Items* $\rightarrow$ *Ratings*. The key insight of CARS is that users' preferences on items may be also a function of the context in which those items are encountered. Incorporating contexts requires that we estimate user preferences using a multidimensional rating function, *Users* $\times$ *Items* $\times$ *Contexts* $\rightarrow$ *Ratings* [1].

In the past decade, several context-aware recommendation algorithms have been developed. By additionally incorporating context information, these algorithms have been demonstrated to be useful to improve recommendation performance in numerous domains, such as e-commerce [28, 15], movies [33, 26, 10], music [3, 16], restaurants [30, 27], travels [39, 8], educational learning [37], mobile applications [4, 6], and so forth. The context variables adopted in those applications are domain-specific ones. And the most widely used context information are the time of the day, the day of the week, and location information which can be easily captured from ubiquitous environment, such as Web logs, mobile devices, sensors.

It is well known that human decision making is subject to both rational and emotional influences [14]. The field of affective computing takes this fact as basic to the design of computing systems [29]. The role of emotions in recommender systems was recognized by the research community as early as 2005 [23], giving rise to research in emotion-based movie recommender systems [18] and the impact of emotions in group recommender systems [23, 11]. This results in the highlight of research on affective recommender systems [34] which have been proved to be useful on recommendation performance in several domains, such as music [22, 32, 9] and movies [7, 25, 18].

Emotional states, accordingly, are also viewed and used as contexts in recommender systems. Shi et al. [33] mined the mood similarity to assist context-aware movie recommendation. Odic, et al. [26] identified the significant contributions by emotional variables compared with other contextual factors in the LDOS-CoMoDa movie rating data. Mood information can also be used for television and video content recommendation [36]. Baltrunas, et al. [3] adopted mood as context to assist context-aware music recommendation. The role of emotions in context-aware recommendation is summarized in [45, 44] which helps additionally discover insights about why and where emotional states play an important role in the recommendation process.

Emotional states are usually dynamic and may change from time to time. Based on the introduction about the affective recommender systems [34], the emotional information in three stages may be useful: entry stage (i.e., before the activity), consumption stage (i.e., during the activity) and exit stage (i.e. after the activity). In this case, emotional reactions can be captured across these three stages. As introduced previously, users may present different emotional reactions, but actually they leave the same or similar ratings on the items. In this paper, we make the first attempt to explore the effect of emotional reactions in the context-aware recommendations.

3. PRELIMINARY

To further discuss the topics in the context-aware recommendation, it is necessary to introduce some terminologies:

Table 1: Sample of a Context-aware Movie Rating Data Set

User	Movie	Rating	Time	Location	Companion
U1	T1	5	Weekday	Home	Kids
U1	T1	3	Weekend	Cinema	Family
U2	T2	3	Weekday	Cinema	Partner
U2	T3	4	Weekday	Home	Kids
U3	T4	2	Weekend	Home	Partner

Table 1 shows an example of context-aware movie data which contains five rating profiles given by three users on four movies in different contextual situations. In our discussions, we will use the term *contextual dimension* to denote the contextual variable, such as "Location", "Time" and "Companion". The term *contextual condition* refers to a specific value in a contextual dimension, e.g. "Home" and "Cinema" are two contextual conditions for the dimension "Location". *Context* or *contextual situation* therefore refers to a combination of contextual conditions, e.g., {Weekday, Home, Kids}.

Next, we introduce the LDOS-CoMoDa movie data [1] which is a data set with multiple contextual dimensions including several emotional variables. We also introduce context-aware matrix factorization which is a popular algorithm in CARS and we use it as a base algorithm in this paper.

3.1 LDOS-CoMoDa Data Set

In the domain of context-aware recommendation, there are very limited number of data sets available for public research, not to mention the data that contains emotional variables. The LDOS-CoMoDa data set [21] introduced below is one of the data sets that was collected from user surveys, and can be used for this type of research in this paper. The data has 2291 ratings (rating scale is 1 to 5) given by 121 users on 1232 items within 12 contextual dimensions. The description of the contextual dimensions and conditions can be described by Table 2.

Table 2: List of Context Information in the LDOS-CoMoDa Data

Dimension	Contextual Conditions
Time	Morning, Afternoon, Evening, Night
Daytype	Working day, Weekend, Holiday
Season	Spring, Summer, Autumn, Winter
Location	Home, Public place, Friend's house
Weather	Sunny / clear, Rainy, Stormy, Snowy, Cloudy
Companion	Alone, Partner, Friends, Colleagues, Parents, Public, Family
endEmo	Sad, Happy, Scared, Surprised, Angry, Disgusted, Neutral
domEmo	Sad, Happy, Scared, Surprised, Angry, Disgusted, Neutral
Mood	Positive, Neutral, Negative
Physical	Healthy, Ill
Decision	Movie choices by themselves or users were given a movie
Interaction	First interaction with a movie, Nth interaction with a movie

Among these 12 contextual dimensions, there are three ones that can be considered emotional dimensions: endEmo, domEmo and mood. "endEmo" is the emotional state experienced at the end of the movie (i.e., emotion in the exit stage). "domEmo" is

[1]LDOS-CoMoDa data set, http://www.ldos.si/comoda.html

the emotional state experienced the most during watching (i.e., emotion in the consumption stage). "mood" is the emotion of the user during that part of the day when the user watched the movie (i.e., emotion in the entry stage). "EndEmo" and "domEmo" contain the same seven conditions: *Sad, Happy, Scared, Surprised, Angry, Disgusted, Neutral*, while "mood" only has three simple conditions: *Positive, Neutral, Negative*.

Context selection is usually performed before we apply any context-aware recommendation algorithms. We'd like to retain the most influential context dimensions, since irrelevant ones may introduce noises in the data and further hamper the recommendation accuracy. Based on the statistical selection method introduced in [26], we only use 7 out of the 12 contextual dimensions in our experiments: time, daytype, location, companion and the three emotional variables.

The three emotional variables (i.e., mood, domEmo and endEmo) describe users' affective states during the user interactions with the movies in terms of three stages respectively: entry stage, consumption stage and exit stage as introduced in [34]. In other words, mood can be viewed as current context before the user starts watching the movie. By contrast, domEmo and endEmo can indicate future emotional states during the user's interactions with the activity of movie watching. These future status can also be viewed as contexts too if we interpret them as user intents. For example, a user is feeling sad now, and he or she wants to select a movie to watch in order to be happy. In this example, "sad" is the current user mood, and "happy" can be viewed as user's future emotional state, such as in the domEmo or endEmo.

3.2 Context-aware Matrix Factorization

One of the most popular context-aware recommendation algorithms is the one built upon matrix factorization, namely, the context-aware matrix factorization (CAMF) approach [5]. There are different variants of CAMF, here we introduce the CAMF_CU approach which incorporate a user-personalized contextual rating bias into matrix factorization. More specifically, the rating prediction function by CAMF_CU can be described by Equation 1.

$$\hat{r}_{uic_1c_2...c_N} = \mu + \sum_{j=1}^{N} B_{u,c_j} + b_i + p_u^T q_i \tag{1}$$

Assume there are totally N contextual dimensions. $c_1c_2..c_N$ is used to denote a contextual situation, where c_1 indicates the value of contextual condition in the 1^{st} context dimension. $\hat{r}_{uic_1c_2...c_N}$ therefore represents the predicted rating for user u on item i in the situation $c_1c_2..c_N$. The prediction function is composed of four components: the global mean rating μ, item rating bias b_i, the aggregated contextual rating bias $\sum_{j=1}^{N} B_{u,c_j}$, and user-item interaction represented by the dot product of a user vector and item vector, $p_u^T q_i$. p_u is the user vector represented by a set of latent factors, and q_i is the item vector represented by the same set of factors. p_u can tell how much the user u likes those latent factors, while q_i indicates how the item i obtains these factors. Therefore, the dot product function is used to estimate how much the user will like this item.

The term B_{u,c_j} is the estimated contextual rating bias for user u in context condition c_j. It is used to denote how user u's rating is deviated in each contextual condition.

$$err = r_{uic_1c_2...c_N} - \hat{r}_{uic_1c_2...c_N} \tag{2}$$

$$\min_{B_*,b_*,p_*,q_*} \sum_{r \in R} \left[\frac{1}{2} err^2 + \frac{\lambda}{2} (\sum_{j=1}^{N} B_{u,c_j}^2 + b_i^2 + ||p_u||^2 + ||q_i||^2) \right] \tag{3}$$

Afterwards, the algorithm is able to learn the corresponding parameters by minimizing the squared errors in prediction. The loss function as shown in Equation 3 is a composition of squared error and regularization terms which are used to alleviate the overfitting problems, where $r_{uic_1c_2...c_N}$ is the real and known rating given by user u on item i in context $c_1c_2..c_N$, and λ is the regularization rate used in the optimization process. By stochastic gradient descent, we are able to learn the parameters iteratively and finally achieve the best performing CAMF_CU model.

CAMF is an effective algorithm and it is able to alleviate the data sparsity to some extent. We choose CAMF_CU because we are going to explore the correlation between users and their emotional reactions, which requires a user-specific context-dependent model. The same thing can also happen to other algorithms which explore intersections or the dependency between users and contexts, such as the CSLIM_CU approach [40].

In the next section, we will introduce how to incorporate the emotional reactions as regularization terms to CAMF_CU.

4. METHODOLOGY

In this section, we introduce our methodology of how to incorporate emotional reactions into context-aware recommender systems.

4.1 Problem Statement

Recall that we assume that the different emotional reactions or expressions can be used to model users' rating behaviors. For example, assume two users gave a high rating on a same tragedy drama movie. One of them indicated his or her emotional state as "happy" when finishing the movie, because this user thought it was a really good movie. But another user may express his or her feeling as "sad" since it is a tragedy movie. The same thing may also happen to the domEmo in addition to the endEmo. The emotional reactions or expressions in this paper, refer to the different values in the dimension domEmo and/or endEmo in the LDOS-CoMoDa data.

Figure 1 presents the distribution of rating counts in each emotional state. Note that "Unknown" indicates the missing value in the LDOS-CoMoDa. We can observe that *Neutral* and *Happy* are the most two common emotional expressions in both domEmo and endEmo.

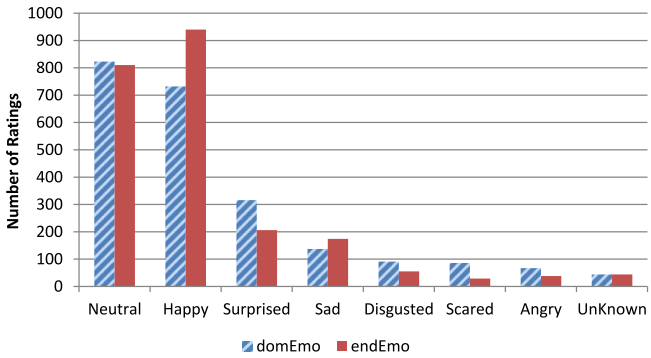


Figure 1: Distribution of Rating Counts in Each Emotional State

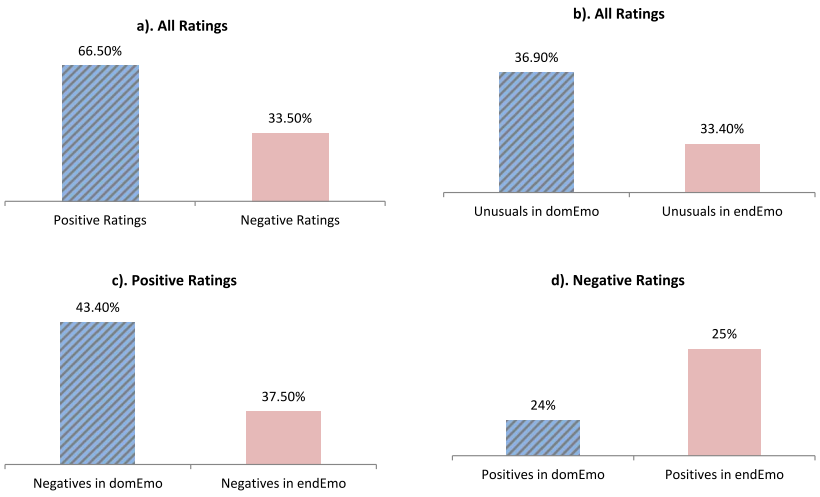


Figure 2: Distribution of Unusual Emotional Reactions in the LDOS-CoMoDa Data

Furthermore, we'd like to learn the *unusual case* to see whether users present different emotional reactions in this data. An unusual case could be two situations: 1). a user leaves a negative rating, but expresses positive emotional states in either domEmo or endEmo; 2). a user gives a positive rating while he finally indicates a negative emotion in either domEmo or endEmo.

To explore these unusual cases, we need to define which ratings and which emotions are positive or negative. In our experiments, we simply view a rating as a positive one if the rating is no less than 4; otherwise, the rating is negative. In terms of the emotional states, we only consider "Happy" and "Surprised" as positive ones, while other emotional states are negative.

A simple statistics about unusual cases in this data can be depicted by Figure 2. First of all, 66.5% of the ratings are positive ones as shown in subfigure a). Based on the subfigure b), we can observe that 36.9% of all the rating records are unusual cases (i.e., the two situations mentioned above) based on the domEmo variable, while it is 33.4% in the endEmo variable. This may tell that domEmo could be more effective and useful than endEmo in modeling users' emotional reactions.

The subfigure c) and d) further describe the two unusual situations among the positive and negative ratings respectively. In the piece of profiles with positive ratings, 43.4% of them are associated with negative emotions in domEmo – many more than the cases in endEmo. It is not surprising, since the theme or the genre of the movie will affect user's dominating emotions during the process of movie watching. For example, users may feel horrible or scared when watching a horrible movie, but finally leave a positive rating since it is a good movie. On the other hand, in terms of the records with negative ratings, there are no significant differences for the unusual cases between domEmo (24%) and endEmo (25%) based on the observations subfigure d). Recall that, there are many more positive ratings than the negative ones in this data. Therefore, it seems that users may express more unusual emotional reactions in domEmo rather than in endEmo. We suspect that the emotional reactions in domEmo may leave more influential impact on our proposed recommendation models.

The underlying assumption in our proposed approach is that user's emotional reactions or expressions on the future emotional states (e.g., domEmo and/or endEmo) can be used to improve recommendations, since they may indicate similar user tastes even if the emotional reactions are different or even opposed. The research problem can be summarized as how to incorporate these emotional reactions into existing recommendation algorithms. More specifically, we want to explore the approach to incorporate them into the CAMF approach. There are three questions we are particularly interested in:

- How to fuse this emotional reactions into CAMF?
- Does it work by providing improvements?
- Which emotional reaction is more effective? The reactions based on domEmo or endEmo?

4.2 Regularization by Emotional Reactions

First of all, how the user reacts on the movies in terms of emotional status is dependent with what type of movies the user is watching. In this case, we additionally use movie genre information in the LDOS-CoMoDa data and aggregated users' ratings for each movie type. A sample of the aggregated data can be shown in Table 3.

Table 3: An Example of Aggregated Rating Matrix

User	Genre	Rating	Time	domEmo	endEmo
U1	Action	5	Weekday	Sad	Happy
U1	Drama	3	Weekend	Sad	Sad
U2	Cartoon	3	Weekday	Happy	Angry
U2	Drama	3	Weekday	Angry	Happy
U3	Action	2	Weekend	Sad	Sad

In Table 3, we replace the column of item by movie genre to construct a new rating matrix. We will use the same 7 contextual dimensions introduced previously. Note that in the LDOS-CoMoDa data we do not know what the movie genre is, since the genre was encoded as numbers in this data.

Afterwards, we can fuse an emotional dimension (either endEmo or domEmo) into the user dimension to create a two-dimensional rating matrix. Let's take the domEmo for example, the converted rating matrix can be described by Table 4.

Specifically, we fuse the values in domEmo into the user column to create new users. The new user is represented by a combination of original user ID and value in the domEmo, and we name those new users as *emotional users*. Meanwhile, we eliminate the other

Table 4: Converted Two-Dimensional Rating Matrix

User, domEmo	Genre	Rating
U1, Sad	Action	5
U1, Sad	Drama	3
U2, Happy	Cartoon	3
U2, Angry	Drama	3
U3, Sad	Action	2

contextual dimensions from the rating matrix. In this case, we can build a matrix factorization model based on this converted two-dimensional rating matrix. And then we are able to calculate the similarity between emotional users based on the cosine similarity of each two vectors which represent emotional users. For example, we can measure how similar the "U1, Sad" to "U2, Angry" based on their co-ratings on the movies with the same genre information.

Theoretically, we can use the item information (e.g., item ID) instead of the movie genre in the rating matrix, but it will increase data sparsity. We use movie genre information only for two reasons: On one hand, using movie genre is based on our assumptions that users' different emotional reactions depend on the movie genre and user's emotional reactions, for example, user may express as happy or sad on a tragedy movie. On the other hand, it is able to alleviate the rating sparsity in the converted two-dimensional rating matrix so that we can obtain more reliable user similarities. We have tried to use item ID, but emotional users have very few co-ratings on the items, which results in worse recommendation performance compared with that when we use genre information only. Note that we use domEmo as an example in Table 4, while we can also have the same process based on the variable endEmo.

In short, the emotional users should be similar if they have similar ratings on the movies with same genre information, even if the original users have different emotional reactions on domEmo or endEmo. For example, the ratings given by "U1, Sad" and "U2, Angry" are all 3-star on the drama movies shown in the Table 4. Therefore, U1 with dominating emotion as "Sad" may share similar user tastes with U2 with dominating emotion as "Angry" to some extent.

Accordingly, we are able to create a regularization term based on the similarity of contextual users. The new loss function can be shown as Equation 4, where β is the regularization rate for the new regularization terms.

$$\min_{B_*,b_*,p_*,q_*} \sum_{r\in R} \left[\begin{array}{l} \frac{1}{2}err^2 + \frac{\lambda}{2}(\sum_{j=1}^{N} B_{u,c_j}^2 + b_i^2 + ||p_u||^2 + ||q_i||^2) \\ +\frac{\beta}{2} \sum_{v,c_{m+}\in K} Sim((v,c_{m+}),(u,c_m)) \times reg\_emo \end{array} \right] \quad (4)$$

$$reg\_emo = (B_{u,c_m} - B_{v,c_{m+}})^2 \quad (5)$$

We will use the same function shown in Equation 1. In addition, we incorporate a new regularization term in Equation 4 compared with the loss function described by Equation 3.

More specifically, we use m to denote the index of an emotional variable (i.e., either domEmo or endEmo). Take domEmo for example, m indicates the position of domEmo in the list of contextual dimensions, thus c_m is used to express user's emotional state in domEmo. According, "u, c_m" is the emotional user (introduced as Table 4), and we use K to denote the top-K nearest neighbor of emotional user "u, c_m" based on the user similarity calculated based on the matrix factorization model built upon the converted two-dimensional rating matrix. Namely, "v, c_{m+}" is one of the identified top-K nearest neighbors. We use c_{m+} to denote the emotional state in domEmo, since it is not necessary to be the same value as c_m. But they should be the contextual condition in the same dimension (i.e., the m^{th} dimension).

As mentioned previously, more similar two emotion users are, their ratings on the items (with same genre) should be similar. In our CAMF_CU model, it can be derived that user's contextual rating deviations in this emotional variable (i.e., the m^{th} contextual variable) should be similar. Namely, B_{u,c_m} and $B_{v,c_{m+}}$ should be very close. We add the squared difference of these two deviations (e.g., Equation 5) as the regularization term in Equation 4.

Additionally, how close the two contextual rating deviations are should be dependent with the similarity of two emotional users. In this case, the regularization term is weighted by the similarity between two emotional users. We name this term as "*emotional regularization term*" in this paper.

Recall that our assumption is that the emotional users should be similar because two difference users have similar ratings even if their emotional reactions are different. It can also tell that the two users actually share something in common, so we assume there should also be a similarity between two users to some extent. Therefore, we are able to additionally incorporate a "*user regularization term*" to build a finer-grained recommendation model, where the loss function can be shown in Equation 6. Again, the user regularization is also weighted by the similarity between two emotional users.

$$\min_{B_*,b_*,p_*,q_*} \sum_{r\in R} \left[\begin{array}{l} \frac{1}{2}err^2 + \frac{\lambda}{2}(\sum_{j=1}^{N} B_{u,c_j}^2 + b_i^2 + ||p_u||^2 + ||q_i||^2) \\ +\frac{\beta}{2} \sum_{v,c_{m+}\in K} Sim((v,c_{m+}),(u,c_m)) \times (reg\_user + reg\_emo) \end{array} \right] \quad (6)$$

$$reg\_user = ||p_u - p_v||^2 \quad (7)$$

Based on those two different loss functions, we are able to build two new CAMF approaches by incorporating the emotional reactions as the regularization terms. We can learn the corresponding parameters based on the gradient decent accordingly. Note that the performance of the models may also depend on the number of K-nearest neighbors used in the algorithm. In our experiments, we set different values to explore the best options in these parameters.

5. EXPERIMENTS

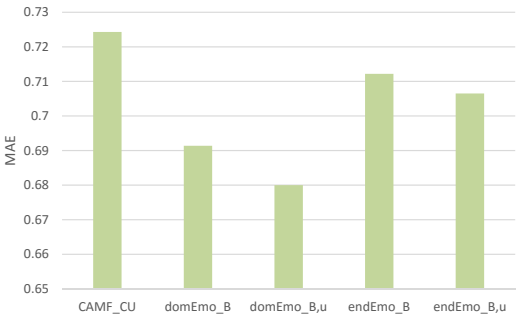
In this section, we introduce our evaluation settings and experimental results, as well as our findings.

5.1 Evaluation Protocols

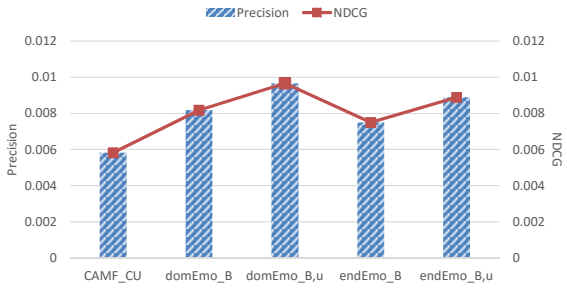
We employ a 5-folds cross-validation on the LDOS-CoMoDa data set. Namely, we split the rating profiles into 5 folds and perform 5 rounds evaluations. For each round, one of the fold will be used as testing set, and the other 4 folds of data will be used as training data. We build our recommendation models based on the training set and evaluate the results according to the ground truth inferred from the testing set.

We use CAMF_CU approach as baseline, and compete its recommendation performance with the CAMF_CU models with different regularization terms. We use the CAMF_CU approach implemented in the open-source toolkit, CARSKit [41], to perform the evaluations.

More specifically, we evaluate the recommendation performance based on the rating prediction and top-10 recommendation tasks. In the rating prediction task, we use mean absolute error (MAE) as evaluation metric. We also further examine the statistical difference of MAE among different algorithms based on paired t-test at a 95% confidence level. In the top-10 recommendation, We



(a) Results on MAE



(b) Results on Top-10 Recommendation

Figure 3: Experimental Results on the Rating Prediction and Top-10 Recommendation Tasks

adopt precision as the relevance metric and Normalized Discounted Cumulative Gain (NDCG) [20] as the ranking metric. More specifically, precision is calculated as the ratio of relevant items selected to the number of items recommended (i.e., 10 in our experiment). NDCG is a measure from information retrieval, where positions are discounted logarithmically.

5.2 Experimental Results

First of all, we present our results based on the rating prediction task in Figure 3(a). We use CAMF_CU to denote the original approach without emotional or user regularization terms. Our approaches introduced in this paper are built upon CAMF_CU approach and they can be generated based on either domEmo or endEmo. We evaluate the performances by them individually. We use "domEmo_B" to represent the model using domEmo for emotional regularization, i.e., c_m denotes the emotional state in domEmo in Equations 4. By contrast, "domEmo_B, u" is used to denote the finer-grained model described in Equation 6 which contains both emotional and user regularization terms. Accordingly, "endEmo_B" and "endEmo_B, u" are the two recommendation models by using endEmo to generate the regularization terms.

Based on the results shown in Figure 3(a), our proposed approaches only using the emotional regularization term can help obtain lower MAE. All of these improvements are statistically significant based on the paired t-test. When we try to use both emotional and user regularization terms, it is able to further improve prediction accuracies. However, the improvement by endEmo_B,u fails the paired t-test compared with the endEmo_B approach. The best performing model in the rating prediction task is domEmo_B,u, where we apply emotional and user regularization terms at the same time, and these regularization terms are generated based on the emotional reactions by domEmo.

We show the top-10 recommendation results based on precision and NDCG in Figure 3(b). The bars present results based on precision at top-10 recommendation, the curve tells the results in NDCG. We can observe similar patterns shown in the rating prediction task: first, we see that the CAMF_CU models with our regularization terms are able to outperform the original CAMF_CU approach in both precision and NDCG. This finding confirms that incorporating emotional regularization terms inferred from users' emotional reactions is helpful to improve performance of context-aware recommendation.

Furthermore, we can observe the finer-grained model with additional user regularization term contributes to obtain more improvements. For example, domEmo_B,u works better than domEmo_B (19.6% improvement on precision, and 18.2% on NDCG), and endEmo_B, u outperforms endEmo_B (30.1% improvement on precision, and 18.5% on NDCG).

As mentioned before, the number of selected neighbors in our models may impact the recommendation performance. We present the impact by the number of neighbors in the finer-grained CAMF_CU approaches with two regularization terms, as shown by the Figure 4. Simply, we vary the number of neighbors from 10 to 80 with an increment of 10 on each step. The best number of neighbors should be around 40 to 50 in this data set. It is essential to examine different number of neighbors to find out the optimal selection for each recommendation model.

Finally, the experimental results help us identify that the domEmo is more useful and effective to be adopted than using endEmo. This finding is consistent with our previous analysis on the unusual cases shown in Figure 2. It makes sense since the emotional status during the process of movie watching may be very different than their emotions at the end. For example, a user may feel horrible if he or she is watching an adventure movie, but finally he or she might feel happy since it is a good movie.

5.3 Discussions

Why emotional reactions or expressions can be reused to improve the recommendation performance? As we mentioned before, one of the potential reasons is that the different emotional reactions are caused by the traits in different user personalities – users may express their emotional states or reactions in different ways. It has been well studied that the emotional expression has strong correlations with user personality, especially in the areas of psychology and social science. For example, the correlation between emotional expression and personality can be used to assist health care [13]. Harker, et al. [17] found that individual differences in positive emotional express were linked to personality stability and development across adulthood. However, there are no applications of using personality inferred from emotional reactions or expressions to further serve real-world applications, such as recommender systems. In this paper, we make our attempts to explore the impacts of emotional reactions or expressions in the recommender systems, especially in the context-aware personalization.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we believe that users may place similar ratings even if they may have different emotional reactions or expressions. We propose to incorporate the corresponding regularization terms in the CAMF_CU approach to assist context-aware recommendation. Our findings based on the experimental results over the LDOS-CoMoDa movie data demonstrate that modeling user's emotional

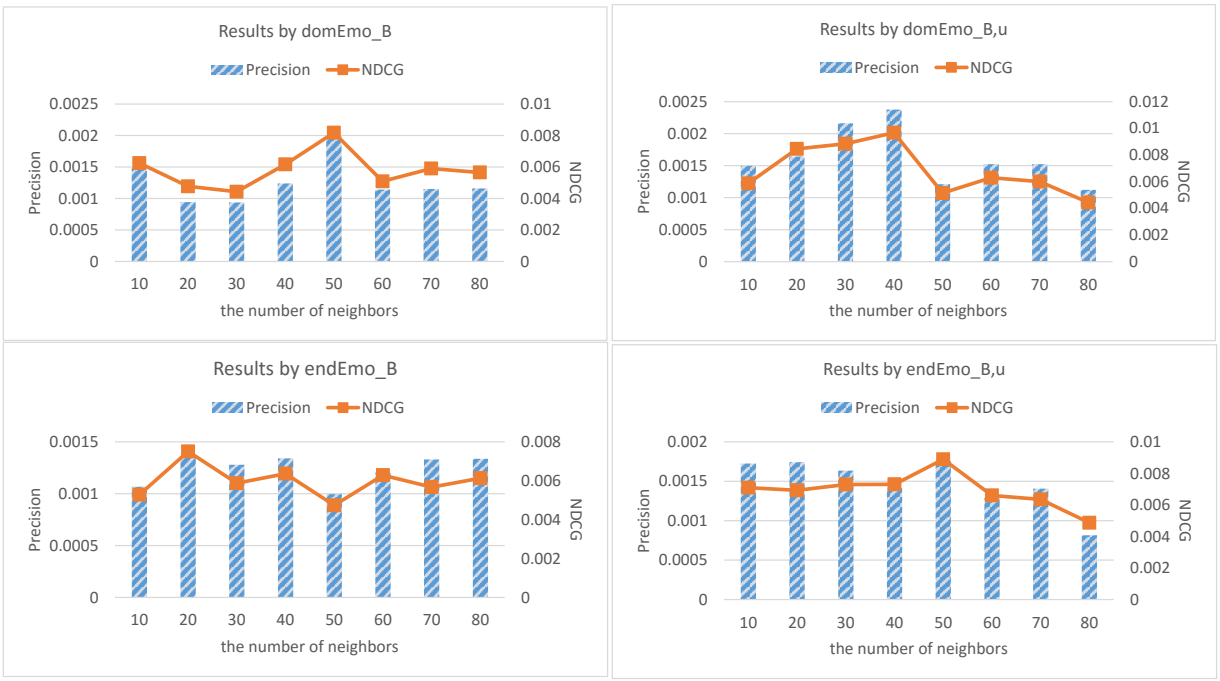


Figure 4: Impact by the Number of Neighbors

reactions is helpful to improve recommendation performance. The results also reveal that domEmo is better than endEmo to generate the regularization terms in this data set. And the finer-grained model by additionally incorporating user regularization is able to offer further improvements.

One of our future work is to incorporate these regularization terms based on different emotional reactions to more context-aware recommendation models. It is interesting to examine the similar approach in the similarity-based context-aware recommendation algorithms [43, 42] so that we can learn the similarities of not only the emotional users but also the emotion themsevles. We will also try to explore the effect of emotional reactions in other applications (such as music) rather than the movie domain.

7. REFERENCES

[1] G. Adomavicius, B. Mobasher, F. Ricci, and A. Tuzhilin. Context-aware recommender systems. *AI Magazine*, 32(3):67–80, 2011.

[2] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, 23(1):103–145, 2005.

[3] L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, A. Aydin, K.-H. Lüke, and R. Schwaiger. Incarmusic: Context-aware music recommendations in a car. In *EC-Web*, volume 11, pages 89–100. Springer, 2011.

[4] L. Baltrunas, B. Ludwig, S. Peer, and F. Ricci. Context relevance assessment and exploitation in mobile recommender systems. *Personal and Ubiquitous Computing*, 16(5):507–526, 2012.

[5] L. Baltrunas, B. Ludwig, and F. Ricci. Matrix factorization techniques for context aware recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 301–304. ACM, 2011.

[6] M. J. Barranco, J. M. Noguera, J. Castro, and L. Martínez. A context-aware mobile recommender system based on location and trajectory. In *Management intelligent systems*, pages 153–162. Springer, 2012.

[7] S. Benini, L. Canini, and R. Leonardi. A connotative space for supporting movie affective recommendation. *Multimedia, IEEE Transactions on*, 13(6):1356–1370, 2011.

[8] M. Braunhofer, M. Elahi, and F. Ricci. Sts: A context-aware mobile recommender system for places of interest. In *UMAP Workshops*. Citeseer, 2014.

[9] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma. Musicsense: contextual music recommendation using emotional allocation modeling. In *Proceedings of the 15th international conference on Multimedia*, pages 553–556. ACM, 2007.

[10] P. G. Campos, I. Fernández-Tobías, I. Cantador, and F. Díez. Context-aware movie recommendations: an empirical comparison of pre-filtering, post-filtering and contextual modeling approaches. In *E-Commerce and web technologies*, pages 137–149. Springer, 2013.

[11] Y. Chen and P. Pu. Cofeel: Using emotions to enhance social interaction in group recommender systems. In *Alpine Rendez-Vous (ARV) 2013 Workshop on Tools and Technology for Emotion-Awareness in Computer Mediated Collaboration and Learning*, 2013.

[12] A. K. Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[13] H. S. Friedman and S. Booth-Kewley. Personality, type a behavior, and coronary heart disease: the role of emotional expression. *Journal of Personality and Social Psychology*, 53(4):783, 1987.

[14] T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press, 2002.

[15] M. Gorgoglione, U. Panniello, and A. Tuzhilin. The effect of context-aware recommendations on customer purchasing behavior and trust. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 85–92. ACM, 2011.

[16] N. Hariri, B. Mobasher, and R. Burke. Context-aware music recommendation based on latenttopic sequential patterns. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 131–138. ACM, 2012.

[17] L. Harker and D. Keltner. Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of personality and social psychology*, 80(1):112, 2001.

[18] A. T. Ho, I. L. Menezes, and Y. Tagmouti. E-mrs: Emotionbased movie recommender system. In *Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell*, pages 1–8, 2006.

[19] R. Hu and P. Pu. Using personality information in collaborative filtering for new users. *Recommender Systems and the Social Web*, page 17, 2010.

[20] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[21] A. KoÅąir. Database for contextual personalization. *ElektrotehniÅąki vestnik*, 78(5):270–274, 2011.

[22] F.-F. Kuo, M.-F. Chiang, M.-K. Shan, and S.-Y. Lee. Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 507–510. ACM, 2005.

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

[24] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[25] Y. Moshfeghi, B. Piwowarski, and J. M. Jose. Handling data sparsity in collaborative filtering using emotion and semantic based features. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 625–634. ACM, 2011.

[26] A. Odic, M. Tkalcic, J. F. Tasic, and A. Košir. Relevant context in a movie recommender system: UsersâĂŹ opinion vs. statistical detection. *ACM RecSys*, 12, 2012.

[27] C. Ono, Y. Takishima, Y. Motomura, and H. Asoh. Context-aware preference model based on a study of difference between real and supposed situation data. *UMAP*, 9:102–113, 2009.

[28] U. Panniello, A. Tuzhilin, M. Gorgoglione, C. Palmisano, and A. Pedone. Experimental comparison of pre-vs. post-filtering approaches in context-aware recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pages 265–268. ACM, 2009.

[29] R. W. Picard and R. Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[30] X. Ramirez-Garcia and M. García-Valdez. Post-filtering for a restaurant context-aware recommender system. In *Recent Advances on Hybrid Approaches for Designing Intelligent Systems*, pages 695–707. Springer, 2014.

[31] P. J. Rentfrow and S. D. Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003.

[32] M.-K. Shan, F.-F. Kuo, M.-F. Chiang, and S.-Y. Lee. Emotion-based music recommendation by affinity discovery from film music. *Expert systems with applications*, 36(4):7666–7674, 2009.

[33] Y. Shi, M. Larson, and A. Hanjalic. Mining mood-specific movie similarity with matrix factorization for context-aware recommendation. In *Proceedings of the workshop on context-aware movie recommendation*, pages 34–40. ACM, 2010.

[34] M. Tkalcic, A. Kosir, and 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*, pages 9–13, 2011.

[35] M. Tkalcic, M. Kunaver, A. Košir, and J. Tasic. Addressing the new user problem with a personality based user similarity measure. In *First International Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems (DEMRA 2011)*, page 106. Citeseer, 2011.

[36] J. Vanattenhoven and D. Geerts. Contextual aspects of typical viewing situations: a new perspective for recommending television and video content. *Personal and Ubiquitous Computing*, 19(5-6):761–779, 2015.

[37] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval. Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335, 2012.

[38] Y. Zheng. A revisit to the identification of contexts in recommender systems. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, pages 133–136. ACM, 2015.

[39] Y. Zheng, R. Burke, and B. Mobasher. Differential context relaxation for context-aware travel recommendation. In *International Conference on Electronic Commerce and Web Technologies*, pages 88–99. Springer, 2012.

[40] Y. Zheng, B. Mobasher, and R. Burke. CSLIM: Contextual SLIM recommendation algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 301–304. ACM, 2014.

[41] Y. Zheng, B. Mobasher, and R. Burke. CARSKit: A Java-based context-aware recommendation engine. In *Proceedings of the 15th IEEE International Conference on Data Mining Workshops. IEEE*, 2015.

[42] Y. Zheng, B. Mobasher, and R. Burke. Integrating context similarity with sparse linear recommendation model. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 370–376. Springer, 2015.

[43] Y. Zheng, B. Mobasher, and R. Burke. Similarity-based context-aware recommendation. In *International Conference on Web Information Systems Engineering*, pages 431–447. Springer, 2015.

[44] Y. Zheng, B. Mobasher, and R. Burke. Emotions in context-aware recommender systems. In Tkalčič, B. Decarolis, M. de Gemmis, A. Košir, and A. Odič, editors, *Emotions and Personality in Personalized Services*, pages 311–326. Springer Berlin Heidelberg, 2016.

[45] Y. Zheng, B. Mobasher, and R. D. Burke. The role of emotions in context-aware recommendation. *Decisions@ RecSys*, 2013:21–28, 2013.

Investigating the Role of Personality Traits and Influence Strategies on the Persuasive Effect of Personalized Recommendations

Gkika Sofia
PhD Student
ELTRUN, E-Business Center
Athens University of Economics and Business
Evelpidon 47-A & Lefkados 33, Room 801, GR-11362, Athens Greece
gkikas@aueb.gr

Skiada Marianna
PhD Student
ELTRUN, E-Business Center
Athens University of Economics and Busines
Evelpidon 47-A & Lefkados 33, Room 801, GR-11362, Athens Greece
mskiada@aueb.gr

Lekakos George
Assosiate Professor
ELTRUN, E-Business Center
Athens University of Economics and Business
Evelpidon 47-A & Lefkados 33, Room 801, GR-11362, Athens Greece
glekakos@aueb.gr

Kourouthanasis Panos
Assistant Professor
ELTRUN, E-Business Center
Athens University of Economics and Business
Evelpidon 47-A & Lefkados 33, Room 801, GR-11362, Athens Greece
pkour@ionio.gr

ABSTRACT

Recommender systems provide suggestions for products, services, or information that match users' interests and/or needs. However, not all recommendations persuade users to select or use the recommended item. The Elaboration Likelihood Model (ELM) suggests that individuals with low motivation or ability to process the information provided with a recommended item could eventually get persuaded to select/use the item if appropriate peripheral cues enrich the recommendation. The purpose of this research is to investigate the persuasive effect of certain influence strategies and the role of personality in the acceptance of recommendations. In the present study, a movie Recommender System was developed in order to empirically investigate the aforementioned questions applying certain persuasive strategies in the form of textual messages alongside the recommended item. The statistical method of Fuzzy-Set Qualitative Comparative Analysis (fsQCA) was used for data analysis and the results revealed that motivating messages do change users' acceptance of the recommender item but not unconditionally since user's personality differentiates the effect of the persuasive strategies.

Keywords

Persuasion, Persuasive Technologies, Personalization, Recommender Systems, Personality, Elaboration Likelihood Model.

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

1. PERSUASIVE MESSAGE PROCESSING

Persuasive Technologies utilize several techniques in order to shape, reinforce or/and change humans' attitudes and behaviours without coercion or deception (Fogg, 2002). On the other hand, Recommender Systems represent a class of personalization technologies that aim to tailor products/information/services according to their users' interests, preferences and needs. Thus, personalized recommendations can significantly strengthen the effect of persuasive interventions due to the inherent influence of personalized communication. Berkovsky et. al. (2012) suggest that most of the extant research examine personalization and persuasive technologies in isolation although "*both personalized and persuasive technologies aim to influence user interactions or the users themselves*", acknowledging "*…the huge untapped potential of personalization to maximize the impact of persuasive applications*" (Berkovsky et. al., 2012).

In information-theoretical terms, persuasion is modeled by the Elaboration Likelihood Model (Petty and Cacioppo, 1986), which suggests that individuals with low motivation or ability may not elaborate the information provided (e.g. through a recommendation) and therefore users' neutral or negative behavioural response in recommendations (expressed in the form of low rating or non-selection of the recommended item) may not depict their actual intention towards the recommended item. In such cases, the utilization of additional peripheral cues (motivating elements) may increase the persuasive effect of recommendations by engaging users to further elaborate the provided information (Fogg, 2009) in order to investigate the potential to adopt the recommendation. In Recommender systems, explanations are typically used to provide users additional information that will support them in their decision making

process and can be eventually utilized as the means to pass users persuasive messages (Tintarev and Masthoff, 2011).

Along the above lines, the first objective of this research is to investigate the persuasive effect of the influence strategies proposed by Cialdini (1993), namely Reciprocation, Consistency, Social Proof, Liking, Authority, Scarcity, which are implemented as persuasive messages in the form of recommendations explanations in a movie recommender system developed for the purposes of this study.

Moreover, previous studies (e.g. Kaptein and Eckles, 2012) suggest that persuasive messages do not always achieve their goal to persuade users. Indeed, if users receive "wrong" messages (i.e. irrelevant or annoying) then negative behavioural responses may be generated. In this context, previous studies (e.g. Halko and Kientz, 2010) have demonstrated the significance of the individual's personality in the (negative or positive) behavioural responses to persuasive messages. Following the above argumentation, the second objective of this study is to examine the role of personality in the acceptance of the recommendations and identify possible differentiations in the users' response on the persuasive strategies that may attributed on their personality type.

In this study, we focus on peripheral cues such as short persuasive messages, developed upon Cialdini's (2001) six influence strategies, presented to user as recommendation explanations. We consider such messages as peripheral cues because they neither affect the quality of argumentation (i.e. how close to the users interests the recommended items are) nor change the recommended item but when users lack of motivation or ability, these peripheral variables influence users by triggering internal heuristic processing rules (Tam and Ho, 2005), which eventually would lead to persuasion

The rest of the paper is organized in five sections. In Section 2 the hypothesis development. Our experiment is presented in Section 3, while in Section 4 the experimental results are discussed. Discussion of the study's findings and a discussion of areas for further research conclude the paper.

2. HYPOTHESIS DEVELOPMENT

2.1 Influence strategies as messages in recommendation explanations

The mainstream of research in Recommender Systems has traditionally focused on designing and developing accurate recommendation algorithms (e.g. Xiao and Benbasat, 2007). More specifically, extant research indicates that the factor that mostly determines the success of a Recommender System is the provision of recommendations that are more close to consumer's preferences. According to the ELM perspective, the accuracy of recommendation algorithms determines the quality of argumentation. In other words, if the recommended item is close to the user preferences, this will eventually lead to persuasion through the central route, i.e. through in-depth processing of the recommendation. ELM suggests that the alternative (peripheral) path may also lead to persuasion if appropriate cues are provided. Such peripheral cues may be implemented as motivating messages in the form of recommendation explanation (Herlocker, 2000).

A recommendation explanation can be considered as any type of additional information accompanying a system's output, having as ultimate goal to persuade users to try or purchase the item that is recommended (Tintarev and Masthoff, 2011). Tintarev and Masthoff (2012) indicate that explanations have an important role in Recommender Systems since an explanation is a mean through which a consumer perceives the value of the recommended item so as to decide whether is close to his/her interests or not. Explanations can operate like motivators and are being used by several systems such as MovieLens (Herlocker et al., 2000) and Social software items (Guy et al., 2009). However, there is no clear indication in extant literature about what would be the content of explanations (i.e. the message passed to users) that can actually lead to persuasion. For example, a description of how the recommendation has emerged (i.e. transparency of recommendations) has been shown to be associated with an increase of trust in recommendations (Herlocker et al., 2000) while still there is no enough empirical evidence that demonstrates what type of messages could lead to persuasion (Halko and Kientz, 2010).

A number of persuasive (or influence) strategies have been proposed in the literature and can be eventually be utilized in the design of persuasive messages. For example, Fogg (2002) describes 42 persuasion strategies and Cialdini (2001) 6 influence strategies (also known as Six Weapons of Influence) In this study, we rely upon Cialdini's influence strategies since they have been broadly used and verified there are evidences that if influence strategies are implemented in a system then they increase its persuasive effect (e.g. Fogg, 2002). According to Cialdini (2001). Cialdini's (2001) influence strategies are the following:

- Reciprocity: humans have the tendency to return favors,
- Commitment or consistency: people's tendency to be consistent with their first opinion,
- Social proof: people tend to do what others do,
- Scarcity: people are inclined to consider more valuable whatever is scarce,
- Liking: people are influenced more by persons they like and
- Authority: people have a sense of duty or obligation to people who are in positions of authority.

Cialdini (1993) suggested that when a compliance professional (e.g. salesperson) uses the above six influence strategies (Reciprocity, Commitment, Social proof, Scarcity, Liking and Authority) in his/her strategy then (s)he managed to influence more successfully the customer to consume a product/service/information. In the same vein, Kaptein et al. (2012) suggests that applying the influence strategies on text messages people get persuaded to reduce snacking consumption. We adopted Cialdini's influence strategies because they have already been tested and validated in other domains such as in e-commerce (Kaptein, 2011), use of credit cards (Shu and Cheng, 2012). They also provide a solid framework in order to investigate the persuasive power of messages as peripheral cues in recommender systems. The above leads to following hypothesis of our study:

H1: Influence strategies (applied as peripheral cues through messages in recommendations explanations) will have a positive persuasive effect on individuals' disposition towards the recommended item.

The examination of the above hypothesis will allow us to demonstrate (if validated) that when the preference matching level of the recommended item is low (i.e. when the recommended item is not close to the user's preferences and interests), then enhancing the recommendation by applying influence strategies in

the form of short explanatory messages, the user will be persuaded to use the recommended item, thus changing his/her original negative behavior towards the recommended item to positive intention to use item.

Influence strategies rely upon different psychological principles that may lead to persuasion and therefore it is expected that they will present different degrees of persuasive effect on the recipients of the respective persuasive messages. Thus, the second hypothesis of our research is:

H2: Influence strategies lead to different degrees of persuasive effect on individuals' disposition towards the recommended item.

2.2 Personality

Kaptein and Eckles (2012) in their study demonstrated that influence strategies do not always lead to persuasion. They indicate that in case a consumer receives a message with 'wrong' principle then this can bring undesired effects. The above suggests that there are also other factors that should be taken into consideration when a persuasive message is used, one of which is individual's personality. A human's personality is defined as 'a dynamic organisation, inside the person, of psychophysical systems that create the persons' characteristic patterns of behaviour, thoughts and feelings' (Allport, 1961, p. 11).

Given that, one of the major aims of a Recommender System is to help consumers in decision making processes, the fact that personality influences how people make their decisions (Nunes et al., 2012), consumer's personality should be taken into consideration when a persuasive message is provided with a recommendation. Indeed, previous studies suggest a relationship between human's preferences and tastes with their personality in different domains such as movies (e.g. Chausson, 2010), music and paintings (Rawlings et al., 2000).

There is a variety of personality taxonomies one of which is Big 5 Dimensions of Personality (John et al., 2008). The personality traits suggested by the Big Five taxonomy are: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. According to psychological research (Jang et al., 2012) the facets for each personality trait are:

- Extraversion: Gregariousness, Assertiveness, Activity, Excitement-Seeking, Positive Emotions, Warmth.
- Agreeableness: Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-Mindedness.
- Conscientiousness: Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, Deliberation.
- Neuroticism: Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, Vulnerability.
- Openness: Ideas, Fantasy, Aesthetics, Actions, Feelings, Values.

The first study that examined message-person congruence effects with a comprehensive model of personality traits is that of Hirsh et al. (2012). Since then message-person congruence effects have been examined in relation to a variety of psychological characteristics (Dijkstra, 2008). Hirsh et al. (2012) demonstrated that persuasive messages are more effective when they are custom-tailored to their interests and concerns. Moreover, Tintarev et al. (2013) demonstrated that people who are characterized from Open to Experience (one of the Big 5 personality traits) tend to prefer diverse recommendations. Additionally, Halko and Kientz (2010) combined persuasive strategies with user's personality using Big Five Dimensions of Personality and the results of their study revealed relationships between individuals' personalities and persuasive technologies which means that not all people are affected from the same persuasive means. Finally, Smith et al. (2016) examined the impact of patients personality on Cialdini's influence strategies in the form of reminders. The research indicated that patient's with high emotional stability seem to be more responsive to all strategies of persuasion, while patients with low agreeableness rated all Cialdini's strategies higher than those with high. Finally, the research demonstrated that the reminders of "Authority" and "Liking" are the most popular.

3. EXPERIMENTAL DESIGN AND PROCEDURE

3.1 Design of Persuasive Explanation

For the execution of the experiment we had first to design the persuasive explanations that would accompany each recommended movie. For this task we followed the methodology proposed by Kaptein et al. (2012). More specifically, a group of three researchers familiar with Persuasive Technology, created thirty (30) textual explanations, i.e. five (5) for each Cialdini's influence strategies. The content of each explanation was developed in order to comply with the main purpose of each principle in the movie domain. For instance, for the influence strategy of Social Proof, the five possible persuasive explanations that were constructed are: (1) The 85% of this research's users rated the recommended movie with four (4) or five (5) stars. (2) The recommended movie is on 'to watch' list of 85% of this research's users. (3) Most of the users with the same age and sex as yours, rated the recommended movie with 4 stars! (4) The recommended movie's video trailer on youtube has more than 550,000 views. (5) The recommended movie's video trailer on youtube has more than 1600 likes and only 200 dislikes.

Seventeen (17) experts in the field of Information Systems and Marketing were invited in order to evaluate each explanation in terms of its compliance with the respective influence strategy. First, a brief presentation of the strategies was given to the evaluators so as to be more familiar with the influence strategies and then they were asked to evaluate the set of persuasive explanations. Each evaluator declared the compliance of each explanation to the respected influence strategy through a 1 to 5 rating scale (from "Completely Disagree" to "Completely Agree"). The persuasive explanation with the highest average was considered as the best-matching explanation for this particular influence strategy.

The six (6) best-matching persuasive explanations (one for each strategy), were chosen for the experiment are the following:

Reciprocity: A Facebook friend, who saw the movie that you suggested him/her in past, recommends you this movie.

Scarcity: The recommended movie will be available to view from 15/1/2014 to 31/1/2014 on cinemas.

Authority: The recommended movie won 3 Oscars!

Social Proof: The 87% of users in this survey rated the recommended movie with 4 or 5 stars!

Liking: Your Facebook friends like this movie.

Commitment: This movie belongs in the kind of movies you enjoy to watch.

3.2 Experiment design and execution

A within subjects experimental design was followed in this research. One of our main concerns in the execution of the experiment was to manage participants' burden by avoiding extensive exposure to treatments and questionnaires (only the psychographic questionnaire consisted of 44 items) while preserving the validity of the experiment. One option to deal with problem was to expose different groups to different cues (i.e. follow a between subjects design). However, this would significantly reduce the sample size within each group and also taking into account the anticipated low number different personality types represented in each of the groups it would have limited our ability to produce valid statistical results. Thus, we selected the within subjects design.

At the first step of the experiment, a set of 20 movies where presented to participants (with no explanations besides the typical information provided by iMDB, such short description of the story, lead actors etc.), where they were asked to state (by checking the appropriate option) whether they have watched each movie and then provide their ratings (in 1-5 scale). Users were explicitly instructed to provide their intention to watch a movie (for all unwatched movies) in the form of a rating. For the movies they had already watched they provided their actual evaluation. Recommendations were drawn from the set of unwatched movies.

The set of 20 movies was randomly selected from a pool of 60 movies from different genres and presented to the participants along with the typical information for each movie (movie's genre, its plot, and the starring actors). The first criterion for the inclusion of a movie in the pool of 60 movies was its genre (action, drama, romance, etc.). In the pool of 60 movies there were at least three movies from each genre, although most of the movies belong to more than one genre. The second criterion was the popularity of the movie, With the term popular movie is meant a movie with high average rating (above 8.0) from a large amount of users (above 1000 users).. Since popular movies are more likely to collect higher ratings while unpopular ones may not be known to the experimental participants (and therefore attract lower ratings), we included in the sample both popular and unpopular movies according to their iMDB ratings. Although that the number of 20 movies was large enough to ensure that at least some of them wouldn't have been watched by the participant, the system was designed to select from the pool of 60 movies and present to participants alternative movies in the extreme case that all 20 movies have been actually watched by the user.

At the second (recommendation) step of the experiment (see Figure 1), the (unwatched) movie for which the participant has expressed the lowest intention to watch (note that if more than one movie was rated with the lowest score, then the recommended movie was selected randomly from the above set of low-rated movies) was presented to the user exactly as the original presentation but enhanced with persuasive explanations. Selecting to present users with the lowest rated movie, is in alignment with our theoretical ELM foundations, which suggest that when the preference matching of the user with respect to the recommended item is low then the peripheral route will be followed. Moreover, this choice enable us to track more easily any changes in the user's intention to watch the movie since in computational terms it is much easier to identify changes in intentions from the lower to the higher levels of the 1-5 scale. It must be noted that the rating expresses the users' intention to watch (or not) the recommended movie is considered in our study as a measure of persuasion (i.e. acceptance of the recommendation), which is operationalized by computing the difference between the original and the final ratings. However, the exact meaning of the "acceptance the recommendation" depends on the business objectives of a site. For example, in some cases (as in e-commerce) the desired behaviour may be to request more information, or to purchase the product and so on.

Figure 1. Second Step of the experiment.

As mentioned above, the recommended movie was enriched with persuasive explanations, based on Cialdini's Principles (i.e. the explanations designed in the first part of the experiment) and the participant was asked to assess the recommended movie in order to examine whether (and which) strategies influenced users in order to change their intention to watch the recommended movie. More specifically, the recommended movie was presented with the same set of information as the first step (title, actors, etc.) while participants were asked to declare their intention to watch the recommended movie, taking into consideration one of the 6 persuasive explanations each time, which were presented as a list below the recommended movie. The order of the persuasive explanations was appeared in a random way to each user but there were the same texts for all of them. For that reason the expressions that were used in the persuasive explanations were in a generic form, e.g. the wording 'the recommended movie' was used instead of the actual title of the recommended movie and so on.

The absolute difference between the original and the final rating was used to measure the persuasive effect. As the "final" rating with respect to the first hypothesis (examining if there are differences before and after the application of the persuasion strategies) we used the highest rating that users provided (independently of the strategy that corresponds to that rating). For the evaluation of the second hypotheses (examining if there are differences among strategies with respect to their persuasive effect), the rating given by the users' as evaluation of each strategy was considered as the "final" rating.

At the third and last step of the experiment participants were asked to complete the psychographic questionnaire that was used to classify users into the Big 5 personality traits. The Big Five Inventory- 44 (BFI) was used, constitutes from 44 questions (John et al., 2008), and is already used in other studies (Bouchard and McGue, 2003; Shiota et al., 2006).

3.3 Sample

The experiment participants were invited through posts in University's Facebook groups (e.g. undergraduate, postgraduate and PhD students) and authors' personal mailing lists to participate in this research. The invitation message was asking recipients to participate in a research in which they would be

asked to rate recommendations provided by an online application as well as to fill in a psychographic questionnaire. The link to access the system was provided and a clear suggestion concerning the anonymity of their participation was included in the message. The invitation did not specify that the research involved movies evaluation. The participants' average scores for the items measuring the personality types in the 44-item psychographic questionnaire are (the standard deviation is included in the parentheses) Extraversion: 3.34 (0.49), Agreeableness: 3.47 (0.42), Conscientiousness: 3.34 (0.42), Neuroticism: 3.30 (0.48), Openness: 3.24 (0.46). The above descriptives showcase that the sample does not exhibit certain personality types more (or less) than others.

In total 117 users participated in our research. 61 (52%) participants of our sample were males while the rest 56 (48%) were females. Additionally, the 46% of the sample was aged between 18 and 24 years old, the 52% was between 25 and 34 years old and the 2% at the age of 35-44 years old.

3.4 Analysis Methodology

This research employs the prescriptions of the fuzzy-set qualitative comparative analysis methodology (fsQCA) to explore which personality traits explain the effectiveness of each persuasion strategy. Opposed to variance-based statistical methods (e.g. structural equation modelling or partial-least squares based regression models) in which the independent variables 'compete' with each other to explain one or more dependent variables, fsQCA treats the hypothesized causal factors as conditions that may be related to the phenomenon under investigation either by themselves or in combination with one another (Rihoux and Ragin,, 2009; Rihoux et al., 2011). Hence, fsQCA does not compute a single, optimal, solution that attributes weights to the independent variables; instead, the methodology proposes multiple alternative solutions, which require the presence or absence of each hypothesized causal factor. This is a fundamental difference from variance-based statistical methods and calls for operationalization of the variables in the dataset.

In effect, fsQCA employs fuzzy set theory and Boolean algebra to evaluate whether the cases in the dataset belong or not in a certain conceptual state. For example, in this research cases may be evaluated in order to assess whether an individual is extravert, open, agreeable, conscious, or neurotic. Likewise, the impact of each persuasion strategy on individuals' attitude change may also be operationalized to capture the degree to which the strategy actually manifested a behavior change. Such operationalizations are captured through fuzzy set membership scores ranging from 0 (non-membership to the set) to 1 (full membership to the set). In-between scores indicate the distance of each case from the outbound scores. The researcher may transform the cases' original values to fuzzy-set membership scores by using specialized fsQCA software. This process is coined with the term 'calibration'. In this research we used fsQCA 2.0 developed by the University of Arizona. The software was also employed throughout the remaining methodology stages.

Fuzzy-set QCA identifies conditions or combinations of conditions that are necessary or sufficient to explain an outcome. In this research, a combination of conditions reflects the personality profile of an individual. Such profile would include specific membership values to each personality trait following the calibration procedure. As such, a value close to 1 in a particular personality trait implies that the individual exhibits this trait. In contrast, membership values close to zero imply that the individual does not exhibit the said personality trait. Necessity of a condition implies that an outcome may not derive without the presence of the condition; nevertheless, the condition alone is not able to produce the outcome. Sufficiency of a condition implies that the condition alone is capable of producing the outcome. In practice, if a solution includes the presence of only one condition (i.e., a solution requires the presence or absence of only one personality trait),, then this condition is sufficient to produce the outcome. To estimate the sufficiency and/ or necessity of hypothesized conditions, fsQCA follows a Boolean minimization process based on truth table analysis. The outcome of this process includes the generic combinations of conditions that are sufficient for the outcome whilst remaining logically true. These are encapsulated in three solutions that differ based on their complexity, named as complex, intermediate, and parsimonious. Of interest is the parsimonious solution, which reduces the causal recipes to the smallest number of conditions possible.

This research explores how individuals' personality traits, in the form of five alternative dimensions, fit with different persuasive strategies. Nevertheless, an individual may not be exclusively categorized under a unique personality trait. Instead, individuals may exhibit elements of multiple traits, which collectively form their personality. Moreover, these personality traits are not fixed within all individuals; a particular persuasive strategy may be perceived as equally appropriate to individuals that exhibit completely dissimilar values on their fundamental personality qualities. As a result, we cannot assume that there is a single, universal, personality profile that explains the impact of a given persuasion strategy, which would call for the application of traditional statistical analysis methods based on regression models, but we need to examine how the different combination of the personality traits interweave in order to explain the suitability of a given persuasion strategy. The modus operandi of fsQCA covers this requirement, thus warrants us to adopt it as our guiding analysis methodology.

4. RESULTS

The first step of our analysis is involved investigating effect of each influence strategy on individuals' attitude towards watching a movie that they, initially, were unmotivated to watch. We performed two different comparisons to examine the persuasive effect of the influence strategies. In the first test, we measured the difference between the maximum of the ratings that each user provided for the six influence strategies and the original rating. The t-test results suggested that on average there are significant differences ($p < .001$) between the original rating and most persuasive (for each user) strategy (original and final ratings average scores: 1.49 and 3.05 respectively with standard deviation 0.50 and 1.23). In the second statistical test, we performed a t-test analysis that compares their initial beliefs and the ones formulated after the application of the strategy. The results suggest that all influence strategies were successful in increasing the likelihood of individuals to watch the movie (Table 1) nevertheless, this increase is marginal in absolute figures..

Table 1: T-test results. All comparisons are significant at p<.001

Influence Strategy	Mean (SD)	T-statistic (Original rating – intention after influence strategy is applied)
Original Rating	1.49(0.50)	n.a
Reciprocity	1.84(0.89)	-4.707 (p<.001)

Scarcity	1.73(0.97)	-2.953 (p<.001)
Authority	2.57(1.16)	-10.941 (p<.001)
Social Proof	2.67(1.17)	-12.349 (p<.001)
Liking	2.07(1.04)	-6.698 (p<.001)

Moreover, a one-way ANOVA test between the attitude changes of individuals for each influence strategy (see Table 2). The results of this analysis indicate that there are statistical differences among the six strategies at the p<.05 level (F= 14.941, p= .000). To probe for differences between the strategies we performed a Games-Howell Post Hoc Test. Based on these results we accept H1.

Table 2: ANOVA results (Sign. < 0.05)

Persuasive Strategy	Reciprocity	Authority	Scarcity	Social Proof	Liking	Consistency
Sign.	**.001**	**.001**	**.006**	**.001**	**.003**	**.007**

H2 was evaluated through the application of fsQCA methodology. We used the five personality traits as possible conditions that influence the acceptance of each influence strategy. As a first step, the prescriptions of fsQCA require for calibration of the cases into membership sets. Calibration was performed using the corresponding function provided by fsQCA 2.0 software. The function demands as input three threshold points; a full-membership value, a non-membership value and a cutoff point. Because the dataset consists of subjective cases, we used cluster analysis following the k-means algorithm (k=3) to calculate the three membership sets. More specifically, high values are correlated with the full-membership set, medium values are correlated with the crossover point set and finally low values are correlates with the non-membership set.

For the independent variables (personality traits) no cluster analysis was conducted due to the fact that the differences among the personality traits' scores were imperceptibly small. Thus, for this case we calculated the independent variables (personality traits) through frequencies with cut points for 4 equal groups, in SPSS. The percentiles that emerged correspond to the full-membership set for the high values, the crossover point set for medium values and finally the non-membership set for low values.

The results of fsQCA indicate 3-7 alternative solutions per influence strategy comprising of alternative combinations of the personality traits that lead to high acceptance of each influence strategy. Black circles indicate the required presence of a personality trait in a solution. White circles indicate the required absence of a personality trait from the solution. Blank cells indicate that in that particular solution, the presence or absence of that personality trait is indifferent. Each solution is accompanied by two additional measurements of fitness, which express the 'predictive power' of each solution, namely the consistency and coverage indexes. Consistency presents how consistent is the empirical evidence with the outcome which is investigated while coverage estimates the proportion of cases that address the outcome which is under investigation.

Table 3 illustrates the results of fsQCA for the Reciprocity influence strategy. The methodology, identified four solutions leading to high influence of an individual by the application of the respective strategy. The results indicate that the absence of even one personality trait is sufficient to individuals in order to be influenced by the Reciprocity strategy

Table 3: fsQCA results for the paths leading to high acceptance of Reciprocity.

	Solutions leading to high acceptance of Reciprocity influence strategy			
Personality Traits	**1**	**2**	**3**	**4**
Extraversion	○			
Agreeableness		○		
Conscientiousness			○	
Openness				○
Neuroticism				
Consistency	0.672	0.636	0.644	0.70
Coverage	0.578	0.624	0.572	0.639
Overall solution consistency	0.611			
Overall solution coverage	0.970			

The methodology identified 6 alternative paths leading to high acceptance of the Scarcity influence strategy. The majority of paths require two personality traits to be present in an individual's personality in order to be influenced by Scarcity strategy (Table 4). For example, individuals that are both agreeable and conscious, but do not exhibit traits of neuroticism are likely to be influenced by the scarcity influence strategy (solution 6).

Table 4: fsQCA results for the paths leading to high acceptance of Scarcity.

	Solutions leading to high acceptance of Scarcity influence strategy					
Personality Traits	**1**	**2**	**3**	**4**	**5**	**6**
Extraversion	○					
Agreeableness	●	○		●		●
Conscientiousness			○	●		●
Openness		●	●	○	●	
Neuroticism					○	○
Consistency	0.797	0.7	0.7	0.87	0.7	0.873
Coverage	0.295	0.416	0.358	0.193	0.376	0.206
Overall solution consistency	0.685					
Overall solution coverage	0.747					

The remaining Tables present the different paths, consisting of combinations of personality traits, which lead to high acceptance of the remaining four influence strategies. These tables may be interpreted as an atypical personality profile of individuals (one per produced fsQCA solution) in order to be influenced by each strategy (Table 5 – Table 8). Similar to the previous solutions, each table should be interpreted as a combination of mandatory personality traits (indicated with black circles) coupled with the mandatory absence of one or more personality traits (indicated with white circles). Hence, each solution represents a unique combination of the personality traits that should exist in order to explain the acceptance of a persuasive strategy.

Table 5: fsQCA results for the paths leading to high acceptance of Authority.

	Solutions leading to high acceptance of Authority influence strategy					
Personality Traits	**1**	2	3	4	5	6
Extraversion	○					
Agreeableness			●	●	○	○
Conscientiousness		○		●		●
Openness	●		○		●	●
Neuroticism		●		○	●	
Consistency	0.598	0.604	0.62	0.674	0.677	0.636
Coverage	0.294	0.303	0.357	0.182	0.252	0.25
Overall solution consistency	0.566					
Overall solution coverage	0.752					

Table 6: fsQCA results for the paths leading to high acceptance of Social Proof.

	Solutions leading to high acceptance of Social Proof influence strategy						
Personality Traits	**1**	2	3	4	5	6	7
Extraversion	○	●	○		●		●
Agreeableness				○	○		
Conscientiousness			○			○	○
Openness	●	○		●		●	
Neuroticism	●	●	○	●	●	●	●
Consistency	0.698	0.645	0.619	0.604	0.581	0.698	0.637
Coverage	0.31	0.25	0.303	0.317	0.250	0.31	0.190
Overall solution consistency	0. 713						
Overall solution coverage	0. 577						

Table 7: fsQCA results for the paths leading to high acceptance of Liking.

	Solutions leading to high acceptance of Liking influence strategy		
Personality Traits	**1**	2	3
Extraversion			○
Agreeableness	○		●
Conscientiousness		○	●
Openness	●		
Neuroticism		●	
Consistency	0.47	0.48	0.64
Coverage	0.41	0.31	0.192
Overall solution consistency	0. 456		
Overall solution coverage	0.643		

Table 8: fsQCA results for the paths leading to high acceptance of Consistency.

	Solutions leading to high acceptance of Liking influence strategy		
Personality Traits	**1**	2	3
Extraversion			○
Agreeableness	○		●
Conscientiousness		○	●
Openness	●		
Neuroticism		●	
Consistency	0.47	0.48	0.64
Coverage	0.41	0.31	0.192
Overall solution consistency	0. 456		
Overall solution coverage	0.643		

5. DISCUSSION

This research emphasizes on two elements of persuasive/ recommender systems. First, we empirically validate that the application of an influence strategy may indeed positively shift the attitude of an individual towards a specific recommended item. Nevertheless, not all influence strategies have the same persuasive effect. We attribute this deviation to the personality traits of the recommender system users. Hence, the second contribution of this study reflects on the development of personality profiles per influence strategy. Each profile, measured as a combination of personality traits that need to be present or absent from the personality mix, reflects the set of traits that fit most with each influence strategy (i.e., individuals sharing the same profile would indeed be persuaded following the application of the respective strategy). It must be noted that an important issue in utilizing recommendation explanations is that persuasive messages may be perceived as promotional ones and therefore impact users' trust in the recommender systems. For this reason we used a control variable measuring (in an 1-5 scale) users' trust in the system, which has shown that no such effect occurred (i.e. no significant differences were found between the trust levels before and after the presentation of the persuasive messages, which was on average 2.96 for the users with low intention to

watch the movie and 3.27 for the users with high intention to watch a movie).

In effect, most studies in the field of recommender systems have primarily focused on the algorithmic perspective through the proposition of algorithms that provide recommendations tailored to users' interests and preferences. In contrast, this study provides insights indicating that the provision of properly selected (i.e. taking into account users' personality) motivating messages have a persuasive effect on users intention to "use" the recommended item, e.g. to watch a movie.

According to the Elaboration Likelihood Model (ELM), when an individual has low motivation (or ability) to process a recommendation then she will not proceed through the central route of persuasion, i.e. he will not thoroughly assess the quality of argumentation in order to get persuaded. Instead, if appropriate peripheral cues are implemented (such as persuasive strategies applied in the form of messages, as suggested in our study) then she will eventually be influenced (i.e. motivated) to elaborate the recommendation following the peripheral route to persuasion. Such peripheral cues act as extra motivating triggers that influence a user by "diverting attention, reallocating cognitive resources, and evoking affective responses and behaviours" (Tam and Ho, 2005).

Current recommendation applications typically disregard items with low degrees of fitness with the users' current interests. The confirmation of the first hypothesis of this study indicate that even for such items, there is strong possibility that they may be favoured by the users if they are presented with the appropriate motivating peripheral cue. Moreover, not all people are influenced from the same persuasive messages. This study provides empirical evidence that there is a relationship between personality and Persuasive Strategies. People with specific combination of personality traits are affected more from particular persuasive messages.

The results of the experiment that was conducted surfaced that motivating messages are not uniformly applied to all recipients of recommendations. Users' personality traits are an important factor that differentiates the effect of influence strategies applied as persuasive explanations. More specifically, a person who is characterized by high extraversion seems to be influenced by all Six Persuasive Strategies. This is reasonable if we take into consideration that they enjoy interacting with the environment whilst such people have the tendency to seek for stimulation (Zhao and Siebert, 2006). Moreover, people with high extraversion have the tendency to be curious, novel, sociable, active, energetic (Costa and McCrae, 1992; Goldberg, 1992), and positive (Watson and Clark, 1997). Along this line, the fact that this type of people favour networking with others (Watson and Clark, 1997) make them more prudent to be influences by "Liking" strategies.

Individuals with high agreeableness are eager to help other people (Costa and McCrae, 1992) while they have the tendency to be kind, generous, fair and unconditional (Goldberg, 1992), so people with high agreeableness tend to be motivated from the "Reciprocity" influence strategy. The fact that people with low agreeableness tend to be suspicious (Digman, 1990).

People with high conscientiousness are dutiful (Major et al., 2006). In other words, they are careful to fulfil obligations, and thus when someone helps them they feel obligated so they become more motivated when a persuasive explanation implementing "Reciprocity" is presented to them. Despite our expectations, humans with low conscientiousness changed their intention to watch the movie influenced by the "Consistency" strategy rather than humans with high conscientiousness. This may be attributed to the fact that individuals with high conscientiousness avoid to take risks because that might make them feel uncertain or cause unexpected delays to their work (James and Mazerolle, 2002; Raja and Johns, 2004).

On the other hand, people with high openness tend to be characterized by creativity, sophistication, and curiosity (Barrick and Mount, 1991). This might explain why in most cases, the trait of openness is absent from the solutions indicated by fsQCA. Finally, individuals with low neuroticism lack confidence. This may explain why the application of the "Social Proof" strategy on neurotics in most of cases depicts low neuroticism and Liking, because they tend to be influenced by people who they like or what the majority says. Additionally, neurotics are characterized by anxiety and typically they do not trust others (Raja and Johns, 2004), so they tend to be consistent with their original thoughts in order to deal with their insecurity and therefore it is expected to get persuaded by the "Consistency" strategy.

The findings of the study must be interpreted taking into account its limitations. The sampling frame (students) and the relatively low sample size restrict the possibility of having an actual representation of the population in the sample in terms of personality types. By extending the experiment, in future research, to a larger sample of users we would also have the opportunity to avoid possible self-selection bias as well as to follow a between subjects design, showing not only more movies to each user but most importantly avoiding the learning effect associated with the presentation of all six strategies to all experiment participants. It must be noted that we tried to control the learning effect bias by showing to users recommendations with persuasive explanations in a random way, i.e. the mix of recommendations representing different persuasive strategies was presented in varying order to each of the participants. It is clear that this study provides insights concerning the movie recommendation domain in which it was applied. The generalization of our findings would be enabled only if this research is extended to other application domains. In our future research plans, besides the extension of our research to other domains (e.g. e-commerce) we aim to investigate additional factors that may influence persuasive communication, as for example the need for cognition, which is a personality variable and reflects people's intrinsic motivation to engage in and enjoy thinking (Cacioppo and Petty, 1982, p. 116).

6. ACKNOWLEDGMENTS

The first author acknowledges the financial support of the Department of Management Science and Technology and the third author the financial support of the Research Center of the Athens University of Economics and Business for the presentation of this work.

7. REFERENCES

[1] Allport, G. W.: Pattern and growth in personality. New York: Holt, Rinehart & Winston (1961).

[2] Barrick, M. R., Mount, M. K.: The Big Five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44, 1-26 (1991).

[3] Berkovsky, S., Freyne, J., Oinas-Kukkonen, H. Influencing Individually: Fusing Personalization and Persuasion. ACM Transactions on Interactive Intelligent Systems, 2(2) (2012).

[4] Cialdini, R. B.: Influence: Science and practice (3rd ed.). New York: HarperCollins (1993).

[5] Cialdini RB. Influence, Science and Practice, Allyn & Bacon, Boston (2001).

[6] Chausson, O.: Who Watches What? Assessing the Impact of Gender and Personality on Film Preferences. myPersonality project, Univeristy of Cambridge (2010).

[7] Costa, P. T., Jr., McCrae, R. R.: Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources (1992).

[8] Digman, J. M.: Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 41, 417-440 (1990).

[9] Fogg, B. J.: Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann (2002).

[10] Fogg, B. J.: A behavior model for persuasive design. Proceeding of Persuasive '09 Proceedings of the 4th International Conference on Persuasive Technology (2009).

[11] Goldberg, L. R.: The development of markers for the Big-Five factor structure. Personality Assessment, 4, 26-42 (1992).

[12] Halko, S., Kientz, J.A.: Personality and persuasive technology: An exploratory study on health-promoting mobile applications. Persuasive technology, pp. 150–161 (2010).

[13] Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. CSCW 2000: 241-250 (2000).

[14] Hirsh, J. B., Kang, S. K., Bodenhausen, G. V.: Personalized Persuasion Tailoring Persuasive Appeals to Recipients' Personality Traits.Psychological science, 23(2), 578–581 (2012).

[15] Jang, K.L.,Livesley,W.J., Vemon, P.A.:Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study. Journal of personality, 64(3), pp. 577–592 (1996).

[16] James, L. R., Mazerolle, M. D.: Personality in work organizations. Thousand Oaks, CA: Sage (2002).

[17] John, O. P., Naumann, L. P., Soto, C. J.: Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (3rd ed., pp. 287–310). NewYork: Guilford Press (2008).

[18] Kaptein, M., Eckles, D.: Heterogeneity in the effects of online persuasion.Journal of Interactive Marketing, 26(3), pp.176–188 (2012).

[19] Kaptein, M., De Ruyter, B., Markopoulos, P., Aarts, E.: Adaptive persuasive systems: a study of tailored persuasive text messages to reduce snacking. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(2) (2012).

[20] Major, D. A., Turner, J. E., Fletcher, T. D.: Linking proactive personality and the big five to motivation to learn and development activity. Journal of Applied Psychology, 91, 927-935. doi:10.1037/0021-9010.91.4.927 (2006).

[21] Nunes, M. A. S. N., Hu, R.: Personality-based Recommender Systems: An Overview. In: Proc. of the 6th ACM Conference on Recommender Systems (RecSys'12), pp. 5-6 (2012).

[22] Petty, R.,Cacioppo, J.: The elaboration likelihood model of persuasion. Advances in experimental social psychology, (1986).

[23] Raja, U., Johns, G.: The impact of personality on psychological contracts. Academy of Management Journal, 47, 350-367 (2004).

[24] Rawlings, D., Barrantes, N., Furnham, A.,: Personality and aesthetic preference in Spain and England: two studies relating sensation seeking and openness to experience to liking for paintings and musi. European Journal of Personality, Volume 14, Number 6, November/December 2000, pp. 553-576(24) (2010).

[25] Rihoux, B., and Ragin, C.C. 2008. Configurational Comparative Methods: Qualitative Comparative Analysis (Qca) and Related Techniques. SAGE Publications, Incorporated.

[26] Rihoux, B., I. Rezsohazy & D. Bol (2011): Qualitative Comparative Analysis (QCA) in Public Policy Analysis: an Extensive Review. German Policy Studies, 7(3), 9–82.

[27] Smith, K.A., Dennis, M., Masthoff, J. Personalizing Reminders to Personality for Melanoma Self-checking. Conference on User Modeling Adaptation and Personalization, p. 85-93. (2016)

[28] Tam, K.Y., Ho, S.Y.: Web personalization as a persuasion strategy: An elaboration likelihood model perspective. Information Systems Research, pp. 271 - 291 (2005).

[29] Tintarev, N., Masthoff J.: Designing and evaluating explanations for recommender systems. Recommender Systems Handbook, Springer, p. 479–510 (2011).

[30] Tintarev, N., Masthoff J.: Evaluating the effectiveness of explanations. User Model User-Adap Inter. pp.399–439 (2012).

[31] Tintarev, N., Dennis, M., Masthoff J Adapting Recommendation Diversity to Openness to Experience: A Study of Human Behaviour. User Modelng, Adaptation and Personalization, vol. 7899, pp.190–202 (2013).

[32] Watson, D., Clark, L. A.: Extraversion and its positive emotional core. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), Handbook of personality psychology (pp. 767–793). San Diego, CA: Academic Press (1997).

[33] Xiao, B., Benbasat, I.: E-commerce product recommendation agents: use, characteristics, and impact. Mis Quarterly, 31(1), pp. 137–209 (2007).

[34] Zhao, H., Seibert, S. E.: The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. Journal of Applied Psychology, 91, pp. 259-271(2006).

Personality in Computational Advertising: A Benchmark

Giorgio Roffo
University of Verona
Department of Computer Science
Giorgio.Roffo@univr.it

Alessandro Vinciarelli
University of Glasgow
School of Computing Science
Alessandro.Vinciarelli@glasgow.ac.uk

ABSTRACT

In the last decade, new ways of shopping online have increased the possibility of buying products and services more easily and faster than ever. In this new context, personality is a key determinant in the decision making of the consumer when shopping. A person's buying choices are influenced by psychological factors like impulsiveness; indeed some consumers may be more susceptible to making impulse purchases than others. Since affective metadata are more closely related to the user's experience than generic parameters, accurate predictions reveal important aspects of user's attitudes, social life, including attitude of others and social identity. This work proposes a highly innovative research that uses a personality perspective to determine the unique associations among the consumer's buying tendency and advert recommendations. In fact, the lack of a publicly available benchmark for computational advertising do not allow both the exploration of this intriguing research direction and the evaluation of recent algorithms. We present the ADS Dataset, a publicly available benchmark consisting of 300 real advertisements (i.e., Rich Media Ads, Image Ads, Text Ads) rated by 120 unacquainted individuals, enriched with Big-Five users' personality factors and 1,200 personal users' pictures.

CCS Concepts

•**Information systems** → **Computational advertising;** *Collaborative search; Test collections;*

Keywords

Recommender Systems, Computational Advertising, Ads Click Prediction, Ads Rating Prediction, Personality Traits, Data Mining

1. INTRODUCTION

Nowadays, online shopping plays an increasingly significant role in our daily lives [10]. Most consumers shop online with the majority of these shoppers preferring to shop online for reasons like saving time and avoiding crowds. Marketing campaigns can create awareness that drive consumers all the way through the process to actually making a purchase online [16]. Accordingly, a challenging problem is to provide the user with a list of recommended advertisements they might prefer, or predict how much they might prefer the content of each advert.

Past studies on recommender systems take into account information like user preferences (e.g., user's past behavior, ratings, etc.), or demographic information (e.g., gender, age, etc.), or item characteristics (e.g., price, category, etc.). For example, collaborative filtering approaches first build a model from a user's past behavior (e.g., items previously purchased and/or ratings given to those items), then use that model to predict items (or ratings for items) that the user may have an interest in by considering the opinions of other like-minded users. Other information (e.g., contexts, tags and social information) have also taken into account in the design of recommender systems [5, 18, 20].

The impact of personality factors on advertisements has been studied at the level of social sciences and microeconomics [2, 9, 35]. Recently, personality-based recommender systems are increasingly attracting the attention of researchers and industry practitioners [6, 15, 33]. Personality is the latent construct that accounts for "*individuals characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns*" [12]. Hence, personality is a critical factor which influences people's behavior and interests. Attitudes, perceptions and motivations are not directly apparent from clicks on advertisements or online purchases, but they are an important part of the success or failure of online marketing strategies. A person's buying choices are further influenced by psychological factors like impulsiveness (e.g., leads to impulse buying behaviors), openness (e.g., which reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has), neuroticism (i.e., sensitive/nervous vs. secure/confident), or extraversion (i.e., outgoing/energetic vs. solitary/reserved) which affect their motivations and attitudes [35].

To the best of our knowledge, the impact of personality factors on advertisements has been largely neglected at the level of *advert recommendation*. There is a high potential that incorporating users' characteristics into recommender systems could enhance recommendation quality and user experience. For example, given a user's preference for some items, it is possible to compute the probability that they are of the same personality type as other users, and, in turn, the probability that they will like new items [24].

Moreover, personality has shown to play an important role also in other aspects of recommender systems, such as implicit feedback, contextual information [21], affective content labeling [34]. With the development of novel techniques for the unobtrusive acquisition of personality (e.g. from social media [7, 28, 29]) this study is meant to contribute to this emerging domain proposing a new corpus which includes questionnaires of the Big-Five (BFI-

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

10) personality model [25], as well as, users' liked/disliked pictures that convey much information about the users' attitudes [7].

The ADS Dataset is a highly innovative collection of 300 real advertisements rated by 120 participants and enriched with the users' five broad personality dimensions, which have been shown to capture most individual differences [4]. The user study is conducted by recruiting a set of test subjects, and asking them to perform several tasks. The subjects included in the corpus were recruited through a public platform purely dedicated to recruiting participants. The process was stopped once the first 120 individuals answered positively. The experimental protocol adopted for the data collection has been designed to capture users' preferences in a controlled usage scenario (see Section 2.1 for further details).

In this work we carry out prediction experiments performing two different tasks: *ad rating prediction* or *ad click prediction*, with the goal in mind to analyze the effect of using personality data for recommending ads. Therefore, we propose Logistic Regression (LR) [8], Support Vector Regression with radial basis function (SVR-rbf) [3], and L2-regularized L2-loss Support Vector Regression (L2-SVR) [8] as baseline systems for recommendation. We then review a large set of properties, and explain how to evaluate systems given relevant properties. We also survey a large set of evaluation metrics in the context of the property that they evaluate, and provide a library within one integrated toolbox.

Summarizing, the contribution of this work is two-fold:

Dataset: we collect and introduce a representative benchmark for computational advertising enriched with affective-like metadata such as personality factors. The benchmark allows to (i) explore the relationship between consumer characteristics, attitude toward online shopping and advert recommendation, (ii) identify the underlying dimensions of consumer shopping motivations and attitudes toward online in-store conversions, and (iii) have a reference benchmark for comparison of state-of-the-art advertisement recommender systems (ARSs). To the best of our knowledge, the ADS dataset is the first attempt at providing a set of advertisements scored by the users according to their interest into the content.

Code library: we present two broad classes of prediction accuracy measures, depending on the task the recommender system is performing: "ad rating prediction" or "ad click prediction", and provide a code library, integrating the evaluation metrics with uniform input and output formats to facilitate large scale performance evaluation. The code library and the annotated dataset are available on the *project page*[1].

The rest of the paper is organized as follows: in Section 2 we present and describe the ADS Dataset. We perform a corpus analysis investigating on the linkages between buying habits, recommendations, and personality. In Section 3, we survey a large set of evaluation metrics in the context of the property that ARSs evaluate. In Section 4 we conduct experiments for each scenario taken into account in this work, investigating on the strengths and weakness of using personality data as features for recommendation. Finally, in Section 5 conclusions are given, and future perspectives are envisaged.

2. CORPUS ANALYSIS

The corpus includes 300 advertisements voted by unacquainted individuals (120 subjects in total. Note, the data collection process is still running). Adverts equally cover three display formats: Rich Media Ads, Image Ads, Text Ads (i.e., 100 ads for each format). Participants rated (from 1-star to 5-stars) each recommended advertisement according to if they would or would not click on it (some examples are shown in the Fig.1). We labeled adverts as "clicked" (rating greater or equal to four), otherwise "not clicked" (rating less than four). The distribution of the ratings across the adverts that were scored by the users turns out to be unbalanced

[1]http://giorgioroffo.it/?ADSdataset

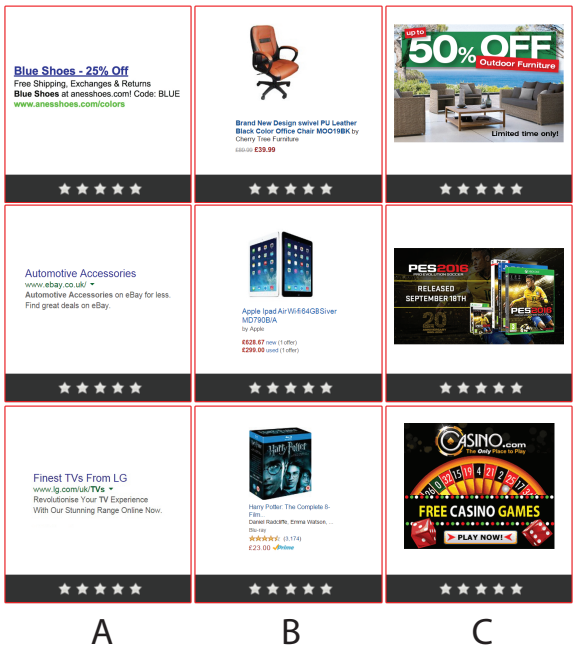


Figure 1: The figure shows three different examples for each display format. (A) Shows Text Ads that received 26.5% of the total amount of clicks. (B) Image Ads (32.7% of clicks), and (C) Rich Media Ads (40.8% of clicks).

Class Labels	Category Name	% Clicks
1	**Clothing & Shoes**	**6.2%**
2	Automotive	3.3%
3	Baby	3.3%
4	Health & Beauty	6.0%
5	Media	6.6%
6	**Consumer Electronics**	**9.2%**
7	**Console & Video Games**	**8.5%**
8	Tools & Hardware	3.0%
9	Outdoor Living	5.6%
10	**Grocery**	**7.3%**
11	Home	4.7%
12	Betting	1.6%
13	Jewelery & Watches	5.9%
14	Musical instruments	3.6%
15	Stationery & Office Supplies	5.4%
16	Pet Supplies	3.1%
17	Computer Software	5.6%
18	Sports	5.0 %
19	Toys & Games	5.1%
20	Social Dating Sites	1.0%

Table 1: ADS Dataset provides a set of 15 real adverts categorized in terms of 20 product/service categories. The most clicked categories are highlighted in green and the less clicked in red.

Group	Type	Description	References
Users' Preferences	Websites, Movies, Music, TV Programmes, Books, Hobbies	Categories of: websites users most often visit (WB), watched films (MV), listened music (MS), watched T.V. Programmes (TV), books users like to read (BK), favourite past times, kinds of sport, travel destinations.	[14, 18, 20]
Demographic	Basic information	Age, nationality, gender, home town, CAP/zip-code, type of job, weekly working hours, monetary well-being of the participant	[20]
Social Signals	Personality Traits	BFI-10: Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN)	[4, 25]
	Images/Aesthetics	Visual features from a gallery of 1.200 *positive / negative* pictures and related meta-tags	[7]
Users' Ratings	Clicks	300 ads annotated with Click / No Click by 120 subjects	[14, 23, 37]
	Feedback	From 1-star (Negative) to 5-stars (Positive) users' feedback on 300 ads	[14, 23, 37]

Table 2: The table reports the type of raw data provided by the ADS Dataset. Data of the first and last group can be considered as historical information about the users in an offline user study.

(4,841 clicked vs 31,159 unclicked).

Advert content is categorized in terms of 20 main product/service categories. For each one of the categories 15 real adverts are provided. Table 1 reports the full list of the categories used with the associated class annotations and the percentage of clicks received. At the category level, the distribution of the ratings results to be balanced (1,229 clicked vs 1,171 unclicked), where a category is considered to be clicked whenever it contains at least one clicked advert.

Inspired from recent findings which investigate the effects of personality traits on online impulse buying [2, 9, 35], and many other approaches based upon behavioral economics, lifestyle analysis, and merchandising effects [2, 19], the proposed dataset supports a trait theory approach to study the effect of personality on user's motivations and attitudes toward online in-store conversions. The trait approach was selected because it encourages the use of scientifically sound scale construction methods for developing reliable and valid measures of individual differences. As a result, the corpus includes the Big Five Inventory-10 to measure personality traits [25], the five factors have been defined as *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*, often listed under the acronyms *OCEAN*.

Recent soft-biometric approaches have shown the ability to unobtrusive acquire these traits from social media [28, 29], or infer the personality types of users from visual cues extracted from their favorite pictures [7] from a social signal processing perspective [36]. While not necessarily corresponding to the actual traits of an individual, attributed traits are still important because they are predictive of important aspects of social life, including attitude of others and social identity.

As a result, the proposed benchmark includes 1,200 spontaneously uploaded images that hold a lot of meaning for the participants and their related annotations: *positive/negative* (see Table 2 for further details). The images are personal (i.e., family, friends etc.) or just images participants really like/dislike. The motivations for labeling a picture as favorite are multiple and include social and affective aspects like, e.g., positive memories related to the content and bonds with the people that have posted the picture. Moreover, they are provided with a set of TAGS describing the content of each of them.

Finally, many other users' preference information are provided. Table 2 lists the raw data provided with the dataset, such as users' past behavior selected from a pre-defined list (e.g., watches movies, listen songs, read books, travel destinations, etc.), demographic information (like age, nationality, gender, etc.). Note, all data is anonymized (i.e., name, surname, private email, etc.), ensuring the privacy of all participants.

For further analyses related to the adverts' quality, this benchmark also provides the entire set of 300 rated advertisements (500 x 500 pixels) in PNG format.

2.1 Participant Recruitment

The subjects involved in the data collection, performed all the steps of the following protocol:

- *Step 1*: All participants have filled in a form providing, anonymously, several information about their preferences (e.g., demographic information, personal preferences).

- *Step 2*: All participants have filled the Big Five Inventory-10 to measure personality traits [25].

- *Step 3*: The participants voted each advert according with if they would or not click on the recommended ad. Ads have been displayed in the same order to all the participants.

- *Step 4*: The participants submitted some images that they like (Positives) and some others that disgust or repulse them (Negatives). Once they have uploaded their images, they also added some TAGS that describe the content of each image.

2.2 The Subjects

This corpus involves 120 English native speakers between 18 and 68. The median of the participants age is 28 (μ=31.7, σ=12.1). Most of the participants have a university education. In terms of gender, 77 are females and 43 males. The percentage distribution of household income within the sample is: 23% less or equal to 11K USD per year, 48% from 11K to 50K USD, 21% from 50K to 85K USD, and 8% more than 85K USD. The median income is between 11K and 50K USD.

In analyzing this complex data, one can observe that users' preferences are not independent of each other, they are likely to be co-expressed. Hence, it is of great significance to study groups of preferences rather than to perform a single analysis. This fact is

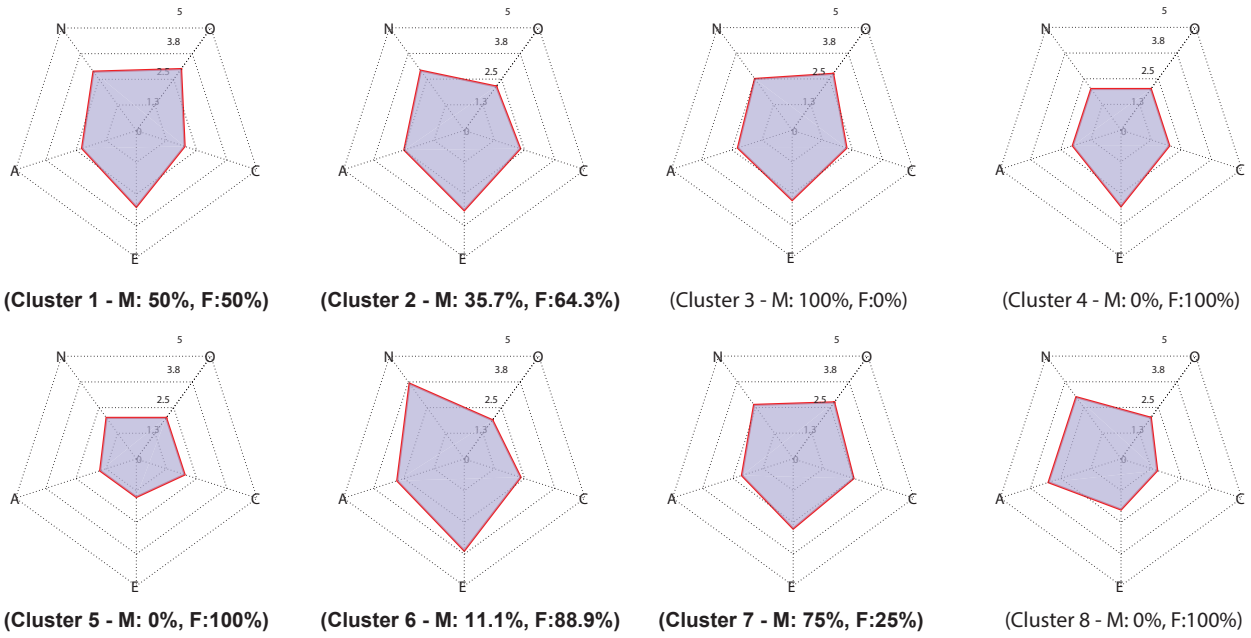


Figure 2: Spider-Diagrams for *O-C-E-A-N* Big-Five traits. The percentage of Males (M) and Female (F) belonging to each cluster is reported. We indicate in bold each instance where a statistacal significant effect (i.e., Pearson correlation at the 5% level) was found between ranks and personality factors.

also true for personality factors, analyzing subsets of data yields crucial information about patterns inside the data. Thus, clustering users' preferences can provide insights into personality of individuals which share the same preferences. We performed a statistical analysis of personality and users' preferences, linking the 5 personality factors and the most favorite users' product categories (i.e., most clicked) by means of the affinity propagation (AP) clustering algorithm [11].

AP is an algorithm that takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars. We calculated a similarity input matrix between each individual u_i considering as feature vectors v_i a binary sequence of click/no-click (i.e., v_i is 1×300). AP exchanges real-valued messages between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. Hence, the number of clusters is automatically detected, and when applied on ADS data, AP grouped the data into 8 different clusters.

Figure 2 illustrates 8 spider-diagrams, one for each cluster. Each diagram shows the average of the big-five factors regarding the subjects within the group (reported in figure as O-C-E-A-N).

Then, we ranked the most clicked categories according with samples within the group in order to compare these two variables by means of correlation obtaining interesting clues.

For instance, let us consider the cluster number 6 where 88.9% of the members are females, and 11.1% males and the average of the group members age is 28. The first 5 most clicked categories are *Baby*, followed by *Consumer Electronics*, *Stationery & Office Supplies*, *Home*, and *Jewelery & Watches*. This group is characterized by high neuroticism (see the diagram in Figure 2.(Cluster 6)), those who score high in neuroticism are often emotionally reactive and vulnerable to stress, high neuroticism causes a reactive and excitable personality, often very dynamic individuals. This group also share the highest levels of extroversion, high extraversion is often perceived as attention-seeking, and domineering.

Cluster 5 shows a subset of individuals which scores low for all the types (see the plot in Figure 2.(Cluster 5)). For instance, those with low openness seek to gain fulfillment through perseverance, and are characterized as pragmatic sometimes even perceived to be dogmatic. Some disagreement remains about how to interpret and contextualize the openness factor. The first 5 most clicked categories are *Clothing & Shoes*, *Health & Beauty*, *Jewelery & Watches*, *Outdoor Living*, and then *Consumer Electronics*. In this case the average of the group members age is 68, and the cluster contains 100% females.

Cluster Id	Avg. Age	r.1	r.2	r.3	r.4	r.5
1	**32**	**6**	**15**	**13**	**19**	**4**
2	**31**	**6**	**5**	**7**	**10**	**17**
3	22	1	7	10	4	6
4	57	6	9	10	2	1
5	**68**	**1**	**4**	**13**	**9**	**6**
6	**28**	**3**	**6**	**15**	**11**	**13**
7	**20**	**7**	**6**	**10**	**11**	**17**
8	52	3	9	19	7	4

Table 3: Top-5 ranked categories. For each cluster the table reports the average age, and the ordered list of the most clicked categories. We indicate in bold each instance where a statistical significant effect (i.e., Pearson correlation at the 5% level) was found between ranks and personality factors.

Cluster 7 is characterized by good levels of conscientiousness that is the tendency to be organized and dependable, aim for achievement, and prefer planned rather than spontaneous behavior. This cluster scores low in agreeableness, which is related to personalities often competitive or challenging people. The openness factor (>2.5) reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has. Interestingly, among the most preferred categories there are *Console & Video Games*, *Consumer Electronics*, *Grocery* and *Computer Software*.

3. EVALUATION METHODOLOGY

Research in the ARS field requires quality measures and evaluation metrics to know the quality of the techniques, methods, and algorithms for predictions and recommendations. In this section we review the process of evaluating an ARS on two main tasks: (i) measuring the accuracy of rating predictions, and (ii) measuring the accuracy of click predictions.

3.1 Scenario 1: Ad Rating Prediction

In most online advertising platforms the allocation of ads is dynamic, tailored to user interests based on their observed feedback. In this first scenario, we want to predict the feedback a user would give to an advert (e.g. 1-star through 5-stars). In such a case, we want to measure the accuracy of the system's predicted ratings. **Root Mean Squared Error (RMSE)** is perhaps the most popular metric used in evaluating the accuracy of predicted ratings. The system generates predicted ratings $\hat{r}_{u,a}$ for a test set T of user-advert pairs (u,a) for which the true ratings $r_{u,a}$ are known. The RMSE between the predicted and actual ratings is given by:

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{(u,a)\in T}(\hat{r}_{u,a} - r_{u,a})^2}. \quad (1)$$

Mean square error (MSE) is an alternative version of RMSE, the main difference between these two estimators is that RMSE penalizes more large errors, and MSE has the same units of measurement as the square of the quantity being estimated, while RMSE has the same units as the quantity being estimated. Therefore, MSE is given by

$$MSE = \frac{1}{|T|}\sum_{(u,a)\in T}(\hat{r}_{u,a} - r_{u,a})^2. \quad (2)$$

Mean Absolute Error (MAE) is a popular alternative, given by

$$MAE = \sqrt{\frac{1}{|T|}\sum_{(u,a)\in T}|\hat{r}_{u,a} - r_{u,a}|}. \quad (3)$$

As the name suggests, the MAE is an average of the absolute errors $err_{u,a} = |\hat{r}_{u,a} - r_{u,a}|$, where $\hat{r}_{u,a}$ is the prediction and $r_{u,a}$ the true value. The MAE is on same scale of data being measured.

3.2 Scenario 2: Ad Click Prediction

In many applications the recommendation system tries to recommend adverts to users in which they may be interested. For example, when items are added to the queue, Amazon suggests a set of adverts on which the user would most probably click. In this case, we are not interested in whether the system properly predicts the ratings of these adverts but rather whether the system properly predicts that the user will click on them (e.g. they perform a conversion). Therefore, we then have four possible outcomes for a recommended advertisement, as shown in Table 4.

	Recommended	Not recommended
Clicked	True-Positive (tp)	False-Negative (fn)
Not clicked	False-Positive (fp)	True-Negative (tn)

Table 4: Classification of the possible result of a recommendation of an advert to a user [22]

We can count the number of examples that fall into each cell in the table and compute the following quantities:

Precision $\frac{tp}{tp+fp}$,

Recall (True Positive Rate) $\frac{tp}{tp+fn}$.

Recall in this context is also referred to as the True Positive Rate (TPR) or *Sensitivity*, and precision is also referred to as positive predictive value (PPV).

Other related measures used include true negative rate and accuracy:

False Positive Rate (1 - Specificity) $\frac{fp}{fp+tn}$,

Accuracy $\frac{tp+tn}{tp+tn+fp+fn}$,

where true negative rate is also called *Specificity*. We can expect a trade-off between these quantities; while allowing longer recommendation lists typically improves recall, it is also likely to reduce the precision. We can compute curves comparing precision to recall, or true positive rate to false positive rate. Curves of the former type are known simply as precision-recall curves, while those of the latter type are known as a Receiver Operating Characteristic or ROC curves. A widely used measurement that summarizes the ROC curve is the **Area Under the ROC Curve (AUC)** [1] which is useful for comparing algorithms independently of application.

When evaluating precision-recall (or ROC curves) for multiple test users, a number of strategies can be employed in aggregating the results, depending on the application at hand. The usual manner in which precision-recall curves are computed in the information retrieval community [13, 27, 31, 32] is to average the resulting curves over users. Such a curve can be used to understand the trade-off between precision and recall (or false positives and false negatives) a typical user would face.

4. EXPERIMENTS AND RESULTS

In this section we show results obtained for the two types of scenarios introduced in Sec. 3. We conduct prediction experiments to explore the strengths and weakness of using personality traits as features for recommendation.

4.1 Evaluated Algorithms

Since a prediction engine lies at the basis of the most recommender systems, we selected some of the most widely used techniques for recommendations and predictions [14], such as Logistic Regression (LR) [8], Support Vector Regression with radial basis function (SVR-rbf) [3], and L2-regularized L2-loss Support Vector Regression (L2-SVR) [8]. These methods have often been based on a set of sparse binary features converted from the original categorical features via one-hot encoding [17, 26]. These engines may predict user opinions to adverts (e.g., a user's positive or negative feedback to an ad) or the probability that a user clicks or performs a conversion (e.g., an in-store purchase) when they see an ad. In Section 4, we evaluate these methods while feeding them with and without features coming from the psychometric traits.

4.2 Experimental Protocol

Let us say $X = \{\bar{x}_1, ..., \bar{x}_N\}$ is the set of observations, where the vectors $\bar{x}_i$ correspond to features coming only from the group "users' preferences" as described in Table 2 and $N = 120$ stands for the number of users involved in the experiment.

A feature is the user's selection from a pre-defined list of choices, hence, for each feature vector one element is 1 and the others are 0. Then, each column vector $\bar{x}_i$ is obtained by stacking the features on top of one another.

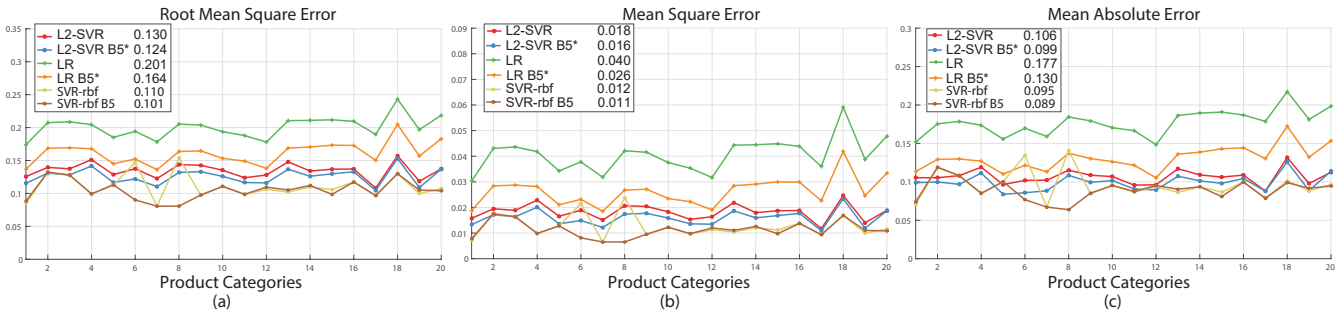


Figure 3: Measuring ratings prediction accuracy: B5 stands for Big-Five features. We indicate with an asterisk each method where B5 features, embedded into a baseline learner, shows a statistical significant effect over the baseline.

Regression is performed over the 20 product categories. The prediction problem is solved using LR, L2-SVR, and SVR-rbf, while feeding them with and without features coming from "personality traits". All experiments were performed using a k-fold approach (k = 10). In k-fold cross-validation, X is randomly partitioned into k's equal sized subsamples (the folds are the maintained the same for each algorithm in comparison). Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k - 1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used only once as the validation data. The k results from the folds can then be averaged to produce a single estimation.

Our experimental protocol includes feature selection, which represents an important pre-processing step given the sparse nature of the input data. It allows to remove many redundant features by reducing the dimensionality of the problem at hand. Hence, the representation above serves as a basis for the feature ranking and selection strategy. Ranking features allow us to detect a subset of cues which is not redundant. Accordingly, we use the training data obtained after the split as input of the infinite feature selection (Inf-FS) [30] algorithm. By construction, the Inf-FS is a graph-based method which exploits the convergence properties of the power series of matrices to evaluate the relevance of a feature with respect to all the other ones taken together. Indeed, in the Inf-FS formulation, each feature is mapped on an affinity graph, where nodes represent features, and weighted edges the relationships between them. In particular, the graph is weighted according to a function which takes into account both correlations and standard deviations between feature distributions. Each path of a certain length l over the graph is seen as a possible selection of features. Therefore, varying these paths and letting them tend to an infinite number permits the investigation of the importance of each possible subset of features.

Finally, the Inf-FS assigns a score to each feature of the initial set; where the score is related to how much the given feature is a good candidate regarding the regression task. Therefore, ranking the outcome of the Inf-FS in descendant order allows us to perform the subset feature selection throughout a *model selection stage*. In this way, we reduce the amount of features, by selecting 75% of the total. The selected features are: the number of favorite websites, T.V. programmes, sports, past times, the most watched movies and most visited websites, where we add the big-five personality traits.

4.3 Exp. 1: Ad Rating Prediction

In this section we report results for rating prediction showing that traces of user's personality can improve the prediction performance of the evaluated methods significantly. Statistical evaluation of experimental results has been considered an essential part of validation of machine learning methods. Given the user u_i, labels are assigned to each category by averaging the votes they gave to the category items such as $u_i = \{y_1, ..., y_{20}\}, y \in [1 - 5]$.

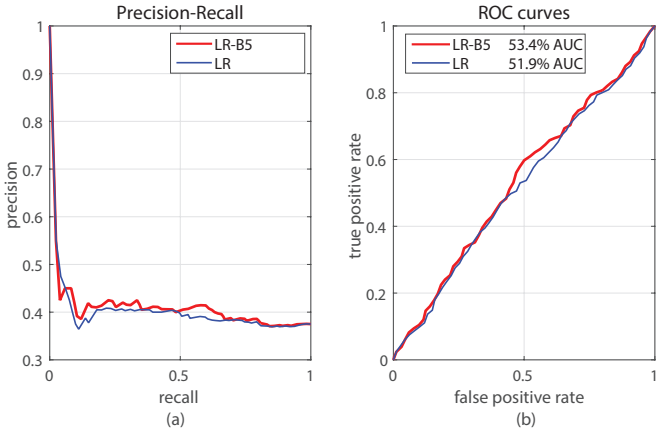


Figure 4: Comparison between LR and LR-B5: Curves show the proportion of preferred items that are actually recommended.

Figure 3 illustrates prediction results in term of RMSE, MSE and MAE plots across the categories. This first analysis shows how personality traits affect prediction performance. In order to assess if the difference in performance is statistically significant, t-tests have been used for comparing the accuracies. This statistical test is used to determine if the accuracies obtained with and without B5 are significantly different from each other (whereas both the distribution of values were normal). The test for assessing whether the data come from normal distributions with unknown, but equal, variances is the *Lilliefors* test.

Results show a statistical significant effect of personality traits while using L2-SVR (p-value < 0.05, Lilliefors Test H=0) and LR (p-value < 0.01, Lilliefors Test H=0).

As for the SVR-rbf, even if improvements in terms of prediction are not significant (B5 against no-B5), it is still interesting to notice the performance loss on categories 6 and 8, where errors go high significantly. In such a case, the B5 features do not seem to have much predictive power, however, they seem to play the role of a reliable stabilizer, but also that of an independent mediator and supporter of the regression process.

4.4 Exp. 2: Ad Click Prediction

This section shows an offline evaluation of click prediction. Along the lines of the previous experiment, a k-fold cross-validation is used. The experiment is performed at the category level, in order to

Method	ROC-AUC	Precision	Recall
L2-SVR	50.5%	39.2%	50.2%
L2-SVR B5	51.4%	39.9%	50.9%
LR	51.9%	40.3%	51.3%
LR-B5*	**53.4%**	**41.2%**	**52.1%**
SVR-rbf	48.3%	36.5%	48.8%
SVR-rb B5	50.1%	38.2%	50.2%

Table 5: Performance for ad click prediction. Big-Five features systematically contribute to the overall performance. The asterisk indicate that the method overcomes all the others.

work on a balanced distribution over the classes (1,229 clicked vs 1,171 not clicked instances), whenever a user showed their interest in a given category (i.e., the category contains at least one clicked advert) we labeled the category as "clicked" (rating greater or equal to four), otherwise "not clicked" (rating less than four). As a result, for each user we obtained a list of 20 labels representing their preference to each category. We computed precision-recall and ROC curves for each user, and then averaged the resulting curves over users. This is the usual manner in which precision-recall (or ROC) curves are computed in the information retrieval community [13, 31, 32]. Such a curve can be used to understand the trade-off between precision and recall and ROC a typical user would face.

Figure 4.(a) reports the precision-recall curves which emphasize the proportion of recommended items that are preferred and recommended. Figure 4.(b) shows the global ROC curves for LR and LR-B5, which emphasize the proportion of adverts that are not clicked but end up being recommended. The LR-B5 curve completely dominates the other curve, the decision about the superior setting for LR is easy.

The Area Under the ROC Curve is calculated as a measure of accuracy, which summarizes the precision recall of ROC curves, we report AUC, precision and recall in terms of the harmonic mean of precision and recall (F-measure) for all the methods in Table 5.

4.5 Discussions and Future Work

In this paper, we conducted a within-subject user study to investigate on the relations between users' personality related to their buying behavior and preferred item categories. A deeper analysis may involve the use of bi-clustering methods. Comparing to traditional clustering methods biclustering is not a blackbox technique. Comprehensibility is one of its main advantages, i.e. it is possible to understand why objects ended up in the same cluster.

It is worth noting that the goal of these experiments is to show how personality traits affect the prediction. In order to improve prediction accuracy, specific feature designing processes are needed so as to represent personality data and to standardize their definitions to be used as input recommender data towards to improve recommendations. In our experiments, we used a set of sparse binary features converted from the original categorical features. Moreover, many other algorithms may be used for this tasks, like the one proposed in [6, 15, 24].

For instance, the personality diagnosis [24] system is a collaborative filtering algorithm, which can be thought of as a hybrid between existing memory- and model-based algorithms. PD is fairly straightforward, maintains all data, and does not require a compilation step to incorporate new data. It is based on a simple and reasonable probabilistic model of how people rate titles.

Most of these recommender systems use to split each test user profile into sets of observed items and hidden items. The former is used as input for each recommender, the latter for performance evaluation. In our experiments, we did not use any information about the previous users' clicks, which turns out to be a more difficult task. We decided on this solution to move the focus of attention on personality data and not on other features like previous clicked ads.

5. CONCLUDING REMARKS

In this paper, we presented the ADS Dataset, a collection of 300 real advertisements rated by 120 unacquainted participants. We conducted a within-subject user study to investigate potential user issues of the personality on their buying behavior and preferred item categories.

The corpus has been collected with the main goal of studying the possible achievable benefits of employing personality traits in modern recommender systems. To obtain stronger and more relevant results for this community, appropriate and high-level features needed to be designed that carry important information for inference. In this paper, we only use raw data as sparse binary features converted from the original categorical features. We used standard techniques for recommending ads in order to show how personality traits affect the prediction, and, at the same time, set a baseline for future work.

We then reviewed a large set of properties, and explain how to evaluate systems given relevant properties. We discuss how to compare ARS based on a set of properties that are relevant for the application. Therefore, we review two main types of experiments in an offline setting, where recommendation approaches are compared with different selections of features (i.e., with and without personality traits) accordingly with our goal. We also discuss how to draw trustworthy conclusions from the conducted experiments.

Future work includes, but is not necessarily limited to, (1) feature engineering and designing for ARSs, represent personality data and standardize their definitions to be used as input recommender data towards to improve recommendations; (2) inference of personality traits and novel approaches for mapping pictures tagged as favorite into personality traits; and (3) identification of the underlying dimensions of consumer shopping motivations and personality factors.

We hope that this work motivates researchers to take into account the use of personality factors as an integral part of their future work, since there is a high potential that incorporating these kind of users' characteristics into ARS could enhance recommendation quality and user experience.

6. REFERENCES

[1] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 1975.

[2] M. Bosnjak, M. Galesic, and T. Tuten. Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach. *Journal of Business Research*, 2007.

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2011.

[4] J. V. Chen, B. chiuan Su, and A. E. Widjaja. Facebook c2c social commerce: A study of online impulse buying. *Decision Support Systems*, 2016.

[5] K. Choi, D. Yoo, G. Kim, and Y. Suh. A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. *Electronic Commerce Research and Applications*, 2012.

[6] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2003.

[7] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 213–222. ACM, 2013.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

[9] B. M. Fennis and A. T. Pruyn. You are what you wear: Brand personality influences on consumer impression formation. *Journal of Business Research*, 2007.

[10] Forrester. Online retail industry in the us will be worth $279 billion in 2015. *TechCrunch*, February 28.

[11] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[12] D. Funder. Personality. *Annual Reviews of Psychology*, 52:197–221, 2001.

[13] D. Harman. Overview of the trec 2002 novelty track. In *Text REtrieval Conference (TREC 2002)*, 2002.

[14] X. He. et al. practical lessons from predicting clicks on ads at facebook. In *Data Mining for Online Advertising*, New York, NY, USA, 2014. ACM.

[15] R. Hu and P. Pu. A Study on User Perception of Personality-Based Recommender Systems. In a. u. l. De, Bra, A. Kobsa, and D. Chin, editors, *User Modeling, Adaptation, and Personalization*, volume 6075 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010.

[16] C. Kim, K. Kwon, and W. Chang. How to measure the effectiveness of online advertising in online marketplaces. *Expert Syst. Appl.*, 2011.

[17] K.-c. Lee, B. Orten, A. Dasdan, and W. Li. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2012.

[18] S. K. Lee, Y. H. Cho, and S. H. Kim. Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. *Information Sciences*, 2010.

[19] J. C. Mowen. *The 3M Model of Motivation and Personality: Theory and Empirical Applications to Consumer Behavior*. Springer US, Boston, MA, 2000.

[20] N. nez Valdéz. et al. implicit feedback techniques on recommender systems applied to electronic books. *Comput. Hum. Behav.*, 2012.

[21] A. Odic, M. Tkalčič, A. Košir, and J. F. Tasič. A.: Relevant context in a movie recommender system: UsersâĂŹ opinion vs. statistical detection. In *In: Proc. of the 4th Workshop on Context-Aware Recommender Systems (2011*.

[22] D. L. Olson and D. Delen. *Advanced Data Mining Techniques*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[23] A. C. P. and V. M. S. Click through rate prediction for display advertisement. *International Journal of Computer Applications*, 2016.

[24] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 473–480, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[25] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 2007.

[26] M. Richardson. Predicting clicks: Estimating the click-through rate for new ads. In *International World Wide Web Conference*. ACM Press, 2007.

[27] G. Roffo, M. Cristani, L. Bazzani, H. Q. Minh, and V. Murino. Trusting skype: Learning the way people chat for fast user recognition and verification. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 748–754, Dec 2013.

[28] G. Roffo, C. Giorgetta, R. Ferrario, and M. Cristani. *Just the Way You Chat: Linking Personality, Style and Recognizability in Chats*, pages 30–41. Springer International Publishing, 2014.

[29] G. Roffo, C. Giorgetta, R. Ferrario, W. Riviera, and M. Cristani. Statistical analysis of personality and identity in chats using a keylogging platform. In *International Conference on Multimodal Interaction*. ACM, 2014.

[30] G. Roffo, S. Melzi, and M. Cristani. Infinite feature selection. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *ACM Conference on Electronic Commerce*. ACM, 2000.

[32] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

[33] M. Tkalčič, U. Burnik, and A. Košir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, 2010.

[34] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic. Affective labeling in a content-based recommender system for images. *IEEE Transactions on Multimedia*, 15(2):391–400, Feb 2013.

[35] C. A. Turkyilmaz, S. Erdem, and A. Uslu. The effects of personality traits and website quality on online impulse buying. 2015. International Conference on Strategic Innovative Marketing.

[36] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.

[37] X. Wang, W. Li, Y. Cui, R. Zhang, and J. Mao. Click-through rate estimation for rare events in online advertising. *Online Multimedia Advertising: Techniques and Technologies*, 2010.

Emotion Elicitation in Socially Intelligent Services: the Intelligent Typing Tutor Study Case

Andrej Košir
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25
Ljubljana, Slovenia
andrej.kosir@fe.uni-lj.si

Marko Meža
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25
Ljubljana, Slovenia
marko.meza@fe.uni-lj.si

Janja Košir
University of Ljubljana, Faculty of Education
Krdeljeva ploščad
Ljubljana, Slovenia
janja.kosir@pef.uni-lj.si

Matija Svetina
University of Ljubljana, Faculty of Fine Arts
Tržaška cesta 2
Ljubljana, Slovenia
matija.svetina@ff.uni-lj.si

Gregor Strle
Scientific Research Centre SAZU
Novi trg 2
Ljubljana, Slovenia
gregor.strle@zrc-sazu.si

ABSTRACT

The paper discusses the challenges of user emotion elicitation in socially intelligent services, based on the experimental design and results of the intelligent typing tutor. Human-machine communication (HMC) of the typing tutor is supported by the continuous real-time emotion elicitation of user's expressed emotions and the emotional feedback of the service, through the graphically rendered emoticons. It is argued that emotion elicitation is an important part of successful HMC, as it improves the communication loop and increases user engagement. Experimental results show that user's valence and arousal are elicited during the typing practice, on average 18% to 25% of the time for valence and 20% to 31% of the time for arousal. However, the efficiency of emotion elicitation varies greatly throughout the use of the service, and also moderately among users. Overall, the results show that emotion elicitation, even via simple graphical emoticons, has significant potential in socially intelligent services.

Keywords

affective computing, emotion elicitation, social intelligence, human-machine communication, intelligent tutoring systems

1. INTRODUCTION

Bridging the gap between modern digital services and the increasing demands and (often insufficient) capabilities of a wide range of users is a challenging task. In recent years, much focus has been given to user adaptation procedures in socially intelligent services, including user modeling, recommender systems, human-machine communication (HMC), among many others [10], [1], [11]. While there have been substantial advances in many of these areas, the state-of-the art technology still lacks satisfactory means to efficiently meet various user needs and/or tailor to their capabilities. As the potential for new users of technology supported services is growing (e.g. groups of elderly users), so is the digital divide [42]. This gap may manifest itself in many forms. It may deprive a particular user group of efficient use of a service (e.g., due to the lack of technological proficiency), it may be limited in scope and only partially attend to users needs (e.g., the use of multiple services for a series of common, integrated tasks), or for some user groups offer no accessibility to a service altogether (e.g., e-banking for the elderly users). In general, it results in frustration and increased cognitive load, requiring significant effort to use a service (e.g. interaction, navigation, finding information, etc.), instead of a service adapting to user needs and capabilities.

One way to address these issues is to establish and sustain efficient (close-to-human) communication level between a user and a service, with HMC at the core of contextualization and adaptation procedures. Whereas natural (human-to-human) communication is innate and in general requires minimal effort for the actors involved to sustain it, HMC is void of both innateness and context, as well as of non-verbal (auditory, visual, olfactory) cues. Thus, for a modern digital service to be successful, it should be capable of expressing minimal social intelligence [45]. Another important and inherent property of natural communication is its continuity in real-time. HMC should be able to exhibit some level of social intelligence by generating and processing social signals in near-real-time.[1] To sustain the feedback loop the user should be at least minimally engaged, with non-verbal (social) signals (such as emotions) elicited at a continuous (minimal delay) rate. Ideally, effective HMC should minimize the user-service adaptation procedures and maximize the engagement and the intended use of a service. In other words, a service is socially intelligent when it is ca-

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

[1]The maximal tolerated delay is about 0.5 seconds.

pable of reading (measuring and estimating) user's social signals (verbal and/or non-verbal communication signals), producing machine generated feedback on these signals, and sustaining and adapting according such HMC.

In general, we believe it is possible to alleviate some of the main obstacles towards more effective user-service adaptation procedures by addressing the following:

- *Non-intrusive user data acquisition.* Some types of user data (e.g., user's emotion state) should be tracked in near-real-time. The problem is users do not like obtrusive data gathering methods (e.g., to repeatedly fill in questionnaires or use wearable sensors in everyday situations). The state-of-the-art techniques for non-intrusive user data acquisition are limited and can not provide sufficient high quality user data for the efficient user-service adaptation procedures;
- *Contextualization.* Contextualization refers to the definition of circumstances relevant for specific user-service adaptation. Effective user adaptation is highly context-sensitive as user involvement, attention and motivation, as well as preferences, are to a large extent context dependent. The emergent technologies of Internet of things (IoT), wearable computing, ubiquitous computing, and others, offer various building blocks to model specific contextualization tasks, however, user interaction data is typically not taken into account;
- *Service functionality and content adaptation for the user.* Ideally, user adaptation procedure is successful when the service is able to adapt to (and improve upon) the user needs and preferences in near real-time. As a result, the adaptation mechanisms of the service need to go beyond generally applicable adaptation procedures to address the specific task-dependent and user-interaction scenarios.

The aim of the paper is to analyze the efficiency of emotion elicitation in a socially intelligent service. The underlying assumption is that emotion elicitation should be an integral part of HMC, as it can greatly improve user-service adaptation procedure. For this purpose, the experiment was conducted using the socially intelligent typing tutor. The tutor is a web-based learning service designed to elicit emotions and thus improve learner's attention and overall engagement in the touch-typing training. Emotion elicitation is utilized together with the notion of positive reinforcement, where the learner is being rewarded for her efforts through the emotional feedback of the service. Moreover, the tutor is able to model and analyze learner's expressed emotions and measure the efficiency of emotion elicitation in the tutoring process.

The paper is structured as follows. Section 2 presents related work, while Section 3 discusses general aspects of emotion elicitation in socially intelligent services and then presents the socially intelligent typing tutor. Section 4 presents the experimental results on emotion elicitation in the intelligent typing tutor. The paper ends with a general conclusion and future work.

2. RELATED WORK

The research and development of a fully functioning socially intelligent service is still at a very early stage. However, various components that will ultimately enable such services are under intensive development for several decades. We briefly present them grouped according to the following subsections.

2.1 Social intelligence, social signals and non-verbal communication cues

There are many definitions of social intelligence applicable in this context [23]. The wider definition used here is by Vernon [44], who defines social intelligence as the person's "ability to get along with people in general, social technique or ease in society, knowledge of social matters, susceptibility to stimuli from other members of a group, as well as insight into the temporary moods or underlying personality traits of strangers". Furthermore, social intelligence is demonstrated as the ability to express and recognize social cues and behaviors [2], [6], including various non-verbal cues (such as gestures, postures and face expressions) exchanged during social interaction [47].

Social signals are extensively being analyzed in the field of human to computer interaction [47], [46], often under different terminology. For example, [33] use the term 'social signals' to define a continuously available information required to estimate emotions, mood, personality, and other traits that are used in human communication. Others [31] define such information as 'honest signals' as they allow to accurately predict the non-verbal cues and, on the other hand, one is not able to control the non-verbal cues to the extent one can control the verbal form. Here, we will use the term social signal.

2.2 Socially intelligent learning services

Several services exist that support some level of social intelligence, ranging from emotion-aware to meta-cognitive. One of the more relevant examples is the intelligent tutoring system AutoTutor/Affective AutoTutor [15]. AutoTutor/Affective AutoTutor employs both affective and cognitive modelling to support learning and engagement, tailored to the individual user [15]. Some other examples include: Cognitive Tutor [7] – an instruction based system for mathematics and computer science, Help Tutor [3] – a meta-cognitive variation of AutoTutor that aims to develop better general help-seeking strategies for students, MetaTutor [9] – which aims to model the complex nature of self-regulated learning, and various constraint-based intelligent tutoring systems that model instructional domains at an abstract level [28], among many others. Studies on affective learning indicate the superiority of emotion-aware over non-emotion-aware services, with the former offering significant performance increase in learning [37], [22], [43].

2.3 Computational models of emotion

One of the core requirements for socially intelligent service is the ability to detect and recognize emotions, and exhibit the capacity for expressing and eliciting basic affective (emotional) states. Most of the literature in this area is dedicated to the affective computing and computational models of emotion [26], [25], [34], which are mainly based on the appraisal theory of emotions [48]. Several challenges remain, most notably the design, training and evaluation of computational models of emotion [20], their critical analysis and comparison, and their relevancy for other research fields (e.g., cognitive science, human emotion psychology), as most computational models of emotion are overly simplistic [12].

2.4 Physiological sensors

The development of wearable sensors enabled the acquisition of user data in near-real-time, as well as the research and estimation of user's internal states (such as emotion and stress level estimation) that started more than a decade ago [5], [4]. Notable advances can also be found in the fields of psychological computing and HCI, with the development of several novel measurement related procedures and techniques. For example, psychophysiological measurements are being employed to extend the communication bandwidth and develop smart technologies [18], along with the design guidelines for conversational intelligence based on the environmental sensors [14]. Several studies deal with human stress estimation [36], workload estimation [30], cognitive load estimation [27], [8], among others, and specific learning tasks related to physiological measurements [49], [21].

2.5 Human emotion elicitation

The field of affective computing has developed several approaches to modeling, analysis and interpretation of human emotions [19]. The most known and widely used emotion annotation and representation model is the Valence-Arousal-Dominance (VAD) emotion space, an extension of Russell's valence-arousal model of affect [35]. The VAD space is used in many human to machine interaction settings [50], [40], [32], and was also adopted in the socially intelligent typing tutor (see section 3.3.2). There are other attempts to define models of human emotions, such as specific emotion spaces for human computer interaction [16], or more recently, models for the automatic and continuous analysis of human emotional behaviour [19]. Recent research on emotion perception argues that traditional emotion models might be overly simplistic, pointing out the notion of emotion is multi-componential, and includes "appraisals, psychophysiological activation, action tendencies, and motor expressions" [38]. Consequently, and relevant to the interpretations of valence in the existing models, some researchers argue there is a need for the "multifaceted conceptualization of valence" that can be linked to "qualitatively different types of evaluations" used in the appraisal theories [39].

Research of emotion elicitation via graphical user interface is far less common. Whereas several studies on emotion elicitation use different stimuli (e.g., pictures, movies, music) [41] and behavior cues [13], none to our knowledge tackle the challenges of graphical user interface design for the purpose of emotion elicitation.

In the intelligent typing tutor, user emotions are elicited by the graphical emoticons (smileys) via the dynamic graphical user interface of the service. The choice of emoticons was due to their semantic simplicity, unobtrusiveness, and ease of continuous measurement – using pictures as a stimuli would add additional cognitive load and likely evoke multiple emotions. This approach also builds upon the results of previous research, which showed that human face-like graphics increase user engagement, that the recognition of emotions represented by emoticons is intuitive for humans, and that emotion elicitation based on emoticons is strong enough to be applicable [17]. The latter assumption is verified in this paper.

3. EMOTION ELICITATION IN SOCIALLY INTELLIGENT SERVICES: THE TYPING TUTOR STUDY CASE

The following sections discuss the role of emotion elicitation in socially intelligent services and its importance for efficient HMC. General requirements and the role of emotion elicitation are discussed in the context of our study case – the intelligent typing tutor. Later sections present the design of the intelligent typing tutor and its emotion elicitation model.

3.1 General requirements for a socially intelligent service

A given service is socially intelligent if it is capable of performing the following elements of social intelligence:

1. Read relevant user behavior cues: human emotions are conveyed via behaviour and non-verbal communication cues such as face expression, gestures, body posture, color of the voice, etc.

2. Analyze, estimate and model user emotions and non-verbal (social) communication cues via computational model: behavior cues are used to estimate user's temporary emotion state. Selected physiological measurements (pupil size, acceleration of the wrist, etc.) are believed to be correlated with user's emotion state and other non-verbal communication cues. These are used as an input to the computational model of user emotions and other non-verbal communication cues.

3. Integrate and model machine generated emotion expressions and other non-verbal communication cues: for example, the notion of positive reinforcement could be integrated into a service to improve user engagement, taking into account user's temporary emotion state and other non-verbal communication cues.

4. Generate emotion elicitation to improve user engagement: continuous feedback loop between user emotion state and machine generated emotion expressions for purpose of emotion elicitation.

5. Context and task-dependent adaptation: adapt the service according to the design goals. For example, in the intelligent typing tutor case study, the intended goal is to improve learner's engagement and progress. The touch-typing lessons are carefully designed and adapt in terms of typing speed and difficulty to meet individual's capabilities, temporary emotion state and other non-verbal communication cues.

Such service is capable of sustaining efficient, continuous and engaging HMC. It also minimizes user-service adaptation procedures. An early-stage example of socially intelligent service is provided below.

3.2 Typing tutor as a socially intelligent service

The overall goal of the socially intelligent typing tutor is to improve the process of learning touch-typing. For this purpose, emotion elicitation is integrated into HMC together with the notion of positive reinforcement, to amplify the attention, motivation, and engagement of the individual

learner. In its current form, the rudimentary model of emotion elicitation utilizes emoticon-like graphics via the graphical user interface of the service, presented to the learner in real-time (see section 3.3). The tutor uses state-of-the-art technology (3.2.1) and is able to model, measure and analyze emotion elicitation throughout the tutoring process.

3.2.1 Architecture and design

Typing tutor's main building blocks consist of:

1. Web GUI: to support typing lessons and machine generated emotion expressions via emoticons (see Fig. 1);
2. Sensors: to conduct physiological measurements and monitor user status (wrist accelerometer, camera, emotion-recognition software to estimate user emotions, eye gaze, pupil size, etc.);
3. Computational model: for measuring user emotions and attention in the tutoring process;
4. Recommender system: for modelling machine generated emotion expressions;
5. Typing content generator: which follows typing lectures designed by the expert.

Real-time sensors are integrated into the service to gather physiological data about the learner. The recorded data is later used to establish the weak ground truth of learner's attention and the efficiency of emotion elicitation. Both are further estimated through the human annotation procedure, based on the carefully designed operational definition and verified using psychometric characteristics. The list of sensors integrated in the tutor includes:

- Keyboard: to monitor cognitive and locomotor errors that occur while typing;
- Video recorder: to extract learner's facial emotion expressions in real-time;
- Wrist accelerometer and gyroscope: to trace the hand movement;
- Eye tracking: to measure pupil size and estimate learner's attention and possible correlates to typing performance.

The intelligent typing tutor is publicly available as a client-server service running in a web browser (http://nacomnet.lucami.org/test/desetprstno_tipkanje). Data is stored on the server for later analyses and human annotation procedures. Such architecture allows for crowd-sourced testing and efficient remote maintenance.

3.3 Emotion elicitation in the intelligent typing tutor

The role of emotion elicitation in the intelligent typing tutor is that of efficient HMC and reward system. The positive reinforcement assumption [29] is used in the design of the emotion elicitation model. Positive reinforcement argues that learning is best motivated by a positive emotional responses from the service when learners ratio of attention over fatigue goes up, and vice versa. Here, machine generated positive emotion expressions act as rewards, with the aim to improve learner's attention, motivation and engagement during the touch-typing practice. The learner is rewarded by a positive emotional response from the service when she invest more effort into practice (the service does not support negative reinforcement). According to the positive reinforcement assumption, the rewarded behaviors will appear more frequently in the future. Negative reinforcement is not used for two reasons: there is no clear indication how negative reinforcement would contribute to the learning experience, and it would require an introduction of additional dimension, making the research topic of the experiment even more complex.

3.3.1 Machine emotion model

The intelligent typing tutor uses emotion elicitation to reward any behavior leading to the improvement of learner's engagement with the service. The rewards come as positive emotional responses conveyed by the emoticon via graphical user interface. The machine generated emotion responses range from neutral to positive (smiley) and act as stimuli for user (learner) emotion elicitation. For this purpose, a subset of emoticons from Official Unicode Consortium code chart (see http://www.unicode.org/) was selected and emoticon-like graphical elements were integrated into the newly designed user interface of the service shown in Fig. 1.

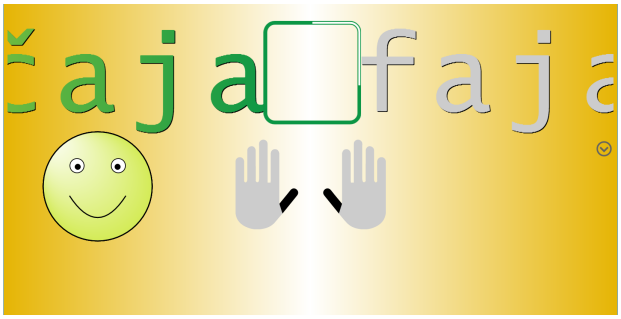


Figure 1: Socially intelligent typing tutor integrates touch-typing tutoring and machine generated emoticons (for emotion elicitation) via its graphical user interface.

Emotional responses are computed according to the learning goals of the tutor. To improve learner's attention and overall engagement in the touch-typing practice, the emotional feedback of the service needs to function in real-time. As mentioned above, the positive reinforcement assumption acts as the core underlying mechanism for modelling machine generated emotions. At the same time such mechanism is suitable for dynamic personalization, similar to the conversational RecSys [24]. In order to implement it successfully, the designer needs to decide on 1. which behaviors need to be reinforced to appear more frequently, and 2. which rewards, relevant for the learner, need reinforcement.

3.3.2 User emotion model

User (learner) emotions are elicited via tutor's graphical user interface, based on the machine generated emotion expressions from (3.3.1). The VAD emotion model is used for representation and measurement of learner elicited emotions, similar to [16]. The VAD dimensions are then measured in real-time by emotion recognition software (see section 4.1).[2]

[2]Here, we only discuss valence Φ_{uV} and arousal Φ_{uA}, the

Two independent linear regression models are used to model user emotion elicitation as a response to the machine generated emoticons. The models are fitted as follows: the measured values of user emotion elicitation for valence and arousal are fitted as dependent variables, whereas the machine generated emotion expression is fitted as an independent variable (Eq.1). The aim is to obtain the models' quality of fit and the proportion of the explained variance in emotion elicitation.

$$\Phi_{uV} = \beta_{1V}\Phi_m + \beta_{0V} + \varepsilon_V, \quad \Phi_{uA} = \beta_{1A}\Phi_m + \beta_{0A} + \varepsilon_A, \quad (1)$$

where Φ_m stands for one dimensional parametrization of the machine emoticon graphics, ranging from 0 (neutral emoticon) to 1 (maximal positive emotion expression). Notations β_{1V} and β_{1A} are user emotion elicitation linear model coefficients, β_{0V} and β_{0A} are the averaged effects of other influences on user emotion elicitation, and ε_V and ε_A are independent variables of white noise.

The linear regression model was selected due to the good statistical power of its goodness of fit estimation R^2. There is no indication that emotion elicitation is linear, but we nevertheless believe the choice of the linear model is justified. The linear model is able to capture the emotion elicitation process, detect emotion elicitation, and provide valid results (see section 4.2). Residual plots (not reported here) show that linear regression assumptions (homoscedasticity, normality of residuals) are not violated.

To further support our argument for emotion elicitation in the intelligent typing tutor, we statistically tested our hypothesis that a significant part of learner's emotions is indeed elicited by the machine generated emoticons. We did this with the null hypothesis testing $H_0 = [R^2 = 0]$ (see section 4.2), which demonstrated good power compared to the statistical tests by some of the known non-linear models.

4. USER EXPERIMENT: THE ESTIMATION OF USER EMOTION ELICITATION

The following sections give an overview of the user experiment and results on emotion elicitation in the intelligent typing tutor.

4.1 User experiment

The experiment consisted of 32 subjects invited to practice touch-typing in the intelligent typing tutor (see 3.2), with the average duration of the typing session approx. 17 minutes (1020 seconds). The same set of carefully designed touch-typing lessons was given to all test subjects. User data was acquired in real-time using sensors (as described in section 3.2), and used as an input to the computational model of machine generated emotion expressions, and recorded for later analysis. For the preliminary analysis presented here, five randomly selected subjects were analysed on the segment of the overall duration of the experiment.[3] The test segment spans from 6 to 11.5 mins (330 seconds) of the experiment.

The test segment used for the analysis is composed of the following steps:[4]

1. Instructions are given to the test users: users are personally informed about the goal and the procedure of the experiment (by the experiment personnel);

2. Setting up sensory equipment, start of the experiment: a wrist accelerometer is put on, the video camera is set on, and the experimental session time recording is started (at 00 seconds);

3. At 60 seconds: machine generated sound disruption of the primary task: "Name the first and the last letter of the word: mouse, letter, backpack, clock";

4. At 240 seconds: machine generated sound disruption of the primary task, "Name the color of the smallest circle", in the figure (Fig 2). This cognitive task is expected to significantly disrupt learner's attention away from the typing exercise;

5. The test segment ends at 330 seconds.

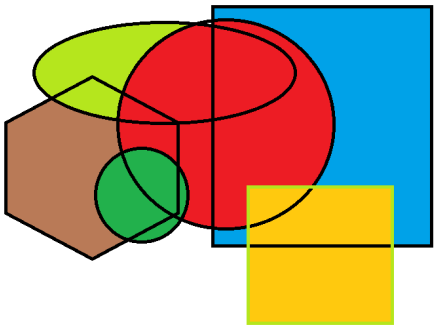


Figure 2: Graphics shown during the second disruption (Step 4) at 240 seconds of the test segment

During the experiment, users' emotion expressions are analyzed using Noldus Observer video analysis software http://www.noldus.com. The recordings are in sync with the machine generated emoticons, readily available for analysis (see next section 4.2).

4.2 Experimental results

The analysis of the experimental data was conducted to measure the effectiveness of emotion elicitation. The x-axis times for all graphs presented below are relative in seconds [s], for the whole duration of the test segment (330 seconds). The estimation is based on the emotion elicitation model (1) fitting. To detect the time when the emotion elicitation is present, we conducted the null hypothesis testing $H_0 = [R^2 = 0]$ at risk level $\alpha = 0.05$. The emotion elicitation is determined as present where the null hypotheses is rejected, and not present otherwise.

An example of valence and arousal ratings for a randomly selected subject is shown in Fig. 3.

The model (1) is fitted using linear regression on the measured data for the duration of the test segment. The data is

two primary dimensions for measuring emotion elicitation.

[3]To simplify the presentation of the experiment results. Note that similar results were found for the remaining subjects.

[4]Due to limited space, the two disruption parts of the experiment (Steps 3. and 4.) are not further discussed.

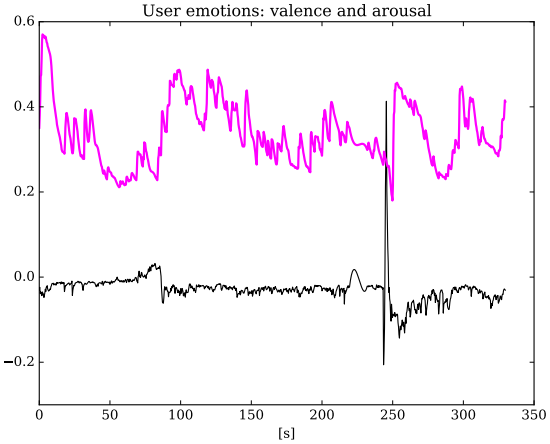


Figure 3: Valence (black line) and arousal (magenta, light line) ratings of learner's emotional state throughout the test segment.

sampled in a non-uniform manner due to the technical properties of the sensors (internal clocks of sensors are not sufficiently accurate, etc.). The data is approximated by continuous smooth B-splines of order 3, according to the upper frequency limit of measured phenomena, and uniformly sampled to time-align data (we skip re-sampling details here).

To fit the regression models the 40 past samples from the current (evaluation) time representing 4 seconds of real-time were used. These two value were selected as an optimum according to competitive arguments for more statistical power (requires more samples) and for enabling to detect time-dynamic changes in the effectiveness of emotion elicitation (requiring shorter time interval leading to less samples). Note that changing this interval from 3 to 5 seconds did not significantly affect the fitting results. Results are given in terms of R_V^2, R_A^2 representing the part of explained variance of valence and arousal when the elicitation is known, and in terms of a p_V, p_A-values testing the null hypothesis regression models $H_{0V} = [R_V^2 = 0]$, $H_{0A} = [R_A^2 = 0]$, respectively. The time dynamics of emotion elicitation is represented by p-values p_A and p_V on Fig. 4.

In order to estimate the effect of emotion elicitation, the percentages were computed on the number of times the elicitation was significant. The analyzed time intervals were uniformly sampled every 2 seconds. The results are shown in Table 1. It turned out that the test interval sampling had no significant impact on the results.

Table 1: Proportion q of the time when the measured emotion elicitation is significant. Notation red. q stands for the reduced efficiency, which is 5% lower than the measured one. Measured for the five selected test subjects.

	Valence		Arousal	
User Id	q %	red. q %	q %	red. q %
1	47.7	45.3	43.2	41.1
2	68.3	65.0	72.2	68.6
3	60.0	57.0	61.3	58.2
4	51.6	49.1	60.6	57.6
5	62.3	59.4	61.9	58.8

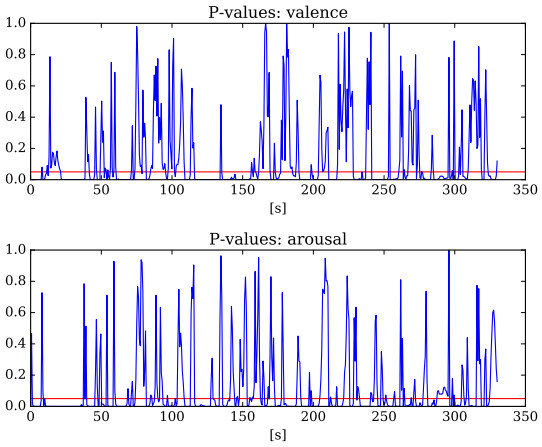


Figure 4: P-values for the null hypothesis testing $H_0 = [R^2 = 0]$ of emotion elicitation for a randomly selected subject, separately for valence (top) and arousal (bottom). The horizontal red line marks the risk level $\alpha = 0.05$, with p-values below the line indicating significant emotion elicitation effect.

We also analyzed the reduced percentages. These are 5% lower than the measured ones, since the significance testing was performed at a risk level $\alpha = 0.05$ and approximately 5% detections are false (type I. errors). Note that Bonferroni correction does not apply here. However, we nevertheless computed the above given percentages using Bonferroni correction and it turned out the percentages drop approximately to one half of the reported values.

The strength of emotion elicitation is shown in the linear regression model R^2 as a function of time (Fig. 5).

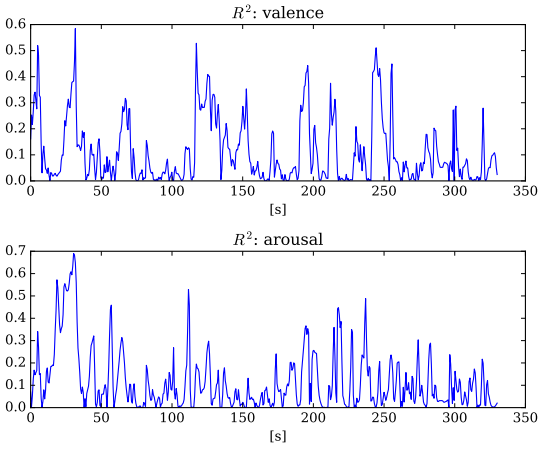


Figure 5: Linear regression model R^2 of emotion elicitation for a randomly selected test subject, separately for valence (top) and arousal (bottom).

The strength of emotion elicitation effect is significant, but also varies highly (Fig. 5). Similar results were detected among all test subjects. However, it is too early to draw any meaningful conclusions on the reasons for high variability at this stage, as many of the potential factors influencing emotion elicitation need further analysis.

To estimate the average strength of emotion elicitation,

the average values of R^2 were computed for the five selected subjects (as in Table 1) – these values are part of the explained variance for learner emotions when the machine generated emotion is known. The average value of R^2 varies across test subjects from 18.3% to 24.5% for valence and 19.7% to 31.4% for arousal, for all time intervals (when significant or non-significant elicitation is present). If we average only over the time intervals when the elicitation is significant, the average value of R^2 varies across test subjects from 32.5% to 39.3% for valence and 36.3% to 44.9% for arousal (see Table 2).

Table 2: Average values for the explained variance for valence and arousal in %: for all time intervals and for the time intervals when emotion elicitation is significant. Measured for the five selected test subjects.

	Valence		Arousal	
User Id	All int.	Signif. int.	All int.	Signif. int.
1	18.3	32.5	19.7	36.3
2	19.4	33.8	27.4	39.2
3	24.5	39.3	31.4	44.9
4	19.8	33.3	23.9	39.9
5	21.7	35.4	26.8	40.2

Observe that there is considerably less variability among the subjects in terms of elicitation strength (average R^2), compared to the proportions of time the elicitation is significant (see Table 1).

5. CONCLUSION AND FUTURE WORK

The paper discussed the efficiency of emotion elicitation in socially intelligent services. The experiment was conducted using the socially intelligent typing tutor. The overall aim of the intelligent typing tutor is to elicit emotions and thus improve learning and engagement in the touch-typing training. Emotion elicitation is utilized together with the notion of positive reinforcement. The tutor is able to model and analyze learner's expressed emotions and measure the efficiency of emotion elicitation in the process. Experimental results show that the efficiency of emotion elicitation is significant, but at times also varies highly for the individual learner and moderately among learners.

Future work will focus on reasons for variations in emotion elicitation by analyzing potential factors, such as the effects of machine generated emotion expressions on emotion elicitation, learner's emotional state, cognitive load, attention, and engagement, among others.

6. REFERENCES

[1] E. B. Ahmed, A. Nabli, and F. Gargouri. A Survey of User-Centric Data Warehouses: From Personalization to Recommendation. *International Journal of Database Management Systems*, 3(2):59–71, 2011.

[2] K. Albrecht. *Social Intelligence: The New Science of Success*. Pfeiffer, 1 edition, February 2009.

[3] V. Aleven, B. Mclaren, I. Roll, and K. Koedinger. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16(2):101–128, 2006.

[4] J. Allanson and S. H. Fairclough. A research agenda for physiological computing. *Interacting with Computers*, 16(5):857–878, 2004.

[5] J. Allanson and G. Wilson. Physiological Computing. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, pages 21–42, 2002.

[6] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111:256–274, 1992.

[7] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2):167–207, 1995.

[8] Y. Ayzenberg, J. Hernandez, and R. Picard. FEEL: frequent EDA and event logging – a mobile social interaction stress monitoring system. *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts CHI EA 12*, page 2357, 2012.

[9] R. Azevedo, A. Witherspoon, A. Chauncey, C. Burkett, and A. Fike. MetaTutor: A MetaCognitive Tool for Enhancing Self-Regulated Learning. *Annual Meeting of the American Association for Artificial Intelligence Symposium on Metacognitive and Cognitive Educational Systems*, pages 14–19, 2009.

[10] P. Biswas and P. Robinson. A brief survey on user modelling in HCI. *Intelligent Techniques for Speech Image and Language Processing SpringerVerlag*, 2010.

[11] J. Bobadilla, F. Ortega, a. Hernando, and a. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[12] J. Broekens, T. Bosse, and S. C. Marsella. Challenges in computational modeling of affective processes. *IEEE Transactions on Affective Computing*, 4(3):242–245, 2013.

[13] J. A. Coan and J. J. Allen, editors. *Handbook of Emotion Elicitation and Assessment (Series in Affective Science)*. Oxford University Press, 1 edition, 4 2007.

[14] D. C. Derrick, J. L. Jenkins, and J. Jay F. Nunamaker. Design Principles for Special Purpose, Embodied, Conversational Intelligence with Environmental Sensors (SPECIES) Agents. *AIS Transactions on Human-Computer Interaction*, 3(2):62–81, 2011.

[15] S. D'mello and A. Graesser. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst.*, 2(4):23:1–23:39, Jan. 2013.

[16] D. C. Dryer. Dominance and valence: a two-factor model for emotion in HCI. In *Proceedings of AAAI Fall Symposium Series on Affect and Cognition*, pages 111 – 117. AAAI Press, 1998.

[17] J. Dunlap, D. Bose, P. R. Lowenthal, C. S. York, M. Atkinson, and J. Murtagh. What sunshine is to flowers: A literature review on the use of emoticons to support online learning. *Emotions, Design, Learning and Technology*, pages 1–17, 2015.

[18] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, 2009.

[19] H. Gunes, B. Schuller, M. Pantic, and R. Cowie.

Emotion representation, analysis and synthesis in continuous space: A survey. *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pages 827–834, 2011.

[20] E. Hudlicka. Guidelines for Designing Computational Models of Emotions. *International Journal of Synthetic Emotions*, 2(1):26–79, 2011.

[21] X. Jiang, B. Zheng, R. Bednarik, and M. S. Atkins. Pupil responses to continuous aiming movements. *International Journal of Human-Computer Studies*, 83:1–11, 2015.

[22] A. C. K. Koedinger. *The Cambridge Handbook of the Learning Sciences*, chapter Cognitive tutors: Technology bringing learning sciences to the classroom, pages 61–78. Cambridge University Press, New York, 2006.

[23] J. F. Kihlstrom; and N. Cantor. Social Intelligence. In R. Sternberg, editor, *Handbook of intelligence*, pages 359–379. Cambridge, U.K.: Cambridge University, 2nd edition, 2000.

[24] T. Mahmood, G. Mujtaba, and A. Venturini. Dynamic personalization in conversational recommender systems. *Information Systems and e-Business Management*, pages 1–26, 2013.

[25] S. Marsella and J. Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2014.

[26] S. Marsella, J. Gratch, and P. Petta. Computational models of emotion. In *Blueprint for Affective Computing (Series in Affective Science)*. Oxford University Press, 2010.

[27] D. McDuff, S. Gontarek, and R. Picard. Remote Measurement of Cognitive Stress via Heart Rate Variability. *Proceedings of 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2957–2960, 2014.

[28] A. Mitrovic, B. Martin, and P. Suraweera. Intelligent tutors for all: The constraint-based approach. *IEEE Intelligent Systems*, 22(4):38–45, 2007.

[29] A. Neuringer. Operant variability: Evidence, functions, and theory. *Psychonomic Bulletin & Review*, 9(4):672–705, 2002.

[30] D. Novak, B. Beyeler, X. Omlin, and R. Riener. Workload estimation in physical human–robot interaction using physiological measurements. *Interacting with Computers*, page iwu021, 2014.

[31] A. Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, 2008.

[32] J. L. Plass, S. Heidig, E. O. Hayward, B. D. Homer, and E. Um. Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction*, 29:128–140, 2014.

[33] V. P. Richmond, J. C. McCroskey, and M. L. Hickson III. *Nonverbal Behavior in Interpersonal Relations (7th Edition)*. Pearson, 7 edition, 4 2011.

[34] L.-F. Rodriguez and F. Ramos. Development of Computational Models of Emotions for Autonomous Agents : A Review. *Cognitive Computing*, pages 351–375, 2014.

[35] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[36] V. Sandulescu, S. Andrews, E. David, N. Bellotto, and O. M. Mozos. Stress Detection Using Wearable Physiological Sensors. *Artificial Computation in Biology and Medicine Lecture Notes in Computer Science*, pages 526–532, 2015.

[37] L. Shen, M. Wang, and R. Shen. Affective e-Learning: Using emotional data to improve learning in pervasive learning environment related work and the pervasive e-learning platform. *Educational Technology & Society*, 12:176–189, 2009.

[38] V. Shuman, E. Clark-Polner, B. Meuleman, D. Sander, and K. R. Scherer. Emotion perception from a componential perspective. *Cognition and Emotion*, 9931(November):1–10, 2015.

[39] V. Shuman, D. Sander, and K. R. Scherer. Levels of valence. *Frontiers in Psychology*, 4(MAY):1–17, 2013.

[40] T. Tijs, D. Brokken, and W. Ijsselsteijn. Creating an emotionally adaptive game. *Lecture Notes in Computer Science*, 5309 LNCS:122–133, 2008.

[41] M. K. Uhrig, N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, and S. Marschall. Emotion Elicitation: A Comparison of Pictures and Films. *Frontiers in psychology*, 7(February):180, 2016.

[42] J. van Dijk. Digital divide research, achievements and shortcomings. *Poetics*, 34(4-5):221–235, 2006.

[43] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé. When are tutorial dialogues more effective than reading? *Cognitive science*, 30:1–60, 2006.

[44] P. E. Vernon. Some characteristics of the good judge of personality. *The Journal of Social Psychology*, 4(1):42–57, 1933.

[45] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function, and automatic analysis: a survey. *Proceedings of the 10th international conference on Multimodal interfaces*, pages 61–68, 2008.

[46] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

[47] A. Vinciarelli and F. Valente. Social Signal Processing: Understanding Nonverbal Communication in Social Interactions. In *Proceedings of Measuring Behavior*, 2010.

[48] T. Wehrle, G. R. Scherer, and N. York. Towards Computational Modeling of Appraisal Theories. *Appraisal*, pages 350–365, 2001.

[49] V. Xia, N. Jaques, S. Taylor, S. Fedor, and R. Picard. Active learning for electrodermal activity classification. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6. IEEE, 2015.

[50] Y.-c. Yeh, S. C. Lai, and C.-W. Lin. The dynamic influence of emotions on game-based creativity: An integrated analysis of emotional valence, activation strength, and regulation focus. *Computers in Human Behavior*, 55:817–825, 2016.

Eliciting Emotions in Design of Games – a Theory Driven Approach

Alessandro Canossa
Northeastern University
360 Huntigton Ave.
Boston, MA
a.canossa@northeastern.edu

Jeremy Badler
Northeastern University
360 Huntigton Ave.
Boston, MA
jbadler@ski.org

Magy Seif El-Nasr
Northeastern University
360 Huntigton Ave.
Boston, MA
m.seifel-nasr@northeastern.edu

Eric Anderson
Northeastern University
360 Huntigton Ave.
Boston, MA
ec.anderson@northeastern.edu

ABSTRACT
As technology becomes more powerful, computer software and game designers have ever-expanding tools available to create immersive, emotional experiences. Until recently, designing emotional experiences was achieved by veteran designers relying on insights from film theory and intuitions developed through years of practice. We propose another approach: leveraging scientific knowledge of emotions to guide the design process. The approach can serve as a resource for fledgling designers trying to break into the field, but will hopefully provide a few new insights for veterans as well. In addition, it may interest emotion researchers and psychologists looking to expand their stimulus repertoire. As a necessary underpinning for the design process, we will discuss several theoretical psychological models of emotions. Classical theories largely treat emotions as basic, universal states that are invariantly evoked by specific stimuli. While intuitive and popular, these theories are not well supported by current evidence. In contrast, a psychological constructionist theory, called the Conceptual Act Theory [6] proposes that emotions are constructed when conceptual knowledge is applied to ever changing affective experiences. The CAT proposes that emotional states can exhibit strong variation across instances and individuals due to differences in situational factors, learning histories and cultural backgrounds. This theory better fits available data, and also provides a framework for modeling emotional changes that vary by situation, person, and culture. The CAT also fits better with the game design process, since it treats users holistically as individuals. During the course of a game, similar to real life, emotions emerge from evaluations of situations and can therefore not be deterministically dictated by a single stimulus. Using the CAT framework, we developed a process to create affective digital game scenarios. Our goal is to give game designers, a scientific framework to better guide the design process.

CCS Concepts
• Human-centered computing → Interaction design → Interaction design process and methods → Scenario-based design

Keywords
Emotion elicitation; Affect; Game design; Personalization; Psychology of affect; Conceptual Act Theory.

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

1. INTRODUCTION
Creating emotionally engaging experiences is an important goal of game design. Game designers and developers use many different design techniques to evoke emotions. The Mechanics, Dynamics, Aesthetics (MDA) model, for example, advocates for the development of mechanics (game rules) that lead to game dynamics (game systems) that achieve aesthetic goals. The goals are defined as states that include: sensation (games as sense-pleasure), fantasy (game as make-believe), narrative (game as drama), challenge (game as obstacle course), fellowship (game as social framework), discovery (game as uncharted territory), expression (game as self-discovery) and submission (game as pastime) [74]. Several game design authors have proposed principles that describe the role of visual design, environment design and other physical properties of games and how they change over time as a way to evoke affect and a general sense of pleasure [75, 76]. The use of writing techniques to develop character and narrative in games that have emotional impact has received attention as well [28]. There have also been several works discussing the development of reward systems to encourage player achievement, competition or collaboration as a way to evoke emotions and sustain engagement (e.g. [78]). Virtual environment researchers have also acknowledged the potential and utility of adopting psychological theories of affect and emotions. One area where emotion theory has been used is in developing computational models of emotion elicitation for creating believable characters. Examples of this work include the Oz project, where the research group used scientific "appraisal" models of emotions [79] to develop expressive believable agents that can inhabit a virtual narrative world [50].

However, top designers see the game experience holistically. Thus the process of evoking emotions arises not just from characters that are expressive or believable, but from the complex interaction of all game elements: lighting, movement, sequences of events and user choices, and from the overall feel of the

environment [70].This necessitates a different kind of theory to guide the design process – an approach that treats the user experience as a whole rather than as separate components (as classical theories of emotion do). The formal research cited above is in many ways the exception; more often than not, designers develop techniques and make choices based on their experience and intuitions. Experience and intuition are difficult to codify. Thus we pose the question: *can a psychological theory of emotion be used in the game design process to enhance the players' emotional experience in a game? If so, how, and what is an optimal emotion theory for this purpose?* We propose that the Conceptual Act Theory [6] can be usefully employed by designers. Taking a holistic view, the theory builds on strong evidence that emotions are not hardwired or invariant entities that can be triggered by specific stimuli. Rather, the CAT proposes that emotional instances are newly created each time they occur from the sum of all stimuli, and vary as a person's internal (i.e., the person's bodily state) and external context changes. The instances also vary across individuals who have different emotion concepts, learning histories and cultural backgrounds. We first outline different psychological theories of emotion and their limitations. We then describe in more detail the CAT, emphasizing in particular features of the theory that are critical to our game design and iterative tuning process, and describing how the theory is different from others currently used by game researchers. Second, we review previous work in creating gaming experiences using emotions. Third, we describe a design process from concept inception to realization, through the example of a game created for research purposes. We conclude the paper briefly discussing our evaluation of the game's usability and playability as well as describing an initial study where we compare the self-reported and peripheral physiological responses of the initial pool of subjects. Last, we discuss our contribution to the game design process, as well as the effectiveness of a new theory of emotions that has not previously been used in the domain of affective computing. We believe the paper will provide promising evidence of the utility of the approach, which may open new research directions in the design of emotional experiences.

2. THEORETICAL FOUNDATION – PSYCHOLOGICAL THEORY OF EMOTIONS AND AFFECT

2.1 Emotions and Affect

A commonly held view of emotions is that there exists a set of discrete, innate and universal emotional states [25, 26, 41, 42, 45, 63]. This set of emotions is often referred to by such English words as anger, sadness, and fear, and are viewed as a natural kind [6]. When boiled down to their fundamental assumptions, basic emotion models make up the dominant scientific paradigm in the psychological study of emotion. Different models emphasize different parts of the process. For instance, one family of theories called "appraisal models", focus on the set of necessary events that trigger emotions [30, 37]. Once an emotion is triggered, the presumed result is an automated set of synchronized changes in response systems that produce the signature emotional response. This view predicts that the experience and perception of emotions are fairly universal, so little variability within or between people would be observed. While intuitive, the 'basic' emotion view is not well supported by the data, variability is the rule rather than the exception. Quantitative reviews of the research have failed to find signatures of emotions in the body [16] or brain [48]. Additionally, evidence is emerging that people from different cultures perceive emotions differently [29] and people within a culture have varied emotional lives [8]. While a complete review of this research is beyond the scope of this chapter, interested readers can consult Barrett et al. [9].

Another way of characterizing emotional states is in terms of their underlying affective dimensions. Two important affective dimensions are valence, the degree of pleasure or displeasure, and activation, the degree of arousal [7, 8, 65, 67]. Together, valence and activation form a unified affective state (Figure 1). Affect is grounded in the physical fluctuations of the body: somatovisceral, kinesthetic, proprioceptive, and neurochemical [7, 59]. Affect is also a central feature in many psychological phenomena, including emotion [7, 8, 20, 65], anticipating the future [31, 32], psychopathology [18, 19], and morality [36, 38]. Affective changes are crucial to the conscious experience of the world around us [24]. People in all cultures around the world seem to have affective experiences [53]. Unlike emotions, affect can be clearly measured in the facial expressions [16], in the voice [66], and in the peripheral nervous system [15, 16]. As a consequence, affect can be thought of as a neurophysiologic barometer of the individual's relationship to an environment at a given point in time, with self-reported feelings as the barometer readings.

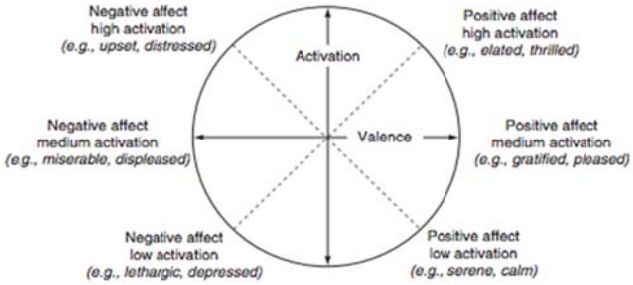


Figure 1. Circumplex model of affect.

2.2 Conceptual Act Model

Using affect as its foundation, the CAT [6] hypothesizes that affective experience becomes a 'real' emotion (fear, anger, etc.) when categorized as such using the emotion concept knowledge of a perceiver. These concepts have been learned from language, socialization, and other cultural artifacts within the person's day-to-day experience. The process of combining incoming sensory input (from the body and from the surroundings) with learned, category knowledge within the perceiver's brain is a normal part of what it means to be conscious. This conceptualizing is instantaneous, ongoing, obligatory, and automatic (meaning, a person will normally not have a sense of agency, effort or control in constructing an emotion). Conceptualizing is rarely due to a deliberate, conscious goal to figure things out. Thus to a person, emotions feel like they just happen. The CAT emphasizes the importance of situations. The conceptual system for emotion is constituted out of past experience, and past experience is largely structured by people within a cultural context. Therefore, the vocabulary of emotion categories that is developed, as well as the population of instances within each category are culturally relative. Such properties integrate the CAT with social construction approaches, positing that interpersonal situations "afford" certain emotions (or certain varieties of an emotion category). As a result, in the CAT emotions (like all mental states) are not assumed to be Platonic, physical types, but instead are treated as abstract, conceptual categories that are populated with variable instances optimized for a particular situation or context.

According to the CAT ,there are at least five sources for the variations that occur in emotional episodes: (1) the behavioral adaptations that serve as initial, affective predictions about how to best act in a particular situation (e.g., it is possible to freeze, flee, fight or faint during fear), (2) the concepts that develop for emotion, (3) the vocabulary used for emotions, (4) the variation in the types of situations that arise in different cultures, and (5) stochastic processes. As a result, there is variation within emotion categories, both within individuals and across people and cultures. Not everyone will experience the same emotion to the same stimulus, and even the same stimulus/person pairing can create different emotions at two different times.

2.3 Utility of the Conceptual Act Model to the game design process

Modern games contain complex, dynamic worlds that are well-suited to the application of the CAT for creating emotional experiences. Several key features of the CAT model are particularly relevant for game design:

1) There is variability in how people will respond to stimuli. This can be due to participants' past experiences, or contextual elements present in the situation that can be interpreted differently by different participants.

2) The context is critical for the experience users will have. A snake may elicit fear in one context, but amusement in another.

3) The sequence of events that lead to a specific situation is important when developing an emotional scenario.

A user's response is not solely determined by current conditions; it is also influenced by the preceding sequence of events. This is an important element of game design. Designers often use a "beat chart" to signify the sequencing of events or beats (single units of action) and their effect on the participant as they go through an experience. Using these three constructs, we will discuss in Section 4 a framework to guide the design process and show how these ideas can aid in developing and designing emotionally engaging scenarios. First, however, we briefly review current theory on designing for emotions.

3. PREVIOUS WORK ON EMOTIONS IN DESIGN

3.1 Computational Models of Emotions

Computer scientists have attempted to model how emotions are elicited by modeling them using digital environments. Marsella, Gratch and Petta [51] summarized several computational models of emotions. Most of these models use classical appraisal theories as the theoretical foundation, with the goal of developing 'emotional' virtual characters used in games and simulation. Computationally-based appraisal models assess events in the surrounding environment, compare them to an internal belief system, and change their emotional state accordingly. For example, the EDA model [33, 35] parameterizes external events in terms of desirability and likelihood of happening, which are then used to map to specific emotions. For example, positive desirability with likelihood < 1 yields hope, while negative desirability with likelihood = 1 yields distress. A good example of customized internal beliefs is the bully agent in the FearNot! system [21], which interprets as desirable another agent having fallen on the floor and crying (having been pushed by the bully); accordingly, a gloating response is produced. Besides appraisal models, three other categories of affective modeling are dimensional, anatomical and rational. Dimensional models do not implement discrete emotional categories, but rather treat emotions as continuous variables (i.e. affect; see Figure 1). For example, WASABI [11] defines different emotions as ranges in arousal-valence-dominance space, appraises the current situation in the same space, and uses the distance between the two to calculate a likelihood of a given facial expression. Anatomical models [4] are built from the ground up based on neuroanatomical data and processes. As such, they tend to be focused on a single emotion (e.g., fear) and have received only limited attention from the computational community. Finally, rational models are in many ways the opposite of anatomical, eschewing psychology almost entirely in favor of a pure artificial-intelligence approach. A good example is Scheutz and Sloman [68], who use the simple affect "hunger" to modify the behavior of intelligent, sensing agents in a world populated by other agents, food and various lethal entities.

Computational models of emotions are often put to use in the broader context of believable characters. Indeed, if computer-controlled agents are ever to appear "human", their ability to realistically express emotion is almost a requirement [10]. Once an agent has selected the appropriate emotion via an affective model, the agent needs to behave accordingly. An agent's emotional state can be conveyed visually by head position and facial expression [5, 22, 23] as well as body posture and movement [1, 3, 17, 60, 61]. The link to cognitively-driven behavior was recognized and exploited early on by the Oz project [72, 73], which developed an expressive artificial intelligence informed by emotional state. More recently, Hudlicka and colleagues [39, 40] modeled affect-induced changes in cognition, such as an increased threat response if the agent is anxious. Many of the researchers computationally modeling emotions use appraisal theory as a theoretical foundation for good reason. Appraisal theories focus on emotion elicitation - exactly what the researchers are attempting to model. For designers, such projects are interesting but leave out an important element: the actual experience of an emotion. Games seek to provide a holistic experience to the player, and since the above models do not include subjective experience they are of limited use to designers. Many game designers have therefore abandoned the use of emotion theory and instead adopted an alternative approach, either (a) creative methods that borrow techniques from other disciplines (e.g., film theory) and rely heavily on intuition, or (b) a more scientific approach where the design is still creative, but is tuned through the iterative process of testing, evaluating outcomes and modifying game variables as needed [2, 55, 54].

3.2 Creating Emotions in Interactive Experiences

Artists, designers, directors and other content creators often seek to evoke or manipulate the emotions of those who experience their work. They are interested in the holistic experience of the user. Many design techniques were documented in the 1960s and 70s, with the rise of film theory as an academic discipline. In films and television [13] as well as advertising [64], visual scenery and ambient light and color play a particularly important role. For example, according to Western cultural norms the color red often evokes violence or passion, while blue is methodical and cold [12]. Games are no exception [62], and may be even more effective conductors of emotion since they provide levels of control and immersion that are impossible using classic techniques [34, 56, 69]. One study [27] asked participants to navigate through versions of a virtual environment that differed only in some visual dimension (color, saturation, brightness or

contrast). It found a measurable effect on physiological signals such as heart rate and body temperature. Aside from visuals, other sensory stimuli such as music and sound [58] and even scent and vibration [56, 69] can also enhance the gameplay experience.

Optimizing a user's sensory experience is not sufficient, however – there are also the underlying story and gameplay itself. There are many narrative techniques that increase the player's emotional connection to the story, such as creating deeper relationships with one or more non-player characters (NPCs), including interesting and multilayered plot elements, and allowing the player to influence the story arc [28]. Even simple, scripted plot elements are sufficient to evoke emotions like joy or anger [71]. NPCs with emotional depth can be implemented using the affective computing methods surveyed earlier (section 3.1). Technical agency, such as giving players control of the game camera, is critical for avoiding frustration in certain games [52]. Even subtle distinctions are important: Leino [46] argues that players are more likely to experience emotions from game content that is integral to play ("undeniable") than purely superficial or aesthetic ("deniable"). Finally, the experience of players can be altered even before they start the game, by priming them to expect a fun or serious simulation for example [49].

Overall, it's necessary to view a game as a gestalt, with visuals and other stimuli, narrative, mechanics, characters and context all working in synergy to maximize the intensity of the user experience [56, 57, 69]. Much of the previous work admirably attempts to codify the intuitions of designers, but is still not driven by psychological theory. This is partly due to the fact that for designers, most theories have focused on stimulus and response while omitting user experience. Furthermore, many psychological theories have assumed that a specific stimulus invariantly causes a specific emotion in all people. This isn't the case - as designers intuitively know. Because CAT does not have the same limitations it can be used to inform design, as we demonstrate in the next section.

4. NEW DESIGN APPROACH USING CAT

We now describe a general method for applying insights from the CAT to the creative design process. To illustrate the method, we concurrently describe how we applied it to develop a short video game that had the explicit purpose of evoking different, robust affective experiences in players. The game was part of a larger project to study individual differences in affective experience and was developed jointly by affective scientists and video game researchers. The game was constructed using the engine, assets and editing tools provided by Fallout New Vegas (Bethesda Softworks). It consisted of four scenarios designed to elicit different affective states, along with a recurring neutral space designed to allow players to return to a relatively quiescent state. The four affective states were chosen to sample the different quadrants of the affective circumplex (see Figure 1). Specific emotions within each quadrant were chosen as target emotions for elicitation. Each scenario included a task for the player to perform and included timing constraints to make the game suitable under restrictive experimental conditions (e.g., fMRI). Navigation paths and player speed were tuned so that each scenario took a minimum of 90 seconds to complete, and a three minute timeout provided an upper bound in case the player did not complete the task. The four scenarios were:

- The Fear Cave (Figure 2A) was a dark, ominous environment with threatening giant insects and rumbling earthquakes. Players were instructed to retrieve a shovel and escape the cave.
- The Calm Valley (Figure 2B) was a peaceful, natural area with trees, flowers and a lake. Players were instructed to retrieve a flower and place it in the middle of the pond.
- The Exciting Casino (Figure 2C) had upbeat music and many lively characters. Players were instructed to pick up a lucky chip and play a slot machine that caused prizes to fall from the ceiling and non-player characters to cheer.
- The Sad Hotel (Figure 2D) was a run-down and somber interior. Players were instructed to fulfill the last wish of a dying man.

The rest area was the Hub (Figure 2E), appearing between each scenario as well as at the beginning and end of the play session. The hub was virtually empty and featured a character in a lab coat ("Doc") that interviewed the player after they completed each area. Doc served as an in-game survey, questioning the player on their affective state. He also led the player through simple psychological tasks, such as counting the number of vowels in a sentence or identifying the item in a picture. Such tasks are frequently used in physiology experiments to bring the subject's signal levels back to baseline.

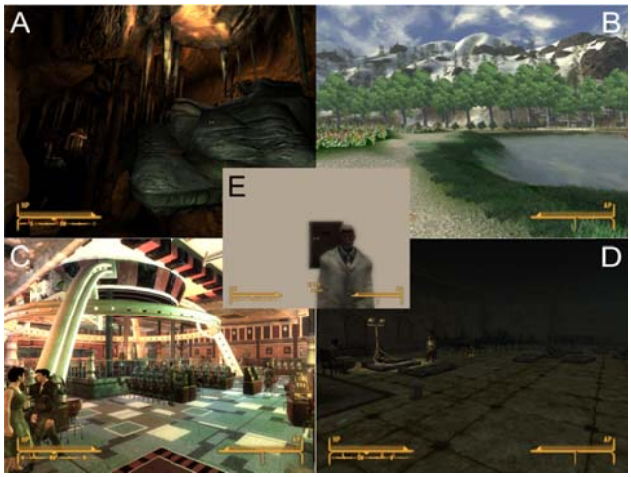


Figure 2. The five scenarios of the emotion-evoking game.

We developed the game by implementing the following methodology for designing affective experiences. It is an iterative, theory-driven process consisting of two phases: A) Affective States Definition and B) Scenario Design and Implementation. At each step the design is evaluated and tuned based on multiple iterations of internal experts' feedback and external testers' validation (Figure 3). The process draws upon the three primary design implications of the CAT (Section 2.3): 1) individual variability, 2) environmental context and 3) sequence of events.

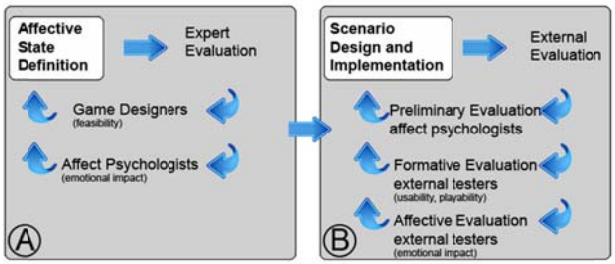


Figure 3. Iterative process for affective state definition and scenario design, implementation and tuning.

A) *Affective State Definition:* In the first phase of the process, the Circumplex model (see Section 2) is used to understand and define the affective design space for the game. At the simplest level, the desired emotions to evoke are simply represented as regions in activation-valence space, which need to be sufficiently large to accommodate individual variability. However, recalling the importance of event sequence, emotions are best plotted as trajectories across the space. In a normal game, the story arc will define the trajectory (Figure 4A) and can be modified as appropriate during this step. For our application, we wanted emotions that were as unambiguous as possible. Therefore, for each scenario we chose a single emotion towards the center of one quadrant (Figure 4B). Each scenario begins after an interval in the neutral center space, transitions to the target emotion, gradually increases in intensity to counteract habituation effects and provide variation, then transitions back to the neutral center to allow a "cool-down" establish a baseline before the next scenario and collecting subjective affective assessment using a five point Likert scale for valence and arousal.

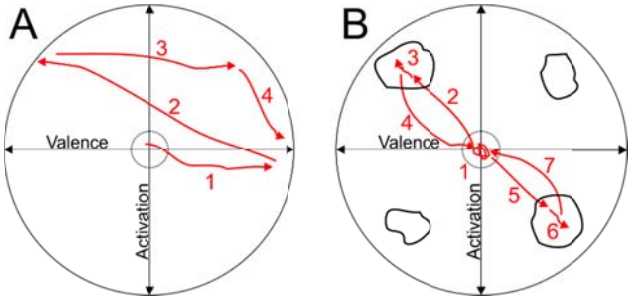


Figure 4. Trajectories in activation-valence space developed during the Affective State Definition phase. A, a typical story arc: 1-exposition, 2-crisis, 3-climax followed by resolution, 4-denouement. B, the first half of our affect game: 1-hub, 3-fear cave, 6-calm valley, 2,4,5,7-transitions. Representations of our four affect targets are also indicated.

Expert evaluation: the evaluation stage assures that all individuals involved in the design process are on the same page with regard to emotional definitions. Given individual variability, it's important that (e.g.) lead designers and level designers share the same goals. Technical limitations of the game engine also need to be accounted for at this stage, since a flawed implementation of an affective state is unlikely to achieve its goal. In our case, we initially selected eight emotional states spread over the four quadrants of activation-valence space, then drafted scenario outlines that targeted each one. The game designers evaluated the scenarios for development feasibility in terms of technology and resources available, while the affective psychologists focused on emotional impact and focus. After several iterations, four final states were selected: Excitement (positive valence, high arousal), Calm (positive valence, low arousal), Sadness (negative valence, low arousal) and Fear (negative valence, high arousal).

B) *Scenario Design and Implementation*: in the second phase of the overall design process, the game is fleshed out based on the states previously identified. The preliminary outlines from the previous phase are expanded into storyboards, then used to construct fully-elaborated scenarios. Appropriate emotional imagery is specified and incorporated into the narrative and game environment with the aid of audiovisual elements used by affective science research: IAPS (International Affective Picture System) [44] and IADS (International Affective Digital Sounds) [14]. These are large sets of standardized, internationally accessible photographs and sounds that include contents from a wide range of semantic categories, developed to provide a set of normative emotional stimuli for experimental investigation of emotion. Since emotional imagery is a continuum of experience to which players will adapt over time, it is important to incorporate narrative and audiovisual variability.

External evaluation: As the game components and levels are transformed from storyboards to fully-realized digital environments, three types of evaluation need to take place. First, the designers and/or affect psychologists need to verify that the constructed environments actually convey the correct emotion, at least for the majority of observers. Only the full, dynamic environment contains the context and event sequence that critically contribute to emotional impact. As an example from our own development process, the team agreed on an initial casino model based on static screenshots, but an early demonstration revealed certain subtle textures that detracted from the overall excitement. The textures were not feasible to replace, so we ended up changing to a different casino model for the final version. The second and third types of evaluation are both conducted by external testers. Playability testing is already familiar to designers, but is particularly important for an emotion-based game because technical issues (e.g., unclear tasks or difficult controls) can frustrate players and completely obliterate any desired emotional impact. In addition, given individual variability, it's vital to assess emotional impact on a number of people as well. Surveys or physiological measurements can be used; for our scenarios we employed both. More details on our formative and affective evaluations will be given in Section 4.1.

Previous work on emotional design (Section 3.2), as well as the CAT theory itself, underscore the necessity of a holistic approach. That is, a game and each contiguous segment within it must be seen as a whole: ambient sounds, music, light properties, colors, tasks to be performed, navigation patterns and shapes collectively contribute to emotion elicitation. The iterative steps of our design process help ensure that as many of the above properties as possible are incorporated, and can act in synergy, as the game is built up from basic emotional states. As mentioned previously, meeting the requirements imposed by our experimental research setup meant a certain degree of compromise, in particular the self-contained scenarios designed to allow for random shuffling and isolate particular affects. However, within each scenario we utilized the CAT-based model, carefully composing the visuals and audio stimuli over time and integrating them with story elements to achieve the targeted affective state results. Visual elements can be glimpsed in Figure 2: the Fear Cave is dark and sinister, the Exciting Casino is bright and active, etc. Audio elements also matched; the Calm Valley featured slow-paced, relaxing music while the Sad Hotel had a much darker soundtrack and pattering rain sounds. Dynamic features were used to counteract adaptation and habituation; for example, the lighting in the Fear Cave becomes progressively more saturated and shifts from a yellow orange tint to red, while the insects progress from large ants to giant scorpions and the earth progressively shakes and rumbles more and more. Two narrative examples are the Casino, with the story arc Enter, Get chip, Play slot machine, Win jackpot, Prizes fall from ceiling, and the Hotel, with arc Enter, See sad people, Find sick man, Bring sick man his diary, Discover sick man is dead before he has a chance to see it.

5. FORMATIVE AND AFFECTIVE EVALUATION

In order to validate our four scenarios, we performed two rounds of testing. The first battery of tests (with 3 subjects, two females, aged 20 and 21 and one male aged 20) was intended as a formative evaluation to assess usability and playability of the experience. Testers were recruited among students enrolled in the Game Design program at Northeastern University. After playing, they were assessed as to how well they could form a mental map of the locations and orient themselves in the designed worlds, as well as whether they could perform the actions required by the scenario. They were also informally queried on their emotional impressions. After testing, the design team made several modifications. For example, the Fear Cave had proved particularly difficult to navigate due to the dim lighting and slightly non-intuitive layout, so to avoid disorientation the floor plan was adjusted and unique light emitters were placed at key junctions.

For the second testing phase, our goal was to assess the effectiveness of the game at evoking affective responses. As mentioned above, there is no biological signature of emotions, though affect is much more reliably measured. Thus we used a multi-measurement approach, with both psychophysiology and retrospective verbal reports. While playing the four scenarios electrodermal activity (EDA) was measured as an index of arousal. EDA and other physiological signals are often used to evaluate some aspect of a user's experience during gameplay [43, 47, 58]. The testers were three subjects (one female/ two male) from the Interdisciplinary Affect Science Lab at Northeastern University who were not familiar with the project and had a variety of experience with video games. The subjects played the game in the physiology experiment space while their EDA was recorded. After completing play, they were asked to report their affective state during the game, questioned about their actions in-game and asked to describe the atmosphere of each scenario.

Although three subjects is too few to provide a reliable analysis of EDA data, the qualitative trends were nonetheless promising: the Fear Cave signals were consistently above baseline, while the Calm Valley and Sad Hotel were consistently below. These findings are consistent with our goal of the Fear Cave eliciting higher arousal and the latter scenario eliciting low arousal. The Exciting Casino was more variable, probably due to the fact that two of the three subjects had trouble executing the scenario task (we have since revised the in-game instructions) causing the onset of frustration that took precedence over any other affective state. Individual differences in game play were also visible in the data; the most striking example was a subject who accidentally removed the clothes from the dead refugee in the Sad Hotel and began laughing uncontrollably, yielding an EDA spike that persisted even after the scenario ended.

Such examples illustrate that subjects who do not successfully complete the task at hand are likely to report very different emotional experiences. They also highlight the fact that aesthetic features alone are not sufficient to guarantee the desired affective state. The tasks to be performed and the action possibilities in each scenario are not just additional elements of the design that can be treated separately, but are a fundamental layer of the whole experience. Additionally, the critical importance of players' interactions with their environment shows how game scenarios rather than video or audio stimuli can achieve deeper emotional impact, an observation of particular interest to affective scientists who wish to study powerful emotions in the lab. During the post-play reports, all of the subjects gave descriptive adjectives that almost exactly matched the development targets. Thus "scary" or "creepy" were used to describe the cave, "calm" or "relaxing" were used for the valley, "exciting" or "fun" the casino, and "depressing" or "sad" the hotel. Our goal is to use these scenarios in future research, modulating both the targeted affect and its intensity to explore individual differences in affective reactivity.

6. CONCLUSIONS

In this paper we discussed a theory driven approach to develop interactive experiences, especially games, which evoke affective responses from users. In particular, we argue for the holistic nature of designing emotional experiences, and thus propose using the CAT as a psychological model of emotions. The CAT acknowledges that individual differences, situational context, past experiences, mindset, and sequence of stimuli jointly influence participants' affect and behavior. Based on the model we developed a generalized, systematic process for designing game scenarios to evoke emotional experiences, and used it to develop our own research tool. We hope this novel approach facilitates a new perspective on theory-driven design and leads to interactive experiences with more varied and vivid emotions within.

7. REFERENCES

[1] Allbeck, J. and Badler, N. (2002). Toward representing agent behaviors modified by personality and emotion. Embodied Conversational Agents at AAMAS, 2, 15-19.

[2] Ambinder, M. (2011). Biofeedback in gameplay, GDC vault

[3] Amaya, K., Bruderlin, A., and Calvert, T. (1996). Emotion from Motion, Graphics Interface '96, pp. 222-229

[4] Armony, J. L., Servan-Schreiber, D., Cohen, J. D., and LeDoux, J. E. (1997). Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. Trends in Cognitive Sciences, 1(1), 28-34.

[5] Arya, A., Jefferies, L., Enns, J., and DiPaola S. (2006). Facial Actions as Visual Cues for Personality, Computer Animation and Virtual Worlds (CAVW) Journal.

[6] Barrett, L. F. (2014). The Conceptual Act Theory: A Précis. Emotion Review, 1–20.

[7] Barrett, L. F. (2006a). Emotions as natural kinds? Perspectives on Psychological Science, 1, 28-58.

[8] Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. Personality and Social Psychology Review, 10, 20–46.

[9] Barrett, L. F., Lindquist, K., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., and Brennan, L. (2007). Of mice and men: Natural kinds of emotion in the mammalian brain? Perspectives on Psychological Science, 2, 297-312.

[10] Bates, J. (1994). The role of emotion in believable agents. Communications of the ACM, 37(7), 122-125.

[11] Becker-Asano, C. and Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. Autonomous Agents and Multi-Agent Systems, 20(1), 32-49.

[12] Bellantoni, P. (2005). If it's purple, someone's gonna die: the power of color in visual storytelling. Taylor and Francis.

[13] Block, B. (2007). The visual story: creating the visual structure of film, TV and digital media. CRC Press.

[14] Bradley, M. M. and Lang, P. J. (1999). International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2). Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

[15] Bradley, M. M. and Lang, P. J. (2000). Measuring emotion: Behavior, feeling, and physiology. In R. D. Lane and L. Nadel (Eds.), Cognitive neuroscience of emotion (pp. 242–276). New York: Oxford University Press.

[16] Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., and Ito, T. A. (2000). The psychophysiology of emotion. In R. Lewis and J. M. Haviland-Jones (Eds.), The handbook of emotion (2nd ed., pp. 173–191). New York: Guilford Press.

[17] Chi, D., Costa, M., Zhao, L., and Badler, N. (2000). The EMOTE model for effort and shape, Proceedings of the 27th annual conference on Computer graphics and interactive techniques, p.173-182.

[18] Davidson, R. J. (2000). Affective style, psychopathology, and resilience: Brain mechanisms and plasticity. American Psychologist, 55, 1196–1214.

[19] Davidson, R. J., Pizzagalli, D., Nitschke, J. B., and Putnam, K. (2002). Depression: Perspec- tives from affective neuroscience. Annual Review of Psychology, 53, 545–574.

[20] Diener, E. (1999). Introduction to the special section on the structure of emotion. Journal of Personality and Social Psychology, 76, 803–804.

[21] Dias, J. and Paiva, A. (2005). Feeling and reasoning: A computational model for emotional characters. In Progress in artificial intelligence (pp. 127-140). Springer Berlin Heidelberg.

[22] DiPaola, S. and Arya, A. (2006). Emotional Remapping of Music to Facial Animation, ACM Siggraph '06 Video Game Symposium Proceedings, 2006.

[23] DiPaola, S. (2013). Face, portrait, mask: using a parameterised system to explore synthetic face space. In Electronic Visualisation in Arts and Culture (pp. 213-227). Springer London.

[24] Duncan, S. and Barrett, L. F. (2007). Affect as a form of cognition: A neurobiological analysis. Cognition and Emotion, 21, 1184–1211.

[25] Ekman, P. (1973). Cross-cultural studies of facial expression. In P. Ekman (Ed.), Darwin and facial expression: A century of research in review (pp. 169–222). New York: Academic.

[26] Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6, 169–200.

[27] El-Nasr, M. S., Morie, J., and Drachen, A. (2010). A scientific look at the design of aesthetically and emotionally engaging interactive entertainment experiences. Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives, 281-307.

[28] Freeman, D. (2003). Creating Emotion in Games: The Art and Craft of Emotioneering. New Riders Publishing, Thousand Oaks, CA, USA.

[29] Gendron, M., Roberson, D., van der Vyver, J. M., and Barrett, L. F (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. Emotion, 14, 251-262.

[30] Gendron, M. and L. F. Barrett (2009). "Reconstructing the past: A century of ideas about emotion in psychology." Emotion Reviews 1(4): 316-339.

[31] Gilbert, D. T., and Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. Journal of Personality and Social Psychology, 82, 503–514.

[32] Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., and Wheatley, T. (1998). Immune neglect: A source of durability bias in affective forecasting. Journal of Personality and Social Psychology, 75, 617–638.

[33] Gratch, J. and Marsella, S. (2001). Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In Proceedings of the fifth international conference on Autonomous agents (pp. 278-285). ACM.

[34] Gratch, J. (2003). Hollywood meets simulation: creating immersive training environments at the ICT. In Virtual Reality, 2003. Proceedings. IEEE (p. 303). IEEE.

[35] Gratch, J. and Marsella, S. (2004). A domain-independent framework for modeling emotion. Cognitive Systems Research, 5(4), 269-306.

[36] Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105–2108.

[37] Gross, J. J. and Barrett, L. F. (2011). "Emotion generation and emotion regulation: One or two depends on your point of view." Emotion Review 3(1): 8-16.

[38] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108, 814–834.

[39] Hudlicka, E. (2009). Affective game engines: motivation and requirements. In Proceedings of the 4th International Conference on Foundations of Digital Games (pp. 299-306). ACM.

[40] Hudlicka, E. and Broekens, J. (2009). Foundations for modelling emotions in game characters: Modelling emotion effects on cognition. In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on (pp. 1-6). IEEE.

[41] Izard, C. E. (1977). Human emotions. New York: Plenum.

[42] Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. Psychological Review, 100, 68–90.

[43] Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., and Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. Journal of Gaming and Virtual Worlds, 3(3), 181-199.

[44] Lang, P.J., Bradley, M.M., and Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

[45] LeDoux, J. E. (1996). The emotional brain: The mysterious underpinnings of emotional life. New York: Simon and Schuster.

[46] Leino, O. T. (2007). Emotions about the Deniable/Undeniable: Sketch for a classification of game content as experienced. In Situated play, proceedings of digra 2007 conference.

[47] Lindley, C. A., Nacke, L., and Sennersten, C. C. (2008). Dissecting Play - Investigating the Cognitive and Emotional Motivations and Affects of Computer Gameplay. In Proceedings of CGAMES. Wolverhampton, UK: University of Wolverhampton.

[48] Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. Behavioral and Brain Sciences, 35, 121-143.

[49] Luigi, D. P., Tortell, R., Morie, J., and Dozois, A. (2006). Effects of priming on behavior in virtual environments. Retrieved August, 8, 2010.

[50] Marsella, S., Gratch, J., and Rickel, J. (2004). Expressive behaviors for virtual worlds. In Life-Like Characters (pp. 317-360). Springer Berlin Heidelberg.

[51] Marsella, S., Gratch, J., and Petta, P. (2010). Computational models of emotion. A Blueprint for Affective Computing-A sourcebook and manual, 21-46.

[52] Martinez, H. P., Jhala, A., and Yannakakis, G. N. (2009). Analyzing the impact of camera viewpoint on player psychophysiology. In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on (pp. 1-6). IEEE.

[53] Mesquita, B. (2003). Emotions as dynamic cultural phenomena. In H. Goldsmith, R. Davidson, and K. Scherer (Eds.), Handbook of the affective sciences (pp. 871–890). New York: Oxford.

[54] Mirza-Babaei, P., Nacke, L. E., Gregory, J., Collins, N., and Fitzpatrick, G. (2013). How does it play better?: exploring user testing and biometric storyboards in games user research. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1499-1508). ACM.

[55] McAllister, G., Mirza-Babaei, P., and Avent, J. (2013). Improving Gameplay with Game Metrics and Player Metrics. In M. S. El-Nasr, A. Drachen, and A. Canossa (Eds.), Game analytics: Maximizing the value of player data. (pp. 621-638). Springer London.

[56] Morie, J. F., Iyer, K., Valanejad, K., Sadek, R., Miraglia, D., Milam, D., ... and Leshin, J. (2003). Sensory design for virtual environments. In ACM SIGGRAPH 2003 Sketches and Applications (pp. 1-1). ACM.

[57] Morie, J., Williams, J., Dozois, A., and Luigi, D. (2005). The Fidelity of "Feel": Emotional Affordance in Virtual Environments 11th International Conference on Human-Computer Interaction, Las Vegas, NV; July 2005.

[58] Nacke, L. E., Grimshaw, M. N., and Lindley, C. A. (2010). More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. Interacting with Computers, 22(5), 336-343.

[59] Nauta, W. (1971). The problem of the frontal lobe: A reinterpretation. Journal of Psychiatric Research, 8, 167–187.

[60] Neff, M. and Fiume, E. (2003). Aesthetic edits for character animation, Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation, San Diego, California

[61] Neff, M. and Fiume, E. (2005). AER: aesthetic exploration and refinement for expressive character animation, Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation, Los Angeles, California

[62] Niedenthal, S. (2008). Complicated shadows: The aesthetic significance of simulated illumination in digital games. Blekinge Technical University.

[63] Panksepp, J. (1998). Affective neuroscience: The foundations of human and animal emotions. New York: Oxford University Press.

[64] Poels, K. and Dewitte, S. (2006). How to capture the heart? Reviewing 20 years of emotion measurement in advertising. DTEW-MO_0605, 1-47.

[65] Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110, 145–172.

[66] Russell, J. A., Bachorowski, J., and Fernandez-Dols, J. M. (2003). Facial and vocal expressions of emotion. Annual Review of Psychology, 54, 329–349.

[67] Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology, 76, 805–819.

[68] Scheutz, M. and Sloman, A. (2001). Affect and agent control: Experiments with simple affective states.

[69] Swartout, W., Hill, R., Gratch, J., Johnson, W. L., Kyriakakis, C., LaBore, C., and Moore, B. (2006). Toward the holodeck: Integrating graphics, sound, character and story. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES.

[70] Totten, C. (2014). An Architectural Approach to Level Design. CRC Press.

[71] Wang, N. and Marsella, S. (2006), Introducing EVG: An Emotion Evoking Game. In 6th International Conference on Intelligent Virtual Agents, Marina del Rey, CA, 2006.

[72] Mateas, M. (1999). An Oz-centric review of interactive drama and believable agents. In Artificial intelligence today (pp. 297-328). Springer Berlin Heidelberg.

[73] Mateas, M. (2001). Expressive AI: A hybrid art and science practice. Leonardo, 34(2), 147-153.

[74] Hunicke, R., LeBlanc, M., and Zubek, R. (2004). MDA: A formal approach to game design and game research, In Proceedings of the Challenges in Games AI Workshop, Nineteenth National Conference on Artificial Intelligence, San Jose, CA.

[75] Kahrs, J., Calahan, S., Carson, D., Poster, S. Pixel Cinematography. Sigghraph 1996

[76] Solarski, C. (2013) The Aesthetics of Game Art and Game Design. Gamasutra, retrieved on 8/8/2016 http://www.gamasutra.com/view/feature/185676/the_aesthetics_of_game_art_and_.php

[77] Adams, E. (2013) Fundamentals of Game Design, New Riders Publishing

[78] Rigby, S., Ryan, R. M. (2011) Glued to Games: How Video Games Draw Us in and Hold Us Spellbound. Praeger Publisher

[79] Ortony, A., Clore, G. L., Collins, A. (1990) The cognitive structure of emotions. Cambridge university press

The Influence of Users' Personality Traits on Satisfaction and Attractiveness of Diversified Recommendation Lists

Bruce Ferwerda
Johannes Kepler University
Altenberger Str. 69
4040 Linz, AT
bruce.ferwerda@jku.at

Mark Graus
Eindhoven University of Technology
IPO 0.20, P.O. Box 513
5600 MB Eindhoven, NL
m.p.graus@tue.nl

Andreu Vall
Johannes Kepler University
Altenberger Str. 69
4040 Linz, AT
andreu.vall@jku.at

Marko Tkalcic
Free University of Bolzano
Piazza Domenicani, 3
39100 Bozen-Bolzano, IT
marko.tkalcic@unibz.it

Markus Schedl
Johannes Kepler University
Altenberger Str. 69
4040 Linz, AT
markus.schedl@jku.at

ABSTRACT

Diversifying recommendations has shown to be a good means to counteract on choice difficulties and overload, and is able to positively influence subjective evaluations, such as satisfaction and attractiveness. Personal characteristics (e.g., domain expertise, prior preference strength) have shown to influence the desired level of diversity in a recommendation list. However, only personal characteristics that are directly related to the domain have been investigated so far. In this work we take personality traits as a general user model and show that specific traits are related to a preference for different levels of diversity (in terms of recommendation satisfaction and attractiveness). Among 103 participants we show that conscientiousness is related to a preference for a higher degree of diversification, while agreeableness is related to a mid-level diversification of the recommendations. Our results have implications on how to personalize recommendation lists (i.e., the amount of diversity that should be provided) depending on users' personality.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI); User models; User studies;**

Keywords

Diversity; Recommender Systems; User-Centric Evaluation; Personality

1. INTRODUCTION

Providing users with a diversified list of recommendations has shown to have positive effects on the user experience. With an abundance of choices available nowadays, providing diversity in the recommendations can counteract on the negative psychological effects that users may experience, such as choice overload and choice difficulties [26]. These negative effects are caused by recommender systems, which are originally designed to output recommendations that are closest to the user's interest. The closer to the user's interest, the higher the accuracy of the recommender system algorithm, but also results in recommendations that are often too similar to each other (e.g., same level of attractiveness to the user). This does not only increase the chance of choice overload and choice difficulties to the user, but also increases the possibility of not covering the full spectrum of the user's interest [3].

Although prior research has shown that recommendation diversity has positive effects on the user experience, differences between diversity needs of users have not been given a lot of attention. Domain expertise and prior choice preferences have shown to play a role in the amount of diversity desired by the user [2, 6, 26]. Others have shown that diversity needs can also be related to cultural dimensions [8, 14]. In this work we consider personality traits as an indicator of satisfaction and attractiveness on differently diversified music recommendation lists.

The use of personality as a general model for users has gained increased interest. Several works revealed personality-based relationships with users' behavior, preferences, and needs (e.g., [10, 15, 25]), how to implicitly acquire personality traits of users from social media trails (e.g., Facebook [1, 4, 12, 20], Twitter [16, 21], and Instagram [11, 13, 24]), and how personality traits can be implemented into a personalized system [7, 9]. With our work we contribute to the personality research by providing more insights into personality-related diversity needs. We found among 103 participants that the conscientiousness and agreeableness personality traits play a role in the desired amount of diversity in a recommendation list. While conscientious participants showed a higher degree of satisfaction and attractiveness with the more diversified recommendations, agreeable participants were more satisfied and found the list more attractive with medium amount of diversity in the recommendations.

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

2. RELATED WORK

The positive effects of recommendation list diversity has been shown by several researchers. Bollen et al. [2] and Willemsen et al. [26] investigated the influence of diversity on movie recommendations and found that diversity has a positive effect on the attractiveness of the recommendation set, the difficulty to make a choice, and eventually on the choice satisfaction. Besides the positive effects of diversification, also personal characteristics play a role on the attractiveness of the diversified recommendation list (e.g., strength of prior preference or domain expertise [2, 23]). Bollen et al. [2] found that expertise in the domain showed a positive effect on the item attractiveness.

The personal characteristics that have been identified so far are domain specific to the kind of recommendations. However, a more general personal characteristic may be present that influences the subjective evaluations with the diversified recommendations. Personality has shown to be an enduring factor, which can relate to one's taste, preference, and interest (e.g., [5, 10, 25]). Chen et al. [5] and Wu et al. [27] showed relationships with personality and preference for diversification based on different movie characteristics (e.g., genre, artist, director). Ferwerda et al. [10] showed that music preferences can be related to the personality of the listener, whereas Tkalcic et al. [25] found relationships between personality traits and the preference of being exposed to certain amounts of multimedia meta-information.

In this work we investigate whether personality traits can be considered a personal characteristic that influences the subjective evaluations of diversified recommendation lists. To this end, we rely on the widely used five-factor model (FFM), which categorizes personality into five general dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [19].

3. DATA PREPARATION & PROCEDURES

We created differently diversified music recommendation lists in order to investigate the influence of personality traits on the subjective evaluation of the recommendation lists. Since we created the recommendation lists off-line, we separated the study in two parts. In the first part participants were recruited and their *complete* Last.fm listening history was crawled in order to create the recommendation lists. After the lists were created, participants from the first part were invited for the second part where they were asked to assess the diversified recommendation lists.

We recruited 254 participants through Amazon Mechanical Turk for the first part of the study. Participation was restricted to those located in the United States with a very good reputation ($\geq$95% HIT approval rate and $\geq$1000 HITs approved) and a Last.fm account with at least 25 listening events. Furthermore, they were asked to fill in the 44-item Big Five Inventory personality questionnaire [19] to measure the FFM. Control questions were asked to filter out fake and careless contributions. A compensation of $1 was provided. We crawled the complete listening history of each participant and aggregated the listening events to represent artist and playcount (i.e., number of times listened to an artist).

In order to prepare the music recommendation lists for each participant, we complemented our data with the LFM-1b dataset [22]. [1] This dataset consists of the complete listening histories of 120,322 Last.fm users from different countries. Since our participants were all located in the United States, we only used the United State users of the LFM-1b dataset to complement our dataset. This resulted in 10,255 additional users, which we also aggregated into artist and playcount for each user. The final dataset consists of user, artist, and artist playcount triplets with a total of 387,037 unique artists for the creation of the recommendation lists.

We used the weighted matrix factorization algorithm of [18] on our final dataset to calculate the recommended items. This algorithm is specifically designed to deal with datasets consisting of implicit feedback (e.g., artist playcounts). We optimized the factorization hyper-parameters by conducting grid-search and picking the setting that yielded the best 5-fold cross-validated mean percentile rank. Specifically, using 20 factors, confidence scaling factor α=40, regularization weight λ=1000 and 10 iterations of alternating least squares, we achieved the best 5-fold cross-validated mean percentile rank of 1.78%. [2] Afterwards we factorized the whole user-artist triplets using this set of hyper-parameters.

The recommended items were diversified as was done in [26] by using the method of [28]. By using the latent features as the basis of diversification instead of additional metadata like genre information (as is done in content-based recommender systems) guarantees that diversity is manipulated in line with user preferences. Previous research demonstrated that this way of diversifying recommendations is perceived accordingly by users [26].

A greedy selection to optimize the intra-list similarity [3] was run on the top 200 recommended artists (i.e., the 200 artists with highest predicted relevance) to maximize the distances between item vectors in the matrix factorization space. This algorithm starts with a recommendation set consisting of the artist with highest predicted relevance. In an iterative fashion items are added to the recommendation set until it contains 10 items.

In each step of the iteration, for each candidate item i the sum of all distances from its item vector to each item vector in the recommendation set is calculated: $c_i = \sum_{j=1}^{z} d(i,j)$, where z is the number of items in the recommendation set and $d(i,j)$ is the Euclidean distance between two item vectors i and j). All candidate items are ranked based on decreasing value of c_i (P_{c_i}) and on predicted relevance (P_{r_i}). A weighting factor β is introduced to balance the trade-off between predicted relevance and diversity. For each candidate item the combined rank is calculated following $w_i^* = \beta * P_{c_i} + (1-\beta) * P_{r_i}$. The item with the highest combined rank is added to the recommendation set and the next step is taken until 10 items are selected.

β was manipulated to achieve different levels of diversification. In the described implementation β=1 corresponds to maximum diversity, β=0 corresponds to maximum predicted relevance. We compared recommendation lists for different values of β in terms of the sum of distances between the latent features scores of items in the recommendation set and their average range. The list for β=0.4 showed to fall halfway between maximum relevance and maximum diversity. Thus, the final β levels for diversification were set at β=0 (low), β=0.4 (medium), and β=1 (high).

After the recommendation lists were created, emails were

[1] Available at http://www.cp.jku.at/datasets/LFM-1b/

[2] See [18] for details on the hyper-parameters and the definition of the mean percentile rank metric.

sent out to all participants to invite them for the second part of the study. We created a login screen so that we could retrieve the personalized recommendation lists for each participant. After the log in, the participant was sequentially presented with a recommendation list for three times, with each time a different level of diversity (i.e., low, medium, or high). The order of presentation was randomized. Each recommended artist was enriched with metadata from Last.fm (i.e., picture, genre, Top-10 songs with the number of listeners and playcounts), which was shown when hovered over the name in the list. Additionally, example songs were provided by clicking on the artist name (new browser screen linked to the artist's YouTube page). Participants were asked to answer questions about perceived diversity, recommendation satisfaction, and recommendation attractiveness [3] before moving on to the next list. These questions needed to be answered for each of the three lists.

After the participant assessed all three recommendation lists, we performed a manipulation check by placing the three lists next to each other (randomly ordered) and asked the participant to rank order the lists by diversity.

There were 103 participants who returned for the second part of the study. We included several control questions to filter out careless contributions, which left us with 100 participants for the analyses. Age: 18-65 (median 28), gender: 54 male, 46 female, and were compensated with $2.

4. RESULTS

4.1 Manipulation Check

A Wilcoxon signed-rank test was used to test the perceived diversity levels of the recommendation lists. Results show an increase of perceived diversity by comparing the low diversity (M=1.28) against the medium (M=2.05, r=.60, Z=10.370, p<.001) and high condition (M=2.65, r=.80, Z=13.784, p<.001). A significant diversity increase was also found between medium and high (r=.45, Z=7.711, p<.001).

4.2 Measures

Items in the questionnaire were assessed using a confirmatory factor analysis (CFA) with repeated ordinal dependent variables and a weighted least squares estimator to determine whether the questions convey the predicted constructs. After deleting questions with high cross-loadings and low commonalities, the model consisting of three constructs showed a good fit: $\chi^2(32)$=108.6, p<.001, CFI=.99, TLI=.98, $RMSEA$=.06. [4] The constructs with their items are shown below (5-point Likert scale; Disagree strongly-Agree strongly). The Cronbach's alpha (α) and the average variance extracted (AVE) of each construct showed good values (i.e., α>.8, AVE>.5), indicating convergent validity. Also, the square root of the AVE for each construct is higher than any of the factor loadings (FL) of the respective construct, which indicates good discriminant validity.

Perceived Diversity (AVE=.723, α=.887):
- The list of artists was varied. (FL=.858)
- Many of the artists in the lists differed from other artists in the list. (FL=.837)
- The artists differed a lot from each other on different aspects. (FL=.855)

Recommendation Satisfaction (AVE=.821, α=.932):
- I am satisfied with the list of recommended artists. (FL=.927)
- In most ways the recommended artists were close to ideal. (FL=.905)
- The list of artist recommendations meet my exact needs. (FL=.885)

Recommendation Attractiveness (AVE=.771, α=.931):
- I would give the recommended artists a high rating. (FL=.874)
- The list of artists showed too many bad items. (FL=-.830)
- The list of artists was attractive. (FL=.914)
- The list of recommendations matched my preferences. (FL=.893)

4.3 Analysis

We used a repeated measures ANOVA in order to investigate the influence of personality traits on the subjective evaluations of the diversified music recommendation lists. Below the results of personality traits on the different subjective evaluations are provided. The effects between diversity levels are all compared against the low diversity condition.

4.3.1 Personality on Perceived Diversity

Results show that Mauchly's test is not violated ($\chi^2(2)$= .115, p=.944), so sphericity can be assumed, and therefore, no correction is needed. The results show that there are no significant main effects of the different personality traits on perceived diversity. However, a general difference in perceived diversity can be assumed (F(2, 22)=51.029, p<.001). Exploring the differences between the levels of diversified recommendation lists show that there is an increase in perceived diversity when comparing the low diversified list against the medium (F(1, 11)=11.596, p<.001) and the high diversified lists (F(1, 11)=31.191, p< .001). This confirms once more that our diversification was effective and was perceived as such by the participants.

4.3.2 Personality on Recommendation Satisfaction

Mauchly's test shows that sphericity is not violated ($\chi^2(2)$= 1.830, p=.401), and therefore no correction is needed. Assessing the effect of the different personality traits on the recommendation satisfaction, the following personality traits show a main effect: conscientiousness (F(4, 22)=2.454, p<.05) and agreeableness (F(4, 22)=3.886, p<.05). Additional analyses by looking at the levels between the diversity levels (i.e., low, medium, and high diversification) show that conscientious participants are increasingly satisfied when provided a higher degree of diversity: medium diversity (F(2, 11)=3.994, p<.05) and high diversity (F(2, 11)=4.036, p<.05). However, the satisfaction differences for agreeable participants show a higher satisfaction for the medium diversification (F(2, 11)=9.660, p<.05) than for the high diversification (F(2, 11)=4.036, p<.05).

[3] Questions measuring perceived diversity and recommendation attractiveness were adapted from [26].

[4] Cutoff values for a good model fit are proposed to be: CFI>.96, TLI>.95, and $RSMEA$<.05 [17].

4.3.3 Personality on Recommendation Attractiveness

Assessing Mauchly's test shows that there is no violating of sphericity ($\chi^2(2)= 1.860$ $p=.395$). Also here, results show main effects for the conscientiousness ($F(4, 22)=3.157$, $p<.05$) and agreeableness ($F(4, 22)=3.469$, $p<.05$) personality traits. By looking at the differences between the levels of diversification, we found similar patterns as with satisfaction. Results show that conscientious participants were increasingly more attracted to more diversified recommendation lists: medium ($F(2, 11)=2.955$, $p<.05$), high ($F(2, 11)=7.866$, $p<.05$). Participants scoring high on the agreeableness personality traits show to be more attracted to the medium ($F(2, 11)=5.933$, $p<.05$) diversified list than to the high ($F(2, 11)=5.314$, $p<.05$) diversified list.

5. CONCLUSION & DISCUSSION

Our results show that certain personality traits (i.e., conscientiousness and agreeableness) are related to the subjective evaluations of diversified recommendation lists. We found that conscientious people judged a higher degree of diversity more attractive and were more satisfied with it, whereas agreeable people showed to have more interest (i.e., list attractiveness and satisfaction) in a medium degree of diversity.

The relationships that we found can be used in personality-based systems as proposed in [7]. With the increased connectedness of applications, such as recommender systems, with social networking sites, users' personality can be acquired without the need of behavioral data in the application (e.g., via Facebook [1, 4, 12, 20], Twitter [16, 21], or Instagram [11, 13, 24]). By identifying relationships with users' personality traits, such as in this work, cross-domain inferences about users' preferences and needs can be made and implemented to provide a personalized experience to users.

6. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF): P25655.

7. REFERENCES

[1] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21:372–374, 2010.

[2] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. Graus. Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on RecSys*, pages 63–70. ACM, 2010.

[3] P. Castells, N. J. Hurley, and S. Vargas. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*, pages 881–918. Springer, 2015.

[4] F. Celli, E. Bruni, and B. Lepri. Automatic personality and interaction style recognition from facebook profile pictures. In *Proceedings of the ACM MM*, 2014.

[5] L. Chen, W. Wu, and L. He. How personality influences users' needs for recommendation diversity? In *Proceeding of CHI'13 EA*. ACM, 2013.

[6] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168. ACM, 2014.

[7] B. Ferwerda and M. Schedl. Enhancing Music Recommender Systems with Personality Information and Emotional States: A Proposal. In *Proceedings of the 2nd Workshop on EMPIRE*, 2014.

[8] B. Ferwerda and M. Schedl. Investigating the relationship between diversity in music consumption behavior and cultural dimensions: A cross-country analysis. In *Proc. of the 1st Workshop on SOAP*, 2016.

[9] B. Ferwerda and M. Schedl. Personality-Based User Modeling for Music Recommender Systems. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2016)*, Riva del Garda, Italy, 2016.

[10] B. Ferwerda, M. Schedl, and M. Tkalcic. Personality & emotional states: Understanding users' music listening needs. *UMAP 2015 Extended Proceedings*.

[11] B. Ferwerda, M. Schedl, and M. Tkalčič. Predicting Personality Traits with Instagram Pictures. In *Proceedings of the 3rd Workshop on EMPIRE*, 2015.

[12] B. Ferwerda, M. Schedl, and M. Tkalčič. Personality Traits and the Relationship with (Non-)Disclosure Behavior on Facebook. In *Companion of the 25th International WWW Conference*, 2016.

[13] B. Ferwerda, M. Schedl, and M. Tkalčič. Using Instagram Picture Features to Predict Users' Personality. In *Proceedings of the 22nd International Conference on MMM*, Miami, USA, January 2016.

[14] B. Ferwerda, A. Vall, M. Tkalčič, and M. Schedl. Exploring Music Diversity Needs Across Countries. In *Proceedings of the 24th International Conference on UMAP*, Halifax, Canada, July 2016.

[15] B. Ferwerda, E. Yang, M. Schedl, and M. Tkalčič. Personality Traits Predict Music Taxonomy Preferences. In *ACM CHI '15 EA*, 2015.

[16] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting Personality from Twitter. In *Proceedings of the 3rd International Conference on SocialCom*, 2011.

[17] L.-t. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

[18] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *ICDM*, 2008.

[19] O. P. John, E. M. Donahue, and R. L. Kentle. The big five inventory: Versions 4a and 54, institute of personality and social research. *UC Berkeley*, 1991.

[20] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman. Automatic Personality Assessment Through Social Media Language. *Journal of Personality and Social Psychology*, 108, November 2014.

[21] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *Proceedings of the 3rd International Conference on SocialCom*, 2011.

[22] M. Schedl. The LFM-1b Dataset for Music Retrieval and Recommendation. In *Proceedings on ICMR*, 2016.

[23] B. Scheibehenne, R. Greifeneder, and P. M. Todd.

What moderates the too-much-choice effect? *Psychology & Marketing*, 26(3):229–253, 2009.

[24] M. Skowron, B. Ferwerda, M. Tkalčič, and M. Schedl. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *Companion Proceedings of the 25th International WWW Conference*, 2016.

[25] M. Tkalcic, B. Ferwerda, D. Hauger, and M. Schedl. Personality correlates for digital concert program notes. *UMAP 2015, Springer LNCS 9146*, 2015.

[26] M. C. Willemsen, B. P. Knijnenburg, M. P. Graus, L. C. Velter-Bremmers, and K. Fu. Using latent features diversification to reduce choice difficulty in recommendation lists. *RecSys*, 11:14–20, 2011.

[27] W. Wu, L. Chen, and L. He. Using personality to adjust diversity in recommender systems. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 225–229. ACM, 2013.

[28] C.-N. Ziegler, S. M. McNee, J. a. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. *WWW*, page 22, 2005.

A Jungian based framework for Artificial Personality Synthesis

David Mascarenas
Los Alamos National Laboratory – Engineering Institute
PO Box 1663 MS T001
Los Alamos, NM, 87544
1-505-665-0881
dmascarenas@lanl.gov

ABSTRACT
The field of computational intelligence has enjoyed much success in developing a variety of algorithms that emulate human cognition. However, a framework to tie these algorithms together in a coherent manner to create a machines that possess the full spectrum of human-like personalities is still needed. To date, research on artificial personality synthesis has focused on using the Big Five model from the field of personality psychology. The overlooked Achilles heel of Big Five (BF) is that it is purely data-driven model, and thus offers only marginal guidance on how a machine with a personality might actually be created. In this work an alternative computational personality framework is presented based on the work of Carl Jung. There are two key insights that suggest a Jungian type-based framework is suitable for synthesizing an artificial personality. First, the cognitive functions which form the building blocks of the Jungian personality model can be mapped to classes of algorithms used to emulate cognition. Second, the Jungian framework suggests that at any given time humans are only using one of the cognitive functions. This suggests that a human personality could be emulated using a state machine with each state implemented using the appropriate class of algorithms.

CCS Concepts
• Human-centered computing→Interaction design theory, concepts and paradigms

Keywords
artificial personality synthesis; Carl Jung; affective computing.

1. INTRODUCTION
The advent of ubiquitous computing has increased interest in techniques for endowing a machine with a human-like personality. Vinciarelli [1] provides an overview of personality computing research which shows that since 2006 a marked increase in research papers with the word "personality," included in the title. Vinciarelli makes one particularly interesting statement with regards to the nature of the work that has been undertaken thus far, "Trait based models are widely accepted in the computing community as well. All of the works surveyed in this article adopt personality traits (the BF in 76 cases out of 81) and, to the best of our knowledge, no other theories were ever adopted in computing oriented research. On one hand, this barely reflects the dominant position of trait based models in personality psychology. On the other hand, trait-based models represent personality in terms of numerical values, a form particularly suitable for computer processing. [1]." This statement raises an interesting question. Why has computational personality research thus far restricted itself to exploring trait-based models of personality? Vinciarelli points out a number of competing personality models including: psychoanalytic, cognitive, and behaviorist and biological. Arguably, personality models such as the biological model may currently not be adequately understood to be implemented on a computer, and as Vinciarelli points out, the numerical nature of trait-based models may be amenable to computer implementation to a degree. However, the nature of the Jungian type-based model also lends itself to implementation on a computer, but to the best knowledge of the author has not been explored to date. This work will outline how Jungian psychological type theory can be used as guidance to synthesize an artificial personality.

This Jungian type-based framework possesses a number of attractive properties. First, a very large fraction of algorithms emulating cognitive processes can be utilized by the framework in a coherent manner. Second, the framework can arguably synthesize the full spectrum of human personalities. Third, Jungian-based personality theory is very popular among laypeople. Modern personality psychologists might argue that this is due to the cognitive bias known as the Barnum effect. From a Turing test perspective however this is irrelevant, and perhaps even an advantage because it would facilitate the illusion of a machine having a human-like personality For practical applications all that really matters is that the machine can convince a human that it has a personality.

2. THE JUNGIAN PERSONALITY TYPE MODEL
Carl Jung is probably most popularly known for introducing the concept of the "introvert," and "extrovert." Jung detailed his theory of psychological types in 1921 [2], [3]. Jung's model for personality is based on the idea of "cognitive functions." Jung identified two fundamentally different kinds of cognitive functions known as "perception," and "judgement. [4]" Perception describes how a person takes in information, and judgment pertains to decision-making. Jung then broke these classes down further. Jung asserted that perception came in two main forms: "sensation" and "intuition." Sensation is focused on physical reality:. It tends to focus on the present and past. Intuition is primarily focused on finding meaning, patterns, or possibilities in information.

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

The tendency is to focus more on the future [5]. Likewise, Jung identified two forms of judgement which he referred to as "thinking," and "feeling." Thinking refers to decision-making processes that focus more on the application of basic truths/principles. It tends to impersonal in the sense that it resists allowing personal value judgements, or the value judgements of others influence decision making. Conversely, "Feeling," puts significant weight on values. These values can be either personal or shared by a community. It tends to prefer decisions that will result in harmony [6]. Jung then further refined his cognitive functions by asserting that each of the four functions has an introverted and extroverted orientation. An introverted orientation implies a tendency towards a person's interior world of thoughts, ideas, feelings and memories. An extroverted orientation focuses on people or experiences external to the self [4]. Jung's clinical observations and reflection ultimately resulted in a total of 8 cognitive functions. For completeness the eight cognitive functions are:

Extroverted Thinking	(Te)
Introverted Thinking	(Ti)
Extroverted Feeling	(Fe)
Introverted Feeling	(Fi)
Extroverted Sensing	(Se)
Introverted Sensing	(Si)
Extroverted Intuition	(Ne)
Introverted Intuition	(Ni)

For the sake of brevity, a full description of Carl Jung's 8 cognitive functions will not be provided in this document. Since Jung initially introduced the concept of cognitive functions, the language used to describe them has evolved as has an understanding of their nature. For the purpose of this work, the cognitive function descriptions provided in [4] will be used throughout this work.

In 1923, Katherine Briggs and her daughter Isabel Briggs Meyers were exposed to the newly available English translation of "Psychological Types" [7], [8] At the time Katherine was in the process of developing her own personality theory motivated partially by a desire to understand the personality of her son-in-law. Upon reading Jung's work they came to the conclusion that Jung's theory was superior to their own and so decided to adopt and refine the Jungian model. Over the next 20 years the mother-daughter team obsessively observed and documented human nature with regards to type. They eventually created a variation on Jung's system with an associated sorting instrument known as the Myers-Briggs Type Indicator (MBTI).

The evolution of Jung's psychological type proposed by Myers and Briggs had a few important characteristics that are worth mentioning. First, Myers and Briggs observed that all people used all of the cognitive functions. The dissimilarity between different types of people was the preference with which they used the cognitive functions. In this model each person uses the cognitive functions in a hierarchical manner with a dominant cognitive function, followed by an auxiliary, tertiary and inferior cognitive functions. Classically, Meyer & Briggs focused on the first four cognitive functions. Furthermore, Myers & Briggs introduced very specific constraints on the hierarchical order the cognitive functions were allowed to assume. One of the constraints Meyers & Briggs introduced was that the hierarchy of cognitive functions had to alternate between introverted and extroverted orientations. It is worth noting that the writing of Carl Jung can be interpreted to indicate a different scheme for ordering the cognitive functions, and competing ordering systems are in existence. For the sake of clarity, the author tends to prefer the ordering system outlined in [4]. However, with regards to artificial personality synthesis, the framework outlined in this work is flexible enough to be adapted to any desired cognitive function ordering scheme.

Ultimately, based on their imposed constraints on cognitive function order, Meyers & Briggs identified 16 unique personality types. These personality types were given four-letter labels. The first letter is either *E* or *I* to indicate an introverted or extraverted orientation of the dominant cognitive function. The second letter is either *S* or *N* to indicate the dominant perception preference of sensation or intuition respectively, the third label is *T* or *F* to indicate the preferred judging cognitive function of either thinking or feeling, and the last letter is either *J* or *P* to indicate whether or not the preferred perception function has an introverted orientation (*J*) or an extroverted orientation (*P*). An example of a personality label from the use of this model would be ENTJ. This would indicate a personality who's dominant function is extroverted thinking with introverted intuition as the auxiliary perceiving function. The constraints imposed by Meyers & Briggs on the cognitive function order would then further specify that the tertiary function is extroverted sensing and the inferior function is introverted feeling.

There is a very widely held misconception that each of the four letters indicates a dimension of personality, like that found in the Big-5 model. It cannot be stressed enough that the four letters used to provide a personality label are in no way representative of dimensions. It is more appropriate the think of each grouping of four letters as a label. The Jungian model is not based on the concept of dimensions in any way. It is based on the concept of cognitive functions and the hierarchical preference with which people with different types of personality use them. Another common misconception is that the Jungian/Meyers & Briggs model implies a binary distinction between -for instance- thinkers and feelers, or judgers and perceivers. The common criticism is that the model implies that a person solely uses only one class of cognitive function or the other. For example, that a person is either a thinker or a feeler. Once again this is not how the theory works. The theory indicates that all people have access to use all the cognitive functions. It is just that people have different orders of preference for different functions. Alternatively, Berens and Nardi [4] explain the preference in terms of energy expenditure. They describe the use of a given cognitive function as requiring the expenditure of more or less energy depending on a person's personality type. Thinking of cognitive function use in terms of energy usage is a very convenient way to guide the selection of cognitive function to use in a given situation because it interfaces well with computational thinking on cost functions in optimization as well as with results in psychology that suggest that self-regulation relies on glucose/energy levels [9].

3. A FRAMEWORK FOR MAPPING JUNGIAN COGNITIVE FUNCTIONS TO ALGORITHMS FOR EMULATING PERSONALITY

The proposed Jungian type-based framework for artificial personality synthesis is based on three key principles.

1. Algorithms that have been developed to emulate cognition (e.g. Principle Components Analysis [10], Artificial Neural Networks, Linear Classifiers, etc.) can be mapped to the cognitive functions that make up the building blocks of the Jungian personality model. These algorithms can be used to implement the cognitive functions.
2. Human personality is inherently serial in nature. Human personality arises from the limitation that humans can only use one cognitive function at any given time. Or at least the human ability to use more than one function at a time appears to be severely limited.
3. The Jungian type-based personality framework allows for the possibility of an individual agent to use any of the 8 cognitive functions at any given time, however, personality emerges from a hierarchical preference for certain cognitive functions over others. The order and magnitude of preference can be selected based on models such as the Myer & Briggs cognitive function orders [4]. They could also possibly be learned from data, or could even be chosen arbitrarily. Initially the authors suggest using the Myers & Briggs cognitive function orders as guidance.

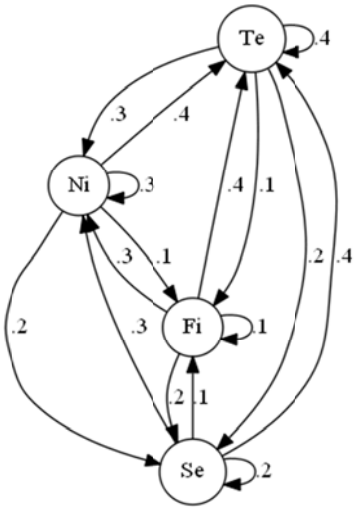


Figure 1 A Markov chain representative of an ENTJ. Cognitive function preference order: Te – Ni – Se – Fi.

Figure 1 shows an implementation of the Jungian type-based personality framework in the form of a Markov chain. The states of the Markov chain are occupied by cognitive functions. The values associated with the edges provide a probability that the artificial personality will transition to each of the alternative cognitive functions. In the case of a machine with an ENTJ personality, a proclivity will exist for the machine to remain in an extroverted thinking (Te) state with occasional transitioning to the intuitive intuition state. It is less common for the ENTJ machine to use extraverted sensing (Se) and it rarely if ever transitions to the introverted feeling (Fi) state. Personality is emulated by executing the algorithms that correspond to the current cognitive function.

It is worth noting that alternatives to the Markov chain representation could be envisioned. For instance, instead of probabilities, the proclivity to transition to another cognitive function could be couched in terms of energy costs associated with the use of a cognitive function for a given personality type as discussed in [4]. Also, please note that only 4 cognitive functions are shown in Figure 1. For completeness all 8 cognitive functions could be included. Only the first 4 functions were used as is specified by the classic Myers and Briggs framework for the sake of clarity. Lower order functions were not included. In initial implementations of this framework it might be preferable to only use the first 2 highest preference cognitive functions for the sake of simplicity. The reason being that the first two cognitive functions are all that is needed to distinguish one personality type from the other 16. Furthermore Jungians tend to believe the first two are the most important.

In order to implement the Jungian type-based personality framework, it is necessary to map algorithms that emulate cognitive processes to the cognitive functions. Table 1. provides a listing of the 8 cognitive functions, a short description of each of the cognitive functions, and a list of possible algorithms that could be used to emulate the cognitive function. This list is by no means exhaustive, and revisions are expected as artificial personality research progresses. The descriptions provided in Table 1. borrow heavily from [4] in order to maintain some uniformity in descriptions.

Table 1: Cognitive functions and associated algorithmic candidates for implementation.

Cognitive Function	Description [4],	Candidate Algorithms
Extroverted Thinking (Te)	Organizing people and things to work efficiently and productively. Organizes the environment and ideas.	• Partially Observable Markov Decision Process • A* • Optimization routines • Linear Programming
Introverted Thinking (Ti)	An internal sense of the essential qualities of something. Noticing the fine distinctions that make it what it is, internal reasoning process of deriving subcategories of sub-classes and sub-principles of general principles.	• Principle Components Analysis • Independent Components Analysis • Sparse Dictionary Learning • Auto Associative Neural Networks

Extroverted Feeling (Fe)	The desire to connect/disconnect with others. Causes response to expressed or unexpressed needs of others. Takes on the feelings of others – Empathy	• Use own embodiment as analog computation based on perception of external affect • Artificial Neural Networks • Resources allocated weighted towards communications/collaboration with other agents (human, machine and otherwise) • State of health evaluation of other agents • Cost functions for optimization designed in such a way that rewards associated with group success outweigh individual rewards associated with individual success • Analysis of how actions will affect group well-being
Introverted Feelings (Fi)	A filter for information that matches what is valued, wanted, or worth believing in. Continual assessment of a given situations with respect to individual values.	• Artificial Neural Networks • Techniques for state of health monitoring of self. • Cost functions for optimization designed in such a way that individual rewards for individual success outweigh group rewards associated with group success • Analysis of how actions will affect individual well-being
Extroverted Sensing (Se)	Use of the concrete senses to become aware of the physical world in detail. An impulse to act on information in order to get immediate results. Active seeking of information until sources of input are exhausted or attention is captured by	• Active learning • Active SLAM • Online Learning • Search based on maximum information gain • PID control
	alternative subject.	
Introverted Sensing (Si)	Storing experiences and information and comparing/contrasting the current situation with similar prior experiences. The similarities/differences are registered as important input.	• Supervised learning • Support Vector Machine Classification • Matched filtering • Narrowband filtering • Autocorrelation • Cross correlation
Extraverted Intuition (Ne)	Cross-contextual, divergent thinking. Generates and explores a host of possible interpretations from a single idea. The ability to entertain a variety of disparate ideas, beliefs and meanings simultaneously while maintaining the possibility that they are all true. Seeing things "as if."	• Search Engines (web) • Compressive Sampling • Random Walk • Genetic Algorithms
Introverted Intuition (Ni)	Lays out how the future might unfold based on unseen trends and signs. Can involve working out complex concepts or systems of thinking or conceiving of symbolic or novel ways to understand things that are universal.	• Simulation/Prediction • Design of Experiment • Autocomplete • System ID • Interpolation/Extrapolation • Bayesian Inference

The introverted and extroverted feeling judging functions merit additional discussion. The feeling judging functions are often associated with emotional response. The author currently likes to think of emotional decision-making in general to be similar to an artificial neural network in the sense that an artificial neural network can often take many complicated inputs and learn relationships between the inputs that can be used to quickly make decisions, however it is not always clear how exactly those decisions are made. The author also prefers to think of the emotional cognitive

functions as being partially the results of an embodied intelligence. Feeling judgements can be thought of as using the body as an analog computer to perform simulations and make decisions. Work by Nummenmaa [10] on mapping the sensation of emotions felt in the human body lends some support for this perspective.

As an example of the application of the framework, a machine endowed with an ENTJ personality using the Markov chain in Figure 1, we might employ the A* algorithm to perform the dominant extroverted thinking (Te) cognitive function to form a plan to move through an environment. After the plans are generated they might be analyzed in the introverted intuition (Ni) state using an appropriate simulation that corresponds to the environment and task of interest. If the simulation verifies that the plan is acceptable the plan may be executed by the extroverted Sensing function. In many cases the machine may totally ignore the introverted feeling state and proceed directly to execution, but when it does go into the introverted feeling state it may use a neural network trained to look out for the machine's own well-being to decide whether or not to actually execute the task based on whether or not the machine "feels," it will advance the machine's self-interests.

It is also worth noting that creating the ability to endow robots with personalities may also have application in human machine teaming applications. Much as diversity in teams of humans tends to lead to better results [4], endowing teams of robots with different personalities may also lead to more robust robotic teaming. For example, a team consisting of humans could be augmented with machines endowed with personalities different from the existing team members in order to enhance the overall team diversity. Teams consisting only of robots could be designed in such a way that individual team members are endowed with complimentary personalities, thus enhancing the overall robustness and performance of the machine team.

4. ADDRESSING SOME CONCERNS WITH THE JUNGIAN MODEL

Since the MBTI was developed more than 10,000 companies, 2500 colleges/universities and 200 U.S. government agencies have used the test [11]. It is estimated that more the 50 million people have used the instrument since 1962. Jungian personality theory has had a great influence on corporate America as well as popular culture. For instance, Carl Jung popularized the terms introvert and extrovert. Despite the popularity of Jungian personality theory in industry and among laypeople, academia tends to discount it. One common criticism against the MBTI is that it lacks test-retest reliability [12]. The current perspective of the author is that this criticism is probably valid. At this time the instrument itself appears to have problems. The reason for the lack of reliability with the current instruments may be that the instrument is based on the analysis of a self-report inventory. This type of instrument may be suitable if a Cartesian model of personality is used, but the Jungian model is better described as a dynamic system. Ultimately enhanced versions of projective tests such as those suggested by Ottley [13], and Brown's [14] work may be better able to characterize human personality. Another common criticism is that some Jungian advocates made the claim that MBTI score distributions assumed bimodal distributions, thus lending support to the misunderstanding of Jungian theory that people fell into one of two groups with respect to each letter in the Meyers-Briggs labels. Some research suggests this bimodal distribution was an artifact of the analysis techniques used [15]. The problem with the original argument was that it was not necessarily respecting the Jungian model as a dynamic system and was making the assumption that the Jungian model could be represented with a four dimensional Cartesian coordinate system. It is not clear what type of distribution a personality consisting of cognitive functions as building blocks should generate when evaluated using a self-report inventory. A central limit theorem argument could be made to suggest it come out as a Gaussian, but to really make a definitive statement a more rigorous analysis should be undertaken. Ultimately when synthesizing an artificial personality, the most important criteria for most applications is that the personality be convincing to humans. As long as the machine can pass a Turing-like test it is probably an acceptable solution. The fact that the Jungian model is so widely popular suggests that it may have a low barrier to acceptance among the majority of the population.

5. JUSTIFICATION FROM A NEUROSCIENCE PERSPECTIVE

A number of interesting results have come out of the psychology and neuroscience communities that lends some support for the idea of using the Jungian type model as outlined in this work. The idea that humans can effectively only use one cognitive function at a time gains support from the results of Watson [16] that suggest 98% of humans are incapable of multitasking. Jack's analysis of fMRI measurements of the human brain suggested that there are physiological constraints on our ability to simultaneously engage two distinct cognitive modes. In this case they found humans could not attend to tasks that require social cognition and physical reasoning simultaneously [17]. Grondin (2015) [18] found neuroanatomical differences between Agentic (achievement-oriented) and Affiliative (sociable) extroverts. From a Jungian perspective the concept of an agentic extroverts corresponds very well with personality types with a dominant extroverted thinking preference (ENTJ, ESTJ) and affiliative extroverts correspond strongly with personality types with a dominant extroverted feeling preference (ENFJ, ESFJ).

6. CONCLUSIONS

Personality from a Jungian perspective can be thought of as a zero sum game. Humans only have limited cognitive resources, and our personality is based on how we tend to choose to use those resources. This proposed framework is particularly attractive because it uses established algorithms as building blocks for personality. Because the building blocks are algorithms, and in some cases learning algorithms, the machine is ultimately able to learn and adapt to experiences. The personality framework provides a genotype so to speak but the ultimate phenotype of the machine depends on the experiences it encounters throughout its span of existence. This framework allows for great diversity in resulting perceived personality phenotype.

An interesting implication of the Jungian type-based personality framework is that it might help in the development of robust, high-performance human-machine teams consisting of members with diverse personalities. Observations on the personality diversity of teams and their

performance suggest that teams consisting of members with diverse personalities tend to perform better [19]. A team consisting of humans could be augmented by machines endowed with personalities that enhance the team's overall personality diversity. Alternatively, teams consisting solely of machines could consist of members who provide each other with different perspectives of the world they are interacting with.

Experience has shown that data science problems often benefit from the use of a combination of many heterogeneous models. The Netflix prize provides a good example of showing the advantages associated with simply combining different approaches [20]. However contemporary computational personality research is dangerously homogeneous in the sense that all computational personality research is currently using a trait-based paradigm [1]. The field would benefit from competing models and approaches. Furthermore, from an engineering/Turing test perspective it really does not matter whether or not the approach used to generate an artificial personality accurately models what is occurring in the human mind. It is only necessary to convince the person interacting with it that it is a human-like personality. In fact, the widespread popularity of Meyer's Briggs in business settings and with the general public suggests that an artificial personality based on the Jungian model may perform well with respect to convincing other people of the machine's personality.

Artificial personality synthesis is at a similar point in development as the airplane was at the time of the Wright Brothers. You do not need to understand how a bird flies in order to build a transatlantic aircraft, and you do not need to understand how the human mind works in order to build a machine that has a recognizable personality. Finally, it is interesting to note that while the Jungian perspective on artificial personality synthesis presented here is very different from the mainstream views on the topic that center on the use of Big-5 theory, this paradigm is not any more controversial than the Freudian paradigm advocated for by Marvin Minsky [21]. The Jungian paradigm presented in this work is worth consideration in future artificial personality synthesis. It is attractive in the sense that it can leverage a variety of existing algorithms to implement the personality. It also has the property that an artificial personality made with this framework should exhibit the full range of human personality as described by Jung.

7. ACKNOWLEDGMENTS

This work was completed under the support of a Los Alamos National Laboratory Early Career Research Award #20140629ECR. The concepts presented in this work were refined with the help of Sebastián A. Zanlongo.

8. REFERENCES

[1] A. Vinciarelli and G. Mohammadi, "A Survey of Persoanlity Computing," *IEEE Transactions on Affective Computing,* vol. 5, no. 3, pp. 273-291, June 2014.

[2] C. G. Jung, Psychologische Typen (Psychological Types), 1 ed., Zurich: Rascher Verlag, 1921.

[3] C. Jung, The Portable Jung, J. Cambell, Ed., New York: The Viking Press, 1976.

[4] L. V. Berens and D. Nardi, Understanding Yourself and Others: An Introduction to the Personality Type Code, Huntington Beach, CA: Telos Publications, August 1, 2004.

[5] The Myers and Briggs Foundation, [Online]. Available: http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/sensing-or-intuition.htm. [Accessed 1 June 2015].

[6] The Myers & Briggs Foundation, [Online]. Available: http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/thinking-or-feeling.htm. [Accessed 1 June 2015].

[7] The Meyers & Briggs Foundation, [Online]. Available: http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/isabel-briggs-myers.htm. [Accessed 1 June 2015].

[8] Center for Applications of Psychological Types, [Online]. Available: http://www.capt.org/mbti-assessment/isabel-myers.htm. [Accessed 1 June 2015].

[9] M. Gailliot, R. Baumeister, C. DeWall, J. Maner, E. Plant, D. Tice, L. Brewer and B. Schmeichel, "Self-control relies on glucose as a limited energy source: willpower is more than a metaphor," *J Pers Soc Psychol,* vol. 92, no. 2, pp. 325-336, February 2007.

[10] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine,* vol. 2, no. 11, pp. 559-572, 1901.

[11] L. Nummenmaa, E. Glerean, R. Hari and J. K. Hietanen, "Bodily maps of emotions," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, no. 2, pp. 646-651, November 27 2013.

[12] L. Cunningham, April 13, 2013. [Online]. Available: http://www.seattletimes.com/business/myers-briggs-personality-test-embraced-by-employers-not-all-psychologists/. [Accessed 1 June 2015].

[13] D. J. Pittenger, "Measuring the MBTI... And Coming Up Short," *Journal of Career Planning and Employment,* vol. 54, no. 1, pp. 48-52, 1993.

[14] A. Ottley, H. Yang and R. Chang, "Personality as a Predictor of User Strategy: How Locus of Control Affects Search Strategies on Tree Visualizations," in *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, Seoul, Korea, 2015.

[15] E. Brown, A. Ottley, J. Zhao, Q. Lin, A. Endert, R. Souvenir and R. Chang, "Finding Waldo: Learning about Users from their Interactions," *IEEE Transactions on Visualization and Computer Graphics (TVCG),* vol. 20, no. 14, December 2014.

[16] T. Bess and H. R.J., "Bimodal score distributions and the Meyers-Briggs Type Indicator: Fact or Artifact," *J Pers Assess,* vol. 78, no. 1, pp. 176-186, February 2002.

[17] J. Watson and D. Strayer, "Supertaskers: Profiles in extraordinary multitasking ability," *Psychonomic Bulletin & Review,* vol. 17, no. 4, pp. 479-485, 2010.

[18] A. I. Jack, A. Dawson, K. Begany, R. L. Leckie, K. Barry, A. Ciccia and A. Snyder, "fMRI reveals reciprocal inhibition between social and physical cognitive domains," *NeuroImage,* vol. 0, pp. 385-401, 2012.

[19] E. Grodin and T. L. White, "The neuroanatomical delineation of agentic and affiliative extraversion," *Cognitive, Affective, & Behavioral Neuroscience,* vol. 15, no. 2, pp. 321-334, June 2015.

[20] D. Wilde, "More Diverse Personalities Mean More Successful Teams," *ASME,* March 2011.

[21] E. V. Buskirk, "How the Netflix Prize was Won," *Wired,* 22 Sept 2009.

[22] M. Minsky, The Emotion Machine, Simon & Schuster, 2006.

A Comparative Analysis of Personality-Based Music Recommender Systems

Melissa Onori
Department of Engineering
Roma Tre University
Via della Vasca Navale, 79
00146 Rome, Italy
melissa.onori@gmail.com

Alessandro Micarelli
Department of Engineering
Roma Tre University
Via della Vasca Navale, 79
00146 Rome, Italy
micarel@dia.uniroma3.it

Giuseppe Sansonetti
Department of Engineering
Roma Tre University
Via della Vasca Navale, 79
00146 Rome, Italy
gsansone@dia.uniroma3.it

ABSTRACT

This article describes a preliminary study on considering information about the target user's personality in music recommender systems (MRSs). For this purpose, we devised and implemented four MRSs and evaluated them on a sample of real users and real-world datasets. Experimental results show that MRSs that rely on purely users' personality information are able to provide performance comparable with those of a state-of-the-art MRS, even better in terms of the diversity of the suggested items.

Keywords

Personality; music recommendation; evaluation

1. INTRODUCTION

Music plays an important role in entertainment and leisure of human beings. With the advent of Web 2.0, a huge amount of music content has been made available to millions of people around the world. This has provided new opportunities for researchers working on music information with the aim of creating new services that support navigation, discovery, sharing, and the development of online communities among users. Music recommender systems (MRSs) aim to predict what people like to listen to. A recent research field in music recommendation explores the possibility of harnessing information on the target user's personality in the recommendation process.

The goal of the research work described in this paper is to assess the potential benefits of such integration. To this end, we implemented and compared with each other different MRSs, three of them based on users' personality inferred from explicit and implicit feedbacks, and one that does not consider users' personality.

2. RELATED WORK

In the research literature, there exist several works that reveal how information about a user's personality can help infer her music preferences and contribute to a more accurate recommendation process [31]. Therefore, several noteworthy MRSs considering the active user's personality have been proposed. Among others, Ferwerda and Schedl [12] propose an approach where users' personality and emotional states are implicitly extracted by analyzing their microblogs on Twitter. The authors make use of the extraction techniques described by Golbeck [14] and Quercia *et al.* [30], also trying to combine them for better predictions. Hu and Pu [18] compare a personality test-based MRS with a classic rating-based one. The authors point out that users are more inclined to results returned from the former. According to Hu and Pu, the active user perceives less effort and less time to use the personality test-based MRS. They further claim that users show a strong intention to use such MRS again and an unexpected surprise in its results, as they feel that the personality-based approach is able to reveal their hidden preferences, thereby improving the recommendation process. Also Tkalčič *et al.* [34] show that recommenders based on Big Five data can outperform rating-based recommenders. In [19], Hu and Pu consider again their previous results, exploring the use of personality tests for creating psychological profiles of user's friends as well. They enable the MRS to generate recommendations for users and their friends too. They also suggest that personality-based MRSs are preferred by no music connoisseurs, which do not know their music preferences in depth.

3. PERSONALITY

Generally speaking, an individual's personality can be defined as a combination of characteristics and qualities that make up the way she thinks, feels, and behaves in different situations [33]. Personality and emotions shape our everyday life, having a strong influence on our tastes [32], decisions [29], purchases [6], and general behavior [7]. It has been shown that people with similar personalities turn out to have similar preferences [8]. However, giving a more rigorous definition of personality can be challenging, so different theories have been formulated to specifically make easier the comprehension of self and others [9]. Each of these theories differently addresses the problem of representing and characterizing the human personality. We are interested in theories that would allow us to differentiate people from each other through measurable traits. The subject of the psychology of personality traits is the study of the psychological differences between individuals and relies on empirical research.

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

Initially, it was studied and defined by Gordon W.Allport [1, 2], which specified 17953 specific traits to describe an individual's personality. Then, particular effort was devoted to the attempt of limiting the number of traits that would otherwise be unmanageable. This led to the definition of the well-known Big Five Model [11]. After several revisions, the Big Five factors were finally labeled as follows [24]:

- *Extraversion*;
- *Agreeableness*;
- *Conscientiousness*;
- *Neuroticism*;
- *Openness (to experience).*

In spite of several criticisms (e.g., [21]), such model is widely adopted in various fields, ranging from Medicine to Business. From the computer science point of view, personality traits include a set of human characteristics that can be modeled and implemented, for example, in personalized services. Prediction of personality traits can be accomplished explicitly (e.g., by administering personality tests), or implicitly (e.g., by monitoring the user's behavior).

3.1 Explicit Acquisition

Nowadays, questionnaires are the most popular method for extracting an individual's personality. They consist of a more or less large number of different questions, which are directly related to the granularity of the traits to be determined. Nunes *et al.* [28] show that the number of items influences the accuracy of measurements of traits. As expected, the higher the number of items, the more accurate the traits extracted. In particular, personality tests based on the Big Five Model are numerous and varied. A reasonable trade-off between accuracy and ease of use is represented by the Big Five Inventory (BFI) [4]. The 44-item BFI has been developed to create a brief questionnaire for efficient and flexible inference of the five factors, without the need to define more individual facets [21].

3.2 Implicit Acquisition

An individual's online behavior has long been the subject of many studies in the social sciences [3, 7, 25]. Results in cognitive psychology show that the general factors of personality can predict the aspects of the Internet use [25]. In fact, personality traits can be reflected in users' actions and ways of surfing the Web [3, 10, 27]. There are also studies that investigate the possibility of inferring the user's personality by user-generated content on social networks such as Facebook and Twitter. For instance, Gao *et al.* derive users' personality traits from their microblogs [13]. Golbeck *et al.* identify users' personality traits by analyzing their Facebook profiles, including peculiarities of language, business, and personal information [15]. Moreover, Golbeck *et al.* [14] and Quercia *et al.* [30] predict users' personality from Twitter, by examining their tweet content and observing their characteristics (e.g., popularity, influential users, etc.). Kosinski *et al.* [23] show that *likes* on Facebook can be used to automatically and accurately predict a set of personal attributes, including personality traits. For instance, the accuracy of prediction of the *Openness* factor is similar to the accuracy that can be obtained through a classic personality test, with the advantage of not having to force the user to answer a significant number of questions. Along this direction, the authors developed the Apply Magic Sauce (AMS)[1] that allows for the prediction of users' personality from the analysis of their activities on Facebook. Such application, developed at the University of Cambridge Psychometrics Centre, relies on over six million social media profiles and determines personality traits through psychometric evaluations, as described in [23]. The model is based on the dataset of myPersonality project [2].

4. USER STUDY

In this section, we describe the dataset, the setup, and the results of the experimental evaluation.

4.1 Dataset

The experimental tests were performed on the Last.fm [3] music listening data kindly provided by the researchers of myPersonality project [22] and Liam McNamara [26]. From this data, we extracted 1,875 Last.fm users with related information about personality tests and listening histories. The user's preferences were inferred from the *playcount* attribute, which denotes how many times the user listened to that particular song. The final value is obtained by normalizing it between 1 and 5.

4.2 Users

The users who took part in the experimental trials were 65, all of them with an active Facebook account. Their characteristics in terms of gender, age, occupation, and education are illustrated in Tables 1, 2, 3, 4, respectively.

Table 1: Gender

Female	Male
27	38

Table 2: Age

0-18	19-24	25-29	30-35	36-45	46-55	56-65
2	25	26	2	5	4	1

Table 3: Occupation

None	Student	Employee	Professional	Housewife
6	35	21	2	1

Table 4: Education

Primary	Secondary	Bachelor	Master	PhD
6	29	18	10	2

[1]http://applymagicsauce.com/
[2]www.mypersonality.org
[3]http://www.last.fm/

4.3 Setup

For presenting the user with the suggested playlists we designed a simple interface that allows for a quick and easy use of the system. Furthermore, we made us of the Spotify APIs [4], which offer a preview of 30 seconds of each song in the playlist. We deemed such time enough for the user to understand whether a given song is to her liking or not. Moreover, listening to the whole playlist is short, thus avoiding that the user will get bored and stop listening to the recommended songs. In this way, the user will be able to express a well-founded opinion.

Each user was required to test all MRSs and evaluate the returned playlists. MRSs were proposed in a random order and with the user completely unaware of their details. Ratings expressed by users in the evaluation phase were related to *novelty*, *serendipity*, *diversity*, *interest*, and *future use*. To this end, each user was asked to provide an assessment in relation to the following five statements:

1. "I found new songs by artists already known to me." *(novelty)*
2. "I found songs by artists that I did not know and, as of now, will begin to listen to." *(serendipity)*
3. "I found songs by artists of different music genres." *(diversity)*
4. "I found the suggested playlist interesting." *(interest)*
5. "I would use this MRS again in the future." *(future use)*

For each of these statements the user could express a numerical value in a Likert 5-point scale (i.e., 1: strongly disagree, 5: strongly agree). In addition, the user could leave a feedback as well.

4.4 Music Recommender Systems

In this section, we introduce the music recommender systems (MRSs) developed as part of our research work.

4.4.1 MRS based on Relations between Explicit Personality and Music Genres

The first MRS acquires information on the target user's personality explicitly, by administering a personality test. We chose the 44-item Big Five Inventory test introduced in Section 3.1, as its length represents an appropriate trade-off between compilation time and results accuracy. Such test is proposed to the target user through a web interface. Once the test is completed, the system analyzes the responses and computes the Big Five factors. In [8], the relations between users' personality types and their preferences in multiple entertainment domains are investigated. The authors derive a set of association rules that connect the Big Five factors with music genres. Based on those rules, this MRS returns the resulting playlist to the user.

4.4.2 MRS based on Explicit Personality and Neighbors

Even the second MRS relies on the user's personality explicitly inferred through the use of the questionnaire. The recommendation mechanism, however, is different. More precisely, this MRS identifies the most similar users to the target one within a dataset containing information related to personality and music habits of a group of Last.fm users. The user u's personality is compared to that of each user v in the dataset by computing the cosine similarity applied to the Big Five factors, which is defined as follows:

$$simp(u,v) = \frac{\sum_{k=1}^{5} p_u^k \times p_v^k}{\sqrt{\sum_{k=1}^{5}(p_u^k)^2}\sqrt{\sum_{k=1}^{5}(p_v^k)^2}} \quad (1)$$

where p_u^k expresses the value of the Big Five factor k of the user u. Based on such values, the system selects the ten Last.fm users most similar to the user u and generates a playlist from their listening histories.

4.4.3 MRS based on Implicit Personality and Neighbors

The implicit personality acquisition can be carried out by analyzing the user's behavior on the Web, especially on social networks. To this end, we used the APIs of the Apply Magic Sauce (AMS) application introduced in Section 3.2. In order to infer the user's personality, AMS analyzes how she assigns *likes* on Facebook. For such reason, the system allows users to login via Facebook. In this way, AMS enters the user profile, extracts the required information, and returns the predicted information, such as age, intelligence, life satisfaction, interest in specific areas, and her personality traits. Based on such features, the MRS identifies the most similar users to the target one within the Last.fm dataset by computing the similarity function 1. From the information related to the music such users listen to, the MRS builds the personalized playlist for the active user. However, this MRS has a drawback: it is necessary that the user has inserted a sufficient number of *likes* in her profile. Otherwise, the AMS application is not able to predict the user's personality and, as a result, the MRS is not able to deliver any playlist.

4.4.4 MRS based on Music Preferences

This MRS does not exploit information about the user's personality, and has been realized as a baseline to be used in the experimental evaluation. The recommender works as follows. The user is presented with a screenshot of the images of ten songs belonging to the Last.fm top track, and is asked to choose her favorites. Alternatively, the user can enter the title of some of her favorite songs. After that, the system leverages the Last.fm APIs to retrieve songs similar to those chosen by the user and includes them in the suggested playlist. Even though the actual algorithm underlying the Last.fm recommender is unknown, it is reasonable to assume that it mostly relies on collaborative filtering and tagging activity.

4.5 Results

Experimental results are shown in Table 5. In the description of the experimental results, the implemented MRSs are denoted as follows:

I: MRS based on relations between explicit personality and music genres;

II: MRS based on explicit personality and neighbors;

IIII: MRS based on implicit personality and neighbors;

IV: MRS based on music preferences.

[4]https://developer.spotify.com/web-api/

Table 5: Results in terms of mean and standard deviation of user ratings

MRS	# of Users	Novelty	Serendipity	Diversity	Interest	Future Use
I	65	2.5 - 1.0	2.5 - 0.8	3.0 - 0.9	3.0 - 0.8	3.4 - 0.8
II	65	2.4 - 0.9	2.6 - 0.8	2.8 - 0.8	3.2 - 0.7	3.3 - 0.8
III	43	2.2 - 0.7	2.2 - 0.6	3.2 - 0.9	2.4 - 0.7	2.8 - 0.9
IV	65	2.9 - 0.8	2.4 - 0.9	1.7 - 0.5	3.5 - 0.6	3.5 - 0.6

The reason for the smaller number of users who experienced the third MRS (i.e., the one based on implicit personality and neighbors) was that not all testers had a sufficient number of *likes* on Facebook to enable the AMS application to predict their personality. It can be noted that the first three systems received very similar assessments, as regards novelty, serendipity, and diversity. Precisely, novelty values are not high, because we asked users if new songs by known artists were in the suggested playlists, not if new artists were in the playlists. Serendipity shows similar values to novelty. Diversity values are quite high, which is obviously positive, since in this way the user can broaden her music knowledge, having a more varied set of possible music listening. The playlist was judged interesting for each system, a bit less for the third one. Users also showed some interest in the reuse of MRSs, a bit less for the third, where the result revealed some skepticism, due to the lower interest in the returned playlist. Different results emerged from the user evaluation of the last system. As expected, the results were higher than the others, except for serendipity (in line with the others) and diversity (lower). In fact, for such MRS the target user directly inserts her preferences. As a result, she was more interested in the suggested playlist and showed higher intention in reusing that recommender. These results may also be related to the difficulty that users appreciate new songs on the first listening and nourish curiosity in music genres different from their usual ones.

5. CONCLUSIONS AND FUTURE WORK

The research work presented here analyzed the effects of integrating the target user's personality in music recommender systems (MRSs). To this end, four different MRSs were developed. Three of them were only personality-based, the fourth did not take into account users' personality at all. The experimental results show that the personality-based ones had performance almost similar to that of a classic MRS. They also prove that the former ones are able to recommend songs with higher diversity than those suggested by the latter one.

This research effort is just beginning, so the possible future developments are manifold. Among others, the extension of the type and number of MRSs to be compared with each other, and the inclusion of music preferences and sentiments extracted from music reviews [16, 17] in the user model. Furthermore, as regards the experimental procedure, we intend to broaden the number of involved users and tested datasets, and to develop a layered evaluation for distinguishing the contributions of the user model from those of the user interface.

6. ACKNOWLEDGMENTS

The authors sincerely thank Michal Kosinski, David Stillwell of the myPersonality project, and Liam McNamara for kindly providing the datasets used in the experimental evaluation.

7. REFERENCES

[1] F. H. Allport and G. W. Allport. *Personality traits: their classification and measurement*. Yale University Press, 1921.

[2] G. W. Allport. Concepts of trait and personality. *Psychological Bulletin*, 24:284–293, 1927.

[3] Y. Amichai-Hamburger and G. Vinitzky. Social network use and personality. *Computers in Human Behaviour*, 26(6):1289–1295, 2010.

[4] V. Benet-Martinez and O. E. John. Los cinco grandes across cultures and ethnic groups: multitrait multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology*, pages 729–750, 1998.

[5] S. Berkovsky, E. Herder, P. Lops, and O. C. Santos, editors. *Late-Breaking Results, Project Papers and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization., Rome, Italy, June 10-14, 2013*, volume 997 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[6] C. Bologna, A. C. De Rosa, A. De Vivo, M. Gaeta, G. Sansonetti, and V. Viserta. Personality-based recommendation in e-commerce. In Berkovsky et al. [5].

[7] B. Caci, M. Cardaci, M. E. Tabacchi, and F. Scrima. Personality variables as predictors of facebook usage. *Psychological reports*, 113(2):528–539, 2014.

[8] I. Cantador, I. Fernández-Tobías, and A. Bellogín. Relating personality types with user preferences in multiple entertainment domains. In Berkovsky et al. [5].

[9] D. S. Cartwright. *Theories and Models of Personality*. W. C. Brown Company, 1979.

[10] F. Celli and L. Polonio. Relationship between personality and interactions in facebook. In X. M. Tu, A. M. White, and N. Lu, editors, *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, chapter 3, pages 41–53. Nova Science Publishers, Inc., 2013.

[11] P. T. Costa and R. R. McCrea. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, Inc., Odessa, Fla. P.O. Box 998, Odessa 33556, 1992.

[12] B. Ferwerda and M. Schedl. Enhancing music recommender systems with personality information and emotional states: A proposal. In I. Cantador, M. Chi, R. Farzan, and R. Jäschke, editors, *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling,*

Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014., volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[13] R. Gao, B. Hao, S. Bai, L. Li, A. Li, and T. Zhu. Improving user profile with personality traits predicted from social media content. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 355–358, New York, NY, USA, 2013. ACM.

[14] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011* [20], pages 149–156.

[15] J. Golbeck, C. Robles, and K. Turner. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 253–262, New York, NY, USA, 2011. ACM.

[16] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti. A sentiment-based approach to twitter user recommendation. In B. Mobasher, D. Jannach, W. Geyer, J. Freyne, A. Hotho, S. S. Anand, and I. Guy, editors, *Proceedings of the Fifth ACM RecSys Workshop on Recommender Systems and the Social Web co-located with the 7th ACM Conference on Recommender Systems (RecSys 2013), Hong Kong, China, October 13, 2013.*, volume 1066 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[17] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti. iscur: Interest and sentiment-based community detection for user recommendation on twitter. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, and G. Houben, editors, *User Modeling, Adaptation, and Personalization - 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings*, volume 8538 of *Lecture Notes in Computer Science*, pages 314–319. Springer, 2014.

[18] R. Hu and P. Pu. A comparative user study on rating vs. personality quiz based preference elicitation methods. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 367–372, New York, NY, USA, 2009. ACM.

[19] R. Hu and P. Pu. A study on user perception of personality-based recommender systems. In P. D. Bra, A. Kobsa, and D. N. Chin, editors, *User Modeling, Adaptation, and Personalization, 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings*, volume 6075 of *Lecture Notes in Computer Science*, pages 291–302. Springer, 2010.

[20] IEEE. *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*. IEEE, 2011.

[21] O. P. John and S. Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, New York, second edition, 1999.

[22] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543–556, Sept. 2015.

[23] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, Apr. 2013.

[24] R. R. McCrae and O. P. John. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2):175–215, 1992.

[25] J. C. Mcelroy, A. R. Hendrickson, A. M. Townsend, and S. M. Demarie. Dispositional factors in internet use: Personality versus cognitive style. *MIS Quarterly*, 31(4):809–820, 2007.

[26] L. McNamara, C. Mascolo, and L. Capra. Media sharing based on colocation prediction in urban transport. In *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, MobiCom '08, pages 58–69, New York, NY, USA, 2008. ACM.

[27] K. Moore and J. C. McElroy. The influence of personality on facebook usage, wall postings, and regret. *Comput. Hum. Behav.*, 28(1):267–274, Jan. 2012.

[28] M. A. Nunes, J. Bezerra, and A. de Oliveira. Personalityml: A markup language to standardize the user personality in recommender systems. *GEINTEC - Gestão, Inovação e Tecnologias*, 2(3):255–273, 2012.

[29] M. A. S. Nunes and R. Hu. Personality-based recommender systems: An overview. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 5–6, New York, NY, USA, 2012. ACM.

[30] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011* [20], pages 180–185.

[31] D. Rawlings and V. Ciancarelli. Music Preference and the Five-Factor Model of the NEO Personality Inventory. *Psychology of Music*, 25(2):120–132, 1997.

[32] P. J. Rentfrow and S. D. Gosling. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.

[33] D. L. Schacter, D. T. Gilbert, D. M. Wegner, and B. M. Hood. *Psychology: Second European Edition*. Worth Publishers. Palgrave Macmillan Limited, 2015.

[34] M. Tkalcic, M. Kunaver, J. Tasic, and A. Košir. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, pages 30–37, 2009.

Recommender System Incorporating User Personality Profile through Analysis of Written Reviews

Peter Potash
Department of Computer Science
University of Massachusetts Lowell
Lowell, Massachusetts
ppotash@cs.uml.edu

Anna Rumshisky
Department of Computer Science
University of Massachusetts Lowell
Lowell, Massachusetts
arum@cs.uml.edu

ABSTRACT

In this work we directly incorporate user personality profiles into the task of matrix factorization for predicting user ratings. Unlike previous work using personality in recommender systems, we use only the presence of written reviews by users. Other work that incorporates text directly into the recommendation framework focuses primarily on insights into products/categories, potentially disregarding important traits about the reviewers themselves. By using the reviews to determine the users' personalities directly, we can acquire key insights into understanding a user's taste. Our ability to create the personality profile is based on a supervised model trained on the MyPersonality dataset. Leveraging a set of linguistics features, we are able to create a predictive model for all Big 5 personality dimensions and apply it to the task of predicting personality dimensions for users in a different dataset. We use Kernelized Probabilistic Matrix Factorization to integrate the personality profile of the users as side-information. Lastly, we show the empirical effectiveness of using the MyPersonality dataset for predicting user ratings. Our results show that combining the personality model's raw linguistic features with the predicted personality scores provides the best performance. Furthermore, the personality scores alone outperform a dimensionality reduction of the linguistics features.

CCS Concepts

•**Human-centered computing** → **Collaborative filtering;** *Empirical studies in collaborative and social computing;* Social networks;

Keywords

Human-Centered Computing; Collaborative Filtering; Recommender Systems; Social Networks

1. INTRODUCTION

Recent work [20, 2, 1] has shown the effectiveness of incorporating user reviews into the matrix factorization framework. Unfortunately, the information derived from the reviews is primarily used to understand items/item categories, as opposed to users. Given that it is the users who provide the reviews, we believe that there could be important information about the reviewers lost in these methodologies. Even if the methodologies were modified slightly to glean insight into the users themselves, the representations learned by these methodologies still require manual inspection to fully understand their meaning. Alternatively, when it comes to understanding users, personality can be an important concept to leverage – the intersection of personality and linguistics dates back decades [8, 33, 14]. Given that personality is a well-researched topic, it is an interpretable aspect to attempt to derive from written reviews. Furthermore, we believe it can be effective side-information that can be used to produce more accurate predictions.

More specifically, we will use the MyPersonality dataset [18] to build a predictive model to attain the Big 5 Personality traits [13] for reviewers (users). The dataset provides status updates from Facebook users along with users' personality scores that are based on the users taking separate psychological tests. Thus, the personality scores in this dataset are grounded in proven psychological research. We will then take advantage of the Kernelized Probabilistic Matrix Factorization (KPMF) framework to incorporate the personality scores as side-information.

To further motivate the idea of personality profile as an added signal for user rating prediction, take as an example the following excerpts from two different movie reviews for the film 'Inception'. Both of the reviewers rated the movie 10 out of 10, but observe how each user begins his/her review. One reviewer writes:

> "My sister has been bothering me to see this movie for more than two months, and I am really glad that she did, because this movie was excellent, E-X-C-E-L-L-E-N-T, EXCELLENT!"

Whereas the other reviewer notes:

> "So far, Christopher Nolan has not disappointed me as a director, and 'Inception' is another good one."

While the two users have given the same numerical rating to the movie, we can obtain deeper insight into the users them-

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

selves by examining what they wrote. The first reviewer appears to be a more casual moviegoer, seeing movies people recommend, and finding pleasure in them. The second reviewer, in contrast, appears to be more of a movie aficionado. The reviewer immediately identifies who the director is, and indicates that he/she is familiar with the director's work. Such an analysis can indicate that their ratings for other items could diverge substantially.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work on matrix factorization, as well as at the intersection of recommender systems and natural language processing (NLP). Section 3 describes the KPMF methodology. In Section 4, we explain how the predictive model for the Big 5 personality traits was built, as well as how it is incorporated as the side-information format for KPMF. Section 5 describes our experimental design for predicting user ratings that incorporate personality. Finally, in Sectons 6 and 7, we present and discuss our results, as well as future research directions based on this work.

2. BACKGROUND

In this section, we will give a brief review of the history of recommender systems using matrix factorization over the course of the past decade, as well as then discuss examples of previous work where NLP methods have been used to create recommender systems.

2.1 Matrix Factorization Systems

The Netflix Challenge that commenced in 2006 marked a seminal event in the field of recommender systems. As [3] notes, The state-of-the art system that Netflix was using, Cinematch, was based on a nearest-neighbor technique. The system used an extension of Pearson's correlation, which the system produced by analyzing the ratings for each movie. The system then uses these correlation values to create neighborhoods for the movies. Finally, the system uses these correlations in multi-variate regression to produce the final rating prediction.

The team that ultimately took home the million dollar prize, however, relied on a fundamentally different technique: latent factors via matrix factorization [17]. Rather than calculating neighborhoods for items and/or users, matrix factorization models users and items as latent vectors. Stacking these vectors into two separate matrices, one for users and one for items, produces the latent matrices that represent users and items. The models predict ratings simply by taking the dot-product of the latent vectors of the user and item for which it is desired, or simply multiplying the two matrices to predict all ratings.

During the course of the Netflix Challenge, researchers developed probabilistic extensions of standard matrix factorization [26, 27] that could adapt well to large, sparse matrices that are generally representative of rating matrices. These models assume a generative process of probability distributions for the latent user/item vectors, as well as the ratings themselves. Our technique for rating prediction follows the methodology of KPMF, detailed by [36]. KPMF builds upon a probabilistic framework and we will explain the model in full detail in Section 3.

2.2 Recommender Systems and NLP

Various researchers have already completed NLP-related tasks in the overall goal of constructing an effective recommender system. [28] combines topic modeling on plot summaries with probabilistic matrix factorization to predict user ratings for movies. Their paper proposes an expanded generative process for rating prediction that can incorporate the models of Correlated Topic Modeling [5] and Latent Dirichlet Association [6]. In similar fashion, [35] combines topic modeling on the text of scientific article with probabilistic matrix factorization in the effort of recommending relevant articles/papers to researchers. In an example of a non-matrix factorization approach, [29] uses sentiment analysis on movie reviews for movie recommendations. Here, the researchers use a recommendation technique more akin to nearest-neighbors by defining a similarity measure among users and items based on how users rate items and how items are rated. Once the similarity is measured, the researchers use the result of the sentiment analysis to produce their final recommendations. In [10], the authors mine users' written reviews to understand both generalized and context-specific user preferences. These two aspects are then combined into a linear regression-based recommendation system. [11] provides a thorough presentation of the intersection between NLP and recommender systems.

In recent years, researchers have established methodologies that integrate the content of text reviews directly into the matrix factorization framework. In [20, 2], the authors fuse together topic modeling with matrix factorization, allowing models to learn representations of users and items, as well as topical distributions related to items and categories. More recently, in [1], the authors add the modeling of distributed language representations to the matrix factorization framework. This allows the authors to learn individual word representations as well as a general language model for the categories in their dataset.

The work that closely resembles ours is that of [25]. In their work, the authors create a personality-based recommender algorithm for recommending relevant online reviews. The authors train their personality model on a corpus of stream-of-consciousness essays, that include an accompanying personality score for each writer [24]. The authors, unfortunately, do not detail what accuracy their personality model scores on a supervised cross-validation of the dataset. Our own efforts to create a classification model from the same data using similar features produced an accuracy below 60%, which we do not deem accurate enough for use in further applications. Once the authors predicted the users' personalities, they clustered the results together in order to provide recommendations for users. While the approach is relevant, the authors are unable to test their recommendations against a gold-standard. Furthermore, in the effort of generating recommendations, matrix factorization has shown to be more accurate than nearest-neighbor approaches.

2.3 Recommender Systems with Personality

Aside from [25], several other researchers have integrated personality profiles into recommender systems. For example, [31] and [22] both use user personality profiles in the process of generating recommendations. However, the important difference between our work and the work of these researchers is that their methodology requires the explicit completion of personality tests by users. The researchers then derive personality scores directly from these tests. Such requirements make it inconceivable to use these systems in

a large-scale, applied nature. Our work is unique in the fact that we derive personality scores purely from an analysis of the users' written reviews. We require no further action from users aside from allowing them to express their opinion through ratings and reviews. Because of this, we contend that our methodology has the potential for large-scale application.

3. MATRIX FACTORIZATION

As we have previously mentioned, we use the technique of KPMF to incorporate the information that we generate by analyzing a given user's written reviews. What we generate from the analysis is a personality profile for a given user. We conjecture that by including this information of user personality in our model, we can ultimately produce more accurate movie ratings. We acknowledge that the choice of KPMF to incorporate side-information into the matrix factorization framework is somewhat arbitrary, and the work of [7, 15] could potentially be used instead.

3.1 KPMF

For the purpose of this paper we will explain the specifics of KPMF. To understand probabilistic matrix factorization in general and how KPMF is unique in this area, we encourage the reader to refer to the previously cited papers. In KPMF, we assume that the dimensions for the latent vectors representing items and users are drawn from a Gaussian Process (GP). Although in this GP we assume a zero mean function, it is the formulation of the covariance function that allows us to integrate side-information into our model. This covariance function – or covariance matrix in our application – dictates a 'similarity' across the the users and/or items. Our notation will follow the notation the original authors provided. Here is the notation we will use:

R — $N \times M$ data matrix
U — $N \times D$ latent matrix for rows of R
V — $M \times D$ latent matrix for columns of R
K_U — $N \times N$ covariance matrix for rows
K_V — $M \times M$ covariance matrix for columns
S_U — $N \times N$ inverse of K_U
S_V — $M \times M$ inverse of K_V
A — number of non-missing entries in R
$\delta_{n,m}$ — indicator variable for rating $R_{n,m}$

The generative process for KPMF is as follows (refer to Figure 1 for plate diagram):

1. Generate $U_{:,d} \sim GP(\mathbf{0}, K_U)$ for $d \in \{1,...,D\}$
2. Generate $V_{:,d} \sim GP(\mathbf{0}, K_V)$ for $d \in \{1,...,D\}$
3. For each non-missing entry $R_{n,m}$, generate $R_{n,m} \sim \mathcal{N}(U_{n,:} V^T_{m,:}, \sigma)$, where σ is constant

The likelihood of the data matrix R given U and V over the observed entries is:

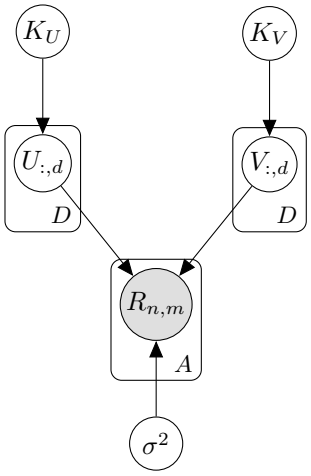


Figure 1: The generative process for KPMF.

$$p(R|U,V,\sigma^2) = \prod_{n=1}^{N} \prod_{m=1}^{M} [\mathcal{N}(R_{n,m}|U_{n,:} V^T_{m,:}, \sigma^2)]^{\delta_{n,m}} \quad (1)$$

Where the prior probabilities over U and V are:

$$p(U|K_U) = \prod_{d=1}^{D} GP(U_{:,d}|\mathbf{0}, K_U) \quad (2)$$

$$p(V|K_V) = \prod_{d=1}^{D} GP(V_{:,d}|\mathbf{0}, K_V) \quad (3)$$

Combining (1) with (2) and (3), the log-posterior over U and V becomes:

$$\begin{aligned} &log\ p(U,V|R,\sigma^2,K_U,K_V) \\ &= -\tfrac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \delta_{n,m}(R_{n,m} - U_{n,:} V^T_{m,:})^2 \\ &- \tfrac{1}{2} \sum_{d=1}^{D} U^T_{:,d} S_U U_{:,d} - \tfrac{1}{2} \sum_{d=1}^{D} V^T_{:,d} S_V V_{:,d} \\ &- Alog\sigma^2 - \tfrac{D}{2}(log|K_U| + log|K_U|) + C \end{aligned} \quad (4)$$

Where $|K|$ is the determinant of K and C is a constant that does not depend on U and V.

3.2 Learning KPMF

To learn the matrices U and V we can apply a MAP estimate to (4). The result is optimizing the following objective function:

$$\begin{aligned} E = &\tfrac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \delta_{n,m}(R_{n,m} - U_{n,:} V^T_{m,:})^2 \\ &+ \tfrac{1}{2} \sum_{d=1}^{D} U^T_{:,d} S_U U_{:,d} + \tfrac{1}{2} \sum_{d=1}^{D} V^T_{:,d} S_V V_{:,d} \end{aligned} \quad (5)$$

[36] provides implementations of both gradient descent and stochastic gradient descent to minimize E. For our experiments we used regular gradient descent, as gradient descent achieved the highest accuracy in the original work and our rating matrix is a manageable size. We will note that in the authors' work, the accuracy of stochastic gradient descent was less than that of regular gradient descent by only

a small margin and its speed was hundreds of times faster.

The partial derivatives for our objective function are the following:

$$\frac{\partial E}{\partial U_{n,d}} = -\frac{1}{\sigma^2} \sum_{m=1}^{M} (R_{n,m} - U_{n,:} V_{m,:}^T) V_{m,d} + \frac{1}{2} \boldsymbol{e}_{(n)}^T S_U U_{:,d} \quad (6)$$

$$\frac{\partial E}{\partial V_{m,d}} = -\frac{1}{\sigma^2} \sum_{n=1}^{N} (R_{n,m} - U_{n,:} V_{m,:}^T) U_{n,d} + \frac{1}{2} \boldsymbol{e}_{(m)}^T S_V V_{:,d} \quad (7)$$

where $\boldsymbol{e}_{(n)}$ represents an N-dimensional vector of all zeros except for the n^{th} index, which is one.

The update equations for U and V are as follows:

$$U_{n,d}^{t+1} = U_{n,d}^t - \eta(\frac{\partial E}{\partial U_{n,d}}) \quad (8)$$

$$V_{m,d}^{t+1} = V_{m,d}^t - \eta(\frac{\partial E}{\partial V_{m,d}}) \quad (9)$$

where η is the learning rate of the algorithm.

This completes our detailing of KPMF. In the next section we describe our approach for creating the covariance matrix for the users, K_U.

4. CREATING PERSONALITY PROFILES

Since we are using KPMF as our recommendation model, any vector representation of the written reviews (for a given user, across all users) would suffice to create K_U. However, it is best to generate covariance across a numeric representation that we can interpret. Since personality scores have a long history of analysis, which we will detail in this section, personality profiles are an optimal representation for K_U. In this section we cover two topics: first, how we create the personality profile for a given user. Second, how we use this personality profile to generate the user covariance matrix.

4.1 MyPersonality

In 2013, [9] held a workshop on computational personality recognition. For this workshop, the organizers released a subset of the data collected by the MyPersonality project [18]. The dataset for the workshop consists of the Facebook activity for 250 users, roughly 10,000 status updates from all users. Along with the status updates, the dataset includes information about the users' social networks. For each user, the dataset includes a personality score as well as a binary classification as to whether the user exhibits a given personality trait. The personality scores/classifications for each user have five dimensions, one for each trait in the Big 5 personality model. The five traits in the model are openness, conscientiousness, extraversion, agreeableness, and neuroticism. Analysis of lexicon and personality has a long-standing tradition [8, 33, 13], and it is [14] who brought the current model to prominence.

The approaches to the dataset in the workshop are varied. [32] focus on predicting a single personality trait, conscientiousness. The authors exploit an analysis of event-based verbs in the status updates to produce features for their model. [34] create an ensemble model for predicting personality traits. In their base model, the authors use most frequent trigrams as features. The authors then use the prediction of the baseline model to generate their final predictions. [12] and [19] have similar approaches: using a general textual analysis combined with social network attributes to create features for their predictive models. However, Markoviki et al. report a higher precision/recall for their model, so we will use their approach to feature selection as the guide for our model for personality prediction.

4.2 Personality Model

In their paper, Markoviki et al. detail a fined-grained feature selection for each personality trait, including social network features. Since, for our recommendation experiment, we will not have social network information, we do not include these features in our model. While most authors who used the MyPersonality data sought to create a classification model for personality prediction, we will predict personality score. We believe having a continuous output from our model will make for a better translation into user covariance. Based on an analysis of correlation between features and personality traits in Markoviki et al., we use the following features in our personality model (and we encourage a review of the original work for a thorough discussion of the effectiveness of these features):

Punctuation Count: We count the frequency of the following punctuation marks in a user's status updates: . ? ! - , <> / ; : [ ] { } ( ) & ' " ?

POS Count: We count the frequency of verbs and adjectives appearing in a user's status updates. We used the POS tagger available in NLTK [4].

Affin Count: We count the frequency of words appearing in a user's status updates that have an emotional valence score between -5 and 5 [21].

"To" Count: We count the number of times the word "to" appears in a user's status updates.

General Inquirer Tags: We process the text using the General Inquirer (GI) tool [30]. This tool has 182 categories for tagging words in a text. We use the frequency of these tags for our feature set.

While Markovikij et al. produced their best results when using a different subset of the GI tags for each personality trait, as well as Affin words only of a particular score, we did not find that this fine-grained breakdown produced the best results for our own experiments. Instead, we use the same feature space for all the personality traits, which included all GI tags and all words with any recorded Affin score. Lastly, all count features are normalized by the total word count (for a given user), and punctuation count is normalized by the total character count.

The personality scores are in a continuous range from 1 to 5 for users in the MyPersonality dataset . Thus, linear regression is a natural choice to train our model. We use the Ridge Regression algorithm available from scikit-learn [23]. Ridge Regression implements standard linear regression with a regularization parameter. The optimization task

is:

$$\min_w \|Xw - y\|_2{}^2 + \alpha\|w\|_2{}^2 \quad (10)$$

Where w is the weight vector, X is the data matrix, y is the vector of scores and α is the regularization parameter. The algorithm in scikit-learn performs automatic cross-validation on the regularization parameter by allowing us to define a list of α's for the input. While the feature space for each personality trait is the same, we train a different model for each trait. To be clear, we are not testing the personality of a single status update, but rather of a given user, which is the amalgamation of his/her status updates.

To test the utility of our models, we divide the set of Facebook users into a 80%/20% training/test split. Also, we normalize the matrices we use in our models by, for each feature dimension, subtracting the mean and dividing by the standard deviation. We randomly shuffle the set of users and record the root-mean-square error (RMSE) of the resulting trained model on the held-out test set. That is, given a predicted personality score for user i, $\hat{y_i}$, and the true personality score y_i, we calculate the RMSE of all users in the test set. Table 1 shows the accuracy of our model averaged across 5 different times shuffling the dataset. This model is compared to a baseline, which is the average user rating for personality scores in the training set. When creating the models that we will apply to predicting personality traits from movie reviews, we included all the Facebook users when training the models.

Personality Trait	Model	Baseline
Extraversion	0.785	0.833
Neuroticism	0.738	0.767
Agreeableness	0.635	0.661
Conscientiousness	0.767	0.799
Openness	0.529	0.563

Table 1: RMSE for personality model trained on Facebook statuses, as well as baseline model.

4.3 User Covariance Matrix

Once we have trained the personality models on the Facebook data we apply it to the movie reviews written by a given user to determine his/her personality profile. We preprocess the movie reviews just as we did for the Facebook data to create the same feature space. The result is a 5-dimensional vector, which we will denote p_i, for user i. For users i and j, we calculate entry i, j of K_U as follows:

$$K_{U_{i,j}} = \frac{CS(p_i,p_j)-\alpha}{\beta-\alpha} * \gamma \quad (11)$$

Where $CS(x, y)$ denotes the cosine similarity between vectors x, y, calculated as follows:

$$CS(x,y) = \frac{xy^T}{\|x\|\|y\|} \quad (12)$$

α and β are minimum and maximum values from our computed cosine similarities, across all possible user pairs:

$$\alpha = \min_{i,j} CS(p_i, p_j) \quad (13)$$

$$\beta = \max_{i,j} CS(p_i, p_j) \quad (14)$$

γ controls the ceiling of the normalization: $K_{U_{i,j}} \in [0, \gamma]$. We set $\gamma = 0.4$. To compute cosine similarity we use the cosine similarity method provided in scikit-learn. Note β will always be 1, as $CS(p_i, p_i) = 1$.

This, however, is not the final covariance matrix we will use in our recommender system. Since all the personality scores are in the range [1,5], the cosine similarity between personality vectors p_i and p_j is very close to one. To accentuate the differences in personality profile, we create a regularized covariance matrix, $\overline{K_U}$, as follows:

$$\overline{K_U} = {K_U}^n \quad (15)$$

Where n is a hyperparameter we hand-tune. The proper value of n can greatly influence the accuracy of the model. We take $\overline{K_U}$ as the covariance matrix in our experiment when we use personality profiles to produce the user covariance matrix, but we still refer to it as K_U to avoid confusion.

5. EXPERIMENTAL DESIGN

Our goal is to integrate the information contained in the reviews written by a user into a recommender system, and in particular, investigate whether user personality, as reflected in the text generated by that user, would allow us to improve the accuracy of predicted ratings. We crawled IMDB to collect a dataset of scores and written reviews for multiple IMDB users. Our dataset consists of 2,087 users and 3,500 movies. Each user has rated/reviewed as little as 4 movies and as many as 210, with 54 being the average number of ratings/reviews for the users. The total rating matrix is 1.55% dense, which reflects the typical sparsity of this type of dataset [16].

We randomly split the ratings by each user into training, evaluation, and test sets, each comprising 3/5, 1/5 and 1/5 of the data, respectively. We randomly shuffle the full set of ratings to produce five different training/evaluation/test splits, and report the results averaged over five runs. We use the ratings from these sets to create the appropriate matrices in our methodology. The training matrix is equivalent to R in our notation.

In all the experiments, we use a diagonal item covariance matrix, K_V. Thus, in our model, we are not assuming any covariance across items. Following the results of Zhou et al. we let $D = 10$ and $\sigma = 0.4$. We use gradient descent to learn the latent matrices U and V. We use the proportional change in RMSE on our evaluation matrix as the stopping criteria for gradient descent. Once the algorithm converges, we calculate the RMSE on our test matrix. When calculating RMSE, we only do so for non-zero entries, i.e. $\delta_{n,m} = 1$.

6. RESULTS

For each run, we train five different models and calculate their RMSE on a held-out test set: (1) KPMF with K_U calculated according to user personality profile, (2) KPMF with K_U calculated using a user's text-generated feature space for (10) as our p vector in equation (11), (3) KPMF with K_U

as a diagonal matrix (no similarity across users), (4) matrix factorization (MF) without trying to optimize U and V according to an objective function, and (5) KPMF with a PCA-reduction of the text-based feature space as p. Aside from providing a tangible vector representation of user reviews, the Big 5 personality model also acts as a guided dimensionality reduction of the textual feature space we use to generate personality scores. Therefore, we have compared the 5-dimensional output of our personality model to the result of using PCA to compute a reduction of the text-based feature space to 5 dimensions. We used the PCA implementation from scikit-learn. The RMSE values averaged over five runs for each model are shown in Table 2. For the purposes of RMSE calculation, the rating values in our data, which were originally 1-10, have been normalized to fall in the interval $[0.1, 1]$.

Model	RMSE
KPMF with Personality	0.2006
KPMF with Personality Model Features	0.1980
KPMF Personality *and* Model Features	**0.1901**
KPMF with Diagonal Matrix	0.2122
KPMF with PCA Feature Reduction	0.2087
MF	0.2262

Table 2: RMSE predicting user ratings.

7. DISCUSSION

As we expected, the KPMF models performed better than the non-optimized MF model, lowering the RMSE by 16.0%, 12.5%, 11.3%, 7.7% and 6.2% respectively. Comparing the KPMF models together, the personality model improves upon the diagonal model by 5.5%, however we see that a more accurate model is achieved by applying the textual personality features directly, and the most effective model uses a combination of the textual features and the predicted personality scores. It is important to note the percent difference along with RMSE, especially when the baseline metric performs well. When comparing the two models of 'dimensionality reduction', the personality model performs better than the PCA-model. This would dictate that the personality scores do capture a stronger signal of user similarity, as opposed to an arbitrary reduction of the raw text features. The the personality scores on their own do not perform as well as the raw textual features. We will discuss shortly a major added benefit for using personality scores, aside from testing accuracy.

One immediate question that arises is whether a more accurate personality predictive model actually does correlate to a more accurate KPMF model when using the personality profile. While our personality predictive model scores reasonably well, it is inconsistent across the personality traits. Future work can have a renewed focus on the MyPersonality data now that the recommendation framework has a solid foundation. Furthermore, as we have previously stated, representing users as personality profiles provides a gateway to a number of interesting analyses relating personality to product recommendation. For example, in our current recommendation model, each personality trait is given equal weight when we use the personality model to generate the covariance matrix. However, it is interesting to imagine a model where each personality trait should be weighted differently. For example, similarity in user conscientiousness might be more important than similarity in user agreeableness when determining overall similarity in user preference. We can create a new variable Q, a 5-by-5 diagonal matrix where each entry $Q_{i,i}$ is the weight for a given personality trait. If we stack the personality vectors to form a $M \times 5$ matrix P, the covariance matrix K_U becomes:

$$K_U = PQP^T \tag{16}$$

We can learn the diagonal entries of Q along with U and V in our model. The final values of Q would provide a novel outcome as to how important each personality trait is for predicting movie ratings. We leave this approach for future work.

8. REFERENCES

[1] A. Almahairi, K. Kastner, K. Cho, and A. Courville. Learning distributed representations from reviews for collaborative filtering. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 147–154. ACM, 2015.

[2] Y. Bao, H. Fang, and J. Zhang. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *AAAI*, pages 2–8, 2014.

[3] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007.

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[5] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[7] G. Bouchard, D. Yin, and S. Guo. Convex collective matrix factorization. In *AISTATS*, volume 13, pages 144–152, 2013.

[8] R. B. Cattell. Personality and motivation structure and measurement. 1957.

[9] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski. Workshop on computational personality recognition: Shared task. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[10] G. Chen and L. Chen. Augmenting service recommender systems by incorporating contextual opinions from user reviews. *User Modeling and User-Adapted Interaction*, 25(3):295–329, 2015.

[11] L. Chen, G. Chen, and F. Wang. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, 25(2):99–154, 2015.

[12] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock. Recognising personality traits using facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*, 2013.

[13] L. R. Goldberg. Language and individual differences: The search for universals in personality lexicons.

Review of personality and social psychology, 2(1):141–165, 1981.

[14] L. R. Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

[15] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang. Consistent collective matrix completion under joint low rank structure. In *AISTATS*, 2015.

[16] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[17] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[18] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[19] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell. Mining facebook data for predictive personality modeling. In *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013)*, 2013.

[20] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[21] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[22] M. A. S. N. Nunes. *Recommender systems based on personality traits*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc, 2008.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[25] A. Roshchina, J. Cardiff, and P. Rosso. A comparative evaluation of personality estimation algorithms for the twin recommender system. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 11–18. ACM, 2011.

[26] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, volume 1, pages 2–1, 2007.

[27] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.

[28] H. Shan and A. Banerjee. Generalized probabilistic matrix factorizations for collaborative filtering. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 1025–1030. IEEE, 2010.

[29] V. K. Singh, M. Mukherjee, and G. K. Mehta. Combining collaborative filtering and sentiment classification for improved movie recommendations. In *Multi-disciplinary Trends in Artificial Intelligence*, pages 38–50. Springer, 2011.

[30] P. J. Stone, D. C. Dunphy, and M. S. Smith. The general inquirer: A computer approach to content analysis. 1966.

[31] M. Tkalcic, M. Kunaver, J. Tasic, and A. Košir. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, pages 30–37, 2009.

[32] M. T. Tomlinson, D. Hinote, and D. B. Bracewell. Predicting conscientiousness through semantic analysis of facebook posts. *Proceedings of WCPR*, 2013.

[33] E. C. Tupes and R. E. Christal. Recurrent personality factors based on trait ratings. Technical report, DTIC Document, 1961.

[34] B. Verhoeven, W. Daelemans, and T. De Smedt. Ensemble methods for personality recognition. In *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, pages 35–38, 2013.

[35] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[36] T. Zhou, H. Shan, A. Banerjee, and G. Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, volume 12, pages 403–414. SIAM, 2012.