Affective Computing and Bandits: Capturing Context in Cold Start Situations

Sebastian Oehme
Munich School of Engineering
Technical University of Munich
Garching, Germany
sebastian.oehme@tum.de

Linus W. Dietz
Department of Informatics
Technical University of Munich
Garching, Germany
linus.dietz@tum.de

ABSTRACT

The cold start problem describes the initial phase of a collaborative recommender where the quality of recommendation is low due to an insufficient number of ratings. Overcoming this is crucial because the system's adoption will be impeded by low recommendation quality. In this paper, we propose capturing context via computer vision to improve recommender systems in the cold start phase. Computer vision algorithms can derive stereotypes such as gender or age, but also the user's emotions without explicit interaction. We present an approach based on the statistical framework of bandit algorithms to incorporate stereotypic information and affective reactions into the recommendation. In a preliminary evaluation in a lab study with 21 participants, we already observe an improvement of the number of positive ratings. Furthermore, we report additional findings of experimenting with affective computing for recommender systems.

KEYWORDS

Recommender systems, affective computing, bandit algorithms

ACM Reference Format:
Sebastian Oehme and Linus W. Dietz. 2018. Affective Computing and Bandits: Capturing Context in Cold Start Situations. In *Proceedings of IntRS Workshop, October 2018 (IntRS'18).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Recommender systems (RS) match items to users, therefore the accuracy of recommendations is highly dependent on the quality of information the system has about these. Collaborative filtering (CF) has frequently been used if the items' characteristics are unknown or it is costly to derive them. CF systems are, however, not suited for scenarios where the user is anonymous and interacts with the RS only for a short period. For example, a smart display inside a fashion store could provide recommendations, however, the interaction will be brief and tentative. In such cold start scenarios, literature suggests including context and stereotypes into the recommendations [1]. If the weather is hot, suggest bathing attire; a male customer will need shorts instead of a bikini. Motivated by this kind of a scenario, we develop an affective RS [13] based on stereotypes derived via computer vision with little user collaboration. Our research was guided by the following questions:

RQ 1: How can stereotypic information be incorporated into a RS?
RQ 2: Can facial classification and affective reactions be a surrogate for explicit feedback?

In the following section, we describe the foundations of our RS: bandit strategies and facial classification using computer vision. Then, an in-depth description of the proposed approach and a preliminary evaluation in a user study follow in Section 3. Finally, we draw our conclusions and point out future work.

2 FOUNDATIONS

Ever since *Grundy* [10], it has been known that using stereotypic information can be used to model users [2] and thereby improve recommendation accuracy. Driven by our research questions, we discuss a combination of two concepts applied for recommender systems: contextual bandits and facial classification using computer vision.

2.1 Bandit Strategies

In real-world applications, recommendations are often linked to a reward. For example, the purpose of recommendations in a shop is to improve revenue by suggesting products to customers that they are more likely to buy. However, calculating the probabilities of a successful recommendation directly is usually not possible due to a lack of information about the customer's taste and the attractiveness of items.

Bandit strategies provide a computational framework that trades off profit-maximization via items that are known to sell well and experimentation with items whose potential is yet to be determined. The terminology stems from the probability theory of gambling [12]. A gambler at a row of one-armed bandits (slot machines) has to decide based on incomplete knowledge: what arm to play, how often to pull and when to play [6]. A bandit recommender engine seeks to find the right balance between experimenting with new recommendations, i.e., *exploration*, and *exploiting* items that are already known to have a high chance of reward. A classic algorithm for handling exploration vs. exploitation is the ε-Greedy algorithm [11]. It chooses with a probability of ε to either exploit the best available arm at the moment or to randomly explore any other arm. In cold start situations, however, a bandit recommender suffers similar limitations as traditional methods, such as collaborative filtering. This can be overcome by adding context information, e.g., demographic information [8] to augment the bandit's choice between exploration or exploitation with more data. These types of bandit strategies are referred to as *contextual bandits*. In contrast to the ε-Greedy algorithm, they incorporate contextual information and are able to choose their action based on the situation. The classic algorithm is the Contextual-ε-Greedy strategy [3]. At each turn, it compares the user's situation (e.g., location, time, social activity) to a set of high-level 'critical situations'. If the situation is critical, the algorithm exploits this by showing items that are known to be well suited and similar. Consequently, it explores other items if

the situation is not critical. It has been shown that the Contextual-ε-Greedy algorithm generally achieves better click-through rates than ε-Greedy algorithms or pure exploration.

In our approach, we propose using facial classification through the use of computer vision to infer age, gender and emotions as contextual information within a contextual bandit algorithm.

2.2 Facial Classification

Computer vision has already been used to improve systems situated in public places. For example, Müller et al. [5] described a system for digital signage. However, this and similar early approaches were ahead of their time: due to low face-detection accuracy, the outcomes of these experiments were not significant. Computer vision-based approaches analyze users' faces frame by frame via facial recognition software during an experimental task such as watching videos. Zhao et al. [15] drew affective cues from users' affection changes. They used emotional changes to segment videos, classified the video's category and then presented recommendations. Tkalčič et al. propose a framework for affective recommender systems, where they distinguish between three phases of user interaction: the entry, consumption, and exit stage [13]. The affective cues drawn while watching content in the consumption stage are compared to the emotional state in the entry phase. The exit stage can simultaneously be the following entry stage when the next item is recommended and the looped process continues. Affective labeling of users' faces has been applied e.g., to RSs [14] and commercials [4], where they show promising results in terms of accuracy and user satisfaction.

The accuracy of classification and the runtime performance of computer vision algorithms have improved over the past years and with YOLO [9], the breakthrough to real time object detection has been achieved. In emotion detection, the state-of-the-art algorithms are closed source and only available using web APIs. Prominent vendors like *Microsoft Face*[1], *Kairos*[2] and *Affectiva*[3] offer RESTful client libraries and respective pricing models. The centralization of this technology to few market players that cloak their algorithms in secrecy should be seen with concern. Nevertheless, it should also be mentioned that such systems improve with the size of the training set and enable researchers to work with this technology without hardware requirements. In our recommender system, we use the *Microsoft Face* service to detect the age, gender and emotions of our test subjects. The *Face Emotion Recognition API* returns continuous values [0;1] for the following emotions: *anger*, *contempt*, *disgust*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise* at a small cost of about €1.40 per 1000 requests.

3 CONTEXTUAL RECOMMENDER MODEL

In our RS, the items are displayed to the user successively. While the user inspects the items, she is observed by a camera whose imagery is continuously analyzed by computer vision. In this section, we first present how we incorporated computer vision into the recommendation task, followed by the experimental setup and our findings.

[1]https://azure.microsoft.com/en-us/services/cognitive-services/face/
[2]https://www.kairos.com/emotion-analysis-api
[3]https://www.affectiva.com/product/emotion-sdk/

Our model extends the approach of Bouneffouf et al. [3] and likewise proceeds in discrete trials $t = 1 \dots T$. At each t, the following tasks are performed:

Task 1: Let U^t be the current user's profile and P the set of other known user profiles. The system compares U^t with the user profiles in P in order to choose the most similar one, U^P:

$$U^P = \underset{U^c \in P}{\operatorname{argmax}}(sim(U^t, U^c)) \tag{1}$$

Our adapted similarity metric is the weighted sum of the similarity metrics for age, gender, and EF, the combination of emotions and feedback. α, β, γ are weights associated with these metrics, defined in the following subsection:

$$sim(U^t, U^c) = \alpha \cdot sim(a^t, a^c) + \beta \cdot sim(g^t, g^c) + \gamma \cdot EF \tag{2}$$

EF, short for emotional feedback, corresponds to the sum of k affective reactions $sim_k(e_k^t, e_k^c) \in [0, 1]$ depending on equal feedback $sim_k(f_k^t, f_k^c) \in \{0, 1\}$ of the current user with respect to other users' profiles. This feedback, called *reward* in the bandit terminology, can be any explicit or implicit feedback to the item, e.g., the user's rating or adding the item to the shopping basket. If the feedback differs for an item, this item's affective reaction will not contribute to the sum, hence it will be 0. EF is normalized to the number of items i which U^t has seen so far.

$$EF = \frac{\sum_k sim_k(f_k^t, f_k^c) \cdot \left(1 + sim_k(e_k^t, e_k^c)\right)}{2i} \tag{3}$$

Task 2: Let M be the set of items, M_t the items seen by the current user U^t and $M_P \in \{M \backslash M_t\}$ the items recommended to the user U^P, but not to U^t. After retrieving M_P, the system displays the next item $m \in M_P$ to U^t while observing the user's affective reactions during presentation.

Task 3: After receiving the user's reward, the algorithm refines its item selection strategy with the new observation: user U^P gives item m_P a binary reward. The expected reward for an item is the average reward per total number of ratings n.

Our adapted Contextual-ε-Greedy recommends items as follows:

$$m = \begin{cases} \underset{M_P}{\operatorname{argmax}}\,(\text{expectedReward}(m)) & \text{if } q > \varepsilon \\ \text{random}\,((M \setminus M_t)) & \text{otherwise} \end{cases} \tag{4}$$

In Equation 4, the random variable q is responsible for the exploration versus exploitation behavior. In our approach it is uniformly distributed over [0,1]. If q is larger than ε, the item with the highest expected reward from $M_P = \{m_1, \dots, m_P\}$ will be selected, which are all items rated by the most similar user. For this at least one unseen and positively rated item by the past user is required. In case all suitable items have been exploited or the current user is the first user and hence no other user profiles exist, the algorithms falls back to exploration, where $random(M)$ selects a random item.

To influence the original ε-Greedy algorithm with contextual information, ε is computed by maximizing Equation 2, the similarity of the current user's profile U^t to the profile U^P of the most similar other user:

$$\varepsilon = 1 - \operatorname{argmax}(sim(U^t, U^c)) \tag{5}$$

3.1 Similarity Measures

The Contextual-ε-Greedy strategy is driven by the stereotypic similarity of the current user to previously seen users. In this first experiment, we used $\alpha = \beta = 0.25$ and $\gamma = 0.5$ as weights for Equation 2.

Gender similarity is binary, due to output of the employed facial classification algorithm. Either it matches, or it does not: $sim(g^t, g^c) \in \{0, 1\}$.

Age similarity is more fuzzy and we have not found an established similarity measure in literature. Therefore, we constructed an ad-hoc similarity measure $sim(a^t, a^c) \in [0, 1]$, which considers age differences of up to 15 years as somewhat similar [7].

Emotional similarity measures the affective response to a displayed item in comparison to the emotional reaction of previous users to it. As previously mentioned, today's computer vision algorithms are capable of detecting several emotions at once. Therefore, it is calculated by the cosine similarity of two emotion vectors, as can be seen in Equation 6.

$$sim(e^t, e^c) = \frac{\sum_{i=1}^{n} \bar{e}_i^t \cdot \bar{e}_i^c}{\sqrt{\sum_{i=1}^{n} (\bar{e}_i^t)^2} \sqrt{\sum_{i=1}^{n} (\bar{e}_i^c)^2}} \tag{6}$$

3.2 Capturing Affective Cues

Microsoft Face analyzes the user's face for age, gender, and up to eight emotions. Experimenting with the computer vision service before the main experiment showed that users tend to express their emotional reactions shortly before requesting the next item and maintain their facial expression for some time when the next item is already shown. We call this *'overflowing emotions'*, as the user's emotional reaction to the previous item overflows to the current item and is then adjusted during the consumption and exit stage. Since we are interested in the actual response to the item *after* the content has been processed, we used the following weighted average over all analyzed frames n as the aggregated metric to emphasize the emotions from the exit stage.

$$\bar{e} = \frac{\sum_{i=1}^{n} 2^i \cdot e_i}{\sum_{i=1}^{n} 2^i} \tag{7}$$

Figure 1 shows the comparison of the mean value to our proposed weighted average. Over the course of three items, the level of

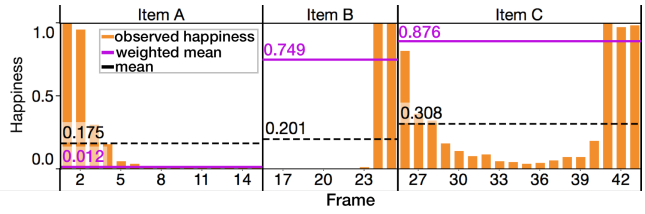

Figure 1: Overflowing Emotions. Happiness Example

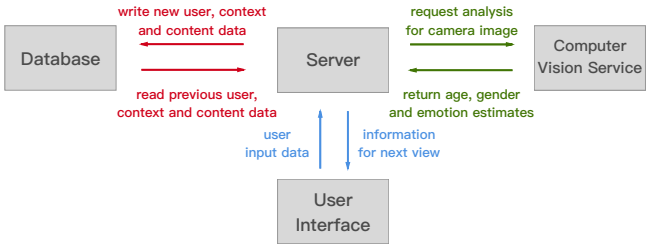

Figure 2: Prototype System Architecture

observed *happiness* is shown in orange for 15 frames in the case of Item A. Since we assume that the important reaction to the content is at the end of the item display period, we are quite satisfied with our weighted mean calculation. Note that we used a sampling rate of one analyzed frame per second.

An alternative would have been to aggregate over the last $p\%$ of the frames. While we think that our measure is more robust, an in-depth analysis of different aggregation strategies is left for future work. Another idea for separating successive content is to show a neutral screen for some time before showing the next. It is, however, unclear what an adequate time is for that, as users tend to show emotions for an unknown duration and may find this delay annoying.

3.3 Prototype and Experiment

To evaluate our approach, we implemented an image recommender prototype using Python. Figure 2 shows the high-level architecture: the core part is a Flask[4] web server that serves web pages with the recommendations based on context information (age, gender, emotions) from the computer vision service and the history of user interactions retrieved from a PostgresSQL[5] database.

To answer our second research question, we compare our variant of the Contextual-ε-Greedy with the traditional ε-Greedy in a controlled lab experiment. The experimental procedure was the following: The participant's task is to rate images. Hoping to evoke a large spectrum of emotions, we used a self-scraped data set of 3000 *memes* from the social web platform 9GAG[6] over the period from January 24, to February 9, 2018. The subject is instructed to take a seat in front of the screen with a webcam, it pointed out that the camera is recording and information is being stored according to local data privacy protection laws. She is asked to view consecutively displayed images and provide feedback for each one in the form of a *'like'* or *'dislike'* rating. The recommendation engine attempts to optimize the amount of positive feedback using our Contextual-ε-Greedy or the baseline ε-Greedy. Each subject is shown 60 images per strategy, which is our independent variable. The order of strategy is selected at random without the subject being aware of this.

We conducted the experiment in April 2018 in Garching with 21 volunteers (11 f / 10 m) affiliated with the Technical University of Munich. The subjects' ages varied between 19 to 31 years with a mean value of 24.09. The dependent variables are the users' feedback

[4] http://flask.pocoo.org
[5] https://www.postgresql.org
[6] https://9gag.com

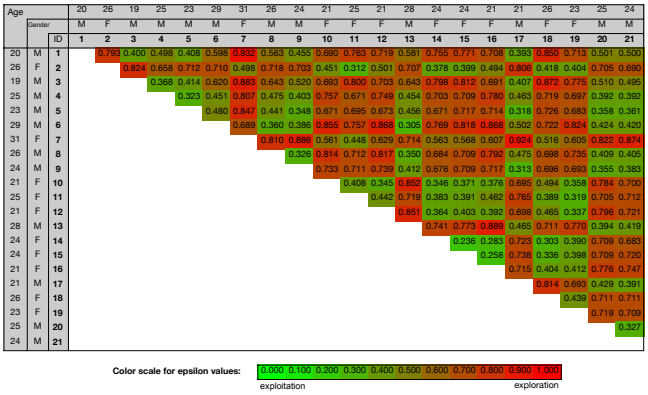

Figure 3: Values of ε Throughout the *Contextual* Experiment

to the item, the detected affective cues from the computer vision service and additional information collected with a questionnaire.

3.4 Evaluation Results

In the convergence analysis of the algorithms, we observe an improvement of the accuracy of time, i.e., the number of positive ratings, in both recommendation strategies. To showcase this, we fit a linear model over the algorithm convergence described in Table 1. Over the course of 21 observations, the Contextual-ε-Greedy starts slightly worse with 46.64% positive rewards; however, it improves faster over time reaching 60.7% at the end of the experiment. Note that the difference between the strategies is not significant and this model should not be used to predict further observations. Clearly, 21 observations with 60 ratings each are not enough for the bandit algorithms to converge.

Table 1: Linear Trend Models of Rewards

Strategy	Linear Equation	f(21)
ε-Greedy	$f_(x) = 0.47754 + 0.0047835 \cdot x$	0.578
Contextual-ε-Greedy	$f(x) = 0.463968 + 0.0068831 \cdot x$	0.607

A closer look into the properties of the Contextual-ε-Greedy algorithm reveals avenues for improvement. Figure 3 depicts the similarity of a participant's stereotypic attributes to the previous subjects. The most similar user pair per column has the lowest ε and was leveraged by the *Contextual* algorithm for recommending the next item (cf. Equation 4). A clearly visible pattern is that the same gender plays a dominant role in the distance measure. Depending on the recommended items, this could be adjusted in future studies.

Further, we notice that the Microsoft Face algorithm mostly detected two emotions. Overall, *happiness* and *neutral* make up 93.65% of the observed emotions, with *neutral* being the dominant emotion. However, as seen in Table 2, positive feedback is more likely if the affective response was *happiness* instead of *neutral*.

Overall, the subjects rated 53.97% of the items positively, although this varied a lot per user, ranging from only 3 positive ratings up to 47 of 60. Also, the experiment showed that the duration of item consumption varies, underlining the need for a dynamic aggregation of the analyzed frame as in Equation 7.

Table 2: Correlation of Emotions with Rating Feedback

Feedback	happiness	neutral	other	n
positive	25.06%	68.90%	6.04%	680
negative	7.24%	86.04%	6.72%	580

4 CONCLUSIONS AND FUTURE WORK

Bandit algorithms provide a robust framework not only for online advertisement, but also for personalized recommendations. The possibility of calibrating the exploration vs. exploitation probabilities using weighted similarity measures is an elegant way for the hybridization of recommendation and active learning. Although computer vision has not yet reached its full potential, it is sufficiently affordable and accurate to experiment with for RS research.

In this paper, we have presented an approach for recommending images using bandit algorithms and computer vision focusing on improving recommendations in the cold start phase. Although our contextual bandit algorithm was not significantly better than the baseline, our work comprises the following contributions: (1) We have developed a practical approach for using information from facial classification within RSs, (2) we presented an adaptation of the Contextual-ε-Greedy suited for incorporating stereotypic information, (3) we developed a strategy with a weighted average to mitigate the overflowing emotions problem, and (4) we have shown using a lab study that by putting the pieces together, an improvement of the recommendation accuracy could be achieved. While this study was conducted with the informed consent of the participants, the unconscious measuring of people's emotions in real-world applications is critical with respect to privacy concerns.

Having realized this prototype based on many assumptions, we can highlight the path for further research: Our post-mortem analysis has shown the necessity of an evidence-based method for adjusting the weights of the hybrid similarity measure. Having identified the *'overflowing emotions'* problem in sequential recommendations, an in-depth analysis thereof would be interesting. Finally, we plan to analyze the long term convergence of our bandit recommender algorithm in a larger field experiment against simpler baselines, e.g., random items, and to investigate the accuracy of emotional classification and its potential impact on performance.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2015. Context-Aware Recommender Systems. In *Recommender Systems Handbook*. Springer, 191–226.
[2] Mohammad Yahya H. Al-Shamri. 2016. User Profiling Approaches for Demographic Recommender Systems. *Knowledge-Based Systems* 100 (2016), 175–187.
[3] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. 2012. A Contextual-Bandit Algorithm for Mobile Context-Aware Recommender System. In *International Conference on Neural Information Processing*. Springer, 324–331.
[4] Il Young Choi, Myung Geun Oh, Jae Kyeong Kim, and Young U. Ryu. 2016. Collaborative Filtering with Facial Expressions for Online Video Recommendation. *International Journal of Information Management* 36, 3 (2016), 397–402.
[5] Juliane Exeler, Markus Buzeck, and Jörg Müller. 2009. eMir: Digital Signs that React to Audience Emotion. In *2nd Workshop on Pervasive Advertising*. 38–44.
[6] John C. Gittins. 1979. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 42, 2 (1979), 148–177.
[7] Sebastian Oehme. 2018. *Utilizing Facial Classification for Improving Recommender Systems.* Bachelor's thesis. Technical University of Munich.

[8] Michael J. Pazzani. 1999. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial intelligence review* 13, 5 (Dec. 1999), 393–408.
[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Conference on Computer Vision and Pattern Recognition (CVPR '16)*. IEEE, 779–788.
[10] Elaine Rich. 1979. User Modeling via Stereotypes. *Cognitive Science* 3, 4 (Oct. 1979), 329–354.
[11] Andrew G. Barto Richard S. Sutton. 1998. *Reinforcement Learning*. MIT Press.
[12] Herbert Robbins. 1985. Some Aspects of the Sequential Design of Experiments. In *Herbert Robbins Selected Papers*. Springer, 169–177.
[13] Marko Tkalčič, Urban Burnik, Ante Odić, Andrej Košir, and Jurij Tasič. 2013. Emotion-Aware Recommender Systems–a Framework and a Case Study. In *ICT Innovations 2012*. Springer, 141–150.
[14] Marko Tkalčič, Ante Odic, Andrej Kosir, and Jurij Tasic. 2013. Affective Labeling in a Content-Based Recommender System for Images. *IEEE Transactions on Multimedia* 15, 2 (Feb. 2013), 391–400.
[15] Sicheng Zhao, Hongxun Yao, and Xiaoshuai Sun. 2013. Video Classification and Recommendation Based on Affective Analysis of Viewers. *Neurocomputing* 119 (2013), 101–110.