User Modeling and User-Adapted Interaction (2019) 29:345–379
https://doi.org/10.1007/s11257-018-9213-x

Affective recommender systems in online news industry: how emotions influence reading choices

Jan Mizgajski[1] · Mikołaj Morzy[2]

Received: 15 December 2017 / Accepted in revised form: 9 November 2018 /
Published online: 19 November 2018

Abstract

Recommender systems have become ubiquitous over the last decade, providing users with personalized search results, video streams, news excerpts, and purchasing hints. Human emotions are widely regarded as important predictors of behavior and preference. They are a crucial factor in decision making, but until recently, relatively little has been known about the effectiveness of using human emotions in personalizing real-world recommender systems. In this paper we introduce the Emotion Aware Recommender System (EARS), a large scale system for recommending news items using user's self-assessed emotional reactions. Our original contribution includes the formulation of a multi-dimensional model of emotions for news item recommendations, introduction of affective item features that can be used to describe recommended items, construction of affective similarity measures, and validation of the EARS on a large corpus of real-world Web traffic. We collect over 13,000,000 page views from 2,700,000 unique users of two news sites and we gather over 160,000 emotional reactions to 85,000 news articles. We discover that incorporating pleasant emotions into collaborative filtering recommendations consistently outperforms all other algorithms. We also find that targeting recommendations by selected emotional reactions presents a promising direction for further research. As an additional contribution we share our experiences in designing and developing a real-world emotion-based recommendation engine, pointing to various challenges posed by the practical aspects of deploying emotion-based recommenders.

This paper or a similar version is not currently under review by a journal or conference. This paper is void of plagiarism or self-plagiarism as defined by the Committee on Publication Ethics and Springer Guidelines.

✉ Mikołaj Morzy
mikolaj.morzy@put.poznan.pl

Jan Mizgajski
mizgajski.jan@gmail.com

[1] AppliedIntelligence.pl, Libelta 34/1, 61-707 Poznan, Poland

[2] CAMIL: Center for Artificial Intelligence and Machine Learning, Faculty of Computing, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

Keywords Recommender systems · Affective computing · Emotion analysis · Emotion mining · News recommendations

1 Introduction

Every user has interacted with a recommender system, oftentimes without being fully aware of the fact, that her actions and choices were invisibly tampered by a recommendation engine. Every purchase of a book using Amazon's "Customers Who Bought This Item Also Bought…" feature, every song on Spotify played as the result of pressing the "Discover" button, every hour wasted on binge-watching Netflix just because that one show has popped up, these are the results of clever algorithms trying to guess human interests, longings, and desires. The Web is all about personalization and ubiquitous recommender systems are tirelessly shaping the universe of possible choices to best match human expectations in countless application domains.

Emotions greatly influence human behaviors and choices (Shiv and Fedorikhin 1999). Emotions are widely recognized as key factors in decision making, in particular, when impromptu decisions are taken [as explained via the interactive influence model of emotion and cognition (Luo and Yu 2015)]. Unfortunately, emotions are also very difficult to operationalize, quantify and measure precisely, which is one of the primary reasons for a relatively small number of previous work on using emotions in recommender systems. In this paper we are filling the gap by presenting the Emotion Aware Recommender System (EARS), a large scale recommender system capable of incorporating human emotions into personalized recommendations of news items.

The idea behind EARS is simple. Users visit a website and read news articles. Under each article a widget is placed which allows users to self-report the emotion triggered by the news article with a single click. There are two primary incentives for users to report their emotional reactions to consumed content. Firstly, a user can compare her emotional reaction to the distribution of emotional reactions of other users. Secondly, the recommender engine can incorporate the information about emotional reaction into its algorithm in order to serve more relevant and engaging recommendations for further reading. Consider a scenario where the news article reports on a controversial statement made by a politician. Let us suppose that after reading the article Ann reports that she is feeling "amused". This is a subtle hint that Ann is supportive of the politician and she would possibly be interested in more articles about the politician. Thus, the algorithm should recommend to Ann articles which were popular among other people who exhibited the same (or similar) emotional reactions to the article about the politician. On the other hand, let us assume that Bill reports feeling "angry" after reading the same article. This could mean that Bill dislikes the politician and serving further contents about the politician is counter-productive. The algorithm should find articles which were often read by other users who also felt "angry", "scared" or "sad" after reading the article about the politician. In other words, incorporating information about emotional reactions to news articles allows the recommendation engine to better estimate the similarity between users, and to provide more targeted recommendations as the result.

This paper focuses on advancing the field of recommender systems by expanding techniques and models borrowed from psychology and affective computing by:

- investigating how different psychological and behavioral concepts can be applied in the context of recommender systems (Oatley and Johnson-laird 1987; Oatley et al. 2006),
- overviewing methods for collecting emotional reactions and assessing strengths and weaknesses of these methods in the context of recommender systems (Calvo and D'Mello 2010; Reisenzein 2010),
- introducing a new, unobtrusive and scalable collection method of emotional reactions.
- discussing different emotion models and how they can be used in recommender systems,

The paper is organized as follows. Section 2 formalizes the notion of a recommender system for online news industry. In Sect. 3 we introduce the Emotion-Aware Recommender System, its main algorithms and formulas, and the underlying multidimensional emotional model. Section 4 gives an overview of the design and implementation of the EARS and its algorithms. In Sect. 5 we report the results of conducted experiments. The paper concludes in Sect. 6 with a brief summary.

2 Recommender systems for online news industry

The main goal of a recommender system is to support users in their (online) decision making process (Jannach et al. 2010). On a very high level, a recommender system simplifies the *discovery* of items and subsequent *selection*, by presenting only a small subset of available items based on some ranking function. The ranking function can take into account:

- preferences and characteristics of the user, either explicit or inferred, and their extrapolation based on similar users,
- social relationships between users,
- context in which the recommendation is presented, for example, a news item that the user is currently reading,
- semantic, statistical or ontological relationships between items,
- characteristics of the recommended item itself, such as popularity, age, or subject,
- various business objectives, for instance, the expected monetary return if the recommendation is successful.

2.1 Recommendation techniques

The introduction of social networks, open ontologies and new techniques to extract metadata and features from textual, image and video content have intensified research and development of new recommendation techniques. The explosion of available approaches renders a comprehensive summary beyond the scope of this paper. Consequently we have settled for an overview of techniques that can be most commonly found in industry-grade recommenders for the media industry. To a more inquisitive reader we recommend Bobadilla et al. (2013), Adomavicius and Tuzhilin (2005), Ricci et al. (2011) that expand on techniques described in this section and introduce additional classes of recommender systems, and we point to Doychev et al. (2014) for a brief overview of challenges particular to news item recommendation domain.

- *Knowledge-based systems*: A knowledge-based (KB) recommender suggests items based on inferences about user's needs and preferences (Burke 2002). Knowledge-based techniques differentiate themselves by having functional knowledge of how a particular item meets a particular user's need. KB recommendations do not have to be constrained to a single user. As pointed by Cremonesi et al. (2010), in scenarios where information is scarce it may be beneficial to create groups that represent the needs of a wider audience or to create general rules about items that, although depersonalized, can leverage the general appeal of an item. One of the biggest strengths of a KB recommender is the fact that it does not need an extensive history of user-item interactions and can be applied when personalized data is not available. This advantage however comes at a price, KB recommenders have static suggestion ability (i.e., they do not learn) and their performance depends on the quality of knowledge engineering, which, consequently, becomes the biggest cost in implementing this recommendation technique. The diversity of scenarios in news item recommendations precludes the usage of KB systems, primarily due to excessive costs of knowledge engineering that would have been involved.
- *Item-based Collaborative Filtering*: Item-based collaborative filtering (IBCF) computes predictions using the similarity between items and not the similarity between users. To predict the rating of the user u on a new item i one calculates the weighted average of past ratings, where the weights are equal to the similarity between the new item and the items already rated by the user u. IBCF systems focus on a single user, i.e., when making a recommendation they are searching for items that are most similar to items which have previously attracted the attention of the user. The underlying assumption is that for each user there exists a history of previously watched content, based on which future relevant recommendations can be computed. Unfortunately, this assumption does not hold in media industry where the majority of users have very short (or non-existent) histories of user-item interactions.
- *User-based Collaborative Filtering*: User-based collaborative filtering (UBCF) exploits information about past behaviors or opinions of an existing user for predicting which items the user might like, be interested in, or what her reaction to particular items might be. UBCF relies on the notion of *similar users* or *neighbors*. The neighborhood of a user is usually determined using a distance measure which is based on the agreements-disagreements of user ratings. There exists a wide variety of distance measures and ways to optimize their performance (Cheung and Tian 2004; Spertus et al. 2005).
- *Hybridization methods*: Hybridization methods try to combine two or more recommendation techniques in a single recommender system. Hybridization either eases specific weaknesses of its constituents, or tries to combine their strengths. Burke (2002), Zanker et al. (2007) and Jannach et al. (2010) have identified several

hybridization schemes, including the *weighted scheme* (where scores of several recommendation techniques are combined to produce a single recommendation set), the *switching scheme* (where an oracle decides which recommender should be used depending on circumstances), the *mixed scheme* (where recommendations from different techniques are combined), and the *cascade scheme* (where a sequence of recommendation techniques works as a chain of filters).

2.2 News items recommendations

Historically, recommender systems have been applied in very different industries, in each facing different problems, constraints and domain specific circumstances. The media industry was one of the first to receive considerable attention from both scientific and business perspectives (Montaner et al. 2003). Here we focus on the industry perspective, highlighting notable incumbents in the space of news recommender systems, distilling a definition of Recommendations as a Service and analyzing the most prevalent challenges in this domain.

In the contemporary scene of news recommenders two distinct approaches seem to have emerged, namely stand-alone recommenders and recommendation widgets or APIs (RaaS). The stand-alone recommender class consists of web-based news aggregators from big tech corporations such as Google News or Yahoo! News, RSS feed aggregators such as Feedly and mobile aggregators such as Flipboard. Recommendations as a Service have received a lot of attention and media coverage after Outbrain was reported to receive almost $100M in Venture Capital funding and AOL announced a $83M acquisition of Gravity. Most notable Outbrain competitors in this space include Taboola, Salithru, Gravity, plista and nrelate.

A clear advantage of stand-alone recommenders is much easier user data management and acquisition. A user signs up for the service and has to log in every time she uses it, this way her data is clearly linked to her profile. The fact that the user has willingly signed up for the recommendations also makes it easier to ask for additional information, because there is a sense of common goal in improving the quality of recommendations. On the other hand, stand alone recommenders are often unable to asses posterior engagement with the recommended item and the publisher offering it. Also, before history-based or machine learning techniques can be used in the recommender, a sufficient user base has to be acquired. This poses a sort of an “chicken and egg” problem and is a major challenge for small companies that wish to offer such a product. In most cases recommendations are introduced as a new feature of the system, so a sufficient value proposition has to be supplied before the product can become a standalone recommendation platform.

In the area of Recommendations as a Service the most important players in the domain of news recommendation include:

- Outbrain: founded in 2006, Outbrain offers recommendations in the form of a widget which is added under or beside an article on a publisher’s page. Widget is composed of two sections, first internal recommendations are shown, leading to more articles from the hosting publisher, then a “From around the web” section is presented leading to promoted articles from other publishers. For each recommendation a picture and a title is presented, the titles of promoted content are further accompanied by their source (the name of the publisher). It is worth noting that promoted recommendations are the primary source of revenue for Outbrain with a business model borrowed from pay-per-click advertising. Publishers who wish to have their content distributed in the "From around the web" section pay for each click on their article. Outbrain uses a combination of content-based recommendations, collaborative filtering, and knowledge-based recommendations that try to understand trends by collecting article visits, click-through rates on recommendations and the social performance (i.e. the number of Facebook shares) of articles. As their Unique Selling Proposition Outbrain claims not only to optimize the click rates on their recommendations, but also to improve the posterior engagement of the user on the site the recommendation has brought her to. This is achieved by using a look-ahead heuristic that takes into account the probability that a user will click on a given recommendation and the probability the user will continue to use the recommendation service after reading the recommended article (Gur 2014).

- Gravity: founded in 2009, Gravity operates on a similar business model and integration process as Outbrain. They differentiate themselves from the competition by relying on the ontology called the *interest graph* derived from DBpedia and other open ontologies and enhanced with large scale Natural Language Processing and data mining. A user is defined by a weighted set of edges linking her with different concepts in the ontology. The final recommendation is provided by a combination of interest-based filtering that relies on the aforementioned ontology and user history, collaborative filtering, information about current trends and trending topics, and social popularity of the content being recommended. Recommendation strategies are further enhanced by undisclosed machine learning techniques focused on assigning the right combination of algorithms per publisher and solving the user cold start problem.

There are several challenges that are specific to news item recommendation domain. These challenges must be addressed by any industry-grade solution and they make news item recommendations a particularly interesting problem.

- Short shelf life of articles: Some articles are only relevant for a short period of time. This problem is especially visible for news publishers. Articles can quickly become irrelevant due to new developments or the shift in interest of the public opinion. This poses a significant constraint on a recommender system, because it has to either be able to distinguish between short lived and evergreen content, be able to recommend items from a pre-filtered set of recent articles on which it may not have a sufficient history, or continuously A/B test different recommendation sets to identify articles that are still relevant.
- Rate at which new articles are added to the system: Larger web publishers with the aid of news agencies and user generated content are often producing hundreds of articles per day. When considering cross-publisher recommendation, the recommender system may be dealing with vast amount of new articles per day. For a recommender system collecting sufficient information about each new item could pose a significant challenge rendering a lot of "out of the box" techniques, which

rely on historical user-article interactions, insufficient for generating relevant recommendations.

- Fluctuation of general interest due to evolving trends: When new trends emerge and the public interest shifts to new topic, the relevancy of certain articles changes, some articles may stop converting because they are "yesterdays news" and other may enjoy a boost in popularity because they are related to trending topics. This poses a complex problem of constantly changing conversion rates for articles, keywords and topics. Recommenders that wish to optimize for click through rates have to exploit these trends proactively or use sufficient feedback loops that quickly identify the repercussions of these trends and adjust their strategy accordingly.
- Perpetual new user problem (cold start): When a recommender system is provided as an add-on to a publisher website, browser cookies become the only way to identify the user (Kille et al. 2014). Forcing the user to identify herself with a different method would dramatically reduce adoption as readers mainly consume the media anonymously, without logging into the publisher's website. Since cookies are volatile, the recommender faces a perpetual new user problem, which also reduces the average length of a user's history. This increases sparsity of the data and reduces the effectiveness of all techniques that are based on user-article interactions history.
- Privacy and opt-out: European law requires all businesses using browser cookies for collection of user data to provide an opt-out option for the user. Recommender system engineer is then faced with the problem of collecting anonymous usage data on articles without linking it to a particular opt-out user. This introduces extra complexity to the algorithm logic and data management.
- User trust: the recommendation widgets are usually displayed on each article, consequently the user sees and preferably interacts with the recommendations very often. Irrelevant and misguiding recommendations erode the trust of the user in the recommender system and diminish long term return on sponsored recommendations. A recommender system must therefore be able to quickly identify these articles and remove them from recommendations.
- Publisher trust: For cross-publisher recommendations maintaining publisher trust is crucial for the business success of a recommender system. In most situations this requires a manual or semi-automated curation of articles available for cross-publisher recommendations that exclude explicit, misguiding or otherwise inappropriate content that may be submitted by third parties. Furthermore certain publishers may wish to exclude certain sources, such as competition or low profile content, to maintain control over their brand and market differentiation strategy.
- Ease of integration: Cost of entry is often a significant factor in business decisions. For Recommendation as a Service it is crucial to maintain a low cost of entry and minimal mandatory involvement from the publisher. It is therefore expected that recommendations are provided in form of a widget, which installation requires only a javascript snippet and an addition of an HTML element to the website. Acquisition and maintenance of article data, such as the tile, image, content, keywords or categories, has to be therefore handled by the recommender system.
- Scale: High profile on-line media publishers are among the most visited websites on the Web. A recommendation is served with each article view resulting in mas-

sive amount of traffic for the recommender system. A sufficient publisher network is a crucial requirement from the advertiser's point of view, when sponsored recommendation business model is considered. Consequently, recommender systems must handle huge volume of traffic and collect massive amounts of data from each website they support. This poses significant scalability challenges for the designers of such systems.

2.3 Implicit signal feedback in news recommendations

In the absence, sparsity or high cost of acquisition of explicit user ratings on the one hand, and the abundance of other implicit signals on the other hand (browsing history, engagement time, tagging, commenting, sharing or positive social feedback), it is obvious that incorporating implicit feedback into the recommendation process is beneficial. There have been many previous attempts to include these implicit feedbacks into news recommendations (Liu et al. 2010). For instance, Ilievski and Roy (2013) introduce a framework to model user interest in individual news items using a taxonomy of hierarchical facets that capture various semantic aspects of a story that might appeal to the user. Lin et al. (2014) focus on implicit social factors, such as opinions of experts and other influential persons, in news items recommendations. They augment traditional content-based recommenders and collaborative filtering recommenders with information diffusion models which help to predict the effect an influential opinion about a news item may have on the relevance of that news item for a particular user. Similar solution has been presented as a demo in Kazai et al. (2016), where user location and her social media feeds have been used as additional features for real time news items recommendation.

Incorporation of implicit feedback into the recommendation workflow is not trivial. At least two problems arise: how to infer item relevance to the user from available signals, and how to assess the similarity between signals. The first problem of inferring item relevance can be partially solved by creating heuristics which incorporate business knowledge and key performance indicators. It is important to stress that the term *signal* is used in this context to refer to any type of implicit or explicit feedback that the user provides after interacting with a news item. Browsing a news item is a weak indicator of positive relevance, since the item headline and image had to be relevant enough for the user to evoke the intention of reading, but it does not carry information about posterior engagement (which really accounts for the quality of the recommendation) and can therefore promote low quality news items with good headlines, also referred to as click-baits. Engagement time can be a strong indicator of relevance or irrelevance when assessed together with the statistical moments of engagement time normalized with respect to the content type (i.e., number of characters for written content). Social actions, such as commenting or sharing on social media, can be regarded as very strong indicators of relevance, because they require additional effort from the user and often reap additional benefits for the content provider — sharing brings additional users, commenting enriches the content and "liking" is a direct way of providing positive feedback. Exiting a site after consuming a news item is generally regarded as a weak irrelevance feedback, the rationale behind this assumption is that a user was dissatisfied with the quality of the news item and the overall experience and has decided to leave the site. Conversely, the user could have left the site because she followed an external link, her search objectives were satisfied or he simply run out of time for browsing, which makes this signal ambiguous.

The problem of assessing the similarity between signals is traditionally solved by converting all signals to a single numeric scale or by heuristic similarity approach (Burke 2002). The first technique has obvious limitations because it is essentially compressing the distance matrix from the multidimensional space, in which different signals can be represented, into a single dimension. Heuristic similarity consists in creating a signal-to-signal similarity matrix and then assessing user similarity on the signal-by-signal basis. The signal-to-signal similarity matrix is rooted in the domain specific knowledge about the signals, for example, it can be based on the multi-dimensional emotion theory.

Given that the similarity between two signals is $d(s_1, s_2)$, the general formula for similarity between two users u and v is:

$$s(u, v) = h(I_u, I_v, I_u \cap I_v) \sum_{i \in I_u \cap I_v} w_i d(s_i^u, s_i^v)$$

where:

- I_u is the set of items for which signals from the user u were recorded,
- I_v is the set of items for which signals from the user v were recorded,
- s_i^u is the signal recorded for the user u on the item i,
- s_i^v is the signal recorded for the user v on the item i,
- w_i is the item weight (Breese et al. 1998),
- $h(I_u, I_v, I_{u \cap v})$ is an overlap regularization function, which purpose is to take into account the degree of overlap between I_u and I_v, since a simple sum may fall short in situations where there are a lot of overlapping items in I_u and I_v, but respective similarity between signals is close to 0.

2.4 Related work

Affective computing is a broad research field that explores detection and interpretation of human emotions and behaviors caused by emotions (Picard 2000; Tao and Tan 2005; Tkalčič et al. 2013b). Until recently, relatively little research has been conducted on using affective features in the design of recommender systems. Most previous work focused on the utilization of personality traits for personalization of recommender systems (Nunes and Hu 2012), with the special focus on the music domain (Andjelkovic et al. 2016; Strle et al. 2016; Wakil et al. 2015; Wang et al. 2015). For instance Andjelkovic et al. (2016) present a recommender system that selects subsequent songs based on the mood and affective features of songs. This approach is significantly different than our approach as the moods and emotions refer to the properties of recommended items, and not to affective states of users. Similar idea of using music affective features is presented in Kaminskas and Ricci (2016), however the work considers a static problem of matching music to places rather than the dynamic problem of music recommendation. A slightly more relevant to our work is the emotion state transformation model which maps human emotional states and their transitions by music (Han et al. 2010). We also track human emotional state changes, but we are not interested in emotion classification, but in improving user engagement with recommended contents.

Since recommender systems play such a vital role in the decision making process, more and more research is conducted on the use of emotions for generating relevant recommendation. Recently, several proposals for affective recommendation frameworks have been presented. González et al. (2007) introduce the concept of Ambient Recommender Systems (an elaboration of concepts presented by Burke (2002), Zheng et al. (2013)) which employ emotional context as one of the dimensions of operation. Unfortunately, this work does not extend beyond a mere high-level description and does not inform the design and development of real world affective recommender systems in any meaningful way. The same can be said of the work by Costa and Macedo (2013). A much more elaborate proposal has been introduced by Tkalčič et al. (2011) [further extended by Tkalčič et al. (2013a)]. The authors introduce a unifying framework for using emotions in user interactions with a recommender system, assuming sequential process consisting of an entry stage, a consumption stage, and an exit stage of interaction. Several factors make this scenario unfeasible for a real world recommender of news items. Firstly, the authors assume that it is possible to measure the affective state of a user at each stage of interaction with the recommender system (the authors advocate the use of the Self Assessment Manikin developed by Bradley and Lang (1994)). Alternatively, the authors discuss the use of other modalities for implicit emotion acquisition: video cameras, speech, EEG, ECG, etc. All these approaches are clearly unfeasible in the real world scenario. Nevertheless, the general framework for including affective features in content-based recommendation systems is relevant in the light of our work (Tkalčič et al. 2010).

2.5 Evaluation of recommender systems

The challenge of providing high-quality recommendations generates an array of follow-up goals from both technical and psychological standpoints. On the one hand, the creators of a recommender system have to find methods that efficiently and effectively exploit available information and knowledge to provide highly relevant recommendations. On the other hand, one must take into account that users rarely act as rational agents and qualitatively asses the utility of recommendations. Efficient implementation of a recommender system has to excel both in user experience and quality of recommendations. Traditionally, the evaluation of recommender systems was performed offline, on a historical data set, and typically used evaluation metrics such as:

- *MAE*: mean average error of predicted ratings to actual, left out ratings,
- *precision* P: the ratio of relevant recommendations to all recommendations,
- *recall* R: the ratio of relevant recommendations to the theoretical maximum number of relevant recommendations,
- F_1: the harmonic mean of precision and recall defined as $2\frac{P\cdot R}{P+R}$.

Unfortunately, offline evaluation is not always accurate. Neither does it consider the appeal of items when presented in the context of recommendations, nor does it account for user interface design factors, such as the number of recommended items or the call to action that presents the recommendations. Finally, offline evaluation is not directly correlated with business objectives, for example, high precision does not necessarily mean an increase in page views for the content provider. Consequently, we have decided to use online evaluation based on click-through rates (CTRs). CTR is the ratio of recommendation sets with a hit (a recommendation which resulted in a click) to all generated recommendation sets. For CTR to be statistically significant, it has to be collected on a sufficiently large sample and over an extended period of time.

3 Affective recommender systems

3.1 Emotion models

Emotion is a subjective, conscious experience characterized by psycho-physiological expressions, biological reactions and mental states (Scherer et al. 1984). Emotions can be elicited in response to some external or internal stimuli. Emotional reactions are a subset of emotions elicited solely by, and in response to, *external* stimulus. Defining and describing emotions is a well established problem (Ekman 1999; Schröeder et al. 2010). Two mainstream approaches of describing the affective state of a user are the *universal emotions model* and the *dimensional model*.

In the universal emotions model each affective state is described as a distinct state or combination of distinct universal emotions, however, consensus has yet to be reached on defining the set of universal emotions. Two most prevalent sets can be attributed to Plutchik (2001), who defines eight basic emotions (joy, trust, fear, surprise, sadness, disgust, anger and anticipation), and Ekman (1999), who define seven basic emotions with different observable facial features (neutral, anger, disgust, fear, happiness, sadness and surprise, see Fig. 1) and eleven additional, universal emotions that do not exhibit such features. More elaborate emotion models include the OCC model by Colby et al. (1989) and the CBDTE model by Reisenzein (2009a, b). The OCC model emphasizes the fact that emotions arise as results of affective reactions to stimuli being judged as either beneficial or harmful to one's concern (Ahmadpour 2014). Emotions in the OCC model depend on person's focus at the time of the stimulus, their concern and their appraisal of the stimulus. The OCC model not only defines types of emotions that can arise, but also provides variables to describe emotions' intensity such as sense of reality, proximity, unexpectedness, and arousal. Although comprehensive and broad, the OCC model was too difficult to implement in a real online news recommender engine. The CBDTE model is based on the Computational Belief-Desire Theory of Emotions and its primary tenet is that emotions not only require the belief about the goodness or badness of a stimulus (the appraisal of the stimulus), but that they also require desires (motivational states) regarding these stimuli. Within the scope of our project this elaborate model of emotions was overwhelming and unnecessary. Besides, it was practically unfeasible to try to extract user desires regarding recom-

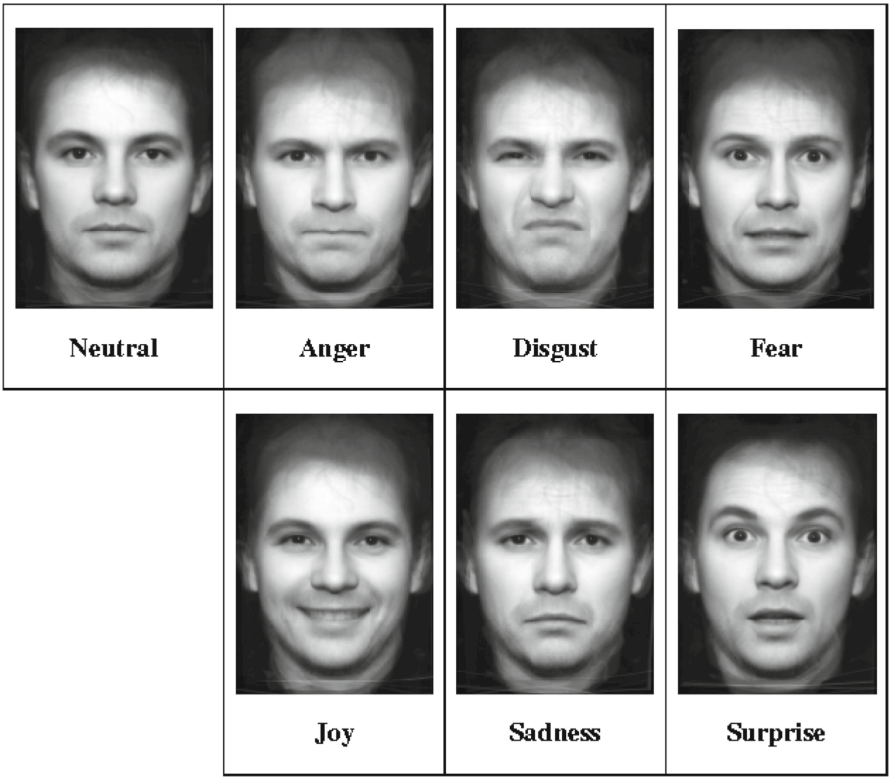

Fig. 1 Ekman's 7 basic emotions and corresponding facial expressions

mended news items and it is highly questionable whether such extraction is at all possible.

The dimensional models describe each affective state as a point in a multidimensional space. Several such models have been introduced over the years: Arnold (1960), Lazarus et al. (1970), Russell (1980), Desmet and Hekkert (2002). Dimensional models assume that affective states are systematically related to each other through a core affect which in turn can be described using a set of dimensions. Usually, these dimensions describe the core affect in terms of valence (pleasant vs unpleasant) and arousal (calm vs excited). With two dimensions a whole circular space of emotions, or circumplex, can be defined around core affects, the model proposed by Desmet and Hekkert (2002) is a good example of this class of emotion models. The model has been developed to help describe emotional reactions to consumer products and it contains 24 emotion terms in eight categories defined by combinations of valence and arousal (excited, excited and pleasant, pleasant, pleasant and calm, etc. After initial attempts to adjust this model to the field of online news recommendations we have decided against it due to the excess of emotions. We have not been able to design an effective emotional reaction collection widget, a crucial element of the entire recommendation engine.

We have decided to use the original model introduced by Mehrabian (1996). The dimensions of this model are pleasure (P), arousal (A) and dominance (D). Pleasure accounts for the pleasantness (or valence) of the emotion, arousal accounts for its strength and dominance defines if the subject feels in control of the experience causing the emotion. From the perspective of a recommender system the multi-dimensional model has an important advantage over the universal emotion model, because it is computational, i.e., its dimensional nature enables the creation of similarity measures between emotions and assessing the difference between reactions of two users in a numerical way. In addition, multi-dimensional emotion model allows for easy aggregation of emotions, i.e., a set of emotional reactions can be meaningfully "averaged" and the resulting emotional reaction would still be useful for the recommendation engine. The universal emotion model does not have this property, as emotions would have to be assessed as categorical variables with only partial order defined between emotions (as in circumplex models). This would limit or complicate the ability to compare user feedback when different emotions are reported. On the other hand, universal emotion models are easy to understand, even for casual users, and can relatively easily be turned into visually attractive widgets for emotional reaction collection. We have decided to get the best of the two worlds by combining both models. The multi-dimensional model is used for machine representation of emotions and for all computations, and the universal emotions model is used to create user-friendly emotional reaction collection method.

3.2 Emotional reaction measurement

Self assessment is a popular, cost efficient and scalable method of collecting affective reactions and states. The method usually consists of a written or visual questionnaire supplied to the subject after some emotional stimuli, for example, during the creation of the International Affective Picture System (IAPS) the researchers used SAM - the self assessment manikin (Fig. 2) to assess pleasure and arousal levels of test subjects after each picture (Lang et al. 2005).

Self assessment is widely accepted in the industry and can take many different forms, from Facebook "likes", through widgets including affective labels, to complex, affect-focused solutions based on the Plutchik's wheel of emotions (Plutchik 2001, see Fig. 3).

From the perspective of recommender systems the scale and unobtrusiveness that can be achieved by self assessment cannot be matched by other techniques. A reaction widget alongside content seems natural and in place both for the user and the publisher. This method is however far from perfect (Dunning et al. 2004) with important drawbacks including:

- priming, which occurs when user's feedback is influenced by the feedback of other users (Sabini et al. 1999),
- biases, which occur when the user gives feedback that differs from the actual experience for some intrinsic reason,

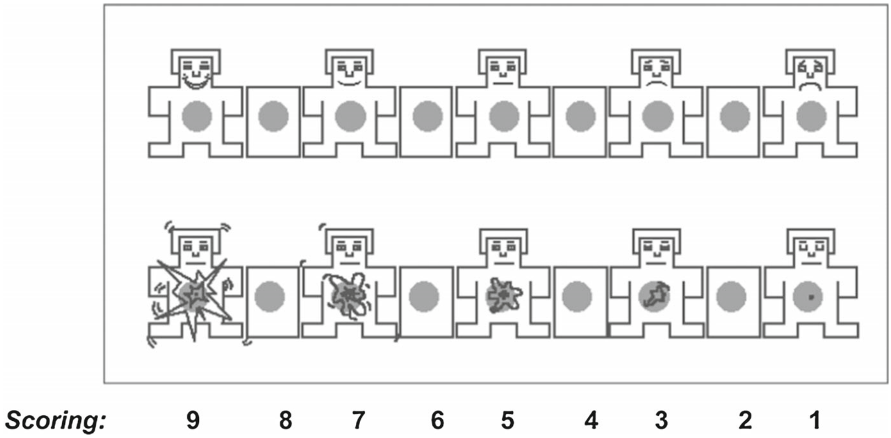

Fig. 2 SAM self assessment manikin used in the IAPS study

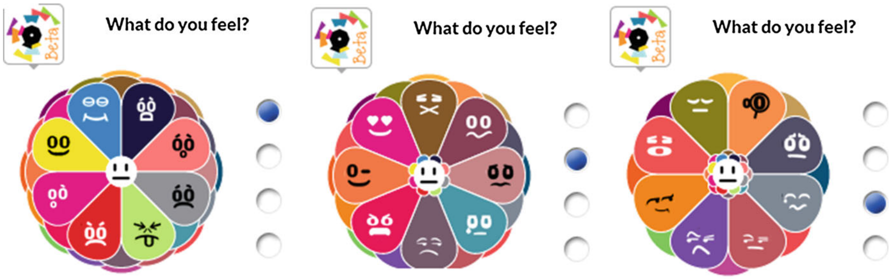

Fig. 3 mySmark emotional feedback widget using different layers of the Plutchik's wheel of emotion

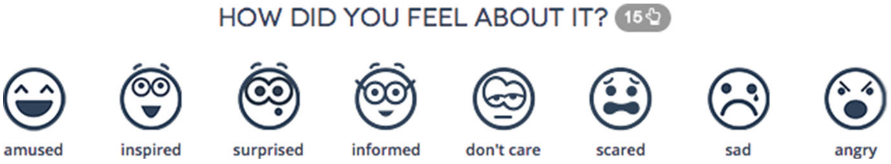

Fig. 4 Emotional reaction collection widget

- low fidelity, which is caused by the low maximum number of choices (affective labels) that can be presented to the user without perplexing her and significantly lowering conversion rates.

For the above reasons the self assessment methodology has to be designed with care in order to minimize the impact of these imperfections while maximizing user engagement.

We have decided to use self-assessment as the preferred method of gathering data on users' emotional reactions to the presented content. We have also selected the dimensional model for use in the machine representation of user emotional reactions. The following principles guided the design of our emotional reaction widget:

- single click: the user has to be able to provide her reaction with a single click of a mouse,
- visual first: the assessment of the choices presented to the user has to be possible without or with little aid from printed information,
- no more than 8 elements: no more than 8 choices can be presented to the user when she is selecting her reaction,
- diversity: choices presented to the user have to be easily distinguishable and encompass a large fraction of the emotional space,
- no priming: reactions of other users cannot be presented before a user has made her own decision,
- incentive: additional incentive should be provided to the user.

A simple widget showing eight predefined emotions (amused, angry, don't care (indifferent), informed, inspired, sad, scared, surprised) was placed directly under each news item and the widget prompted the user to pick her emotional reaction (see Fig. 4). Available choices were selected based on the universal primary emotions identified by Ekman (1999). Six emotions have been mapped directly to possible responses. Disgust was omitted to make room for a positive, highly arousing emotional reaction "inspired". This design choice has been dictated by practical reasons. Firstly, media publishers generally avoid contents which is openly disgusting (not to be confused with highly controversial contents). Secondly, inspiration is a desirable feature of online contents, because it tends to prolong the visit of the user. Among these seven reactions, four have negative valence (angry, scared, sad, don't care), and three have positive valence (amused, surprised, inspired). In order to reduce the skewness of the set of choices towards negative emotions, "informed" reaction was added. It represents positive utility and has a clear business interpretation in the context of news recommendations. Finally, "don't care" reaction was visually designed to be associated with boredom and account for negative interest and utility. For simplicity we refer to all eight concepts as emotional reactions, although many researchers would classify some of them (e.g., indifferent, informed, surprised) as cognitive states rather than emotional reactions.

The eight emotional reactions were mapped to the PAD space using the Affective Norms for English Words (ANEW) proposed by Bradley and Lang (1999). Figure 5 depicts the position of each emotional reaction in the PAD space. The choice of the PAD model is deliberate as PAD dimensions allow us to distinguish between emotional reactions elicited by news items presented to users. While selecting "pleasure" and "arousal" seems non-controversial in the context of news items, we need to explain why "dominance" has been added as a dimension. As can be seen in Fig. 5, P and A dimensions alone put pairs of emotional reactions close together (angry-scared, inspired-amused), suggesting much more similarity between these reactions. The addition of the D dimension allows to properly differentiate both between angry-scared reaction, as well as inspired-amused reaction. It is important to note here that an affective recommender system cannot use only a discrete emotional model or multidimensional emotional model. Discrete models are very convenient for emotional reaction measurement self-assessment (users quickly identify an emotion), but they cannot be used for computational purposes since they do not provide a method to

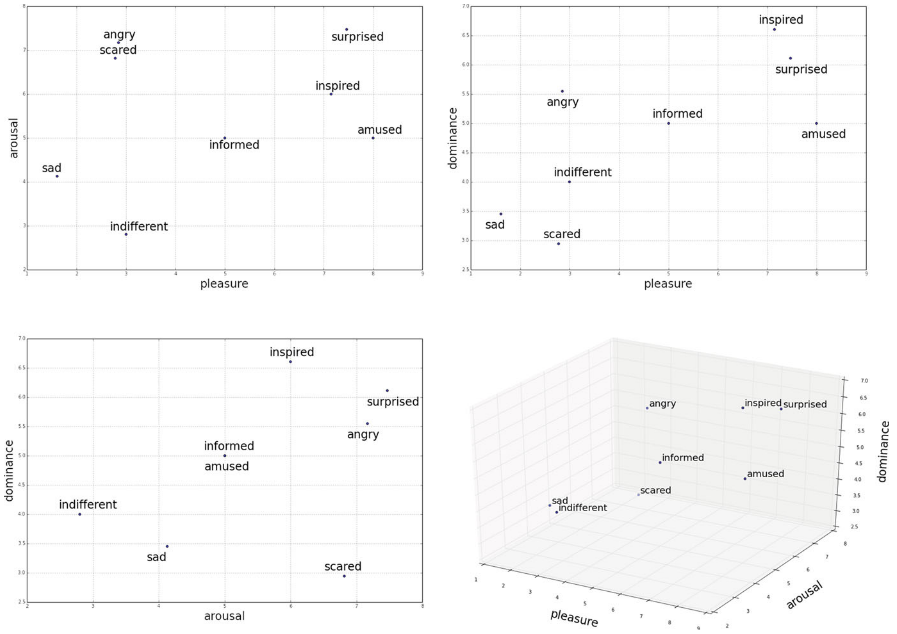

Fig. 5 Position of selected emotional reactions in the PAD space

measure the distance between emotions. Multidimensional models on the other hand are easy to incorporate into computations, but are very unintelligible to users. When designing an affective recommender system one has to provide a solid and firm mapping from the discrete emotional space to the multidimensional emotional space.

3.3 Affective extensions of recommendation techniques

3.3.1 Affective item features

Collecting emotional reactions opens a new spectrum of features that can be derived based on the aggregated user feedback. We focus on the first two statistical moments (mean value, standard deviation) of the pleasure (μ_p, σ_p), arousal (μ_a, σ_a) and dominance (μ_d, σ_d). These values allow us to define affective item features, i.e., characteristics of news items derived from emotional reactions exhibited by users after interacting with these news items. We introduce four such affective item features, namely *controversy*, *diversity*, *pleasantness* and *unpleasantness*. The statistical moments are computed based on the mappings for each emotional reaction to the PAD space and aggregated across all reactions submitted for the item i. The affective item features are defined as follows:

$$controversy(i) = \mu_a + 2\sigma_p \qquad (1)$$

The formula for *controversy* spans from the fact that controversial topics usually elicit extreme emotions characterized by high level of activation (arousal) and that they polarize people into supporters, who have positive valence (pleasure) towards the issue, and opponents, who have negative valence. Looking at Figure 5 we see that items which score high on the controversy elicit emotions ranging from angry and scared to surprised and inspired. If an item elicits low arousal (with emotions such as indifference or sadness), such item will score low on controversy. If, however, an item elicits some arousal, and, in particular, if there is large variability in arousal (with emotions ranging from fear to anger), such an item will score high on controversy.

$$diversity(i) = \sigma_p^2 + \sigma_a^2 + \sigma_d^2 \tag{2}$$

Diversity measures the disagreement of user emotional reactions in any of the PAD dimensions and, because it is designed to be used for ranking items, the disagreement is amplified (by taking the variance instead of standard deviation) to provide more discriminating rankings. In other words, the more disagreement there is in the assessment of pleasure, arousal, and dominance of emotions elicited by an item, the higher the item scores on the diversity scale. The main reason why this affective feature is quadratic and not linear is the fact that we want to score these items which really elicit very diversified emotional reactions.

$$pleasantness(i) = \mu_p + \mu_d + \frac{1}{\sigma_p} \tag{3}$$

Pleasantness maximizes pleasure and dominance, the inverse of the standard deviation of pleasure is added to promote items that are more uniformly pleasant to all users who submitted their emotional reactions. Again, looking at Fig. 5 one finds that items which score high on pleasantness should elicit emotions such as inspiration, surprise, and amusement.

$$unpleasantness(i) = \frac{1}{\mu_p} + \frac{1}{\mu_d} + \frac{1}{\sigma_p} \tag{4}$$

Unpleasantness minimizes pleasure and dominance, the inverse of the standard deviation of pleasure is added to promote items that are more uniformly unpleasant to all users who submitted their emotional reactions. In the PAD space emotions characterized by low pleasure and low dominance include sadness, scare, indifference, and, to a lesser extent, anger. Therefore, items which elicit these emotions would score high on the unpleasantness scale.

It should be noted that the above definitions are somewhat arbitrary and other aspects of user emotional reactions could be measured. These four features, however, cover a large spectrum of user emotional reactions and describe items well, in particular, in the context of news items recommendations, where controversy, diversity, and pleasantness play important role.

3.3.2 Emotional reaction similarity

To fully exploit the advances in emotion modeling in the context of recommender systems based on collaborative filtering, the notion of similarity between emotional reactions has to be introduced. Without it the emotions would have to be treated as categorical ratings and only equal reactions would be comparable and useful for calculating similarity metrics between users or items. This would obviously greatly increase the sparsity of the data, hindering the performance of a recommender. Fortunately, the multi-dimensional model of emotions enables the calculation of distances between discrete emotional reactions which, in turn, makes it possible to asses their similarity.

We define the similarity between emotional reactions e_i and e_j as follows:

$$esim(e_i, e_j) = \frac{|\hat{d} - d(e_i, e_j)|}{\hat{d}} \tag{5}$$

where:

- $d(e_i, e_j)$ is the distance between emotional reactions e_i and e_j in the PAD space according to a given metric (e.g. euclidean, cityblock),
- $\hat{d}$ is the average of the distances between all pairs of considered emotional reactions.

If the distance in the PAD space between emotional reactions e_i and e_j is comparable with the average distance between any pair of emotional reactions $\hat{d}$, the similarity of emotional reactions e_i and e_j is very low. If, however, the distance between emotional reactions e_i and e_j is significantly smaller than the average distance $\hat{d}$, the similarity between emotional reactions e_i and e_j will be close to 1. This approach assumes an even distribution of emotional reactions in the PAD space, an assumption that holds true for our set of eight available emotional reactions. It has to be noted that a concentration of a large fraction of reactions would skew the mean distance between emotional reactions and make our similarity measure inadequate. It is therefore important to verify the distribution of emotional reactions prior to using the emotional reaction similarity measure.

3.3.3 Affective similarity for UBCF

Users may enjoy news items for different reasons. They may be interested in controversial topics and highly polarized opinions, they may be seeking latest breaking news, they may browse gossip websites looking for funny news items on celebrities, they may actively search for news items that are related to scary, sensational, or inspiring content. Since the spectrum of possible emotional reactions to a news item is much broader than the spectrum allowed by simple binary or Likert scales, it is possible to segment users into much more meaningful distinctive groups based on their expressed emotional reactions to presented news items. We have developed an affective similarity measure for UBCF (user-based collaborative filtering) which is based on the similarity of user emotional reactions to co-rated items. Given the similarity between two emotional reactions $esim(e_i, e_j)$ and given the emotional reaction of the user u to the item i denoted as $e_{u,i}$, the affective similarity between two users u and v is defined as:

$$affsim(u, v) = \sum_{i \in I_u \cap I_v} esim(e_{u,i}, e_{v,i}) \frac{\left(1 + log \frac{|U|}{|U_{(i)}|}\right)^2}{\sqrt{|\{j : j \in I_v \wedge e_{v,j} = e_{v,i}\}|}} \tag{6}$$

where:

- I_u is the set of items for which emotional reactions of the user u were recorded,
- I_v is the set of items for which emotional reactions of the user v were recorded,
- $U_{(i)}$ is the set of users who submitted their emotional reactions to the item i,
- U is the set of all users.

Let us scrutinize the formula in detail. The second part of the formula resembles the well-known term-frequency inverse document frequency (TF-IDF) measure from information retrieval. This factor is used to assign weights for each pair of overlapping emotional reactions. The value of the nominator increases for items that were rated only by a handful of users, and reaches its minimum value of 1 for items which were rated by all users. The value of the denominator controls for the self-assessment bias of the user v by counting the number of items that were rated by the user v identically as the item i. In theory, this part of the formula is spurious, but its presence is dictated by practical concerns. In practice, many users tend to click on a small subset of emotional reactions available in our widget, and not taking this into consideration could significantly skew the results. From the theoretical standpoint the inverse item reaction frequency gives more weight to reactions on niche items. This is dictated by the assumption that niche items are more likely to better discriminate between user preferences. The first part of the formula identifies all items for which emotional reactions of users u and v were collected, and aggregates the emotional reaction similarity weighting each emotional reaction similarity by the inverse item reaction frequency.

Affective similarity allows to find users who react similarly to a given user, and consequently, based on their reactions, it allows to predict the reaction of the given user to a particular item. This opens up an interesting possibility of recommending items according to the expected emotional reaction of the user and the likelihood that a given item will elicit that particular emotional reaction. We call this recommendation technique the affective user-based collaborative filtering (AUBCF). In the AUBCF items are ranked according to the number of users who responded to the given item with a similar emotional reaction. Each reaction is weighted according to the emotional reaction similarity (Eq. 5), and the affective similarity (Eq. 6). The item score is then adjusted based on:

- The frequency with which a given emotional reaction was submitted for the item in question (*item reaction type frequency norm*), the purpose of this factor is to increase the score of less popular items,
- The number of emotional reactions which similar users have submitted so far, the purpose of this factor is to increase the weight of emotional reactions of less

expressive users (the assumption is that these users have a higher stimulation threshold and thus their emotional reactions carry more weight).

Given a user u, an item i, and an expected reaction e, the affective score of the item i is given by:

$$score(u, i, e) = \sum_{v \in \Lambda_u \cap U_{(i)}} affsim(u, v) esim(e, e_{v,i}) \frac{(1 + log \frac{|I|}{|I_v|})^2}{\sqrt{|\{j : j \in U_{(i)} \wedge e_{v,j} = e\}|}} \tag{7}$$

where:

- Λ_u is the set of top n users most similar to the user u according to $affsim()$,
- $U_{(i)}$ is the set of users who submitted an emotional reaction to the item i,
- $affsim(u, v)$ is the affective similarity between users u and v,
- $e_{v,i}$ is the emotional reaction of the user v to the item i,
- $esim(e, e_{v,i})$ is the similarity between the emotional reaction submitted by the user v to the item i and the expected reaction e,
- I_v is the set of items to which the user v submitted an emotional reaction,
- I is the set of all items.

Let us analyze this formula in more detail. The last part of the formula is analogous to the last part of the Eq. 6, the nominator increases its value for items, for which only a small set of users recorded their emotional reactions. The denominator compensates for user rating bias by punishing items rated by many users with the expected emotional reaction e. The first part of the formula identifies top n users who are affectively most similar to the user u (with the neighbor size parameter n being the hyperparameter of the recommender) and who have submitted their emotional reactions to the item i. Next, we measure how exactly similar are those users to the user u and how similar are their emotional reactions to the item i with respect to the target emotional reaction e. This allows us to predict the propensity of the user u to react to the item i with the expected emotional reaction e.

For example, let us assume that the news item i concerns a gossip about a popular actor. Given a target user u, we want to know how likely is this item to elicit emotions of sadness, amusement, inspiration, etc. in the user u. For each of the eight available emotional reactions we do the following: we find the set of users who submitted their emotional reaction to the item i and we select those users, who are emotionally most similar to the user u (this is measured using the $affsim()$ function). Then, we take their emotional reactions to the item i (this is measured using the $esim()$ function) and we compare these reactions with the expected global reaction. If, for example, this gossip about the actor elicited amusement among the majority of users, and the set of users most affectively similar to the user u also finds this news item amusing, then the score of the item i for the user u is diminished (the user is expected to be amused by the news item). However, if the predominant emotional reaction to the item i is amusement, but the majority of users most affectively similar to the user u find this item sad (in other words, if the group of users who have similar emotional reactions to the user u has different emotional reaction to the item i than the global population), the score of the item i for the user u in the context of emotional reaction ”sad” increases.

Table 1 Number of users who record emotional reactions

# reactions	# users	% of users with reactions	% of all users
At least 1	65,000	100	2.2
Less than 3	53,600	87	1.9
Exactly 1	45,980	76	1.7
10 or more	1567	2.6	0.06
20 or more	723	1.1	0.026

Equation 7 is the fundamental formula used by our affective recommender system. It allows to find items which are the most relevant for a given user, taking into consideration affective feedback collected from all users.

4 Architecture and implementation

4.1 Data and architecture overview

During the course of six months our widget collected over 13,000,000 impressions and over 160,000 reactions from 2,700,000 unique users over a set of 85,000 news items. Over 1,500,000 impressions have been generated by users who either clean their cookies after the visit, or who have visited the site only once. 25% of users have left more than one but less than seven impressions, 25% of users have left more than seven impressions, and 5.5% of the users have left more than 80 impressions. Table 1 presents statistics regarding the collection of emotional reactions. Over 65,000 users have recorded their emotional reactions, 87% have recorded less than three emotional reactions, and 76% have recorded just one emotional reaction. Only 1567 users have recorded ten or more emotional reactions and 723 users have recorded more than twenty emotional reactions. As we can see, despite very careful design of the emotional reaction widget, a very small percentage of users who visit news websites records their emotional reactions, which poses a significant challenge to emotion-aware recommender systems.

The traffic came from three websites: epoznan.pl, radiomerkury.pl and culturowo.pl, with over 90% of traffic coming from epoznan.pl, which is the top 400th most visited website in Poland. The most popular news item gathered 70,000 impressions and the most emotional item have triggered over 1600 emotional reactions. More than 2,000,000 recommendation sets were served, each consisting of four recommended news items.

This traffic required a robust architecture capable of serving recommendations in real time and taking advantage of new user feedback as it arrived. We have used the high performance, in-memory data structure server Redis and the Elasticsearch NoSQL database. To provide a convenient way for asynchronous computation and increase future scalability of the solution we have used the distributed task queue Celery. The backbone of the entire system was implemented in Python using the Django framework.

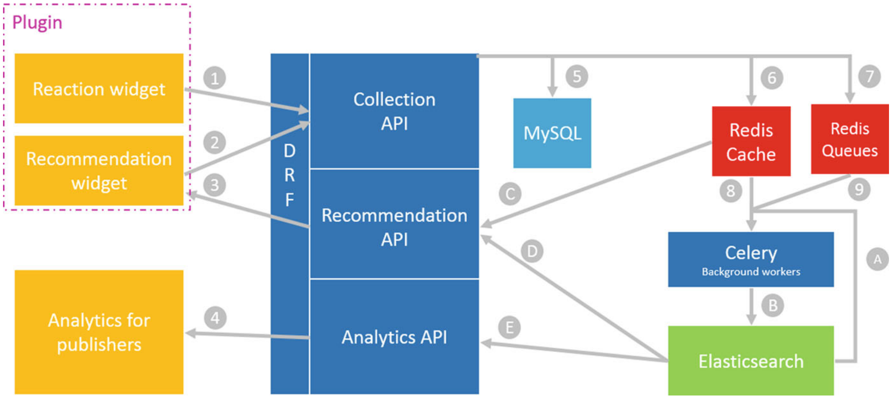

Fig. 6 Architecture overview of the EARS

Figure 6 depicts the high level components and the communication between them. The Collection API processes actions reported by the Reaction Widget (1), such as article impressions, user emotional reactions, and Facebook shares of user emotional reactions. The Recommendation widget also reports recommendation clicks to the Collection API. All actions are then stored into MySQL (5). Synchronously the actions are added to user and item caches (6), serialized to JSON and added to insert queues (7). The profiles of users and items that participated in these actions are marked as “dirty” and added to update sets (7) to be refreshed by background workers. Insert queues are then periodically consumed by Celery workers (9) and batch inserted into Elasticsearch (B). Profiles marked as dirty are also periodically consumed (in batches) by Celery workers (9) that pull data from Redis cache and aggregate action data in Elasticsearch (A) to construct up-to-date profiles, which are inserted (B) into Elasticsearch. The Recommendation API then uses data stored in Redis cache (C) to construct recommendation queries that are run against user and item profiles stored in Elasticsearch (D). The recommendations are then served to the Recommendation widget (3) and displayed to the user. Actions indexed in Elasticsearch are aggregated (E) to provide general and item-wise statistics (4) to Analytics for publisher.

4.2 Algorithms

All algorithms were implemented in Elasticsearch with the use of the Elasticsearch Query DSL (Domain Specific Language) and Lucene expressions. Lucene expressions allow to run arbitrary computations on a single document, over a set of documents, in parallel and distributed over any number of machines. This architecture allowed us to express a lot of algorithms in a declarative way and write simple extensions for capabilities that were not available out of the box. The system was designed to handle multiple forms of hybridization and to support collaborative filtering, knowledge-based and content-based recommendation techniques. Knowledge-based recommenders were implemented as Elasticsearch queries enhanced with Lucene expressions over statistical profiles of users.

With query DSL and robust item profiles it was easy to express algorithms like:

- recommend items not older than 14 days,
- from the domain the user is currently viewing,
- for which the dominating emotional reaction is different that that of the article hosting the recommendation,
- and was reported by more than 40% of people who gave their reaction,
- excluding items already seen by user.

We have essentially dissected a simple knowledge-based algorithm into individual parts that can be expressed in the Elasticsearch query DSL enhanced with Lucene expressions. Collaborative filtering recommendations were also implemented by querying user profiles with query terms determined by the list of news items visited by the user for whom we were serving a recommendation. After finding the list of similar users, news item profiles were searched for the second time. News items returned by the second query were served as recommendations. Affective user similarity was implemented analogously, but instead of a single query, multiple queries were combined to find similar users. Finally, we have employed three hybridization techniques, namely switching, cascade and weighted hybridization. The switching technique was used to diversify recommendations, increase reliability and give a convenient way to test new algorithms. Each algorithm was assigned a very fast validation procedure, that determined if it were suitable in a given situation. For example an algorithm could require a user to have a minimum history of 3 impressions. Algorithms which passed the validation stage entered the pool of available algorithms. Then a candidate algorithm was drawn using roulette selection based on manually assigned algorithm priority. Furthermore each algorithm had a *complexity level*. In case of failure of the selected algorithm, all algorithms with equal complexity level were excluded from the pool and a new candidate was drawn. Lowest complexity algorithms were guaranteed to work every time (since they did not involve personalization) and hence a recommendation procedure never failed. Switching was implemented as a part of application logic in the Recommendation API. Cascade hybridization in our system boiled down to the use of filters. An arbitrary number of filters could be applied before the scoring algorithm was applied, this enabled algorithms such as collaborative filtering that recommended items with a given dominating reaction or limiting the set of recommendation candidates to articles published in the last 14 days. A filter could also be based on an arbitrary expression. Re-scoring was also available although it was not used throughout the experiments. Cascade hybridization was implemented with the use of Elasticsearch query DSL. Finally, weighted hybridization was implemented as a two phase procedure. Since each algorithm was essentially a complex Elasticsearch query that could involve an arbitrary functional expressions, there was no way of knowing what the maximum score of the query was before running it. Therefore the first phase involved running each recommendation query separately to asses the maximum score. Then an expression that combined the queries and assigned weights was composed using respective maximum scores to scale each constituting score.

It should be noted that, although we have implemented an industry-grade recommendation engine containing a hybridization pipeline of state-of-the-art recommenders (content-based, collaborative filtering and knowledge-based), the design and development of the most effective engine was not the main goal of this research. Instead, we were trying to assess the usefulness and benefit of using affective item features in news items recommendations. Thus, in order not to skew the results by using a non-realistic setting, we have decided to conduct our experiments on live traffic. As the result, the baseline of the experiment was very strong in terms of the click through rate (CTR). Had we compared our solution with pure content-based, collaborative filtering, or knowledge-based recommender, our results would be much stronger, but such setting would have been unrealistic and over optimistic. We strongly believe that our experimental setup is the most accurate representation of the real world conditions in which affective recommender system would have to compete.

5 Experiments and results

The experiments were performed on live traffic for the epoznan.pl news website. All recommendations were performed in real time. Tests were run on an Ubuntu 14.04 LTS machine with 14 cores of an Intel® Xeon® Processor E5-2630 v2 (15M Cache, 2.60 GHz), 100 GB of SSD and 72 GB of RAM. 95% of recommendation requests finished in under 2000 ms with median request time at 500 ms. The set of similar users was computed on every request and so were the resulting recommendations. As the evaluation metrics we have decided to use click-through rates (CTRs). CTR is the ratio of recommendation sets with a hit (a recommendation which resulted in a click) to all generated recommendation sets. For CTR to be statistically significant, it has to be collected on a sufficiently large sample and over an extended period of time.

5.1 Experiment 1: affective user similarity and hybridization with affective item features

In the first experiment we focused on investigating the performance of affective user similarity (Eq. 5) and exploring hybridization opportunities of both traditional user-based collaborative filtering (UBCF) and user-based collaborative filtering with affective user similarity (AUBCF, Sect. 3.3.3). We have investigated 12 variations of algorithms varying user history length ($[3, 4]$, $[5, 9]$, $[10, \infty)$), which resulted in the total of 36 algorithm configurations. When computing similarity between emotions (Eq. 5) we have used both cityblock and euclidean distances. For UBCF and AUBCF we have investigated their performance both without hybridization and with weighted hybridization using pleasantness, unpleasantness and controversy (Sect. 3.3.1). These affective item features were included in the hybridization algorithm by putting preferable weights on news items which had these features. Each algorithm operated on a candidate set of news items from the epoznan.pl domain that were not older than 14 days and not previously seen by the user. Each algorithm used all available history to asses similarity between users and selected the neighborhood of 30 similar users. Each knowledge-based algorithm used additional pre-filtering which limited candidate sets to news items which received at least 20 emotional reactions. The experiment run for two weeks and over 300,000 recommendation sets were served, with approximately

Table 2 Click-through rate for user-based collaborative filtering recommender w.r.t. user history length

User history length	Successes	Total	CTR (average)
UBCF [3,4]	735	29,572	2.3–2.7% (2.5%)
UBCF [5,9]	445	18,764	2.1–2.6% (2.4%)
UBCF [10, ∞]	400	19,177	1.9–2.3% (**2.1%**)

30% of the traffic directed to the baseline algorithm. The baseline algorithm was a random selection of 4 news items from the set of 12 most popular news items not older than 14 days and not previously seen by the user. The baseline click-through rate (CTR) was 0.02156, a very high baseline threshold which poses a significant challenge to a recommender system. Detailed results are presented in the Appendix, below we discuss the aggregated results.

The first interesting aggregated observation is presented in Table 2. The table presents the comparison of pure UBCF algorithms depending on the length of user history, the second column reports on the number of successful clicks, the third column reports on the number of recommendation sets shown to users, and the fourth column contains the click-through rate (minimum, maximum and average). Scenario in which the average CTR is statistically significantly different from the baseline (as measured by the *p* value of the two-sample test of averages) is marked with bold face. One thing clearly stands out in the results. The CTR for pure UBCF with no affective features for users with the history larger than 10 impressions is lower than for users with shorter histories (and the difference is statistically significant. For users with shorter histories the average CTR is between 2.4% and 2.5% (the baseline CTR is 2.2%), but the difference in these averages is not statistically significant. While counter-intuitive at first (larger histories usually yield higher precision and recall), this phenomenon could be explained by the way power users differ from casual users. It is probable that these users in their reading sessions adhere to a certain routine which involves coming back to the home page or category page and screening through the list of available news items, and recommendations disrupt this natural flow. No analogous, statistically significant decrease in CTR was observed for AUBCF. For the above reason we have decided to focus on the limited subset of samples and report the cases for all available history only when it outperformed the baseline in a statistically significant way.

Table 3 presents aggregate results for two types of affective recommender systems: pure user-based collaborative filtering systems augmented with affective item features and full affective user-based collaborative filtering systems. The interpretation of columns is the same as in the case of Table 2. For instance, UBCF with user histories shorter than 10 items with the preference for news items with high pleasantness (second row of Table 3) has produced 12,108 recommendation sets, and these recommendations resulted in 330 clicks. The average success rate for this algorithm was 2.7%, and the average improvement was 26%, which was statistically significantly better than the baseline (the *p* value of 0.0004 allows to reject the null hypothesis of no statistical difference in the average CTR). When examining the results we quickly notice that AUBCF is consistently outperformed by the baseline algorithm, with approximately 21% lower CTR. The hybridization and adding affective item features does

Table 3 Click-through rate for affective user-based collaborative filtering recommenders

Algorithm	Successes	Total	CTR (average)	p value	Improvement (%)
Baseline	1847	85,666	2–2.3% (2.2%)	–	–
UBCF [3, 9] + pleasantness	330	12108	2.4–3.1% (**2.7%**)	0.0004	26
UBCF [3, 9]	277	11,924	2–2.7% (2.3%)	0.8	7.7
AUBCF [3, 9]	220	12,849	1.4–2% (**1.7%**)	0.0063	− 21
UBCF [3, 9] + controversy	295	12,192	2.1–2.8% (2.4%)	0.32	12
UBCF [3, 9] + unpleasantness	278	12,112	2–2.7% (2.3%)	0.09	6.5
UBCF [3, ∞] + pleasantness	463	16,998	2.3–2.9% (**2.6%**)	0.0055	19

not improve the performance of AUBCF. A reasonable conclusion is that investigated affective similarity is not suitable to be combined with collaborative filtering. This can be explained by the fact that AUBCF has two orders of magnitude smaller history (it is based on emotional reactions) and a much smaller candidate user set at its disposal when it comes to finding a set of similar users. Consequently, similar users identified by affective similarity measure have less predictive power in extrapolating user impressions than a competing set based on larger history of impressions. AUBCF is also not affected by hybridization with no statistically important differences between pure and hybridized variations. No difference was recorded between using euclidean and cityblock distance measures for assessing emotional reaction similarity in AUBCF. Interestingly, no statistically important improvement over baseline was recorded for UBCF despite its 7.7% higher CTR or its hybrids combined with unpleasantness ranking (6.6% higher CTR) and controversy ranking (12.5% higher CTR). This suggests that UBCF may serve as a valuable component of a hybrid strategy (it performs better than the baseline), but it is not capable of solving the recommendation problem for the media industry as a standalone technique.

In our experiment one hybridization stands out as a clear winner. A significant improvement of 26% for users with history of less than 10 impressions and 19% regardless of user history was recorded for a hybrid of UBCF and pleasantness ranking. The effectiveness of pleasantness as a part of the hybrid algorithm based on UBCF proves that emotions can be effectively used to improve performance of recommender systems. Let us recall that pleasantness has been defined to maximize pleasure and dominance dimensions of the multi-dimensional space. In our model emotional reactions that have high pleasantness are the reactions of being inspired, surprised, and amused. In our opinion these results strongly suggest that positive, elevating news items are far more captivating and engaging. People seem to have strong preference towards uplifting contents and tend to actively seek news with positive contents.

As we have discovered, using affective user similarity did not produce results surpassing the baseline. This result can lead to six possible conjectures:

- using Affective Norms for English Words (ANEW) Bradley and Lang (1999) as a proxy for assessing pleasure, arousal, and dominance is inaccurate,
- the proposed emotional reaction similarity measures do not capture the actual relationships between emotional reactions, which results in an inaccurate assessment of similarity between two users and hence an inaccurate selection of neighbors,

- the proposed formula for affective user similarity measure is defective,
- the scarcity of emotional reaction feedback as compared to implicit feedback (impressions) outweighs the utility of AUBCF when compared to UBCF as it significantly reduces the number of users available for comparison (only 7,800 with more than two emotional reactions, as opposed to 600,000 with more than seven impressions),
- emotional similarity does not produce neighborhoods that can be efficiently used to extrapolate user interest and hence the probability of consuming a recommendation.
- news items tend to have very short life spans and they loose their appeal quickly, independent of emotions they elicit. In order for affective similarity (either affective similarity between users or affective item feature similarity) to influence the recommendation process, some minimum threshold of emotional reactions must be recorded for a given item and a given emotional reaction. It is possible that, despite the fact that we have deployed our recommender on a live and popular website, the sheer volume of traffic was too low to collect enough emotional reactions and the results would have become visible only after deployment on a much more popular website.

We look forward to investigating the first three conjectures in future work. At this stage it seems that in the context of news items recommendations, using affective features in combination with pure user-based collaborative filtering is more promising than trying to assess the similarity between users based on the similarity of their emotional reactions.

5.2 Experiment 2: targeting by expected reaction

The objective of the second experiment was to investigate the performance of the affective user based collaborative filtering with targeting by expected reaction (AUBCF-WTBER) and its hybrids with controversy-based ranking. The idea behind the experiment is simple: if the analysis of user's history reveals that a certain emotional reaction is predominant among users who are similar to a given user, we try to elicit that particular emotional reaction by showing news items which often result in that emotional reaction. For instance, if Ann's reading patterns make her most similar to Bill and Crystal, and for Bill and Crystal the most common emotional reaction is "inspired", we will recommend to Ann news items which have inspired many users. We have tested targeting by 7 out of 8 available emotional reactions (excluding the indifferent reaction) using both euclidean and cityblock emotion similarity distance metrics. This resulted in the total of 28 investigated algorithm variations. Each algorithm operated on a candidate set of news items from the epoznan.pl domain which were not older than 14 days and have not been previously seen by the user. Each algorithm used all available history to asses similarity between users and to select the neighborhood of 30 most similar users. Each knowledge-based algorithm used additional pre-filtering which limited candidate sets to news items that received at least 20 emotional reactions.

The experiment run for one week and over 110,000 recommendation sets were served, with approximately 45% of the traffic being directed to the baseline algorithm. The baseline algorithm was a random selection of 4 news items from the set of 12 most popular news items not older than 14 days and not previously seen by the user. The baseline CTR was 0.0252, a very powerful baseline especially for casual users. Please note that the baseline CTR in Experiment 2 was significantly higher than in Experiment 1, presenting a demanding challenge for the recommender system. Raw results of the experiment are presented in the Appendix in Table 5, here we present our interpretation of the results. The first conclusion is that a pure AUBCF-WTBER suffers from significantly reduced coverage as it served approximately 5 times fewer recommendation sets than its hybridized counterparts despite equal traffic allocation. This means that the algorithm has failed to provide a sufficient number of recommendations in approximately 80% of cases. Because of this reduced coverage no statistically significant improvement over the baseline could be established. Nevertheless, we note that click-through rates for certain emotional reactions, namely “amused”, “sad” and “surprising”, are very promising, despite the challenging baseline threshold. This result partially validates our previous findings regarding the utility of pleasantness. All three emotional reactions, for which we observe an improvement in click-through rates, are placed at the extreme values of the pleasure dimension of the PAD model. This leads us to believe that both affective item features related to pleasure and emotional reactions related to pleasure play a crucial role in the process of discovering and consuming news items.

Another consequence of the reduced coverage was a poorer performance of hybridized variations. A plausible explanation for this is that reduced coverage produced small or empty candidate sets from AUBCF-WTBER in most of the cases, essentially reducing the algorithm to a controversy ranking. Differences between target reactions are not assessed, because a small number of trails per target for AUBCF-WTBER and skewed results for hybrids makes such evaluation misleading.

The second experiment proved inconclusive whether AUBCF-WTBER can provide better results than the baseline algorithm, but it highlighted an important issue of reduced coverage that significantly limits the commercial utility of the method. We are looking forward to testing AUBCF-WTBER over a more extensive period of time to asses its potential for users for whom a sufficient set of candidate recommendations exists.

6 Summary and conclusions

This paper presents the practical evaluation of the Emotion Aware Recommender System on live traffic from various news sites. Since affective recommender systems are a new field, our research attempts to provide a comprehensive overview of approaches one may consider, and challenges one may expect when attempting to leverage emotions in the field of content recommendations. Let us recall the four main goals of our work:

- to investigate how different psychological and behavioral concepts can be applied in the context of recommender systems,
- to investigate methods for collection of emotional reactions and assess their usability in the context of recommender systems,
- to introduce a new, unobtrusive and scalable collection method of emotional reactions.
- to discuss different emotion models and how they can be used in recommender systems,

With regard to the first goal we have investigated the competitive landscape of recommender systems for media publishers and key challenges and limitations such as short shelf life of articles for news publishers, fluctuations of general interest due to trends, perpetual new user problem or user trust. Subsequently we have described state-of-the-art techniques used in recommender systems and highlighted their key advantages and weaknesses from the perspective of news items recommendations. We have focused on user based collaborative filtering and how it can be applied to arbitrary types of signals given that a suitable notion of similarity between signals is defined. We have also considered knowledge-based recommenders which we had identified as a technique most suitable for using emotional features that can be derived from user feedback.

The second and third goals were achieved by careful investigation of available emotional models (both discrete and multidimensional) and user experience factors. We have designed a widget for emotional reaction collection and we have developed a mapping from the discrete emotional model presented to end users to the multidimensional emotional model used by algorithms. We have also presented the architecture of an industry-grade recommendation service capable of serving real world workloads, outlining engineering challenges posed by the productization of affective recommender systems in the online media industry.

As for the fourth goal, we have proposed three techniques that can be used to augment recommender systems with affective features. First, we have introduced affective item features such as controversy, diversity, pleasantness, and unpleasantness in order to improve the computation of similarity between users. We have formulated the affective user similarity, a similarity measure for user-based collaborative filtering computed from emotional reactions reported by users. We have also introduced two emotional reaction similarity metrics that use the dimensional model of emotions. Based on affective user similarity we have introduced a novel method for recommending items—user-based collaborative filtering with targeting by expected reaction.

We have performed two experiments on live traffic. In the first experiment we have compared traditional user-based collaborative filtering with a recommender using our affective user similarity and we have hybridized both techniques with algorithms based on ranking items according to affective item features. We have found that a hybrid of user-based collaborative filtering and pleasantness-based ranking has consistently outperformed all other algorithms. The second experiment investigated the potential of affective user-based collaborative filtering with targeting by expected reaction. This experiment has proven to be inconclusive due to significantly reduced coverage, but the results strongly suggest that targeting by selected emotional reactions (amused, sad, surprised) leads to the improvement of the click-through rates.

We would like to conclude this paper with a brief comment of the experimental results. Although only some of the combinations of affective recommendations show statistically significant improvement, we note an important fact. Due to technical and legal reasons we were not able to run the experiment for a longer period of time, increasing the sample size as the result. It is a well-known phenomenon that any effects measured by statistical tests are harder to detect in smaller samples. In order to increase the statistical power of the test, we would have to collect more data (which, unfortunately, we could not have done). On a closer inspection of Table 3 one finds that all affective recommenders improve the baseline, though not enough to reject the null hypothesis. If we were to attribute the success of using pleasantness in recommendations to either chance or some latent factor, we would not expect other recommenders to be any different from the baseline. It is possible that simply by increasing the sample size (in other words, by having a larger coverage for each recommendation scenario) we could improve the power of the test and show that affective recommenders work in general, despite a highly competitive baseline. Unfortunately, this remains our conjecture as we cannot prove it.

Acknowledgements The development of the emotion reaction widget and the implementation of the collection service has been financially supported by the European marketing data science company Analyx (www.analyx.com). The authors wish to express their gratitude to Analyx for continuous support and kind permission to publish the scientific results of the project. The authors also wish to thank anonymous reviewers, whose detailed and thorough comments and remarks have significantly improved the quality of the manuscript.

Appendix

In the Appendix we present raw results of experiments conducted on affective user similarity, the usefulness of affective item features in collaborative filtering, and recommending items based on expected user emotional reactions. Table 4 presents raw results of Experiment 1 (Sect. 5.1). The first column indicates whether affective user similarity (Eq. 6) has been used. The second column indicates whether emotional reaction similarity (Eq. 5) has been used, and if so, which distance measure has been used. The next three columns indicate whether affective item features have been used. The sixth column represents the user history depth used when generating recommendations. Finally, the last three columns report on the click-through rate, the number of recommendations, and the number of clicks, respectively. The baseline CTR was 0.02156, results above the baseline are marked with the bold face.

Table 5 presents the raw results of Experiment 2 (Sect. 5.2). The first column represents the distance measure used for computing emotional reaction similarity (Eq. 5), the second column indicates if targeting by emotion applied only to most

Table 4 Experiment 1 : A—whether affective item features have been used, B—which distance function was used to compute emotional reaction similarity, C—use of unpleasantness score feature, D—use of the pleasantness score feature, E—use of the controversy score feature, F—minimum length of user history, scenarios where the average CTR is above the baseline are marked with bold face

A	B	D	E	F	G	CTR	Count	Clicked
No	None	No	No	No	3	**0.023708**	7297	173
					5	**0.022477**	4627	104
					10	0.019570	4701	92
				Yes	3	**0.024823**	7493	186
					5	**0.023196**	4699	109
					10	**0.023913**	4809	115
			Yes	No	3	**0.027052**	7430	201
					5	**0.027576**	4678	129
					10	**0.021721**	4880	106
		Yes	No	No	3	**0.023803**	7352	175
					5	**0.021639**	4760	103
					10	0.018174	4787	87
Yes	Cityblock	No	No	No	3	**0.023704**	1350	32
					5	0.016643	9073	151
					10	0.013799	1087	15
				Yes	3	0.016506	7270	120
					5	0.014254	37953	541
					10	0.013930	4738	66
			Yes	No	3	0.011707	7346	86
					5	0.013545	4725	64
					10	0.013319	4805	64
		Yes	No	No	3	0.014045	7405	104
					5	0.011782	4753	56
					10	0.016187	4757	77
	Euclidean	No	No	No	3	0.012546	1355	17
					5	0.018674	1071	20
					10	0.016807	1190	20
				Yes	3	0.013691	7304	100
					5	0.014255	4700	67
					10	0.013860	4762	66
			Yes	No	3	0.013203	7271	96
					5	0.015926	4772	76
					10	0.015106	4634	70
		Yes	No	No	3	0.014511	7236	105
					5	0.014021	4636	65
					10	0.010434	4696	49

Table 5 Experiment 2 : A—distance function used to compute emotional reaction similarity, B—use of the controversy score feature, scenarios where the average CTR is above the baseline are marked with bold face

A	B	Target	CTR	Count	Clicked
Cityblock	No	Amused	0.012238	572	7
		Angry	0.005455	550	3
		Informed	0.016245	554	9
		Inspired	0.005618	534	3
		Sad	**0.026168**	535	14
		Scared	0.024955	561	14
		Surprised	0.013311	601	8
	Yes	Amused	0.013113	3508	46
		Angry	0.013845	3467	48
		Informed	0.008525	3519	30
		Inspired	0.010888	3490	38
		Sad	0.007595	3555	27
		Scared	0.010373	3567	37
		Surprised	0.011478	3572	41
Euclidean	No	Amused	**0.027027**	592	16
		Angry	0.009328	536	5
		Informed	0.009276	539	5
		Inspired	0.014085	568	8
		Sad	**0.031193**	545	17
		Scared	0.017668	566	10
		Surprised	**0.025862**	580	15
	Yes	Amused	0.010687	3462	37
		Angry	0.008197	3538	29
		Informed	0.013498	3556	48
		Inspired	0.009033	3432	31
		Sad	0.009249	3460	32
		Scared	0.009305	3439	32
		Surprised	0.012307	3494	43

controversial news items, or all news items. Third column indicates the target emotional reaction. The last three columns report on the click-through rate, number of served recommendation sets, and the number of clicks, respectively. The baseline CTR was 0.0252, results above the baseline are marked with the bold face.

References

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

Ahmadpour (2014) OCC model: application and comparison to the dimensional model of emotion. In: Proceedings of the 5th Kanesi Engineering and Emotion Research, International Conference, Linköping, Sweden, 11–13 June, pp. 607–617 (2014)

Andjelkovic, I., Parra, D., O'Donovan, J.: Moodplay: interactive mood-based music discovery and recommendation. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, pp. 275–279. ACM (2016)

Arnold, M.B.: Emotion and Personality. Columbia University Press, New York (1960)

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl.-Based Syst. **46**, 109–132 (2013)

Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **1**(25), 49–59 (1994)

Bradley, M.M., Lang, P.J.: Affective norms for English words (anew): instruction manual and affective ratings. Technical report, Citeseer (1999)

Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp. 43–52 (1998)

Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adap. Inter. **12**(4), 331–370 (2002)

Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)

Cheung, K.W., Tian, L.F.: Learning user similarity and rating style for collaborative recommendation. Inf. Retr. **7**(3–4), 395–410 (2004)

Colby, B.N., Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Contemp. Sociol. **18**(6), 957 (1989)

Costa, H., Macedo, L.: Emotion-based recommender system for overcoming the problem of information overload. In: International Conference on Practical Applications of Agents and Multi-agent Systems, pp. 178–189. Springer (2013)

Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 39–46. ACM (2010)

Desmet, P.M.A., Hekkert, P.: The basis of product emotions. In: Green, W., Jordan, P. (eds.) Pleasure with products, beyond usability, pp. 60–68. Taylor & Francis, London (2002)

Doychev, D., Lawlor, A., Rafter, R., Smyth, B.: An analysis of recommender algorithms for online news. In: CLEF 2014 Conference and Labs of the Evaluation Forum: Information Access Evaluation Meets Multilinguality, Multimodality and Interaction, 15–18 Sept 2014, Sheffield, UK, pp. 177–184 (2014)

Dunning, D., Heath, C., Suls, J.M.: Flawed self-assessment: implications for health, education, and the workplace. Psychol. Sci. Public Interest **5**(3), 69–106 (2004)

Ekman, P.: Basic emotions. In: Dalgleish, T., Power, M. (eds.) Handbook of Cognition and Emotion, pp. 45–60. John Wiley & Sons, Chichester, England (1999)

González, G., De La Rosa, J.L., Montaner, M., Delfin, S.: Embedding emotional context in recommender systems. In: Proceedings—International Conference on Data Engineering, pp. 845–852 (2007)

Gur, Y.: Sequential optimization in changing environments: theory and application to online content recommendation services. Ph.D. thesis, Columbia University (2014)

Han, B.J., Rho, S., Jun, S., Hwang, E.: Music emotion classification and context-based music recommendation. Multimedia Tools Appl. **47**(3), 433–460 (2010)

Ilievski, I., Roy, S.: Personalized news recommendation based on implicit feedback. In: Proceedings of the 2013 International News Recommender Systems Workshop and Challenge, pp. 10–15. ACM (2013)

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, Cambridge (2010)

Kaminskas, M., Ricci, F.: Emotion-based matching of music to places. In: Tkalčič, M., De Carolis, B., de Gemmis, M., Odić, A., Košir, A. (eds.) Emotions and Personality in Personalized Services, pp. 287–310. Springer, Berlin (2016)

Kazai, G., Yusof, I., Clarke, D.: Personalised news and blog recommendations based on user location, Facebook and Twitter user profiling. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1129–1132. ACM (2016)

Kille, B., Brodt, T., Heintz, T., Hopfgartner, F., Lommatzsch, A., Seiler, J.: Overview of CLEF NewsREEL 2014: news recommendation evaluation labs (2014)

Kille, B., Lommatzsch, A., Turrin, R., Sereny, A., Larson, M., Brodt, T., Seiler, J., Hopfgartner, F.: Overview of CLEF News REEL 2015: News Recommendation Evaluation Lab. In: Working Notes of CLEF 2015 – Conference and Labs of the Evaluation forum, Toulouse, France, September 8–11 (2015)

Lang, P.J., Bradley, M.M., Cuthbert, B.N., et al.: International affective picture system (IAPS): affective ratings of pictures and instruction manual. NIMH, Center for the Study of Emotion & Attention (2005)

Lazarus, R.S., Averill, J.R., Opton Jr., E.M.: Toward a cognitive theory of emotions. In: Arnold, M. (ed.) Feelings and Emotions, pp. 207–32. Academic, New York (1970)

Lin, C., Xie, R., Guan, X., Li, L., Li, T.: Personalized news recommendation via implicit social experts. Inf. Sci. **254**, 1–18 (2014)

Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, pp. 31–40. ACM (2010)

Luo, J., Yu, R.: Follow the heart or the head? The interactive influence model of emotion and cognition. Front. Psychol. **6**, 573 (2015)

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

Montaner, M., López, B., De La Rosa, J.L.: A taxonomy of recommender agents on the internet. Artif. Intell. Rev. **19**(4), 285–330 (2003)

Nunes, M.A.S., Hu, R.: Personality-based recommender systems: an overview. In: Proceedings of the Sixth ACM Conference on Recommender Systems, pp. 5–6. ACM (2012)

Oatley, K., Johnson-laird, P.N.: Towards a cognitive theory of emotions. Cognit. Emot. **1**(1), 29–50 (1987)

Oatley, K., Keltner, D., Jenkins, J.M.: Understanding Emotions, 2nd edn. Blackwell Publishing, Malden (2006)

Picard, R.W.: Affective Computing. MIT Press, Cambridge (2000)

Plutchik, R.: The nature of emotions. Am. Sci. **89**(4), 344–350 (2001)

Reisenzein, R.: Emotional experience in the computational belief-desire theory of emotion. Emot. Rev. **1**(3), 214–222 (2009a)

Reisenzein, R.: Emotions as metarepresentational states of mind: naturalizing the belief–desire theory of emotion. Cognit. Syst. Res. **10**(1), 6–20 (2009b)

Reisenzein, R.: Broadening the scope of affect detection research. IEEE Trans. Affect. Comput. **1**(1), 42–45 (2010)

Ricci, F., Rokach, L., Shapira, B.: Introduction to Recommender Systems Handbook. Springer, Berlin (2011)

Russell, J.A.: A circumplex model of affect. J. Personal. Soc. Psychol. **39**(6), 1161–1178 (1980)

Sabini, J., Cosmas, K., Siepmann, M., Stein, J.: Underestimates and truly false consensus effects in estimates of embarrassment and other emotions. Basic Appl. Soc. Psychol. **21**(3), 223–241 (1999)

Scherer, K.R., Scherer, K.R., Ekman, P.: On the nature and function of emotion: a component process approach. Approaches Emot. **2293**, 317 (1984)

Schröeder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C., Zovato, E.: Emotion markup language (emotionml) 1.0. W3C Working Draft, vol. 29, pp. 3–22 (2010)

Shiv, B., Fedorikhin, A.: Heart and mind in conflict: the interplay of affect and cognition in consumer decision making. J. Consum. Res. **26**(3), 278–292 (1999)

Spertus, E., Sahami, M., Buyukkokten, : Evaluating similarity measures: a large-scale study in the Orkut social network. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 678–684. ACM (2005)

Strle, G., Pesek, M., Marolt, M.: Towards user-aware music information retrieval: emotional and color perception of music. In: Tkalčič, M., De Carolis, B., de Gemmis, M., Odić, A., Košir, A. (eds.) Emotions and Personality in Personalized Services, pp. 327–353. Springer, Berlin (2016)

Tao, J., Tan, T.: Affective computing: a review. In: Tao, J., Tan, T., Picard, R.W. (eds.) Affective Computing and Intelligent Interaction, pp. 981–995. Springer, Berlin (2005)

Tkalčič, M., Burnik, U., Košir, A.: Using affective parameters in a content-based recommender system for images. User Model. User-Adap. Inter. **20**(4), 279–311 (2010)

Tkalčič, M., Košir, A., Tasić, J.: Affective recommender systems: the role of emotions in recommender systems. In: Proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems, pp. 9–13. Citeseer (2011)

Tkalčič, M., Burnik, U., Odić, A., Košir, A., Tasič, J.: Emotion-aware recommender systems—a framework and a case study. In: ICT Innovations 2012, pp. 141–150. Springer (2013a)

Tkalčič, M., Odić, A., Košir, A., Tasić, J.: Affective labeling in a content-based recommender system for images. IEEE Trans. Multimedia **15**(2), 391–400 (2013b)

Wakil, K., Ali, K., Bakhtyar, R., Alaadin, K.: Improving web movie recommender system based on emotions. Int. J. Adv. Comput. Sci. Appl. **6**(2), 218–226 (2015)

Wang, J.C., Yang, Y.H., Wang, H.M.: Affective music information retrieval. arXiv preprint arXiv:1502.05131v1 (2015)

Zanker, M., Aschinger, M., Jessenitschnig, M.: Development of a collaborative and constraint-based web configuration system for personalized bundling of products and services. In: Web Information Systems Engineering—WISE 2007, pp. 273–284. Springer (2007)

Zheng, Y., Mobasher, B., Burke, R.D.: The role of emotions in context-aware recommendation. Decisions RecSys **2013**, 21–28 (2013)

Jan Mizgajski received his Ms. (2015) degree from Poznan University of Technology, Poland. His research interests focus on machine learning and operations research and their applications in natural language processing, information retrieval, recommender and assistant systems, affective computing and advertising technology. He acts as an independent machine learning, data engineering and product ownership consultant organizing and leading expert teams in creating products that leverage new scientific developments in his research focus areas. He founded and hosts meet.ml—a machine learning and big data meetup in Poznan, Poland.

Mikołaj Morzy is an Associate Professor in the Institute of Computing at Poznan University of Technology, Poland. He received his Ph.D. (2004) and D.Sc. (2010) degrees from Poznan University of Technology. His research interests focus on machine learning and its applications in recommender systems, complex network systems, and social networks. He is currently holding the post of the vice-dean for science at the Faculty of Computing. He is also the member of the Editorial Board of PLOS One. In the past he has been working at the Westfalische-Wilhelms Universität Münster (Germany) and Loyola University New Orleans (USA).