Using Emotional Context from Article for Contextual Music Recommendation

Chih-Ming Chen[1], Ming-Feng Tsai[1], Jen-Yu Liu[2], Yi-Hsuan Yang[2]

[1]Department of Computer Science & Program in Digital Content and Technology
National Chengchi University
Taipei 11605, Taiwan
{g10018, mftsai}@cs.nccu.edu.tw

[2]Research Center for Information Technology Innovation
Academia Sinica
Taipei 11564, Taiwan
{ciaua, yang}@citi.sinica.edu.tw

ABSTRACT

This paper proposes a context-aware approach that recommends music to a user based on the user's emotional state predicted from the article the user writes. We analyze the association between user-generated text and music by using a real-world dataset with <user, text, music> tripartite information collected from the social blogging website LiveJournal. The audio information represents various perceptual dimensions of music listening, including danceability, loudness, mode, and tempo; the emotional text information consists of bag-of-words and three dimensional affective states within an article: valence, arousal and dominance. To combine these factors for music recommendation, a factorization machine-based approach is taken. Our evaluation shows that the emotional context information mined from user-generated articles does improve the quality of recommendation, comparing to either the collaborative filtering approach or the content-based approach.

Categories and Subject Descriptors

I.m [**Computing Methodologies**]: Miscellaneous

Keywords

Emotion-based music recommendation, Listening context

1. INTRODUCTION

Music usually carries people's emotions, and people sometimes express their feelings by writing articles while listening to music. In light of this obervation, we propose to employ the emotional context information manifested in user-generated articles to build a context-aware music recommendation system. Although emotion has been shown a useful cue for matching songs to documents according to the audio and text content [3], less work has been done on using emotional context features mined from user-generated articles to improve the quality of music recommendation.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502170.

There have been many studies on contextual recommendation. For example, Jiang *et al.* [8] developed a novel way to represent social networks with multiple relational domains and employed the hybrid random walk techniques to learn the pattern of users' preference. Kailong *et al.* [4] performed tweet recommendations by a collaborative ranking method that captures personal interests. In the KDDCup 2012 competition, Chen *et al.* [5] combined a variety of models by blending different features. Their result shows that adopting more diverse information helps cover more possible targets, suggesting that contextual information could be a good information source for recommendation. Many studies [7, 12, 15, 18] have tried to model the behavior of music listening via various contextual factors such as time, location and weather.

This paper attempts to model the relationship between user-generated text and the music listening behavior. To this end, we adopt factorization machine (FM) [13], an instance of matrix factorization (MF)-based algorithms, as our learning framework for incorporating large number of features extracted from heterogeneous sources, such as audio content features and contextual features extracted from text. The audio features are collected from the EchoNest website (`http://echonest.com/`), including the loudness, mode, and tempo, danceability of music. The text features include *term frequency* and *inverse document frequency*. In addition, the ANEW affective lexicon [2] is also used to generate affective features that characterize the user's emotion state from text. Text information has been shown useful for information filtering [10, 14, 16], but its application to music recommendation is relatively less studied.

The quality of the dataset is important for such a study. Instead of using a dataset collected in a controlled environment, we compile a dataset by crawling LiveJournal (`http://www.livejournal.com/`), a social blog website, as it contains rich contextual information that is entered by users spontaneously in their day-to-day lives [19]. The dataset contains 225,652 listening records from 19,596 users and 30,260 songs. From this dataset, we are able to know which song a user wants to listen to, given an article he or she writes in a real-world context.

We compare the performance of different learning algorithms and different features including CB ones and context-based ones. The result shows that, while the hybrid recommendation approach based on CF+CB performs better than the one using CF-based only, adding contextual emotion features from text further improves the recommendation quality remarkably. The mean average precision (mAP) reaches

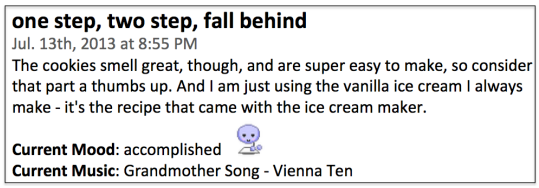

Figure 1: A sample post from LiveJournal.

0.5026, as opposed to 0.0578 for a random baseline, 0.3817 and 0.4708 for CF and CF+CB, respectively.

2. FACTORIZATION MACHINE

Factorization machine is a variant of MF-based methods [13]. Below we briefly describe the main idea of the technique. The FM is generally defined as:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \hat{w}_{ij} x_i x_j \,, \qquad (1)$$

where w_0 learns the global bias, w_i learns each weight of features x_i, and $\hat{w}_{ij}$ models the interaction of each pair of features. Instead of using single parameter for each interaction, FM factorizes it as the dot product of two vectors:

$$\hat{w}_{ij} = \sum_{f=1}^{\kappa} v_{if} v_{jf} \,, \qquad (2)$$

where κ is the model complexity. This way allows high-quality parameters estimated by higher-order interactions under sparsity. Note that all features in FM are transferred into indicator variables, which can be incorporated with context information. Unlike the generic MF models, FM not only keeps the information on user-item matrix, but also incorporates the interactions between pairs via various features. With FM, it is possible of modeling contextual information and providing context-aware rating predictions by using factorized interaction parameters. Features can be easily embedded in FM, so our goals is to investigate the effect of different information, including content-based and context-based, on recommendation. For more details of FM, please refer to [13].

3. DATASET AND FEATURES

LiveJournal is a well-known blog website where users can write blogs, or online diaries. The users are able to listen to a song and label a mood tag that reflects his or her emotional state while writing an article, as exemplified in Figure 1. In our experiments, we consider only the users with more than 10 articles. The collected data contains 19,596 users, 225,652 articles, and 30,260 songs. For the experimental settings, we first split the users into two sets according to the following 80/20 rule: we keep the full listening history of the 80% users and the half of listening history for the remaining 20% users as the training data, and the missing half of the remaining 20% users as the test data. Given a set of user's articles and their listening history, we attempt to recommend the songs that the user will listen to.

There are six different features for experiments, including content-based features and context-based features. Below we describe these features in detail.

Table 1: Some examples of the ANEW lexicon

Description	Valence	Arousal	Dominance
dream	6.73	4.53	5.53
eat	7.47	5.69	5.60
favor	6.46	4.54	5.67
good	7.47	5.43	6.41
hate	2.12	6.95	5.05

3.1 Content-based Feature

Content-based features generally refer to the characteristics of a recommended item. Considering the abundant information within music, we use 53 audio features to represent various perceptual dimensions of music, including danceability, loudness, mode, and tempo. They are extracted by using the EchoNest API (`http://developer.echonest.com/`), a commonly used audio feature extraction tool developed in the field of music information retrieval [17].

3.2 Context-based Feature

As for user-generated articles, two features are extracted to describe the user mood — mood tags (MOOD) and VAD. The former considers the mood tags labeled by users directly, whereas the latter tries to infer the emotional states of a user from his/her article, which is easier to obain than mood tags.

To this end, we also evaluate the case when we count the occurrence of terms in an article to characterize the content of the article. Considering that the high dimension of different words of the second feature may incur the so-called "curse of dimensionality" problem, the second feature can be considered as a result of dimension reduction of the word counts. Specifically, we convert the text of articles to an emotional word vector by using the lexicon of Affective Norms for English Words (ANEW) [2], which provides a set of normative emotional ratings for English words. The emotional words are rated by *valence* (or pleasantness; positive/negative affective states), *activation* (or arousal; energy and stimulation level) and *dominance* (or potency; a sense of control or freedom to act), the fundamental emotion dimensions found by psychologists [6]. Table 1 shows some words with their valence, arousal, and dominance ratings in the ANEW lexicon. By representing the affective content of an article in the 3-D space spanned by these emotion dimensions, we are able to obtain a compact yet informative representation of texts.

Specifically, we only leave the words which can be found in the ANEW lexicon. There are totally 2,476 emotional words in ANEW, and there are about 3% of articles discarded, because they have no ANEW words. Each emotional word is weighted by Term-Frequency Inverse-Document-Frequency (TFIDF) measure, which helps enhance the significance of terms with high weight and occurs rarely in the whole corpus.

$$tf(t,d) = \frac{f(w,d)}{\max\{f(w,d) : w \in d\}} \,, \qquad (3)$$

$$idf(t,d) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \,, \qquad (4)$$

where D is the total number of articles. Each term in vector is scored by $tf(t,d) \times idf(t,d)$. After the TFIDF weighting, we can get the Valence, Arousal, and Dominance (VAD)

Table 2: The feature sets considered in this work. *Cb* denotes the content-based feature that are extracted from songs, and *Cx* denotes the context-based feature that are extracted from user-generated articles

Label	Attribute	#Unique Indices	Type
U	User ID	19,596	–
S	Song title	30,260	–
CB	Audio features	53	Cb
MOOD	Mood tag	132	Cx
TFIDF	Words of article	2,476	Cx
VAD	VAD of articles	3	Cx

values of an article by a weighted summation of the VAD values of the emotional words occur in the article. For example, for the sentence "*I had a dream last night, I was eating a marshmallow*," the VAD values are 14.2, 10.22, and 11.13, respectively, according to Table 1. Note that in our experiments all the words are stemmed, and the values are normalized to a scale from 0 to 1. Table 2 summaries all the features used in this work along with their number of unique indices and notations.

4. EXPERIMENTAL RESULTS

In our experiments, two metrics are used for evaluating the recommendation performance: mAP and recall. Although mAP can take all retrieved items into account, we only focus on the top-k result. For each user, let $P(k)$ denotes the precision at cut-off k, in our experiments k is set to 10:

$$AP(u,o) = \frac{\sum_{p=1}^{k} P(k) \times r_{uo(p)}}{I(u)}\,, \tag{5}$$

where $o(p) = i$ means the item i is ranked at position p in the order list o, and r_{ui} means whether the user u has listened to song i or not$(1 = yes, 0 = no)$. The truncated mean average precision ($mAP@k$) is the mean of the average precision scores:

$$mAP@k = \frac{\sum_{u=1}^{U} AP(u,o)}{U}\,, \tag{6}$$

where U is the total number of target users. Recall, the fraction of listened songs that are recommended, are calculated as follows:

$$Recall = \frac{|\{Correct\ Songs\}| \cap |\{Returned\ Top\ k\ Songs\}|}{|\{Correct\ Songs\}|}\,. \tag{7}$$

High recall means that most of the listened songs have been recommended.

4.1 CF-based Recommendation

Our first evaluation focuses on the use of CF information only for music recommendation. We compare FM with the following three famous CF methods:

- **User-based CF** [1]: This method weights all users with respect to their similarity to each other, and selects a subset of users (who are highly similar to the target user) as neighbors. It predicts the rating of specific song based on the neighbors' ratings. Let $S(u)$ be the set of songs that are chosen by the user u. The similarity between user u and user v is calculated by following formula:

$$s_{uv} = \frac{S(u) \cap S(v)}{|S(u)|^{\alpha}|S(v)|^{1-\alpha}}\,, \tag{8}$$

  where $\alpha \in [0,1]$ is a parameter to tune.

- **Item-based CF** [1]: This method is similar to the user-based CF method. It computes the similarity between songs and scores a song based on user's listening history. The song similarity is calculated as follows:

$$s_{ij} = \frac{U(i) \cap U(j)}{|U(i)|^{\alpha}|U(j)|^{1-\alpha}}\,, \tag{9}$$

  where $U(i)$ the set of the users who have listened to the song i.

- **SVD++** [9]: This method is an extended version of SVD-based latent factor models by integrating implicit feedback into the model. In specific, the prediction formula can be the following:

$$r_{ui} = \mu + b_u + b_i + q_i^T \left(p_u + \frac{1}{\sqrt{|N(u)|}} \times \sum_{j \in N(u)} y_i \right)\,, \tag{10}$$

  where $N(u)$ is the set of implicit information, μ is the global mean rating, b_u is a scalar bias for user u, b_i is a scalar bias for item i, p_u is a feature vectors for user u, q_i is a feature vector for item i.

Table 3: Evaluation result of CF-based algorithms

Model	mAP@10	Recall
Randomize	0.0578	0.1656
User-based CF	0.3668	0.4748
Item-based CF	0.3093	0.5115
SVD++	0.3506	0.4844
FM	**0.3817**	**0.5216**

Table 3 lists the preliminary results of mAP@10 and recall. As the table shows, the performance of all the implemented methods, except for the random baseline, appears to be reasonable, achieving about 0.30 to 0.38 in terms of mAP. The item-based approach already gets a good recall in the task, but the FM model generate a more effective recommendation than it. Among the four methods, FM obtains the highest performance, which shows that FM can be a competitive framework for this task. We therefore focus on the use of FM hereafter.

4.2 FM with Content-based Recommendations

Next, we evaluate the performance of content-based recommendation. As the first two rows of Table 4 show, the hybrid CF+CB method (i.e., U + S + CB) outperforms the CF-based one (i.e., U + S) by a great margin. The CF-based approach usually suffers from the so-called "cold start" problem, which occurs when new items and new users are considered. As the experiments show, when the content-based features (i.e., audio information) are added, the quality of recommendation is improved from 0.3817 to 0.4708.

Table 4: Performance of factorization machine with different feature combinations

Features	mAP@10	Recall
U + S	0.3817	0.5216
U + S + CB	0.4708	0.6185
U + S + MOOD	0.4159	0.5628
U + S + TFIDF	0.4212	0.5643
U + S + VAD	0.4483	0.5905
U + S + CB + VAD	0.4901	0.6397
U + S + CB + MOOD + VAD	**0.5026**	**0.6540**

4.3 FM with Content-based and Context-based Recommendations

We evaluate the performance of context-based recommendation by using MOOD and VAD, both represent the users' mood. As shown in the third and fourth row of Table 4, the performance of adding the MOOD feature is improved from 0.3817 to 0.4159 in terms of mAP@10. This result shows that the contextual users' mood information indeed improves the performance of recommendation. With the another contextual VAD feature from user-generated context, the performance is even higher, with the mAP@10 attaining 0.4483. This result implies that the VAD feature provides more emotional information of the user context, which might not be easily captured by mood tags or words only.

Finally, we evaluate the hybrid model that combines the two contextual features and the content-based features (i.e., U + S + CB + MOOD + VAD). As the last row of Table 4 shows, this hybrid model greatly outperforms the content-based method, achieving 0.50 and 0.65 in terms of mAP@10 and recall, respectively. The performance differences between the hybrid model and the CF-based or content-based models are all significant under the two-tailed t-test (p-value< 0.001). On the other hand, we also provide an experimental result that without using the user-provided Mood tags (i.e., U + S+ CB + VAD). By comparing it to purely hybrid CF+CB method, we still see great performance improvement. In sum, the experimental results suggest that the contextual information mined from user-generated articles improves the quality of music recommendation.

5. CONCLUSIONS

In this paper, we have described a music-recommendation approach that combines the listening history and content-based audio information with the contextual emotion information mined from user-generated articles. By using the factorization machine technique, the propose system enhances the quality of music recommendation when evaluating on a real-world dataset. Instead of using simple word counts of the articles the users write as the context feature, we find it more informative to use affective contextual text information. For future work, we plan to use more sophisticated sentimental-analysis techniques, such as Sentimental Latent Dirichlet Association [11], to extract advanced contextual features for music recommendation. We also plan to use the grouping techniques of factorization machine to further improve its optimization process.

6. ACKNOWLEDGMENTS

This work was partially supported by the National Science Council of Taiwan via the grants NSC 100-2218-E-004-001, 101-2221-E-004-017, 102-2221-E-001-004-MY3, and a research grant from KKBOX.

7. REFERENCES

[1] F. Aiolli. A preliminary study of a recommender system for the million songs dataset challenge. In *Proc. ECAI*, 2012.

[2] M. Bradley and P. J. Lang. Affective norms for english words ANEW: Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, Univ. Florida, 1999.

[3] R. Cai et al. MusicSense: contextual music recommendation using emotional allocation modeling. In *Proc. ACM MM*, pages 553–556, 2007.

[4] K. Chen et al. Collaborative personalized tweet recommendation. In *Proc. ACM SIGIR*, pages 661–670, 2012.

[5] T. Chen et al. Combining factorization model and additive forest for collaborative followee recommendation. In *KDD CUP*, 2012.

[6] J. Fontaine et al. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050, 2007.

[7] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang. Social contextual recommendation. In *Proc. ACM CIKM*, pages 45–54, 2012.

[8] M. Jiang et al. Social recommendation across multiple relational domains. In *Proc. ACM CIKM*, pages 1422–1431, 2012.

[9] Y. Koren. Collaborative filtering with temporal dynamics. In *Proc. ACM KDD*, pages 447–456, 2009.

[10] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proc. ISMIR*, 2007.

[11] F. Li et al. Sentiment analysis with global topics and local dependency. In *Proc. AAAI*, 2010.

[12] H. Ma et al. Probabilistic factor models for web site recommendation. In *Proc. ACM SIGIR*, pages 265–274, 2011.

[13] S. Rendle. Factorization machines with libfm. *ACM Trans. Intelligent Systems and Technology*, 3(3):57:1–57:22, 2012.

[14] M. Schedl, T. Pohle, P. Knees, and G. Widmer. Exploring the music similarity space on the web. *ACM Trans. Information System*, 29(3):14:1–14:24, 2011.

[15] Y. Shi et al. TFMAP: optimizing MAP for top-n context-aware recommendation. In *Proc. ACM SIGIR*, pages 155–164, 2012.

[16] B. Sriram et al. Short text classification in twitter to improve information filtering. In *Proc. ACM SIGIR*, pages 841–842, 2010.

[17] D. Tingle, Y. E. Kim, and D. Turnbull. Exploring automatic music annotation with acoustically-objective tags. In *Proc. ACM MIR*, pages 55–62, 2010.

[18] J. Weston, C. Wang, R. Weiss, and A. Berenzweig. Latent collaborative retrieval. In *Proc. ICML*, pages 9–16, 2012.

[19] Y.-H. Yang and J.-Y. Liu. Quantitative study of music listening behavior in a social and affective context. *IEEE Trans. Multimedia*, 15(6), 2013.