A LARGE IN-SITU DATASET FOR CONTEXT-AWARE MUSIC RECOMMENDATION ON SMARTPHONES

Yuan-Ching Teng, Ying-Shu Kuo, and Yi-Hsuan Yang

Research Center for IT Innovation, Academia Sinica
{spencer,hank5925,yang}@citi.sinica.edu.tw

ABSTRACT

Context-based services have received an increasing attention due to the prevalence of sensor-rich mobile devices such as smartphones. The idea is to recommend information that would be of interest to a user according to the user's surround context. Although remarkable progress has been made, relatively little research has been made to contextualize music playback based on a large-scale dataset of real-life listening records. This paper presents our recent endeavor in collecting 5,502 real-life listening records with context annotation using Android smartphones in-situ. The user-provided context annotation contains labels selected from 10 user activity categories and 10 user mood categories. Moreover, we also compute a rich set of sensor features to capture the context at which the users listen to music, encompassing location, time, acceleration, proximity, etc. Our evaluation shows that with such context information we are able to significantly improve the performance of music recommendation, using factorization machine as the recommendation engine.

1. INTRODUCTION

Driven by the rapid progress in mobile sensing and computing, an emerging body of research work has been made to analyze body-worn sensor data to interpret the context of a user and thereby realize context-aware applications [1].

Given that people's short-term music preference is usually influenced by the context of listening, recent years have witnessed an increasing body of research work on context-aware music recommendation (CAMR) [2, 3]. For example, some users prefer upbeat music when being in a gym and soothing music while studying. With a CAMR system, music lovers are free from the time-consuming process of manually creating playlists or folders for different circumstances, and the need to specify their musical preferences as queries. CAMR is in particular desirable for users of online streaming or radio services, as digital music libraries are usually in the scale of millions.

In light of the above observations, this paper presents a smartphone-based system for CAMR using a novel dataset collected *in-situ* during a user's daily life. To this end, we surveyed participants using an experience-sample methodology [4], asking participants to annotate (i.e., "tag") their *activity* and *mood* anytime as long as they are about to start a new session of music listening (cf. Figure 1). We record both the listening history and sensor data available from the built-in multimodal sensors of smartphones right after the action of tagging, so as to build a tripartite relationship among user context, sensor data, and music listening behavior. The participants are free to listen to any song they like. A total number of 5,502 tripartite records are collected from 41 participants, each spent three weeks using a HTC smartphone and the Android app we developed and installed on the smartphone. The dataset contains the listening profile related to 15,381 songs in total. This represents the largest in-situ dataset we are aware of for CAMR.

Figure 1: User interface for annotating activity (left) and mood (right).

To better understand and model user behavior, a large-scale dataset collected in real-life, music listening context is needed. Based on this dataset, we conduct a preliminary study of the performance of music recommendation using context information obtained from either user-provided tags or sensor features extracted from the sensor data. To leverage both listening profiles and context information, the Factorization Machine [5], a matrix factorization-based algorithm, is adopted. Our evaluation shows that the use of context information significantly improves the performance

of music recommendation for this in-situ, possibly noisy dataset.

2. RELATED WORK

An emerging body of research work has been made to contextualize music recommendation (i.e., CAMR) using context information such as location [6], time [7], weather [8], walking pace [9], text of the viewed web page [10], to name a few. However, as suggested in a recent survey [2], few existing studies on CAMR are based on the multimodal sensors of a smartphone. Moreover, existing work suffers from weaknesses such as limited evaluation, high user effort to input information, and the difficulty to be implemented in everyday settings [2].

The work most relevant to ours is by Wang *et al.* [11], who proposed a smartphone-based CAMR system that recommends songs for the following six high-level everyday user activities: *working*, *studying*, *running*, *sleeping*, *walking*, and *shopping*. They developed a *music-context* model that estimates the activities (among the aforementioned six) for which a song is suitable using audio content features, and a *sensor-context* model that predicts the activity of a user according to sensor features. Music recommendation is carried out by two stages — predict the user's activity and then select songs suitable for that particular activity.

Our work differs from this prior art mainly in three aspects. First, we utilize an *in-situ* dataset that records the sensor data and music listening behavior at the same time, whereas in [11] the sensor data and music listening profile were collected separately. Second, our dataset contains *5,502 tripartites* of user context, sensor and music collected from 41 users, whereas Wang *et al.* used only *60 bipartites* of user context and sensors from 10 users. Finally, instead of dividing music recommendation into two stages, we integrate the information of listening profiles and user context by a unified, matrix factorization framework.

3. DATA

3.1. Data collection

As there is no standard taxonomy of music listening context, we first conducted a survey to understand the user behavior and preference in music listening. Specifically, we designed a questionnaire and posted it on several Internet forums for music lovers. In the end we collected the response of 275 subjects (63% male; 83% students; 44% listen to music more than 3 hours per day), 79% of which expressed high interest in the idea of context-aware music recommendation. From the response of "when do you usually use mobile devices for music listening" and "what are the most frequently experienced moods when listening to music", we select the 10 most popular activity classes and 10 mood tags to form our taxonomy, as shown in Figure 2.

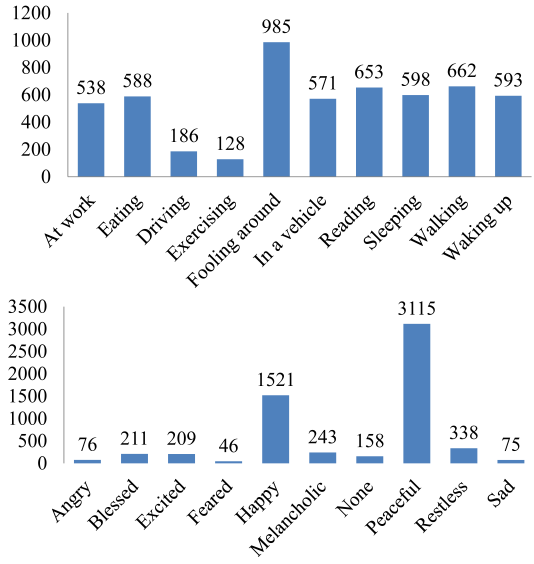

Figure 2: Number of records associated with each activity tag (upper) and mood tag (lower).

To construct a novel in-situ dataset for CAMR, we invited subjects who left their email addresses in the questionnaire to participant our user study. We finally recruited 41 healthy participants (41% male) with no neurological, hearing, or psychological problems. We developed an application (i.e., app) on Android platform to collect real data from the subjects' lives. The participants were asked to carry a HTC Desire S smartphone we purchased for music listening in their daily lives for a period of three weeks. Four sessions of experiment were conducted during June to September, 2012, as we have only 12 HTC Desire S smartphones. Below are the specific procedures we adopted for collecting the dataset.

- Each participant is asked to listen to music using the app we adapted from Google's built-in music player. With this program, they are asked to annotate the underlying activity and mood of music listening, using a grid-like graphic user interface (see Figure 1) with 10 buttons sorted in an order resembling our daily routine. They annotate activity first and then mood.
- The participants are motivated with cash award and free access to 3G network during the experiments. The participants are encouraged to provide more tags to earn more money.

The number of records associated with each activity and mood tag of the final dataset is shown in Figure 2. Please note that the distribution is not balanced across categories. For example, we have less examples of 'driving' or 'exercising' tags, but more 'fooling around' for activity tags. The distribution of mood tags is in particular unbalanced; more

than half of them are associated with 'peaceful.' We conceive such distribution reflects real-world user behavior and do not further apply filtering or post-processing to make the dataset balanced.

3.2. Sensor data processing

Funf Open Sensing Framework [12] is a framework on Android platform developed at the MIT Media Lab. It nicely takes care of issues related to privacy, security, and power consumption, We integrated the Funf into our music player for collecting the sensor data in an non-obtrusive manner. Right after a participant tags the listening context, we record the following sensor data for a period of three minutes:

- *AccelerometerSensorProbe* that measures magnitude of the tri-axial acceleration applied to the device.
- *LightSensorProbe* that detects the ambient light level.
- *ProximitySensorProbe* that measures how far the front of the device (i.e., smartphone) is from an object (e.g., hand or bag). Most implementations outputs a zero distance if something is against the face of the device, and a nonzero distance otherwise.
- *LocationProbe* (2-D) that records the most accurate location (i.e., longitude and latitude) available for the device, given device limitations and respecting battery life.
- *MagneticFieldSensorProbe* that detects magnitude of the local magnetic field. Can be used as to detect the earth's magnetic field, but may be distorted by metal objects around device.
- *OrientationSensorProbe* (3-D) measures azimuth (angle between the magnetic north direction and the y-axis), pitch (rotation around the x-axis), and roll (rotation around the y-axis), measured in degrees.
- *RotationVectorSensorProbe* (3-D) that represents the orientation of the device as a combination of an angle and an axis, in which the device has rotated through an angle around each of the three axes.
- *TimeProbe* that records the hour of the day (0 to 23).

The mean, standard deviation, minimum, and maximum of the probes (except for TimeProbe) during the three minutes are considered as the sensor features in our evaluation. Although more advanced features can be computed [13], we opt for starting with simple ones for this preliminary study. A total number of 49 ($12\times4+1$) features are computed.

4. EXPERIMENT

This section presents a preliminary evaluation of context-based recommendation using the dataset we just describe. As there is no real rating scores associated with the listening records, we consider the user–item–context triplets in our listening records as positive rating records, and randomly sample equivalent number of songs a user does not listen to given a context from the user's personal music collection (also user–item–context triplets) as negative rating records. That is to say, as we know which songs a user listens to at a specific context, we assume that the user considers the other songs in his/her personal collection not appropriate for that listening context.

Our evaluation considers the 28 subjects (among 41) who have more than 200 (positive) listening records as active subjects. We randomly pick 4/5 of the data from these active subjects as the training data, and leave the remaining as the test set. Moreover, the listening records of the remaining 13 users are also utilized as the test set.

4.1. System implementation

A recommendation algorithm is at the core of a recommendation system. We use the factorization machine (FM) algorithm implemented in the libFM toolkit [5] as our recommendation engine, for its excellent performance in many collaborative filtering recommendation tasks. As an instance of the matrix factorization algorithms, FM has the flexibility of mimicking many other factorization models by feature engineering [5]. The inference of model parameters is approached by Markov Chain Monte Carlo (MCMC), a Bayesian inference algorithm.

To assess the contribution of different features for CAMR, we concatenate user-item listening records with different types of context information. The features we consider include:

- Collaborative filtering-based (*CF*): The binary rating (1/0) of each user-item (i.e., song) pair.
- Sensor data (*Sensor*): the 49-D sensor feature vector described in Section 3.2.
- Activity tag (*Activity*): a 10-D binary vector of the user-provided activity tags, with only one of the elements being one (because a user annotates his/her context with only one of the tags in our program).
- Mood tag (*Mood*): another 10-D binary vector for mood tags constructed in a similar way.

Because the output of libFM are numerical values, we evaluate the performance of recommendation in terms of area under the receiver operating characteristic curve (AUC) and root mean square error (RMSE). AUC measures the ability of a retrieval system to rank positive examples above negative examples, scoring them on a scale from 0.5 (chance level) to 1 (perfect ranking). While larger AUC indicates better performance, smaller RMSE is better.

Feature	AUC	RMSE
Random	0.498	0.577
CF	0.628	0.484
CF + Activity	**0.682**	**0.469**
CF + Mood	0.641	0.484
CF + Activity + Mood	0.679	0.472
CF + Sensor	0.625	0.485
CF + Sensor + Activity	0.639	0.481
CF + Sensor + Mood	0.619	0.489

Table 1: Performance of different features

4.2. Result

As Table 1 shows, the accuracy of the conventional CF approach reaches 0.628 for AUC and 0.484 for RMSE. The performance difference between CF and the random baseline is significant (p-value < 0.001) under the two-tailed t-test. This result shows the FM is also effective for a dataset collected in-situ. From 1 we also see that the use of activity tags further improves the AUC and RMSE to 0.682 and 0.469, respectively, showing that the short-term music preference is indeed dependent on the underlying activity of music listening. The performance difference between CF and CF+Activity is also significant (p-value < 0.001).

However, we see that the use of either sensor features or mood tags does not offer additional gain. The performance of CF+Sensor is on par with CF, whereas the performance of CF+Activity+Mood is on par with CF+Activity. This result leads to the following three hypothesis. First, the sensor features we compute are too primitive to capture the context of the user. Deeper processing on the sensor data might be needed. Second, while the listening behavior is related to user activity, the relationship between listening behavior and user mood is relatively weak. Third, the discriminative power of the mood tags might have been hindered by the sever imbalance of the mood distribution (i.e., Figure 2). Future work is needed to validate these hypothesis.

5. CONCLUSION

In this paper, we have presented a novel, large-scaled, in-situ dataset consisting of different listening history, sensor data with usage and mood annotations by 41 subjects from their real life for three weeks. We have also shown that context information does help improve the performance of CAMR. However, the best performance is obtained when we use the activity tag directly, which requires user input and therefore is not that practical. To address this issue, we are currently experimenting with more advanced sensor features and machine learning algorithms to build a sensor-based automatic user activity classifier.

6. ACKNOWLEDGMENTS

This work was supported by the National Science Council of Taiwan under contract NSC 101-2221-E-001-017 and by the Academia Sinica Career Development Award.

7. REFERENCES

[1] D. Roggen, S. Magnenat, M. Waibel, and G. Troster, "Wearable computing: Designing and sharing activity-recognition systems across platforms," *IEEE Robotics & Automation Magazine*, pp. 83–95, 2011.

[2] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2–3, pp. 89–119, 2012.

[3] F. Ricci, "Context-aware music recommender systems," in *Int. Works. Advances in Music Information Research, in conjunction with WWW*, 2012, pp. 865–866.

[4] R. Larson and M. Csikszentmihalyi, "The experience sampling method," *New Directions for Methodology of Social and Behavioral Science*, vol. 15, pp. 41–56, 1983.

[5] S. Rendle, "Factorization machines with libFM," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 57:1–57:22, May 2012.

[6] M. Kaminskas, I. Fernández-Tobías, F. Ricci, and I. Cantador, "Knowledge-based music retrieval for places of interest," in *Int. Works. Music Information Retrieval with User-Centered & Multimodal Strategies*, 2012, pp. 19–24.

[7] J.-H. Su, H.-H. Yeh, P. S. Yu, and V. S. Tseng, "Music recommendation using content and context information mining," *IEEE Intelligent Systems*, vol. 25, no. 1, pp. 16–26, 2010.

[8] S. Reddy and J. Mascia, "Lifetrak: music in tune with your life," in *Workshop on Human-centered Multimedia*, 2006, pp. 25–34.

[9] G. T. Elliott and B. Tomlinson, "Personalsoundtrack: contextaware playlists that adapt to user pace," in *ACM CHI Extended Abstracts on Human Factors in Computing*, 2006, pp. 736–741.

[10] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma, "MusicSense: Contextual music recommendation using emotional allocation modeling," in *Proc. ACM Multimedia*, 2007, pp. 553–556.

[11] X. Wang, D. Rosenblum, and Y. Wang, "Context-aware mobile music recommendation for daily activities," in *Proc. ACM Multimedia*, 2012, pp. 99–108.

[12] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, 2011, [Online] `http://funf.org/`.

[13] P. Lukowicz, O. Amft, D. Roggen, and J. Cheng, "On-body sensing: From gesture-based input to activity-driven interaction," *IEEE Computing*, vol. 43, no. 10, pp. 92–96, 2010.