Hybrid Retrieval Approaches to Geospatial Music Recommendation

Markus Schedl
Department of Computational Perception
Johannes Kepler University, Linz, Austria
markus.schedl@jku.at

Dominik Schnitzer
Austrian Research Institute for Artificial Intelligence, Vienna, Austria
dominik.schnitzer@ofai.at

ABSTRACT

Recent advances in music retrieval and recommendation algorithms highlight the necessity to follow multimodal approaches in order to transcend limits imposed by methods that solely use audio, web, or collaborative filtering data. In this paper, we propose hybrid music recommendation algorithms that combine information on the *music content*, the *music context*, and the *user context*, in particular, integrating location-aware weighting of similarities. Using state-of-the-art techniques to extract audio features and contextual web features, and a novel standardized data set of music listening activities inferred from microblogs (`MusicMicro`), we propose several multimodal retrieval functions.

The main contributions of this paper are (i) a systematic evaluation of mixture coefficients between state-of-the-art audio features and web features, using the first standardized microblog data set of music listening events for retrieval purposes and (ii) novel geospatial music recommendation approaches using location information of microblog users, and a comprehensive evaluation thereof.

Categories and Subject Descriptors

Information systems [**Information search and retrieval**]: Music retrieval; Human-centered computing [**Collaborative and social computing**]: Social media mining

1. INTRODUCTION

The field of Music Information Retrieval (MIR) is seeing a paradigm shift, away from system-centric perspectives towards user-centric approaches [3]. In this vein, incorporating user models and addressing user-specific demands in music retrieval and music recommendation systems is becoming more and more important.

We present several approaches that combine *music content*, *music context*, and *user context* aspects to build a hybrid music retrieval system [12]. Music content and music context are incorporated using state-of-the-art feature extractors and corresponding similarity estimators. The user context is addressed by taking into account *musical preference* and *geospatial data*, using a standardized collection of listening behavior mined from microblog data [11].
We make use of the best feature extraction and similarity computation algorithms currently available to model *music content* and *music context*. We then integrate these similarity models as well as a *user context* model into a novel user-aware music recommendation approach that encompasses all three modalities important to human music perception [12].

The main contributions of this paper are: (i) a systematic evaluation of combining audio- and web-based state-of-the-art approaches to music similarity measurement and (ii) two approaches to incorporate geospatial information into music recommendation algorithms.

The remainder of the paper is organized as follows. Section 2 details the acquisition of the raw music (meta-)data, which serves as input to the feature extraction and data representation techniques presented in Section 3. In Section 4, we construct different hybrid (music content and music context) models and systematically evaluate their mixture coefficients. Section 5 then proposes two methods to incorporate geospatial information into music recommendation models. These extended models are evaluated and compared to the respective models without geospatial data and to a random baseline. Section 6 briefly reviews related literature. Eventually, Section 7 draws conclusions and points to further research directions.

2. DATA ACQUISITION

The only standardized public data set of microblogs, as far as we are aware of, is the one used in the TREC 2011 and 2012 Microblog tracks[1] [4]. Although this set contains approximately 16 million tweets, it is not suited for our task as it is not tailored to music-related activities, i.e. the amount of music-related posts is marginal.

We hence have to acquire multimodal data sets of *music items* and *listeners*, reflecting the three broad aspects of human music perception (*music content*, *music context*, and *user context*) [12]. Whereas the *music content* refers to all information that is derived from the audio signal itself (such as ryhthm, timbre, or melody), the *music context* covers contextual information that cannot be derived from the actual audio with current technology (e.g., meaning of song lyrics, background of a performer, or co-listening relationships between artists). The *user context* encompasses

[1] `http://trec.nist.gov/data/tweets`

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

all information that are intrinsic to the listener. Examples range from musical education to spatiotemporal properties to physiological measures to current activities.

User Context.
Only very recently a data set of music listening activities inferred from microblogs has been released [11]. It is entitled `MusicMicro` and is freely available[2], fostering reproducibility of social media-related MIR research. This data set contains about 600,000 listening events posted on `Twitter`[3]. Each event is represented by a tuple <*twitter-id, user-id, month, weekday, longitude, latitude, country-id, city-id, artist-id, track-id*>, which allows for spatiotemporal identification of listening behavior.

Music Content.
Based on the lists of artist and song names in the `MusicMicro` collection, we gather snippets of the songs from `7digital`[4]. These serve as input to the music content feature extractors (cf. Section 3).

Music Context.
To capture aspects of human music perception which are not encoded in the audio signal, we extract music-related web pages that represent such contextual information. Following the approach suggested in [13], we retrieve the top 50 web pages returned by `Bing`[5] for queries comprising the artist name[6] and the additional keyword "music", to disambiguate the query for artists such as "Bush" or "Kiss".

In summary, we gathered raw data covering each of the three categories of perceptual music aspects [12]: *music content* (audio snippets), *music context* (related web pages), and *user context* (user-specific music listening events with spatiotemporal labels).

3. DATA REPRESENTATION

To represent the *music content*, we use state-of-the-art audio music feature extractors proposed in [7]. These algorithms won three times in a row (since 2010) the annually run benchmarking activity *Music Information Retrieval Evaluation eXchange* (MIREX): "Audio Music Similarity and Retrieval" task[7]. They hence constitute the reference in music feature extraction for similarity-based retrieval tasks. More precisely, we extract the auditory music features proposed in [7], which combine various rhythmic features derived from the audio signal, e.g., "onset patterns" and "onset coefficients" (note onsets), with timbral features, e.g., "Mel Frequency Cepstral Coefficients" (coarse description of the amplitude envelop). The eventual output is pairwise similarity estimates between songs, which are later aggregated to the artist level.

We again employ a state-of-the-art technique to obtain features reflecting the *music context*. To describe the music items at the artist level, we follow the approach proposed in [13]. In particular, we model each artist by creating a "virtual artist documents", i.e. we concatenate all web pages retrieved for the artist. In accordance with findings of [10], we then use a dictionary of music-related terms (genres, styles, instruments, and moods) to index the resulting documents. From the index, we compute term weights according to the best feature combination found in the large-scale experiments of [13]: `TF_C3.IDF_I.SIM_COS`, i.e. computing term weight vectors and artist similarity estimates according to Equations 1, 2, and 3, respectively for *tf*, *idf*, and *cosine similarity*; $f_{d,t}$ represents the number of occurrences of term t in document d, N is the total number of documents, $\mathcal{D}_t$ is the set of documents containing term t, F_t is the total number of occurrences of term t in the document collection, $\mathcal{T}_d$ is the set of distinct terms in document d, and W_d is the length of document d.

$$tf_{d,t} = 1 + \log_2 f_{d,t} \tag{1}$$

$$w_t = 1 - \frac{n_t}{\log_2 N}, \quad n_t = \sum_{d \in \mathcal{D}_t} \left(-\frac{f_{d,t}}{F_t} \log_2 \frac{f_{d,t}}{F_t} \right) \tag{2}$$

$$S_{d_1,d_2} = \frac{\sum_{t \in \mathcal{T}_{d_1,d_2}} (w_{d_1,t} \cdot w_{d_2,t})}{W_{d_1} \cdot W_{d_2}} \tag{3}$$

3.1 Availability of the Data Sets

All components of the data set used in this paper are publicly available to allow researchers reproduce the results reported. The sole exception is the actual audio content of the songs under consideration. We cannot share them due to copyright restrictions. However, we provide identifiers by means of which corresponding 30-second-clips can be downloaded from `7digital`. If you are interested in the data sets, please contact the first author.

4. HYBRID MUSIC RETRIEVAL

One main research question is how to ideally combine audio and web features for music retrieval. Although quite a few MIR researchers suggest such a combination [2, 1, 3, 12, 5], a systematic evaluation of combining state-of-the-art audio and web similarity estimators is still missing, hence provided here.

4.1 Experimental Setup

In a preprocessing step, we aggregate the audio features on the artist level, as they are computed on single tracks. To obtain audio similarities $asim(i, j)$ between two artists i and j, we compute the minimum of the distances between all pairs of tracks by i and j as the minimum yielded the best results in preliminary experiments similar to the ones described later in this section. Web similarities $wsim(i, j)$ are already defined on the artist level. Both, audio and web similarites, are normalized using the global distance scaling method Mutual Proximity [14].

Linear combinations of web similarities and audio similarities yield a hybrid similarity function $sim(i, j)$ between artists i and j. It is given in Equation 4, where ξ is the mixture coefficient, i.e., the weight of the audio part, different values of which we systematically evaluate.

$$sim(i,j) \; = \; \xi \cdot asim(i,j) \; + \; (1-\xi) \cdot wsim(i,j) \tag{4}$$

As *gold standard* we use genre information and assess retrieval performance via the overlap between the genres assigned to the query artist and those assigned to his K nearest

[2] `http://www.cp.jku.at/musicmicro`
[3] `http://www.twitter.com`
[4] `http://www.7digital.com`
[5] `http://www.bing.com`
[6] Please note that issuing queries at the song level is not reasonable, as doing so typically yields only very few results.
[7] `http://www.music-ir.org/mirex/wiki/2012:Audio_Music_Similarity_and_Retrieval`

ξ	$K=1$	$K=3$	$K=5$
web only – 0.00	**.5829**	**.5753**	**.5774**
.05	.6421	.6280	.6257
.15	.6432	**.6286**	**.6261**
.25	**.6433**	.6275	.6258
.35	.6430	.6275	.6257
.45	.6408	.6266	.6252
.55	.6394	.6259	.6244
.65	.6379	.6255	.6232
.75	.6368	.6234	.6221
.85	.6330	.6202	.6188
.95	.6215	.6083	.6059
audio only – 1.00	**.5436**	**.5302**	**.5247**

Table 1: Overlap scores for different mixture coefficients ξ between web and audio features.

neighbors according to the similarity function under investigation. This is a standard evaluation approach in MIR. We gather genre information by (i) retrieving the top tags for each artist via the `Last.fm` API[8] and (ii) using the top 20 main genres from `allmusic`[9] to index the sets of tags.

To evaluate retrieval performance, we use a Jaccard-like overlap measure, shown in Equations 5 and 6, where i is the query artist, $Genres_i$ is the set of genres assigned to i, K is the number of i's nearest neighbors to consider, and A is the number of all artists in the data set. The range of the performance measures is $[0, 1]$, i.e., they are 1.0 if the genres of the seed artist i's K nearest neighbors perfectly overlap with those of i.

$$overlap_i = \frac{1}{K} \cdot \sum_{j=1\ldots K} \frac{|Genres_i \cap Genres_j|}{|Genres_i|} \quad (5)$$

$$overlap = \frac{1}{A} \cdot \sum_{i=1\ldots A} overlap_i \quad (6)$$

4.2 Results

Performance scores for the hybrid retrieval function for different mixture coefficients ξ are shown in Table 1, together with results for a random baseline. Although using only web features ($\xi = 0.0$) yields better results than using audio only ($\xi = 1.0$), adding a small amount of content features to web features (or vice versa) boosts performance considerably. Adding a small amount of a complementary similarity component thus proves highly beneficial. Overall, values of ξ around 0.15 perform best. We hence use Equation 7 as hybrid (audio and web features) music model (MU) for subsequent experiments.

$$sim(i, j) = 0.15 \cdot asim(i, j) + 0.85 \cdot wsim(i, j) \quad (7)$$

5. MUSIC RECOMMENDATION MODELS

Building recommendation systems requires a user model. In our case, each user u is modeled by the set of artists $UM(u)$ he listened to. Based on this simple model, we implement the following recommendation strategies: (i) the hybrid music retrieval model (MU) elaborated in the previous section and (ii) a standard collaborative filtering (CF) model. In the MU model, the hybrid music similarity function (Equation 7) is used to determine the artists closest to $UM(u)$, which are then recommended. In the CF model, the users closest to u are determined (using the Jaccard index between the user models), and the artists listened to by these nearest users are recommended. For comparison, we further implemented a random baseline model (BL) that randomly picks K users from the filtered user set (via the parameter τ, see below) and recommends the artists they listened to. To integrate *geospatial information* into the CF model, we first compute a centroid of each user u's geospatial listening distribution $\mu_u[\lambda, \varphi]$[10]. We then use the normalized geodesic distance $gdist(u, v)$ (Equation 8) between the seed user u and each other user v to weight the distance based on the user models. To this end, we propose two different weighting schemes: linear weighting and weighting according to a Gaussian kernel around $\mu_u[\lambda, \varphi]$. We eventually obtain a geospatially modified user similarity $sim(u, v)$ by adapting the Jaccard index between $UM(u)$ and $UM(v)$ via geospatial, linear or Gauss weighting, according to Equation 9 (GEO-Lin) or Equation 10 (GEO-Gauss), respectively. We recommend the artists listened to by u's nearest users v. Table 2 summarizes all investigated recommendation algorithms.

Abbreviation	Description
BL	random baseline
MU	hybrid music model (Equation 7)
CF	collaborative filtering model
CF-GEO-Lin	CF model: geospatial user weighting using linear spatial distances
CF-GEO-Gauss	CF model: geospatial user weighting weighting using a Gauss kernel

Table 2: Overview of recommendation models.

$$gdist(u, v) = \arccos\big(\sin(\mu_u[\varphi]) \cdot \sin(\mu_v[\varphi]) + \cos(\mu_u[\varphi]) \cdot \cos(\mu_v[\varphi]) \cdot \cos(\mu_u[\lambda] - \mu_v[\lambda])\big) \cdot \max(gdist)^{-1} \quad (8)$$

$$sim(u, v) = J(UM(u), UM(v)) \cdot gdist(u, v)^{-1} \quad (9)$$

$$sim(u, v) = J(UM(u), UM(v)) \cdot \exp(-gdist(u, v)) \quad (10)$$

5.1 Experimental Setup

In order to ensure sufficient artist coverage of users, we evaluate our models using different thresholds τ for the minimum number of unique artists a user must have listened to in order to include him in the experiments. We vary τ between 50 and 150 using a step size of 10. Denoting as U_τ the number of users in the `MusicMicro` data set with equal or more than τ unique artists, we perform U_τ-fold leave-one-out cross-validation for each value of τ.

5.2 Results

Figure 1 shows accuracies for $K = [3, 5]$ nearest neighbors and $\tau = [50 \ldots 150]$. We can see that all approaches significantly outperform the random baseline. Comparing the MU approach with the CF approaches, it is evident that CF generally works better for data sets with high numbers of users (smaller τ), while content-based MU outperforms CF

[8] `http://www.last.fm/api`

[9] `http://www.allmusic.com`

[10] It is common to denote longitude by λ and latitude by φ.

when the number of users is restricted. This finding suggests a combination of MU and CF, which will be addressed as part of future work. As for geospatial weighting, a similar observation comparing the linear weighting with the Gauss weighting can be made. The more active the users (higher τ), the better the performance of the linear weighting approach, and the worse the Gauss kernel approach. An explanation for this may be that very frequent users of `Twitter` typically live in agglomerations, whereas occasional twitterers live in less densely populated areas. For these users in rural areas, a Gauss weighting is seemingly beneficial as very nearby users frequently know each other and share common music tastes (which is not true for highly populated areas). The models that integrate geospatial information outperform the standard CF model for high τ values, indicating again that this kind of information is beneficial for "power users", who typically live in densely populated areas.

6. RELATED WORK

Specific related work on geospatial music retrieval is very sparse, probably due to the fact that geospatially annotated music listening data is hardly available. Among the few works, Park et al. [6] use geospatial positions and suggest music that matches a selected environment, based on aspects such as ambient noise, surrounding, or traffic. Raimond et al. [8] combine information from different sources to derive geospatial information on artists, aiming at locating them on a map. Zangerle et al. [15] use a co-occurrence-based approach to map tweets to artists and songs and eventually construct a music recommendation system. However, they do not take location into account.

On a more general level, this work relates to context-based and hybrid recommendation systems, a detailed review of which is unfortunately beyond the scope of the paper. A comprehensive elaboration, including a decent literature overview, can be found in [9].

7. CONCLUSIONS AND OUTLOOK

We presented the first *systematic evaluation of hybrid music retrieval approaches* (combining the currently best performing audio/music content and web/music context features), using a recently published, standardized data set of music listening activities mined from microblogs. Experiments showed that a linear mixture coefficient of 0.15 for the audio part and 0.85 for the web component performed best, overall. Interestingly, adding only a very small amount of audio-based information to web features (or vice versa) considerably improves results.

To the best of our knowledge, this is also the first work that *integrates geospatial information into music recommendation algorithms*. Experiments indicate that including geospatial information is particularly beneficial for music recommendation when users listen to many different artists. The collaborative filtering approach (CF) outperforms the hybrid music retrieval model (MU) when the data set comprises a high number of users who listen to less artists, overall.

Future work will include considering more diverse data about the user context, such as demographics, listening time (hour of day, working day versus weekend), or gender. In addition, we plan to combine the MU and the CF models, including geospatial weighting. As a further usage scenario, we target users frequently traveling around the world and wanting to listen to music tailored to their current location, but also complying to their music taste. We will look into adapting our approaches accordingly.

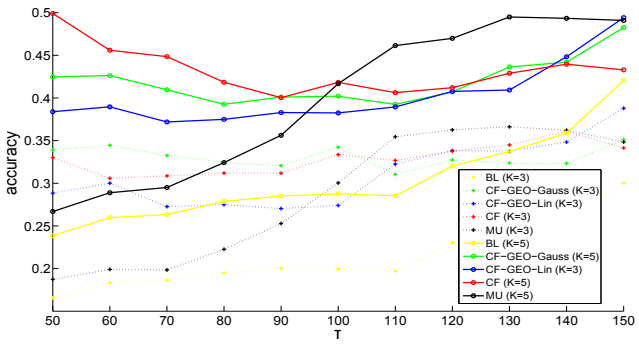

Figure 1: Accuracy plots for different values of K and τ.

Acknowledgments

This research is supported by the Austrian Science Fund (FWF): P22856, P24095, P25655, and the EU FP7: 601166.

8. REFERENCES

[1] D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra. Unifying Low-Level and High-Level Music Similarity Measures. *IEEE Transactions on Multimedia*, 13(4):687–701, Aug 2011.

[2] E. Coviello, A. B. Chan, and G. Lanckriet. Time Series Models for Semantic Music Annotation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1343–1359, Jul 2011.

[3] C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic. The Need for Music Information Retrieval with User-centered and Multimodal Strategies. In *Proc. MIRUM*, Scottsdale, AZ, USA, 2011.

[4] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis, and D. McCullough. On Building a Reusable Twitter Corpus. In *Proc. SIGIR*, Portland, OR, USA, 2012.

[5] B. McFee and G. Lanckriet. Heterogeneous Embedding for Subjective Artist Similarity. In *Proc. ISMIR*, Kobe, Japan, 2009.

[6] S. Park, S. Kim, S. Lee, and W. S. Yeo. Online Map Interface for Creative and Interactive MusicMaking. In *Proc. NIME*, Sydney, Australia, 2010.

[7] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On Rhythm and General Music Similarity. In *Proc. ISMIR*, Kobe, Japan, 2009.

[8] Y. Raimond, C. Sutton, and M. Sandler. Automatic Interlinking of Music Datasets on the Semantic Web. In *Proc. WWW: LDOW Workshop*, Beijing, China, 2008.

[9] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[10] M. Schedl. #nowplaying Madonna: A Large-Scale Evaluation on Estimating Similarities Between Music Artists and Between Movies from Microblogs. *Information Retrieval*, 15:183–217, June 2012.

[11] M. Schedl. Leveraging Microblogs for Spatiotemporal Music Information Retrieval. In *Proc. ECIR*, Moscow, Russia, 2013.

[12] M. Schedl and A. Flexer. Putting the User in the Center of Music Information Retrieval. In *Proc. ISMIR*, Porto, Portugal, 2012.

[13] M. Schedl, T. Pohle, P. Knees, and G. Widmer. Exploring the Music Similarity Space on the Web. *ACM Transactions on Information Systems*, 29(3), Jul 2011.

[14] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research*, 13:2871–2902, October 2012.

[15] E. Zangerle, W. Gassler, and G. Specht. Exploiting Twitter's Collective Knowledge for Music Recommendations. In *Proc. WWW: #MSM Workshop*, Lyon, France, 2012.