Mining Microblogs to Infer Music Artist Similarity and Cultural Listening Patterns

Markus Schedl
Department of Computational Perception
Johannes Kepler University
Altenberger Straße 69
A-4040 Linz, Austria
markus.schedl@jku.at

David Hauger
Department of Computational Perception
Johannes Kepler University
Altenberger Straße 69
A-4040 Linz, Austria
david.hauger@jku.at

ABSTRACT

This paper aims at leveraging microblogs to address two challenges in music information retrieval (MIR), *similarity estimation* between music artists and inferring typical *listening patterns* at different granularity levels (city, country, global). From two collections of several million microblogs, which we gathered over ten months, music-related information is extracted and statistically analyzed. We propose and evaluate four co-occurrence-based methods to compute artist similarity scores. Moreover, we derive and analyze culture-specific music listening patterns to investigate the diversity of listening behavior around the world.

Categories and Subject Descriptors

I.7.m [**Document and Text Processing**]: Miscellaneous—*microblog mining*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*; H.3.3 [**Information Storage and Retrieval**]: Information Modeling—*co-occurrence analysis*

General Terms

Algorithms, Measurement

Keywords

social media mining, music information retrieval, similarity measurement, evaluation

1. INTRODUCTION

The large and ever growing amount of user-generated content in today's social media platforms constitutes a tremendously wealthy, albeit noisy source for various data mining tasks. In particular, microblogging has encountered a considerable gain in popularity during the past few years as it provides an easy way for everyone to report on activities and thoughts. Today's most popular microblogging service `Twitter` [5] has more than 200 million registered users [3] who are creating a billion posts every week [1] (as of March/April 2011). `Twitter` thus represents a rich data source for text-based information extraction (IE) and information retrieval (IR).

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

Music is omnipresent on the (social) web as it plays an important role in many human lives. Everybody enjoys listening to his favorite tunes, and many people share their opinions about songs, artists, or latest album releases. Some even share their own versions of favored music videos. Digital music distribution and consumption are also important economic factors, which is demonstrated by the current success of music streaming services such as `Spotify` [4].

Given the importance both social media and music consumption play for many people, this work addresses the following research questions:

1. Can we create a similarity measure for music artists based on microblogs?
2. Are there typical listening patterns encoded in tweets? If so, do they differ among different places in the world? What can we learn about differences between countries or cities?

Elaborating music similarity measures that reflect resemblance perceived by humans is one of the big challenges in music information retrieval (MIR). These similarity measures enable applications such as music recommender systems [10, 16], automated playlist generators [36, 38], or intelligent user interfaces to music collections [37, 32]. Computational features for music similarity calculation can be broadly categorized into *content-based*, *music context-based*, and *user context-based*. While content-based feature extraction techniques derive the representation of a music item from the audio signal itself [15], music context-based approaches make use of data that is not encoded in the signal [44], for instance, the performer's political background, the meaning of a song's lyrics, images of album covers, or co-occurrence information derived from playlists. Both content-based and music context-based techniques to model musical similarity are relatively well researched in terms of publication numbers, although they are still far from being highly accurate. In contrast, feature extraction and similarity measurement approaches that take into account the user's context are relatively sparse throughout the scientific literature. An overview of user context features likely to be useful for MIR tasks is presented in [45].

The work at hand is one of the first to tackle the intersection between music context and user context since microblog data is available at the level of individual users, but can also be aggregated to model similarity at the city-, country-, or global level. The similarity measurement models proposed here are therefore capable of reflecting music perception at

different scopes, which is a promising path to follow when aiming at creating personalized retrieval models [53].

The remainder of the paper is organized as follows. Section 2 reviews related literature on microblog mining and text-based similarity. Section 3 reports on the acquisition of music-related, geospatialized tweets and presents results of preliminary statistical analyses. Our approaches to infer similarity between artists from microblogs are presented in Section 4, together with an evaluation. In Section 5 we then address the second research question, which is analyzing geographical differences in listening patterns around the world. Eventually, Section 6 summarizes the main findings and points to some future research directions.

2. RELATED WORK

Related literature basically falls into two groups: text-based similarity measurement and microblog mining. Whereas the former has a long tradition, ranging back several decades, the latter is a rather young research field.

2.1 Text-based Similarity Measurement

There exists a wide range of literature on modeling text documents according to the bag-of-words principle using vector space representations, e.g. [9, 40, 35]. Since elaborating on all publications related to the discipline of text-IR is out of this article's scope, we focus on work dealing with text-IR in the context of music and multimedia retrieval on the Web, as this context is closely related to the work at hand.

Text data in the multimedia domain generally constitutes *context information* or *contextual data*, opposed to content-based features directly extracted from the media items. Deriving term feature vectors from Web pages for the purpose of music artist similarity estimation was first undertaken in [22]. Cohen and Fan automatically extract lists of artist names from Web pages, which are found by querying Web search engines. The resulting pages are then parsed according to their DOM tree, and all plain text content with minimum length of 250 characters is further analyzed for occurrences of entity names. Term vectors of co-occurring artist names are then used for artist recommendation. Using artist names to build term vector representations, whose term weights are computed as co-occurrence scores, is an approach also followed later in [54, 46]. In contrast to Cohen and Fan's approach, the authors of [54, 46] derive the term weights from search engine's page count estimates and suggest their method for artist recommendation.

Automatically querying a Web search engine to determine pages related to a specific topic is a common and intuitive task, which is therefore frequently performed for data acquisition in IE research. Examples in the music domain can be found in [52, 26], whereas [19, 20, 31] apply this technique in a more general context.

Building term feature vectors from term sets other than artist names is performed in [52], where Whitman and Lawrence extract different term sets (unigrams, bigrams, noun phrases, artist names, and adjectives) from up to 50 artist-related Web pages obtained via a search engine. After downloading the pages, the authors apply parsers and a part-of-speech (POS) tagger [14] to assign each word to its suited test set(s). An individual term profile for each artist is then created by employing a version of the *TF·IDF* measure. The overlap between the term profiles of two artists, i.e., the sum of weights of all terms that occur in both term profiles, is then used as an estimate for their similarity.

Extending the work presented in [52], Baumann and Hummel [11] introduce filters to prune the set of retrieved Web pages. First, they remove all Web pages with a size of more than 40 kilobytes (after parsing). They also try to filter out advertisements by ignoring text in table cells comprising more than 60 characters, but not forming a correct sentence. Finally, Baumann and Hummel perform keyword spotting in the URL, the title, and the first text part of each page. Each occurrence of the initial query parts (artist name, "music", and "review") contributes to a page score. Pages that score too low are filtered out.

Knees et al. present in [30] an approach similar to [52]. Unlike Whitman and Lawrence who experiment with different term sets, Knees et al. use only one list of unigrams. For each artist, a weighted term profile is created by applying a *TF·IDF* variant. Calculating the similarity between the term profiles of two artists is then performed using the cosine similarity. Knees et al. evaluate their approach in a genre classification setting using as classifiers k-Nearest Neighbor (kNN) and Support Vector Machines (SVM) [50].

Other approaches derive term profiles from more specific Web resources. In [17], for example, Celma et al. propose a music search engine that crawls audio blogs via RSS feeds and calculates *TF·IDF* features. Hu et al. in [27] extract *TF*-based features from music reviews gathered from `Epinions.com` [25]. In [42] Schedl extracts user posts associated with music artists from the microblogging service `Twitter` and models term profiles using term lists specific to the music domain. Although one of the goals (artist similarity measurement) and the data source (microblogs) in [42] resemble the work at hand, [42] bases the similarity computation on *TF·IDF* representations of music artists, whereas the approaches reported in this paper derive a similarity estimate from co-occurrence information.

2.2 Microblog Mining

With the advent of microblogging a huge, albeit noisy data source became available. Literature dealing with microblogs can be broadly categorized into works that study human factors or properties of the Twittersphere and works that exploit microblogs for information extraction and retrieval tasks.

As for the former, Teevan et al. [49] analyze query logs to uncover differences in search behavior between users of classical Web search engines and users looking for information in microblogs. They found that `Twitter` queries are shorter and more popular than `bing` [12] queries on average. Furthermore, microblogs are more often sought for people, opinions, and breaking news. In terms of query formulation, reissuing the same query can be more frequently observed in microblog search. In Web search, by contrast, modifying and extending a query is very popular.

Java et al. [28] study network properties of the microblogosphere as well as geographical distributions and intentions of `Twitter` users. The authors report that `Twitter` is most popular in North America, Europe, and Asia (Japan), and that same language is an important factor for cross-connections ("followers" and "friends") over continents. Employing the *HITS* algorithm [29] on the network of "friend"-relations, Java et al. further derived user intentions from structural properties. They identified the fol-

lowing categories: information sharing, information seeking, and friendship-wise relationships. Analyzing the content of Twitter posts, the authors distilled the following intentions: daily chatter, conversations, sharing information/URLs, and reporting news.

In a recent study, Kwak et al. perform a topological analysis of the Twitter network [33]. The authors report a low level of reciprocity, i.e., only 22% of the connections between users are bidirectional. The average path length was found to be only four, which is surprisingly small for a network the size of the Twittersphere and considering the directional network structure. Moreover, a moderate level of homophily, i.e., a higher likelihood for connections between similar people than between dissimilar people, was discovered when measuring similarity in terms of geographic location and user popularity. In addition, Kwak et al.'s study indicates that information diffusion after the first retweet is very fast.

Work related to content mining of microblogs includes the following: Cheng et al. propose a method to localize Twitter users based on spatial cues ("local" words) extracted from their tweets' content [18]. To this end, in a first step several classifiers are trained to identify words with a strong geospatial meaning. In order to deal with the sparsity in the distribution of these cues, different smoothing approaches, e.g., taking into account neighboring cities when constructing the term representation of a city, are applied subsequently. In an experiment conducted on a set of tweets posted within the USA, Cheng et al.'s approach placed more than a half of the users within a 100-mile-radius of their correct location.

Making use of the fact that tweets are a good source for up-to-date information and breaking news, Dong et al. propose in [23] an approach to identify fresh URLs in Twitter posts. To this end, the authors investigate content-based features extracted from the tweets, an authority score computed for each user, and Twitter-specific statistical features, such as number of retweets or number of users that replied to a message containing a tiny URL. They show that these features can be used to improve both recency ranking and relevance ranking in real-time Web search. Another work that aims at improving ranking can be found in [24]. Duan et al. propose a novel ranking strategy for tweet retrieval. To this end, they investigate different feature sets, including content-based features, Twitter-specific features, and authority scores of users (followers, retweeters, mentioners). Using a learning to rank algorithm, the authors found that the best-performing features are authority scores, length of a tweet, and whether the tweet contains a URL.

An approach to classifying tweets can be found in [48]. Sriram et al. describe each tweet by an eight-dimensional feature vector comprising the author of the post and seven binary attributes indicating, for example, occurrence of slang words, currency and percentage signs, or the use of capitalization and repeated characters. Sriram et al.'s feature set outperformed the standard bag-of-words approach using a Naïve Bayes classifier to categorize tweets into the five classes news, events, opinions, deals, and private messages.

Armentano et al. present in [7] a recommender system that suggests potentially interesting users to follow based on the similarity between tweets posted by the seed user and tweets posted by a set of candidate users. To this end, the authors create and investigate different user profiles, for example, modeling the seed user via term frequencies of his/her aggregate posts or of all of his/her followees. Related to Armentano et al.'s work, Weng et al. aim at identifying influential twitterers for a given topic [51]. To this end, they apply *Latent Dirichlet Allocation* (LDA) [13] to their corpus of tweets. Subsequently, topical similarity between twitterers is computed as the Jensen-Shannon divergence between the distribution of the latent topics of the respective users. Further taking into account the link structure, Weng et al. propose a ranking function for influential twitterers in each topic. Similar to [7], Weng et al. evaluate their approach in a recommendation setting.

Microblogs have also been exploited for the purpose of event and trend detection. Sakaki et al. propose semantic analysis of tweets to detect earthquakes in Japan in real-time [39]. A more general approach to automatically detect events and summarize trends by analyzing tweets is presented by Sharifi et al. [47]. Another work on trend detection is [43], where Schedl exploits tweets for spatio-temporal popularity estimation of music artists. Sankaranarayanan et al. aim at capturing tweets that report on breaking news [41]. They cluster the identified tweets according to their $TF \cdot IDF$ weights and cosine similarity. Furthermore, each cluster is assigned a set of geographic locations using both spatial clues in the tweets themselves and explicit location information as indicated by the twitterers.

3. DATA ACQUISITION AND ANALYSIS

Between May 2010 and March 2011 we crawled Twitter for the hashtag #nowplaying (or its equivalent #NP) in the postings since this hashtag has already been successfully used in the context of spatio-temporal popularity estimation in [43]. Our crawls were restricted to tweets with geospatial information and to all cities $>$ 500,000 inhabitants (790 cities around the world were gathered from World Gazetteer [6]). We included tweets within a radius of 50 kilometers around the city center's coordinates.
Between November 2010 and March 2011 we gathered a second data set, focusing on tweets including #itunes, since this hashtag is frequently used among users of Apple's iTunes and related programs. There also exists a popular plug-in for Apple's social network Ping that automatically tweets iTunes listening activities using this very hashtag.

We were able to retrieve 9,928,817 tweets for #nowplaying and 725,486 tweets for #itunes, respectively. We will henceforth refer to these data sets simply as #nowplaying and #itunes. If not explicitly mentioned, our analysis was conducted on the larger data set (#nowplaying).

There exist, of course, tweets about music that do neither include the hashtag #nowplaying, nor #itunes. Elaborating a more general, text-based tweet classifier that detects microblogs about music will be part of future work.

3.1 Preprocessing

In order to use the crawled tweets for further processing we had to map the tweets to artist names, which raises some challenges. In a first attempt we tried to use a dictionary-based text matching algorithm. However, this approach led to many mis-classifications as parts of song titles or artist names had been identified as artists, which was especially true for common speech terms and first names. Additionally, there is no fixed format for Twitter postings and several tweets featuring the hashtags under consideration also contain comments. Stopping [9] is not an option in this case

Table 1: Top 10 cities (number of tweets).

#nowplaying		#itunes	
city	tweets	city	tweets
New York	126,952	New York	13,603
London	96,801	London	9,813
São Paulo	79,317	Los Angeles	9,030
Los Angeles	73,834	San Francisco	5,787
Amsterdam	66,021	San Jose	5,605
Guarulhos	58,453	Chicago	4,413
Osasco	57,512	Birmingham	3,869
São Bernardo	56,946	Toronto	3,363
Rotterdam	55,113	Hamilton	3,279
Mexico City	52,618	Baltimore	3,245

Table 2: Top 10 countries (number of tweets).

#nowplaying		#itunes	
country	#tweets	country	#tweets
Brazil	725,389	USA	78,460
USA	673,839	Japan	30,932
Japan	458,558	Mexico	23,047
Mexico	419,584	Brazil	16,390
Indonesia	284,082	UK	15,134
South Korea	251,132	Canada	11,266
China	183,178	South Korea	8,652
UK	128,744	Australia	5,119
Netherlands	121,134	China	4,492
Venezuela	110,336	Germany	3,157

since it might introduce erroneous information into the artist names.

To improve our matching algorithm we identified a number of common patterns in the tweets, including "*songtitle* by *artistname*", "*artistname – songtitle*", "*#artistname*", etc., and we matched the potential artist names against a list of 110,588 known artists. The artist set is publicly available.[1]

We were able to identify 31,328 unique artists in 4,237,430 of the 9,928,817 tweets, each artist appearing between 1 and 38,335 (Rihanna) times ($\mu = 144.89$, $\sigma = 907.96$, $median = 12$) in the `#nowplaying` data set. From the `#itunes` data set we extracted 13,002 artists from 220,641 tweets ($min = 1$, $max = 5{,}416$ (The Beatles), $\mu = 17.52$, $\sigma = 94.54$, $median = 3$).

We were able to retrieve data from 766 (603) different cities in 127 (107) countries. Tables 1 and 2 show the top-10 cities and countries, respectively, in terms of the number of postings.

3.2 Distribution of Play Counts

To assess whether the distribution of play counts, that is the total number of each artist's occurrences in the data set, follows a power-law $p(x) \sim x^{-\alpha}$ for $x \geq x_{min}$, we employ the approach presented in [21], using *Maximum Likelihood Estimation* (MLE). The estimated parameters of the power-law model (significantly well) fitted to the data are as follows:

`#nowplaying`: $\alpha = 2.10$, $x_{min} = 697$

`#itunes`: $\alpha = 2.01$, $x_{min} = 16$

[1] `http://www.cp.jku.at/people/schedl/datasets.html`

Figure 1 visualizes the distribution of the playcounts for both data sets (`#nowplaying` and `#itunes`).

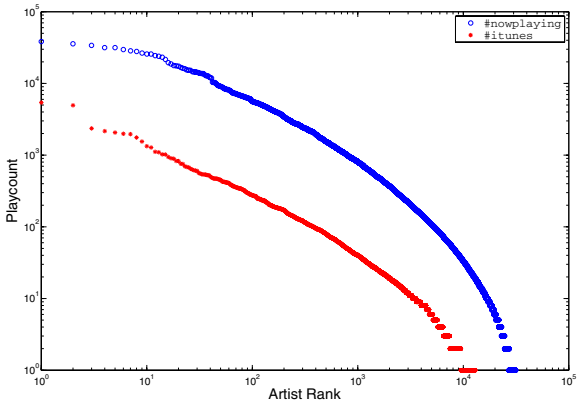

Figure 1: Distribution of artist playcounts.

4. SIMILARITY ESTIMATION

In order to estimate artist similarities we computed a co-occurrence matrix X for artists listened to by the same user. Element $x(i,j)$ denotes the co-occurrence count between artists i and j. For each artist we found between 0 and 16,999 co-occurring artists ($\mu = 1{,}392$, $\sigma = 2{,}040$, $median = 450$) for `#nowplaying` and between 0 and 3,832 co-occurring artists ($\mu = 69.91$, $\sigma = 154.52$, $median = 15$) for `#itunes`. Due to space limitations, we will report on our similarity estimation experiments for the `#nowplaying` data set only.

4.1 Normalization

For the normalization of the co-occurrence matrices, which eventually yields item-to-item similarity matrices, we evaluated four different approaches (the first one simply computes the relative frequency, algorithms two and three make use of a popularity correction factor [52]):

1. $sim(i,j) = \frac{x(i,j)}{occ(i)}$
2. $sim(i,j) = \frac{x(i,j)}{occ(i)} \cdot \left(1 - \frac{|occ(i) - occ(j)|}{\max_k occ(k)}\right)$
3. $sim(i,j) = \frac{x(i,j)}{min(occ(i), occ(j))} \cdot \left(1 - \frac{|occ(i) - occ(j)|}{\max_k occ(k)}\right)$
4. $sim(i,j) = \frac{x(i,j)}{\sqrt{occ(i) \cdot occ(j)}}$

with $occ(i)$ being the number of occurrences of artist i within the set of tweets.

4.2 Evaluation

In a first step we evaluated the four artist similarity measures in a similar-artist-prediction setting, using the list of top-ranked similar artists from `last.fm` [34] as ground truth. As some artists appeared too infrequently in the data sets to create a reliable similarity predictor, we had to define a

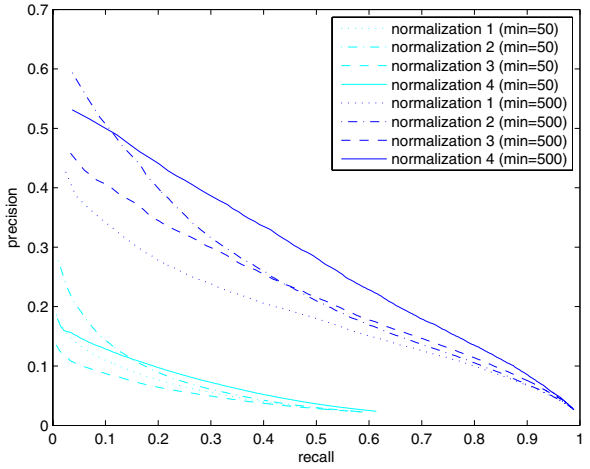

Figure 2: Precision-recall-curves.

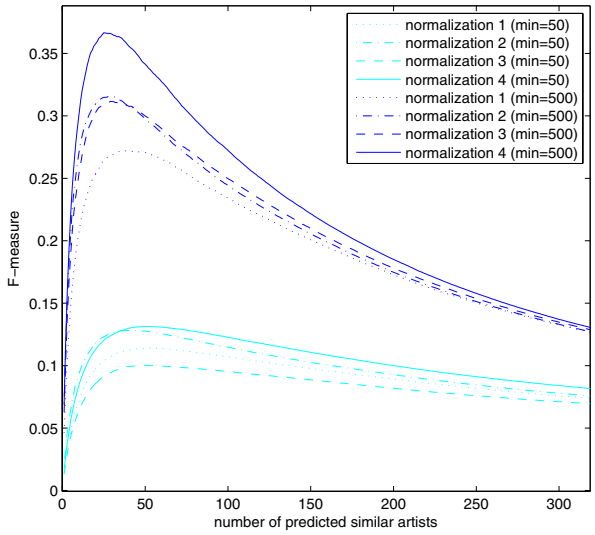

Figure 3: F-measure for number of predicted artists.

Table 3: R-precision for artist overlap.

normalization method	$\geq$ 50 tweets (6,885 artists)	$\geq$ 500 tweets (1,524 artists)
normalization 1	0.1482	0.2919
normalization 2	0.1678	0.3381
normalization 3	0.1121	0.3282
normalization 4	0.1621	0.3930

Table 4: Best F-measures for artist overlap.

normalization method	$\geq$ 50 tweets		$\geq$ 500 tweets	
	pred	F	pred	F
normalization 1	54	0.1141	37	0.2722
normalization 2	39	0.1282	29	0.3163
normalization 3	49	0.1002	29	0.3116
normalization 4	52	0.1314	25	0.3666

minimum number of occurrences. 8,122 artists appeared in at least 50 `#nowplaying` tweets, 1,524 in at least 500 `#nowplaying` tweets. We were able to retrieve our ground truth for 6,885, respectively 1,137 of these artists. For the set with a minimum of 50 tweets per artist, `last.fm` returned between 2 and 93 similar artists ($\mu = 27.20$, $\sigma = 20.46$, $median = 22$), for the smaller set between 2 and 71 similar artists ($\mu = 23.74$, $\sigma = 15.45$, $median = 21$).

The R-precision [8] for the overlap in Table 3 shows that the normalization algorithms 2 and 4 work best. However, comparing the two data sets we see that the algorithms perform differently. A closer look at the precision-recall-curves in Figure 2 shows that symmetric algorithms (3 and 4) have a flatter curve, i.e., they lose less in precision while the number of predictions increases. Figure 3 shows that for the smaller data set (i.e., the more popular artists) symmetric algorithms perform relatively better, especially for an increasing number of predictions. Among the symmetric algorithms 4 always performs better than 3. For the asymmetric algorithms we can see that the popularity factor in algorithm 2 always leads to better results than the relative frequency alone. The maximum F-measures achieved are listed in Table 4 for both thresholds (50 and 500) for the minimum number of artist occurrences. The column *pred* indicates the number k of predicted similar artists, for which this maximum F-measure was achieved.

In a next step we evaluated the rank proximity, predicting the k most similar artists (l=number of artists in the ground truth), and defining precision and recall equivalents using a weighted rank proximity as follows:

$$prec(k) = \frac{1}{k} \cdot \sum_{i=1}^{k} \left(1 - \frac{|r_{pred}(a_i) - r_{gt}(a_i)|}{max(r_{pred}(a_i), r_{gt}(a_i))}\right)$$

$$rec(k) = \frac{1}{l} \cdot \sum_{i=1}^{k} \left(1 - \frac{|r_{pred}(a_i) - r_{gt}(a_i)|}{max(r_{pred}(a_i), r_{gt}(a_i))}\right)$$

If a predicted artist is not found in the ground truth, we set a penalty of 0. Therefore, in contrast to the standard definitions of precision and recall, the true positives in our rank-proximity-formulation do not count as "one" if there is an overlap, but only if the rank is correct as well. By dividing the absolute rank difference by the maximum rank in ground truth and prediction, we put a stronger penalty on errors of top-ranked items. From Figures 4 and 5 we see that normalization algorithm 4 again dominates over the others and that the symmetric algorithms benefits from a lower number of artists.

5. EXTRACTING GEOSPATIAL LISTENING PATTERNS

The second objective of this work relates to extracting geospatial listening patterns from microblogs, analyzing if and in which way they differ among different parts of the world and at different granularity levels, and eventually, interpret these (presumable) differences. In the work at hand,

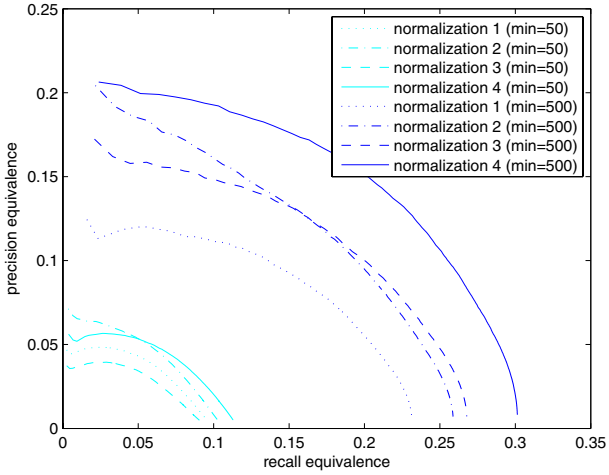

Figure 4: Precision-recall equivalences using rank proximity.

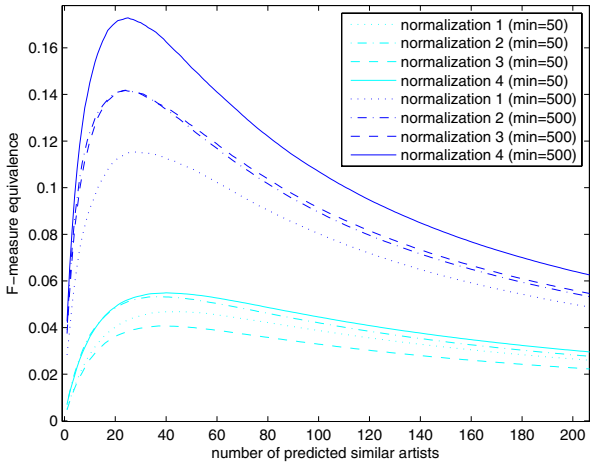

Figure 5: F-measure equivalence using rank proximity for number of predicted artists.

we infer such listening patterns at the level of cities and countries.

To determine typical listening patterns, we use genre information about artists to describe each city/country via a *genre distribution vector*.[2] Aggregating artists on the genre level is necessary because the country-artist-matrix is too sparse to allow for analyzing the listening patterns on the artist level for most countries. Using the set of `allmusic`'s [2] 18 major genres[3], we retrieve the genre of each artist a as the main genre given by `allmusic`'s artist page.

We define the *listening pattern* for a city or country c as the relative frequencies music of each genre is listened to by users within c. The elements of the 18-dimensional genre distribution vector $\boldsymbol{g}^c$ for a city or country c are computed as

$$g_i^c = \frac{\sum_{a \in G_i} occ_{a,c}}{\sum_{a \in A} occ_{a,c}} \qquad i = 1 \ldots 18$$

where G_i denotes the set of artists assigned to genre i, $occ_{a,c}$ is the number of microblogs indicating listening behavior of artist a in city or country c, and A is the set of all artists. We use the relative frequency to account for different intensities of microblogging activity, depending on the scope of interest. Nevertheless, we discard cities or countries for which not enough data is available to derive a reliable listening pattern. To this end, we require at least 100 artist-user pairs in the data set to make a prediction, i.e., $\sum_{a \in A} occ_{a,c} \geq 100$.

In order to investigate to which extent the listening patterns differ among different cities or countries c, we calculate the standard deviation of their genre distribution vectors σ^c over all genres (from the global mean genre distribution, taking the arithmetic mean over the 18 dimensions to come up with a single number). This allows us to determine the most and least representative (or mainstream) twittering population with respect to the average music listener on a global scale.

The results at the country level are illustrated for data sets `#nowplaying` and `#itunes` in Figures 6 and 7, respectively. The first bar in each genre group represents the global mean genre distribution, the subsequent two bars represent the countries with lowest standard deviations, i.e., the most mainstream countries, and the last two bars represent the countries whose twitterers have most particular listening behaviors. When deriving listening patterns on the level of individual cities, the most and least mainstream cities are depicted in Figures 8 and 9, respectively, for sets `#nowplaying` and `#itunes`.

We further compute global genre distribution vectors $\boldsymbol{g}^{\#np}$ and $\boldsymbol{g}^{\#it}$ for both data sets `#nowplaying` and `#itunes`, respectively. Comparing these with the genre distribution vector of the ground truth $\boldsymbol{g}^{GT}$ – which is given by the relative frequencies of artists in each genre among the total number of artists in the data set – reveals some interesting differences between the user group that tends to use `#nowplaying` and the group that tends to use `#itunes` to tweet listening behavior.[4] Figure 10 depicts the genre distribution vectors $\boldsymbol{g}^{GT}$, $\boldsymbol{g}^{\#np}$, and $\boldsymbol{g}^{\#it}$. It shows, for instance, that electronica, rap, and rock are significantly more popular genres among the `#itunes` tweeters, compared to their relative share in the ground truth. The same holds for latin, rap, rnb, and rock when focusing on the microbloggers that use `#nowplaying` to communicate listening activities. On the other hand, almost all other genres are significantly less popular among both user groups. The ℓ^2 distance between $\boldsymbol{g}^{GT}$ and $\boldsymbol{g}^{\#np}$ is also larger (0.0996) than between $\boldsymbol{g}^{GT}$ and $\boldsymbol{g}^{\#it}$ (0.0796), which indicates a slightly less average music taste for the `#itunes` listeners than for the `#nowplaying` tweeterers.

[2]The approach is not restricted to the use of genres; other term sets may feature moods, instruments, or styles.

[3]The used genres are avantgarde (av), blues (bl), celtic (ce), classical (cl), country (co), easylistening (ea), electronica (el), folk (fo), gospel (go), jazz (ja), latin (la), newage (ne), rap (ra), reggae (re), rnb (rn), rock (ro), vocal (vo), and world (wo).

[4]The second one presumably consists foremost of listeners who use `Apple`'s `iTunes`.

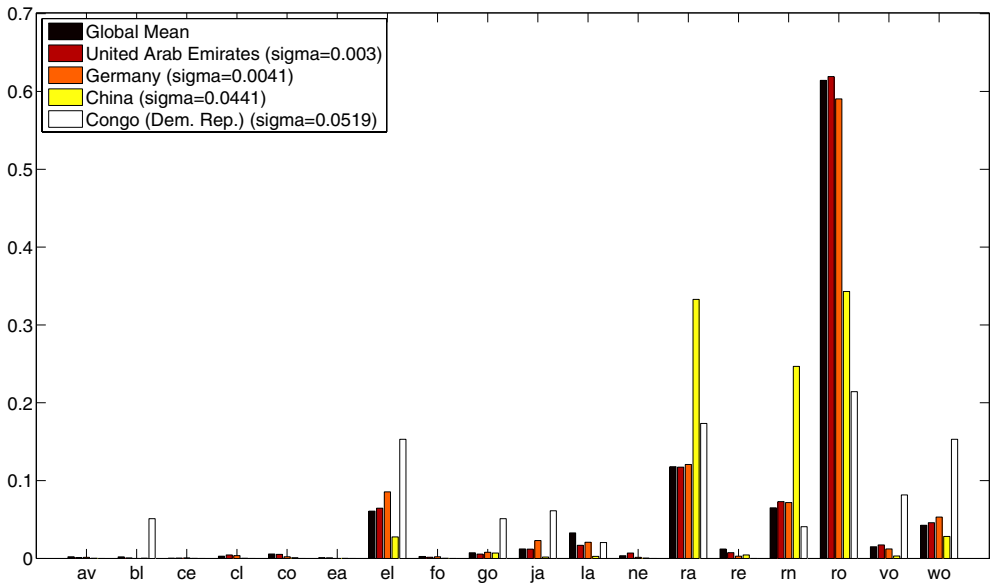

Figure 6: Genre distributions for countries with most and least representative listening behavior, using data set `#nowplaying`.

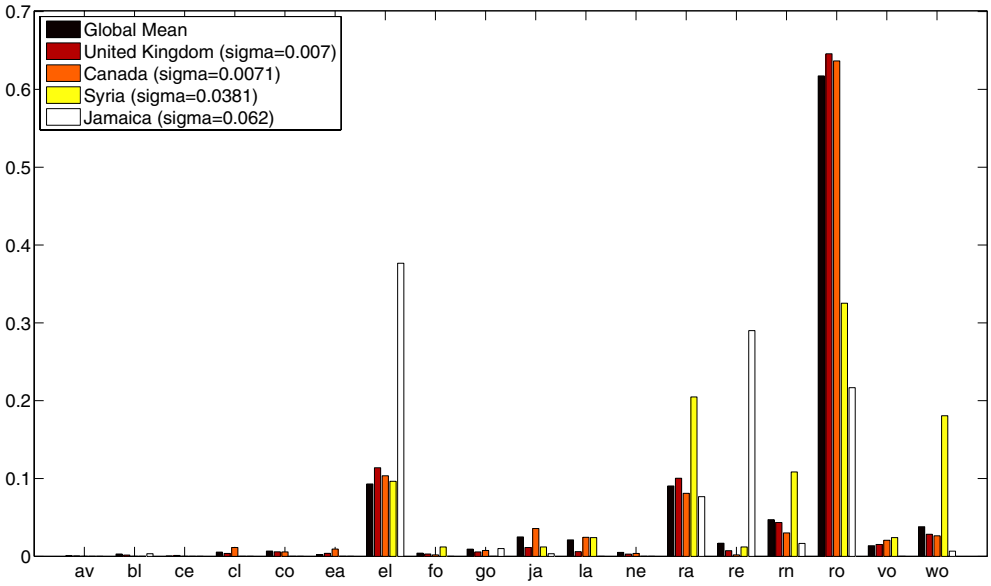

Figure 7: Genre distributions for countries with most and least representative listening behavior, using data set `#itunes`.

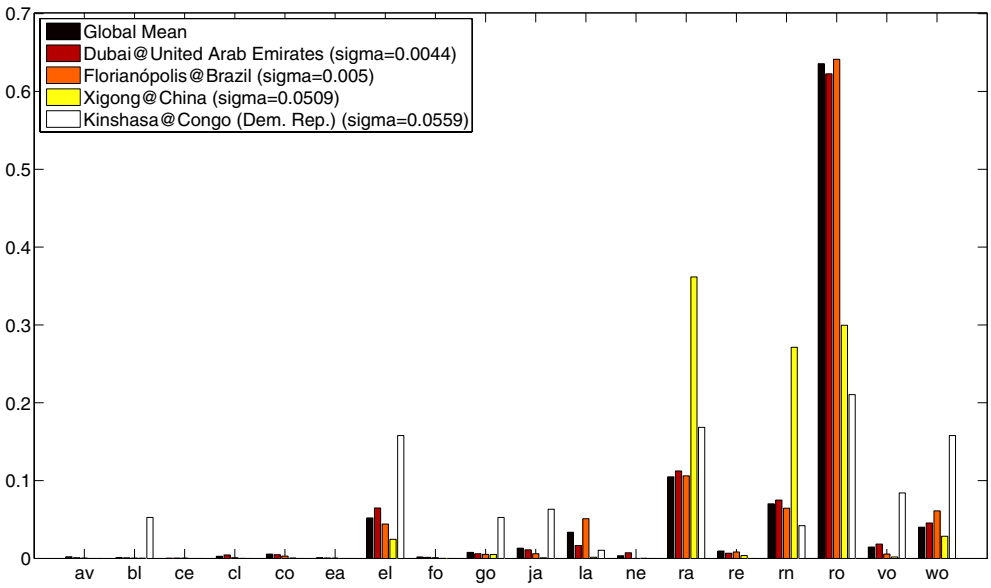

Figure 8: Genre distributions for cities with most and least representative listening behavior, using data set `#nowplaying`.

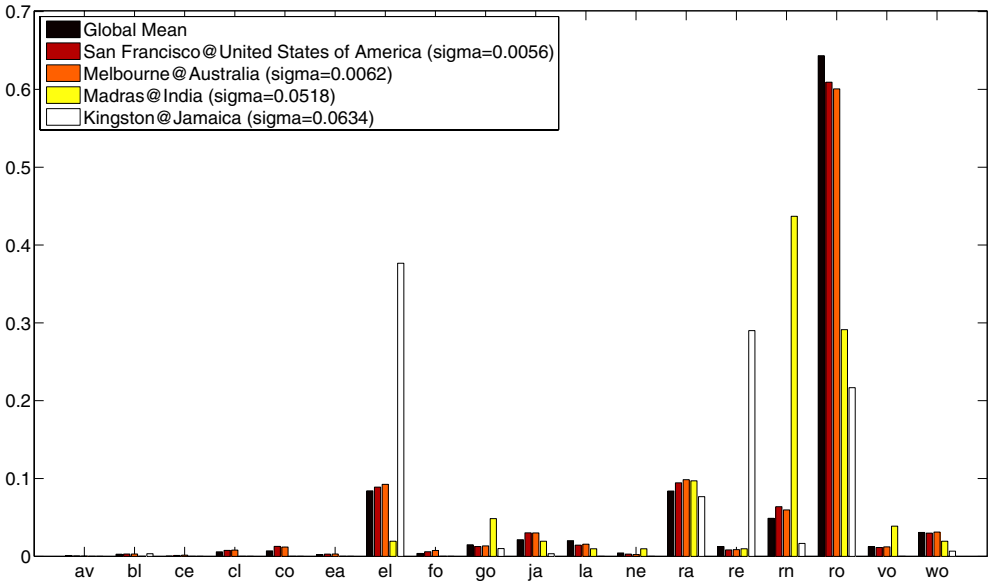

Figure 9: Genre distributions for cities with most and least representative listening behavior, using data set `#itunes`.

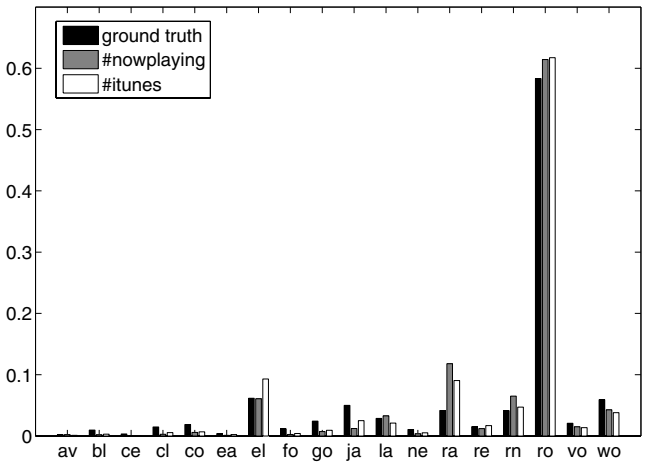

Figure 10: Overall genre distributions given by the ground truth, the `#nowplaying` and the `#itunes` sets.

6. CONCLUSIONS AND FUTURE WORK

We reported on mining social microblogs for music-related information and showed that `Twitter` postings can be used to derive similarity measures for artists. We compared different normalization approaches against similarity information from `last.fm` for their overlap and their weighted rank proximity. In addition, we derived music listening patterns, represented as genre distribution vectors, at different granularities (city, country, global). We found that these listening patterns vary strongly for different cities and countries. Moreover, the global patterns revealed a considerable difference between the data sets `#nowplaying` and `#itunes` in relation to the mean global genre distribution given by the ground truth.

We are currently preparing the feature data (in addition to the meta data) for publication as we believe it represents a valuable source for the research community. Both preprocessed data sets (`#nowplaying` and `#itunes`) will be available shortly.[5] As part of future work, we will investigate more elaborate disambiguation techniques for song and artist names to improve accuracy of the similarity estimators. We believe that preprocessing could be improved by adding track information, e.g., for the tweet "Satellite – Lena", both "Lena" and "Satellite" exist as known artist names, and only additional track information can help to identify "Lena" as the artist and "Satellite" as the track title and not vice versa.

We will further investigate how similarity information derived from microblog data compares to similarity estimation techniques that exploit other data sources, such as the audio signal, music-related web pages, or song lyrics.

Moreover,we plan to account for temporal dynamics in the microblog data, which matches nicely our ultimate research aim to elaborate personalized and user-aware music retrieval and discovery systems that take into account different levels of personalization (individual, peer group, city, country) [45].

The research at hand is also closely related to trend detection from social media sources. Given the large interest the music industry and individual performers have to monitor their success, we hence foresee more work leveraging microblog data for this purpose.

[5] `http://www.cp.jku.at/people/schedl/datasets.html`

7. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Funds (FWF): P22856-N23.

8. REFERENCES

[1] http://blog.twitter.com/2011/03/numbers.html (access: Nov 2011).

[2] http://www.allmusic.com (access: Jan 2010).

[3] http://www.huffingtonpost.com/2011/04/28/twitter-number-of-users_n_855177.html (access: Nov 2011).

[4] http://www.spotify.com (access: Nov 2011).

[5] http://www.twitter.com (access: Nov 2011).

[6] http://www.world-gazetteer.com (access: Oct 2010).

[7] M. G. Armentano, D. Godoy, and A. A. Amandi. Recommending Information Sources to Information Seekers in Twitter. In *Proc. IJCAI: International Workshop on Social Web Mining*, Barcelona, Spain, Jul 2011.

[8] J. A. Aslam and E. Yilmaz. A Geometric Interpretation and Analysis of R-precision. In *Proc. CIKM*, Bremen, Germany, Oct–Nov 2005.

[9] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[10] L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, K.-H. Lüke, and R. Schwaiger. InCarMusic: Context-Aware Music Recommendations in a Car. In *Proc. EC-Web*, Toulouse, France, Aug–Sep 2011.

[11] S. Baumann and O. Hummel. Using Cultural Metadata for Artist Recommendation. In *Proc. WEDELMUSIC*, Leeds, UK, Sep 2003.

[12] http://www.bing.com (access: January 2011).

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Machine Learning Research*, 3, Mar 2003.

[14] E. Brill. A Simple Rule-Based Part of Speech Tagger. In *Proc. ANLP*, Trento, Italy, Mar–Apr 1992.

[15] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proc. IEEE*, 96, Apr 2008.

[16] O. Celma. *Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, Berlin, Heidelberg, Germany, 2010.

[17] O. Celma, P. Cano, and P. Herrera. SearchSounds: An Audio Crawler Focused on Weblogs. In *Proc. ISMIR*, Victoria, Canada, Oct 2006.

[18] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In *Proc. CIKM*, Oct 2010.

[19] P. Cimiano, S. Handschuh, and S. Staab. Towards the Self-Annotating Web. In *Proc. WWW*, New York, NY, USA, May 2004.

[20] P. Cimiano and S. Staab. Learning by Googling. *ACM SIGKDD Explorations Newsletter*, 6(2), 2004.

[21] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Reviews*, 51, Nov 2009.

[22] W. W. Cohen and W. Fan. Web-Collaborative Filtering: Recommending Music by Crawling The Web. *WWW9 / Computer Networks*, 33(1–6), 2000.

[23] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the Essence: Improving Recency Ranking Using Twitter Data. In *Proc. WWW*, Raleigh, NC, USA, Apr 2010.

[24] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H. Shum. An Empirical Study on Learning to Rank of Tweets. In *Proc. COLING*, Beijing, China, Aug 2010.

[25] http://www.epinions.com/music (access: Aug 2007).

[26] G. Geleijnse and J. Korst. Web-based Artist Categorization. In *Proc. ISMIR*, Victoria, Canada, Oct 2006.

[27] X. Hu, J. S. Downie, K. West, and A. Ehmann. Mining Music Reviews: Promising Preliminary Results. In *Proc. ISMIR*, London, UK, Sep 2005.

[28] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. WebKDD and SNA-KDD*, San Jose, CA, USA, Aug 2007.

[29] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5), 1999.

[30] P. Knees, E. Pampalk, and G. Widmer. Artist Classification with Web-based Data. In *Proc. ISMIR*, Barcelona, Spain, Oct 2004.

[31] P. Knees, T. Pohle, M. Schedl, and G. Widmer. A Music Search Engine Built upon Audio-based and Web-based Similarity Measures. In *Proc. SIGIR*, Amsterdam, the Netherlands, Jul 2007.

[32] P. Knees, M. Schedl, T. Pohle, and G. Widmer. An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. In *Proc. ACM Multimedia*, Santa Barbara, CA, USA, Oct 2006.

[33] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc. WWW*, Apr 2010.

[34] http://last.fm (access: January 2010), 2010.

[35] H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal*, October 1957.

[36] B. McFee and G. Lanckriet. The Natural Language of Playlists. In *Proc. ISMIR*, Miami, FL, USA, Oct 2011.

[37] E. Pampalk and M. Goto. MusicRainbow: A New User Interface to Discover Artists Using Audio-based Similarity and Web-based Labeling. In *Proc. ISMIR*, Victoria, Canada, Oct 2006.

[38] T. Pohle, P. Knees, M. Schedl, E. Pampalk, and G. Widmer. "Reinventing the Wheel": A Novel Approach to Music Player Interfaces. *IEEE Transactions on Multimedia*, 9:567–575, 2007.

[39] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proc. WWW*, May 2010.

[40] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Comm. ACM*, 18(11), 1975.

[41] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: News in tweets. In *Proc. SIGSPATIAL*, Seattle, WA, USA, Nov 2009.

[42] M. Schedl. On the Use of Microblogging Posts for Similarity Estimation and Artist Labeling. In *Proc. ISMIR*, Utrecht, the Netherlands, Aug 2010.

[43] M. Schedl. Analyzing the Potential of Microblogs for Spatio-Temporal Popularity Estimation of Music Artists. In *Proc. IJCAI: International Workshop on Social Web Mining*, Barcelona, Spain, Jul 2011.

[44] M. Schedl. *Music Data Mining*, chapter Web-Based and Community-based Music Information Extraction. CRC Press/Chapman Hall, 2011.

[45] M. Schedl and P. Knees. Personalization in Multimodal Music Retrieval. In *Proc. AMR*, 2011.

[46] M. Schedl, P. Knees, and G. Widmer. A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In *Proc. CBMI*, Riga, Latvia, Jun 2005.

[47] B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In *Proc. NAACL HLT*, Jun 2010.

[48] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short Text Classification in Twitter to Improve Information Filtering. In *Proc. SIGIR*, Geneva, Switzerland, Jul 2010.

[49] J. Teevan, D. Ramage, and M. R. Morris. #TwitterSearch: A Comparison of Microblog Search and Web Search. In *Proc. WSDM*, Hong Kong, China, Feb 2011.

[50] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[51] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-Sensitive Influential Twitterers. In *Proc. WSDM*, New York, NY, USA, Feb 2010.

[52] B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proc. ICMC*, Göteborg, Sweden, Sep 2002.

[53] G.-R. Xue, J. Han, Y. Yu, and Q. Yang. User Language Model for Collaborative Personalized Search. *ACM Transactions on Information Systems*, 27(2), Feb 2009.

[54] M. Zadel and I. Fujinaga. Web Services for Music Information Retrieval. In *Proc. ISMIR*, Barcelona, Spain, Oct 2004.