A Survey of Music Similarity and Recommendation from Music Context Data

PETER KNEES and MARKUS SCHEDL, Johannes Kepler University Linz

In this survey article, we give an overview of methods for music similarity estimation and music recommendation based on music context data. Unlike approaches that rely on *music content* and have been researched for almost two decades, *music-context*-based (or *contextual*) approaches to music retrieval are a quite recent field of research within music information retrieval (MIR). *Contextual data* refers to all music-relevant information that is not included in the audio signal itself. In this article, we focus on contextual aspects of music primarily accessible through web technology. We discuss different sources of context-based data for individual music pieces and for music artists. We summarize various approaches for constructing similarity measures based on the collaborative or cultural knowledge incorporated into these data sources. In particular, we identify and review three main types of context-based similarity approaches: *text-retrieval-based approaches* (relying on web-texts, tags, or lyrics), *co-occurrence-based approaches* (relying on playlists, page counts, microblogs, or peer-to-peer-networks), and approaches based on *user ratings* or listening habits. This article elaborates the characteristics of the presented context-based measures and discusses their strengths as well as their weaknesses.

Categories and Subject Descriptors: A.1 [**Introductory and Survey**]; H.5.5 [**Information Interfaces and Presentation (e.g., HCI)**]: Sound and Music Computing; I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Algorithms

Additional Key Words and Phrases: Music information retrieval, music context, music similarity, music recommendation, survey

ACM Reference Format:

Knees, P. and Schedl, M. 2013. A survey on music similarity and recommendation from music context data. ACM Trans. Multimedia Comput. Commun. Appl. 10, 1, Article 2 (December 2013), 21 pages.
DOI: http://dx.doi.org/10.1145/2542205.2542206

1. INTRODUCTION

Music information retrieval (MIR), a subfield of multimedia information retrieval, has been a fast growing field of research during the past two decades. In the bulk of MIR research so far, music-related information is primarily extracted from the audio using signal processing techniques [Casey et al. 2008]. In discovering and recommending music from today's ever growing digital music repositories, however, such content-based features, despite all promises, have not been employed very successfully in large-scale systems so far. Indeed, it seems that collaborative filtering approaches and music

This research is supported by the Austrian Science Fund (FWF): P22856-N23 and P25655.

Authors' address: Dept. of Computational Perception, Johannes Kepler University, Altenberger Str. 69, 4040 Linz, Austria; email: {peter.knees, markus.schedl}@jku.at.

1551-6857/2013/04-ART2
DOI: http://dx.doi.org/10.1145/2542205.2542206

information systems using *contextual metadata*[1] have higher user acceptance and even outperform content-based techniques for music retrieval [Slaney 2011].

In recent years, various platforms and services dedicated to the music and audio domain, such as `Last.fm`,[2] `MusicBrainz`,[3] or `echonest`,[4] have provided novel and powerful, albeit noisy, sources for high level, semantic information on music artists, albums, songs, and other entities. Likewise, a noticeable number of publications that deal with such kind of music-related, contextual data have been published and contributed to establishing an additional research field within MIR.

Exploiting context-based information permits, among others, automatic tagging of artists and music pieces [Sordo et al. 2007; Eck et al. 2008], user interfaces to music collections that support browsing beyond the regrettably widely used *genre - artist - album - track* hierarchy [Pampalk and Goto 2007; Knees et al. 2006b], automatic music recommendation [Celma and Lamere 2007; Zadel and Fujinaga 2004], automatic playlist generation [Aucouturier and Pachet 2002; Pohle et al. 2007], as well as building music search engines [Celma et al. 2006; Knees et al. 2007; Barrington et al. 2009]. Furthermore, because context-based features stem from different sources than content-based features and represent different aspects of music, these two categories can be beneficially combined in order to outperform approaches based on just one source, for example, to accelerate the creation of playlists [Knees et al. 2006a], to improve the quality of classification according to certain metadata categories like genre, instrument, mood, or listening situation [Aucouturier et al. 2007], or to improve music retrieval by incorporating multimodal sources [Zhang et al. 2009]. Although the proposed methods and their intended applications are highly heterogeneous, they have in common that the notion of *music similarity* is key.

With so many different approaches being proposed over the last decade and the broad variety of sources they are being built upon, it is important to take a snapshot of these methods and impose structure to this field of academic interest. Even though similarity is just one essential and reoccuring aspect and there are even more publications which exploit contextual data outside of this work, MIR research on music context similarity has reached a critical mass that justifies a detailed investigation.

The aim of this survey paper is to review work that makes use of contextual data by putting an emphasis on methods to define similarity measures between artists or individual tracks.[5] Whereas the term, *context*, is often used to refer to the user's context or the usage context, expressed through parameters such as location, time, or activity (cf. Wang et al. [2012] and Schedl and Knees [2011]), context here specifically refers to the music context which comprises of information on and related to music artists and pieces. The focus of this article is on similarity which originates from this context of music, more precisely, on information primarily accessible through web technology.

In the remainder of this article, we first, in Section 2, give a general view of content-based and context-based data sources. We subsequently review techniques for context-based similarity estimation that can be categorized into three main areas: *text-retrieval-based*, *co-occurrence-based*, and *user-rating-based* approaches. For text-retrieval approaches, we further distinguish between approaches relying on web-texts, collaborative tags, and lyrics as data sources. This is discussed in Section 3. For co-occurrence approaches, we identify page counts, playlists, microblogs, and peer-to-peer-networks as potential sources (Section 4). Finally, in Section 5, we review approaches that make use of user

[1] This kind of data is also referred to as "cultural features", "community metadata", or "(music) context-based features".

[2] `http://www.last.fm`.

[3] `http://www.musicbrainz.org`.

[4] `http://the.echonest.com`.

[5] A first endeavour to accomplish this has already been undertaken in Schedl and Knees [2009].

Table I. A Comparison of Music-content- and Music-context-based Features

	Content-based	*Context-based*
Prerequisites	Music file	Users
Metadata required	No	Yes
Cold-start problem	No	Yes
Popularity bias	No	Yes
Features	Objective	Subjective
	Direct	Noisy
	Numeric	"Semantic"

ratings and users' listening habits by applying collaborative filtering methods. For each of the presented methods, we describe mining of the corresponding sources in order to construct meaningful features as well as the usage of these features for creating a similarity measure. Furthermore we aim to estimate potential and capabilities of the presented approaches based on the reported evaluations. However, since evaluation strategies and datasets differ largely, a direct and comprehensive comparison of their performances is not possible. Finally, Section 7 summarizes this work, discusses pros and cons of the individual approaches, and gives an outlook on possible directions for further research on context-based music information extraction and similarity calculation.

2. GENERAL CONSIDERATIONS

Before examining existing approaches in detail, we want to discuss general implications of incorporating context-based similarity measures (cf. Turnbull et al. [2008]), especially in contrast to content-based measures. The idea behind content-based approaches is to extract information directly from the audio signal, more precisely, from a digital representation of a recording of the acoustic wave, which needs to be accessible. To compare two pieces, their signals are typically cut into a series of short segments called frames which are optionally transformed from the time-domain representation into a frequency-domain representation, for example, by means of a Fourier transformation. Thereafter, feature extraction is performed on each frame in some approach-specific manner. Finally, the extracted features are summarized for each piece, for example, by statistically modeling their distribution. Between these summarizations, pairwise similarities of audio tracks can be computed. A comprehensive overview of content-based methods is given by Casey et al. [2008].

In contrast to content-based features, to obtain context-based features it is not necessary to have access to the actual music file. Hence, applications like, for instance, music information systems, can be built without any acoustic representation of the music under consideration by having a list of artists [Schedl 2008]. On the other hand, without meta-information like *artist* or *title*, most context-based approaches are inapplicable. Also, improperly labeled pieces and ambiguous identifiers pose a problem. Furthermore, unless one is dealing with user ratings, all contextual methods depend on the existence of available metadata. This means that music not present within the respective sources is virtually inexistent, as may be the case for music from the long-tail ("popularity bias") as well as for up-and-coming music and sparsely populated (collaborative) data sources ("cold start problem"). To sum up, the crucial point is that deriving cultural features requires access to a large amount of unambiguous and non-noisy user generated data. Assuming this condition can be met, community data provides a rich source of information on social context and reflects the "collective wisdom of the crowd" without any explicit or direct human involvement necessary. Table I gives a brief comparison of content- and context-based feature properties.

3. TEXT-BASED APPROACHES

In this section, we review work that exploits textual representations of musical knowledge originating from web pages, user tags, or song lyrics. Given this form, it seems natural to apply techniques originating from traditional Information Retrieval (IR) and Natural Language Processing (NLP), such as the bag-of-words representation, TF·IDF weighting (e.g., Zobel and Moffat [1998]), Latent Semantic Analysis (LSA) [Deerwester et al. 1990], and Part-of-Speech (PoS) Tagging (e.g., Brill [1992] and Charniak [1997]).

3.1 Web-Text Term Profiles

Possibly the most extensive source of cultural data are the zillions of available web pages. The majority of the presented approaches use a web search engine to retrieve relevant documents and create artist term profiles from a set of unstructured web texts. In order to restrict the search to web pages relevant to music, different query schemes are used. Such schemes may comprise of the artist's name augmented by the keywords `music review` [Whitman and Lawrence 2002; Baumann and Hummel 2003] or `music genre style` [Knees et al. 2004]. Additional keywords are particularly important for artists whose names have another meaning outside the music context, such as “50 Cent”, “Hole”, and “Air”. A comparison of different query schemes can be found in Knees et al. [2008].

Whitman and Lawrence [2002] extract different term sets (unigrams, bigrams, noun phrases, artist names, and adjectives) from up to 50 artist-related pages obtained via a search engine. After downloading the web pages, the authors apply parsers and a PoS tagger [Brill 1992] to determine each word's part of speech and the appropriate term set. Based on term occurrences, individual term profiles are created for each artist by employing a version of the well-established TF·IDF measure, which assigns a weight to each term t in the context of each artist A_i. The general idea of TF·IDF is to consider terms more important which occur often within the document (here, the web pages of an artist), but rarely in other documents (other artists' web pages). Technically speaking, terms that have a high *term frequency* (TF) and a low *document frequency* (DF) or, correspondingly, a high *inverse document frequency* (IDF) are assigned higher weights. Equation (1) shows the weighting used by Whitman and Lawrence, where the term frequency $tf(t, A_i)$ is defined as the percentage of retrieved pages for artist A_i containing term t, and the document frequency $df(t)$ as the percentage of artists (in the whole collection) who have at least one web page mentioning term t.

$$w_{simple}(t, A_i) = \frac{tf(t, A_i)}{df(t)}. \tag{1}$$

Alternatively, the authors propose another variant of weighting in which rarely occurring terms, that is, terms with a low DF, also should be weighted down to emphasize terms in the middle IDF range. This scheme is applied to all term sets except for adjectives. Equation (2) shows this alternative version where μ and σ represent values manually chosen to be 6 and 0.9, respectively.

$$w_{gauss}(t, A_i) = \frac{tf(t, A_i)e^{-(log(df(t))-\mu)^2}}{2\sigma^2}. \tag{2}$$

Calculating the TF·IDF weights for all terms in each term set yields individual feature vectors or term profiles for each artist. The *overlap* between the term profiles of two artists, that is, the sum of weights of all terms that occur in both artists' sets, is then used as an estimate for their similarity (Eq. (3)).

$$sim_{overlap}(A_i, A_j) = \sum_{\{\forall t | w(t,A_i)>0, w(t,A_j)>0\}} w(t, A_i) + w(t, A_j). \tag{3}$$

For evaluation, the authors compare these similarities to two other sources of artist similarity information, which serve as ground truth (similar-artist-relations from the online music information system All Music Guide (AMG)[6] and user collections from OpenNap,[7] cf. Section 4.4). Remarkable differences between the individual term sets can be made out. The unigram, bigram, and noun phrase sets perform considerably better than the other two sets, regardless of the utilized ground truth definition.

Extending the work presented in Whitman and Lawrence [2002], Baumann and Hummel [2003] introduce filters to prune the set of retrieved web pages. They discard all web pages with a size of more than 40kB after parsing and ignore text in table cells if it does not comprise of at least one sentence and more than 60 characters in order to exclude advertisements. Finally, they perform keyword spotting in the URL, the title, and the first text part of each page. Each occurrence of the initial query constraints (artist name, “music”, and “review”) contributes to a page score. Pages that score too low are filtered out. In contrast to Whitman and Lawrence [2002], Baumann and Hummel [2003] use a logarithmic IDF weighting in their TF·IDF formulation. Using these modifications, the authors are able to outperform the approach presented in Whitman and Lawrence [2002].

Another approach that applies web mining techniques similarly to Whitman and Lawrence [2002] is presented in Knees et al. [2004]. Knees et al. [2004] do not construct several term sets for each artist, but operate only on a unigram term list. A TF·IDF variant is employed to create a weighted term profile for each artist. Equation (4) shows the TF·IDF formulation, where n is the total number of web pages retrieved for all artists in the collection, $tf(t, A_i)$ is the number of occurrences of term t in all web pages retrieved for artist A_i, and $df(t)$ is the number of pages in which t occurs at least once. In the case of $tf(t, A_i)$ equaling zero, $w_{ltc}(t, A_i)$ is also defined as zero.

$$w_{ltc}(t, A_i) = (1 + \log_2 tf(t, A_i)) \cdot \log_2 \frac{n}{df(t)}. \tag{4}$$

To calculate the similarity between the term profiles of two artists A_i and A_j, the authors use the cosine similarity according to Eq. (5) and (6), where T denotes the set of all terms. In these equations, θ gives the angle between A_i’s and A_j’s feature vectors in the Euclidean space.

$$sim_{cos}(A_i, A_j) = \cos\theta \tag{5}$$

and

$$\cos\theta = \frac{\sum_{t\in T} w(t, A_i) \cdot w(t, A_j)}{\sqrt{\sum_{t\in T} w(t, A_i)^2} \cdot \sqrt{\sum_{t\in T} w(t, A_j)^2}}. \tag{6}$$

The approach is evaluated in a genre classification setting using k-Nearest Neighbor (k-NN) classifiers on a test collection of 224 artists (14 genres, 16 artists per genre). It results in accuracies of up to 77%.

Similarity according to Eqs. (4), (5), and (6) is also used in Pampalk et al. [2005] for clustering of artists. Instead of constructing the feature space from all terms contained in the downloaded web pages, a manually assembled vocabulary of about 1,400 terms related to music (e.g., genre and style names, instruments, moods, and countries) is used. For genre classification using a 1-NN classifier (performing leave-one-out cross validation on the 224-artist-set from Knees et al. [2004]), the unrestricted term set outperformed the vocabulary-based method (85% vs. 79% accuracy).

Another approach that extracts TF·IDF features from artist-related web pages is presented in Pohle et al. [2007]. Pohle et al. [2007] compile a data set of 1,979 artists extracted from AMG. The TF·IDF vectors are calculated for a set of about 3,000 tags extracted from Last.fm. The set of tags is constructed

[6] now named allmusic; http://www.allmusic.com.

[7] http://opennap.sourceforge.net.

by merging tags retrieved for the artists in the collection with `Last.fm`'s most popular tags. For evaluation, k-NN classification experiments with leave-one-out cross validation are performed, resulting in accuracies of about 90%.

Additionally, there exist some other approaches that derive term profiles from more specific web resources. For example, Celma et al. [2006] propose a music search engine that crawls audio blogs via RSS feeds and calculates TF·IDF vectors. Hu et al. [2005] extract TF-based features from music reviews gathered from `Epinions`.[8] Regarding different schemes of calculating TF·IDF weights, incorporating normalization strategies, aggregating data, and measuring similarity, [Schedl et al. 2011] give a comprehensive overview of the impact of these choices on the quality of artist similarity estimates by evaluating several thousand combinations of settings.

3.2 Collaborative Tags

As one of the characteristics of the so-called Web 2.0—where web sites encourage (even require) their users to participate in the generation of content—available items such as photos, films, or music can be labeled by the user community with tags. A tag can be virtually anything, but it usually consists of a short description of one aspect typical to the item (for music, for example, genre or style, instrumentation, mood, or performer). The more people who label an item with a tag, the more the tag is assumed to be relevant to the item. For music, the most prominent platform that makes use of this approach is `Last.fm`. Since `Last.fm` provides the collected tags in a standardized manner, it is a very valuable source for context-related information.

Geleijnse et al. [2007] use tags from `Last.fm` to generate a "tag ground truth" for artists. They filter redundant and noisy tags using the set of tags associated with tracks by the artist under consideration. Similarities between artists are then calculated via the number of overlapping tags. Evaluation against `Last.fm`'s similar artist function shows that the number of overlapping tags between similar artists is much larger than the average overlap between arbitrary artists (approximately 10 vs. 4 after filtering).

Levy and Sandler [2007] retrieve tags from `Last.fm` and `MusicStrands`, a web service (no longer in operation) that allows users to share playlists,[9] to construct a semantic space for music pieces. To this end, all tags found for a specific track are tokenized like normal text descriptions and a standard TF·IDF-based document-term matrix is created, that is, each track is represented by a term vector. For the TF factor, three different calculation methods are explored, namely weighting of the TF by the number of users that applied the tag, no further weighting, and restricting features to adjectives only. Optionally, the dimensionality of the vectors is reduced by applying Latent Semantic Analysis (LSA) [Deerwester et al. 1990]. The similarity between vectors is calculated via the cosine measure, cf. Eq. (6). For evaluation, for each genre or artist term, each track labeled with that term serves as query, and the mean average precision over all queries is calculated. It is shown that filtering for adjectives clearly worsens the performance of the approach and that weighting of term frequency by the number of users may improve genre precision (however, it is noted that this may just artificially emphasize the majority's opinion without really improving the features). Without LSA (i.e., using the full term vectors) genre precision reaches 80%, and artist precision 61%. Using LSA, genre precision reaches up to 82%, and artist precision 63%. The approach is also compared to the web-based term profile approach by Knees et al. [2004]—cf. Section 3.1. Using the full term vectors in a 1-NN leave-one-out cross validation setting, genre classification accuracy touches 95% without and 83% with artist filtering.

[8] `http://www.epinions.com`.

[9] `http://music.strands.com`.

Nanopoulos et al. [2010] extend the “two-dimensional” model of music items and tags by including the dimension of users. From this, a similar approach as in Levy and Sandler [2007] is taken by generalising the method of *singular value decomposition* (SVD) to higher dimensions.

In comparison to web-based term approaches, the tag-based approach exhibits some advantages, namely a more music-targeted and smaller vocabulary with significantly less noisy terms and availability of descriptors for individual tracks rather than just artists. Yet, tag-based approaches also suffer from some limitations. For example, sufficient tagging of comprehensive collections requires a large and active user community. Furthermore, tagging of tracks from the so-called “long tail”, that is, lesser known tracks, is usually very sparse. Additionally, also effects such as a “community bias” may be observed. To remedy some of these problems, recently, the idea of gathering tags via games has arisen [Turnbull et al. 2007; Mandel and Ellis 2007; Law et al. 2007]. Such games provide some form of incentive—be it just the pure joy of gaming—to the human player to solve problems that are hard to solve for computers, for example, capturing emotions evoked when listening to a song. By encouraging users to play such games, a large number of songs can be efficiently annotated with semantic descriptors. Another recent trend to alleviate the data sparsity problem and to allow fast indexing in a “semantic” space is automatic tagging and propagation of tags based on alternative data sources, foremost low-level audio features [Sordo et al. 2007; Eck et al. 2008; Kim et al. 2009; Shen et al. 2010; Zhao et al. 2010].

3.3 Song Lyrics

The lyrics of a song represent an important aspect of the semantics of music since they usually reveal information about the artist or the performer such as cultural background (via different languages or use of slang words), political orientation, or style of music (use of a specific vocabulary in certain music styles).

Logan et al. [2004] use song lyrics for tracks by 399 artists to determine artist similarity. In the first step, Probabilistic Latent Semantic Analysis (PLSA) [Hofmann 1999] is applied to a collection of over 40,000 song lyrics to extract N topics typical to lyrics. In the second step, all lyrics by an artist are processed using each of the extracted topic models to create N-dimensional vectors for which each dimension gives the likelihood of the artist’s tracks to belong to the corresponding topic. Artist vectors are then compared by calculating the L_1 distance (also known as Manhattan distance) as shown in Eq. (7)

$$dist_{L_1}(A_i, A_j) = \sum_{k=1}^{N} \left|a_{i,k} - a_{j,k}\right|. \tag{7}$$

This similarity approach is evaluated against human similarity judgments, that is, the “survey” data for the *uspop2002* set [Berenzweig et al. 2003], and yields worse results than similarity data obtained via acoustic features (irrespective of the chosen N, the usage of stemming, or the filtering of lyrics-specific stopwords). However, as lyrics-based and audio-based approaches make different errors, a combination of both is suggested.

Mahedero et al. [2005] demonstrate the usefulness of lyrics for four important tasks: language identification, structure extraction (i.e., recognition of intro, verse, chorus, bridge, outro, etc.), thematic categorization, and similarity measurement. For similarity calculation, a standard TF·IDF measure with cosine distance is proposed as an initial step. Using this information, a song’s representation is obtained by concatenating distances to all songs in the collection into a new vector. These representations are then compared using an unspecified algorithm. Exploratory experiments indicate some potential for cover version identification and plagiarism detection.

Other approaches do not explicitly aim at finding similar songs in terms of lyrical (or rather semantic) content, but at revealing conceptual clusters [Kleedorfer et al. 2008] or at classifying songs into genres [Mayer et al. 2008] or mood categories [Laurier et al. 2008; Hu et al. 2009]. Most of these approaches are nevertheless of interest in the context of this article, as the extracted features can also be used for similarity calculation. Laurier et al. [2008] strive to classify songs into four mood categories by means of lyrics and content analysis. For lyrics, the TF·IDF measure with cosine distance is incorporated. Optionally, LSA is also applied to the TF·IDF vectors (achieving best results when projecting vectors down to 30 dimensions). In both cases, a 10-fold cross validation with k-NN classification yielded accuracies slightly above 60%. Audio-based features performed better compared to lyrics features, however, a combination of both yielded best results. Hu et al. [2009] experiment with TF·IDF, TF, and Boolean vectors and investigate the impact of stemming, part-of-speech tagging, and function words for soft-categorization into 18 mood clusters. Best results are achieved with TF·IDF weights on stemmed terms. An interesting result is that in this scenario, lyrics-based features alone can outperform audio-based features. Beside TF·IDF and Part-of-Speech features, Mayer et al. [2008] also propose the use of rhyme and statistical features to improve lyrics-based genre classification. To extract rhyme features, lyrics are transcribed to a phonetic representation and searched for different patterns of rhyming lines (e.g., AA, AABB, ABAB). Features consist of the number of occurrences of each pattern, the percentage of rhyming blocks, and the fraction of unique terms used to build the rhymes. Statistical features are constructed by counting various punctuation characters and digits and calculating typical ratios like average words per line or average length of words. Classification experiments show that the proposed style features and a combination of style features and classical TF·IDF features outperform the TF·IDF-only-approach.

In summary, recent scholarly contribution demonstrates that many interesting aspects of context-based similarity can be covered by exploiting lyrics information. However, since new and ground breaking applications for this kind of information have not been discovered yet, the potential of lyrics analysis is currently mainly seen as a complementary source to content-based features for genre or mood classification.

4. CO-OCCURRENCE-BASED APPROACHES

Instead of constructing feature representations for musical entities, the work reviewed in this section follows an immediate approach to estimate similarity. In principle, the idea is that the occurrence of two music pieces or artists within the same context is considered to be an indication for some sort of similarity. As sources for this type of similarity we discuss web pages (and as an abstraction page counts returned by search engines), microblogs, playlists, and peer-to-peer (P2P) networks.

4.1 Web-Based Co-Occurrences and Page Counts

One aspect of a music entity's context is related web pages. Determining and using such music-related pages as data source for MIR tasks was probably first performed by Cohen and Fan [2000]. Cohen and Fan automatically extract lists of artist names from web pages. To determine pages relevant to the music domain, they query `Altavista`[10] and `Northern Light`.[11] The resulting HTML 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. This procedure allows for extracting co-occurring artist names which are then used for artist recommendation. This article reveals, unfortunately, only

[10] `http://www.altavista.com`.

[11] `Northern Light` (`http://www.northernlight.com`), formerly providing a meta search engine, in the meantime has specialized on search solutions tailored to enterprises.

a few details on the exact approach. As ground truth for evaluating their approach, Cohen and Fan exploit server logs of downloads from an internal digital music repository made available within AT&T's intranet. They analyze the network traffic for three months, yielding a total of 5,095 artist-related downloads.

Another sub-category of co-occurrence approaches does not actually retrieve co-occurrence information, but relies on page counts returned to search engine requests. Formulating a conjunctive query made of two artist names and retrieving the page count estimate from a search engine can be considered an abstraction of the standard approach to co-occurrence analysis.

Into this category fall Zadel and Fujinaga [2004], who investigate the usability of two web services to extract co-occurrence information and consecutively derive artist similarity. More precisely, the authors propose an approach that, given a seed artist as input, retrieves a list of potentially related artists from the Amazon web service Listmania!. Based on this list, artist co-occurrences are derived by querying the Google Web API[12] and storing the returned page counts of artist-specific queries. Google is queried for "artist name i" and for "artist name i"+"artist name j". Thereafter, the so-called "relatedness" of each Listmania! artist to the seed artist is calculated as the ratio between the combined page count, that is, the number of web pages on which both artists co-occur, and the minimum of the single page counts of both artists, cf. Eq. (8). The minimum is used to account for different popularities of the two artists.

$$sim_{pc\_min}(A_i, A_j) = \frac{pc(A_i, A_j)}{\min(pc(A_i), pc(A_j))}. \tag{8}$$

Recursively extracting artists from Listmania! and estimating their relatedness to the seed artist via Google page counts allows the construction of lists of similar artists. Although the paper shows that web services can be efficiently used to find artists similar to a seed artist, it lacks a thorough evaluation of the results.

Analyzing Google page counts as a result of artist-related queries is also performed in Schedl et al. [2005]. Unlike the method presented in Zadel and Fujinaga [2004], Schedl et al. [2005] derive complete similarity matrices from artist co-occurrences. This offers additional information since it can also predict which artists are not similar.

Schedl et al. [2005] define the similarity of two artists as the conditional probability that one artist is found on a web page that mentions the other artist. Since the retrieved page counts for queries like "artist name i" or "artist name i"+"artist name j" indicate the relative frequencies of this event, they are used to estimate the conditional probability. Equation (9) gives a formal representation of the symmetrized similarity function

$$sim_{pc\_cp}(A_i, A_j) = \frac{1}{2} \cdot \left(\frac{pc(A_i, A_j)}{pc(A_i)} + \frac{pc(A_i, A_j)}{pc(A_j)} \right). \tag{9}$$

In order to restrict the search to web pages relevant to music, different query schemes are used in Schedl et al. [2005] (cf. Section 3.1). Otherwise, queries for artists whose names have another meaning outside the music context, such as "Kiss," would unjustifiably lead to higher page counts, hence distorting the similarity relations.

Schedl et al. [2005] perform two evaluation experiments on the same 224-artist-data-set as used in Knees et al. [2004]. They estimat the homogeneity of the genres defined by the ground truth by applying the similarity function to artists within the same genre and to artists from different genres. To this end, the authors relate the average similarity between two arbitrary artists from the same genre

[12] Google no longer offers this Web API. It has been replaced by several other APIs, mostly devoted to Web 2.0 development.

to the average similarity of two artists from different genres. The results show that the co-occurrence approach can be used to clearly distinguish between most of the genres. The second evaluation experiment is an artist-to-genre classification task using a k-NN classifier. In this setting, the approach yields in the best case (when combining different query schemes) an accuracy of about 85% averaged over all genres.

A severe shortcoming of the approaches proposed in Zadel and Fujinaga [2004] and Schedl et al. [2005] is that they require a number of search engine requests that is quadratic in the number of artists, for creating a complete similarity matrix. These approaches therefore scale poorly to real-world music collections.

Avoiding the quadratic computational complexity can be achieved with the alternative strategy to co-occurrence analysis described in Schedl [2008, Chapter 3]. This method resembles Cohen and Fan [2000], presented in the beginning of this section. First, for each artist A_i, a certain amount of top-ranked web pages returned by the search engine is retrieved. Subsequently, all pages fetched for artist A_i are searched for occurrences of all other artist names A_j in the collection. The number of page hits represents a co-occurrence count, which equals the document frequency of the artist term "A_j" in the corpus given by the web pages for artist A_i. Relating this count to the total number of pages successfully fetched for artist A_i, a similarity function is constructed. Employing this method, the number of issued queries grows linearly with the number of artists in the collection. The formula for the symmetric artist similarity equals Eq. (11).

4.2 Microblogs

The use of microblogging services, `Twitter`[13] in particular, has considerably increased during the past few years. Since many users share their music listening habits via `Twitter`, it provides a valuable data source for inferring music similarity as perceived by the Twittersphere. Thanks to the restriction of tweets to 140 characters, text processing can be performed in little time, compared to web pages. On the downside, microbloggers might not represent the average person, which potentially introduces a certain bias in approaches that make use of this data source.

Exploiting microblogs to infer similarity between artists or songs is a very recent endeavor. Two quite similar methods that approach the problem are presented in Zangerle et al. [2012] and Schedl and Hauger [2012]. Both make use of `Twitter`'s streaming API[14] and filter incoming tweets for hashtags frequently used to indicate music listening events, such as *#nowplaying*. The filtered tweets are then sought for occurrences of artist and song names, using the `MusicBrainz` data base. Microblogs that can be matched to artists or songs are subsequently aggregated for each user, yielding individual listening histories. Applying co-occurrence analysis to the listening history of each user, a similarity measure is defined in which artists/songs that are frequently listened to by the same user are treated as similar. Zangerle et al. [2012] use absolute numbers of co-occurrences between songs to approximate similarities, while Schedl and Hauger [2012] investigate various normalization techniques to account for different artist popularity and different levels of user listening activity. Using as ground truth similarity relations gathered from `Last.fm` and running a standard retrieval experiment, Schedl and Hauger identify as best performing measure (both in terms of precision and recall) the one given in Eq. (10), where $cooc(A_i, A_j)$ represents the number of co-occurrences in the listening histories of same users, and $oc(A_i)$ denotes the total number of occurrences of artist A_i in all listening histories

$$sim_{tw\_cooc}(A_i, A_j) = \frac{cooc(A_i, A_j)}{\sqrt{oc(A_i) \cdot oc(A_j)}}. \tag{10}$$

[13] `http://www.twitter.com`.

[14] `https://dev.twitter.com/docs/streaming-apis`.

4.3 Playlists

An early approach to derive similarity information from the context of a music entity can be found in Pachet et al. [2001], in which radio station playlists (extracted from a French radio station) and compilation CD databases (using CDDB[15]) are exploited to extract co-occurrences between tracks and between artists. The authors count the number of co-occurrences of two artists (or pieces of music) A_i and A_j on the radio station playlists and compilation CDs. They define the co-occurrence of an entity A_i to itself as the number of occurrences of T_i in the considered corpus. Accounting for different frequencies, that is, popularity of a song or an artist, is performed by normalizing the co-occurrences. Furthermore, assuming that co-occurrence is a symmetric function, the complete co-occurrence-based similarity measure used by the authors is given in Eq. (11)

$$sim_{pl\_cooc}(A_i, A_j) = \frac{1}{2} \cdot \left[\frac{cooc(A_i, A_j)}{oc(A_i, A_i)} + \frac{cooc(A_j, A_i)}{oc(A_j, A_j)} \right]. \quad (11)$$

However, this similarity measure can not capture indirect links that an entity may have with others. In order to capture such indirect links, the complete co-occurrence vectors of two entities A_1 and A_2 (i.e., a vector that gives, for a specific entity, the co-occurrence count with all other entities in the corpus) are considered and their statistical correlation is computed via Pearson's correlation coefficient shown in Eq. (12)

$$sim_{pl\_corr}(A_i, A_j) = \frac{Cov(A_i, A_j)}{\sqrt{Cov(A_i, A_i) \cdot Cov(A_j, A_j)}}. \quad (12)$$

These co-occurrence and correlation functions are used as similarity measures on the track level and on the artist level. Pachet et al. [2001] evaluate them on rather small data sets (a set of 12 tracks and a set of 100 artists) using similarity judgments by music experts from Sony Music as ground truth. The main finding is that artists or tracks that appear consecutively in radio station playlists or on CD samplers indeed show a high similarity. The co-occurrence function generally performs better than the correlation function (70%–76% vs. 53%–59% agreement with ground truth).

Another work that uses playlists in the context of music similarity estimation is Cano and Koppenberger. Cano and Koppenberger [2004] create a similarity network via extracting playlist co-occurrences of more than 48,000 artists retrieved from Art of the Mix[16] in early 2003. Art of the Mix is a web service that allows users to upload and share their mix tapes or playlists. The authors analyze a total of more than 29,000 playlists. They subsequently create a similarity network where a connection between two artists is made if they co-occur in a playlist.

A more recent paper that exploits playlists to derive artist similarity information [Baccigalupo et al. 2008] analyses co-occurrences of artists in playlists shared by members of a web community. The authors look at more than 1 million playlists made publicly available by MusicStrands. They extract from the whole playlist set the 4,000 most popular artists, measuring the popularity as the number of playlists in which each artist occurred. Baccigalupo et al. [2008] further take into account that two artists that consecutively occur in a playlist are probably more similar than two artists that occur farther away in a playlist. To this end, the authors define a distance function $d_h(A_i, A_j)$ that counts how often a song by artist A_i co-occurs with a song by A_j at a distance of h. Thus, h is a parameter that defines the number of songs in between the occurrence of a song by A_i and the occurrence of a

[15] CDDB is a web-based album identification service that returns, for a given unique disc identifier, metadata like artist and album name, tracklist, or release year. This service is offered in a commercial version operated by Gracenote (http://www.gracenote.com) as well as in an open source implementation named freeDB (http://www.freedb.org).

[16] http://www.artofthemix.org.

song by A_j in the same playlist. Baccigalupo et al. [2008] define the distance between two artists A_i and A_j as in Eq. (13), where the playlist counts at distances 0 (two consecutive songs by artists A_i and A_j), 1, and 2 are weighted with β_0, β_1, and β_2, respectively. The authors empirically set the values to $\beta_0 = 1,\ \beta_1 = 0.8,\ \beta_2 = 0.64$

$$dist_{pl\_d}(A_i, A_j) = \sum_{h=0}^{2} \beta_h \cdot [d_h(A_i, A_j) + d_h(A_j, A_i)]. \tag{13}$$

To account for the popularity bias, that is, very popular artists co-occurring with a lot of other artists in many playlists and creating a higher similarity to all other artists when simply relying on Eq. (13), the authors perform normalization according to Eq. (14), where $\widehat{dist_{pl\_d}}(A_i)$ denotes the average distance between A_i and all other artists, that is, $\frac{1}{n-1} \cdot \sum_{j \in X} dist_{pl\_d}(A_i, A_j)$, and X the set of $n-1$ artists other than A_i

$$dist_{|pl\_d|}(A_i, A_j) = \frac{dist_{pl\_d}(A_i, A_j) - \widehat{dist_{pl\_d}}(A_i)}{\left|max\left(dist_{pl\_d}(A_i, A_j) - \widehat{dist_{pl\_d}}(A_i)\right)\right|}. \tag{14}$$

Unfortunately, no evaluation dedicated to artist similarity is conducted.

Aizenberg et al. [2012] apply collaborative filtering methods (cf. Section 5) to the playlists of 4,147 radio stations associated with the web radio station directory ShoutCast[17] collected over a period of 15 days. Their goals are to give music recommendations, to predict existing radio station programs, and to predict the programs of new radio stations. To this end, they model latent factor station affinities as well as temporal effects by maximizing the likelihood of a multinomial distribution.

Chen et al. [2012] model the sequential aspects of playlists via Markov chains and learn to embed the occurring songs as points in a latent multidimensional Euclidean space. The resulting generative model is used for playlist prediction by finding paths that connect points. Although the authors only aim at generating new playlists, the learned projection could also serve as a space for Euclidean similarity calculation between songs.

4.4 Peer-to-Peer Network Co-Occurrences

Peer-to-peer (P2P) networks represent a rich source for mining music-related data since their users are commonly willing to reveal various kinds of metadata about the shared content. In the case of shared music files, file names and ID3 tags are usually disclosed.

Early work that makes use of data extracted from P2P networks comprises of Whitman and Lawrence [2002], Ellis et al. [2002], Logan et al. [2003], and Berenzweig et al. [2003]. All of these papers use, among other sources, data extracted from the P2P network OpenNap to derive music similarity information. Although it is unclear whether the four publications make use of exactly the same data set, the respective authors all state that they extracted metadata, but did not download any files, from OpenNap. Logan et al. [2003] and Berenzweig et al. [2003] report having determined the 400 most popular artists on OpenNap from mid 2002. The authors gather metadata on shared content, which yields about 175,000 user-to-artist relations from about 3,200 shared music collections. Logan et al. [2003] especially highlight the sparsity in the OpenNap data, in comparison with data extracted from the audio signal. Although this is obviously true, the authors miss noting the inherent disadvantage of signal-based feature extraction, that extracting signal-based features is only possible when the audio content is available. Logan et al. [2003] then compare similarities defined by artist co-occurrences in OpenNap collections, expert opinions from AMG, playlist co-occurrences from Art of the Mix, data

[17] http://www.shoutcast.com.

gathered from a web survey, and audio feature extraction via MFCCs, for example, Aucouturier et al. [2005]. To this end, they calculate a "ranking agreement score", which basically compares the top N most similar artists according to each data source and calculates the pair-wise overlap between the sources. The main findings are that the co-occurrence data from OpenNap and from Art of the Mix show a high degree of overlap, the experts from AMG and the participants of the web survey show a moderate agreement, and the signal-based measure has a rather low agreement with all other sources (except when compared with the AMG data).

Whitman and Lawrence [2002] use a software agent to retrieve from OpenNap a total of 1.6 million user-song entries over a period of three weeks in August 2001. To alleviate the popularity bias of the data, Whitman and Lawrence [2002] use a similarity measure as shown in Eq. (15), where $C(A_i)$ denotes the number of users that share songs by artist A_i, $C(A_i, C_j)$ is the number of users that have both artists A_i and A_j in their shared collection, and A_k is the most popular artist in the corpus. The right term in the equation downweights the similarity between two artists if one of them is very popular and the other not

$$sim_{p2p\_wl}(A_i, A_j) = \frac{C(A_i, A_j)}{C(A_j)} \cdot \left(1 - \frac{\left|C(A_i) - C(A_j)\right|}{C(A_k)}\right). \tag{15}$$

Ellis et al. [2002] use the same artist set as Whitman and Lawrence [2002]. Their aim is to build a ground truth for artist similarity estimation. They report extracting from OpenNap about 400,000 user-to-song relations and covering about 3,000 unique artists. Again, the co-occurrence data is compared with artist similarity data gathered by a web survey and with AMG data. In contrast to Whitman and Lawrence [2002], Ellis et al. [2002] take indirect links in AMG's similarity judgments into account. To this end, Ellis et al. propose a transitive similarity function on similar artists from the AMG data, which they call "Erdös distance". More precisely, the distance $d(A_1, A_2)$ between two artists A_1 and A_2 is measured as the minimum number of intermediate artists needed to form a path from A_1 to A_2. As this procedure also allows deriving information on dissimilar artists (those with a high minimum path length), it can be employed to obtain a complete distance matrix. Furthermore, the authors propose an adapted distance measure, the so-called "Resistive Erdös measure", which takes into account that there may exist more than one shortest path of length l between A_1 and A_2. Assuming that two artists are more similar if they are connected via many different paths of length l, the Resistive Erdös similarity measure equals the electrical resistance in a network (cf. Eq. (16)) in which each path from A_i to A_j is modeled as a resistor whose resistance equals the path length $|p|$. However, this adjustment does not improve the agreement of the similarity measure with the data from the web-based survey, as it fails to overcome the popularity bias, in other words, that many different paths between popular artists unjustifiably lower the total resistance

$$dist_{p2p\_res}(A_i, A_j) = \left(\sum_{p \in Paths(A_i, A_j)} \frac{1}{|p|}\right)^{-1}. \tag{16}$$

A recent approach that derives similarity information on the artist and on the song level from the Gnutella P2P file sharing network is presented in Shavitt and Weinsberg [2009]. They collect metadata of shared files from more than 1.2 million Gnutella users in November 2007. Shavitt and Weinsberg restrict their search to music files (.mp3 and .wav), yielding a data set of 530,000 songs. Information on both users and songs are then represented via a 2-mode graph showing users and songs. A link between a song and a user is created when the user shares the song. One finding of analyzing the resulting network is that most users in the P2P network shared similar files. The authors use the data gathered for artist recommendation. To this end, they construct a user-to-artist matrix V, where $V(i, j)$ gives the number of songs by artist A_j that user U_i shares. Shavitt and Weinsberg then perform direct clustering on V using the k-means algorithm [MacQueen 1967] with the Euclidean distance metric. Artist recommendation is then performed using either data from the centroid of the cluster to which the seed user U_i belongs or by using the nearest neighbors of U_i within the cluster to which U_i belongs.

In addition, Shavitt and Weinsberg also address the problem of song clustering. Accounting for the popularity bias, the authors define a distance function that is normalized according to song popularity, as shown in Eq. (17), in which $uc(S_i, S_j)$ denotes the total number of users that share songs S_i and S_j, and C_i and C_j denote, respectively, the popularity of songs S_i and S_j, measured as their total occurrence in the corpus.

$$dist_{p2p\_pop}(S_i, S_j) = -log_2 \left(\frac{uc(S_i, S_j)}{\sqrt{C_i \cdot C_j}} \right) \tag{17}$$

Evaluation experiments are carried out for song clustering. The authors report an average precision of 12.1% and an average recall of 12.7%, which they judge as quite good considering the vast amount of songs shared by the users and the inconsistency in the metadata (ID3 tags).

5. USER RATING-BASED APPROACHES

Another source from which to derive contextual similarity is explicit user feedback. Approaches utilizing this source are also known as *collaborative filtering* (CF). To perform this type of similarity estimation typically applied in recommender systems, one must have access to a (large and active) community and its activities. Thus, CF methods are often to be found in real-world (music) recommendation systems such as Last.fm or Amazon.[18] [Celma 2008] provides a detailed discussion of CF for music recommendation in the long-tail with real-world examples from the music domain.

In their simplest form, CF systems exploit two types of similarity relations that can be inferred by tracking users' habits: item-to-item similarity (where an item could potentially be a track, an artist, a book, etc.) and user-to-user similarity. For example, when representing preferences in a user-item matrix S, where $S_{i,j} > 0$ indicates that user j likes item i (e.g., j has listened to artist i at least once or j has bought product i), $S_{i,j} < 0$ that j dislikes i (e.g., j has skipped track i while listening or j has rated product i negatively), and $S_{i,j} = 0$ that there is no information available (or neutral opinion), user-to-user similarity can be calculated by comparing the corresponding M-dimensional column vectors (where M is the total number of items), whereas item-to-item similarity can be obtained by comparing the respective N-dimensional row vectors (where N is the total number of users) [Linden et al. 2003; Sarwar et al. 2001]. For vector comparison, cosine similarity (see Eq. (6)) and Pearson's correlation coefficient (Eq. (12)) are popular choices. For example, Slaney and White [2007] analyze 1.5 million user ratings by 380,000 users from the Yahoo! music service[19] and obtain music piece similarity by cosine comparing normalized rating vectors over items.

As can be seen from this formulation, in contrast to the text and co-occurrence approaches reviewed in Sections 3 and 4, respectively, CF does not require any additional metadata describing the music items. Due to the nature of rating and feedback matrices, similarities can be calculated without the need to associate occurrences of metadata with actual items. Furthermore, CF approaches are largely domain independent and also allow for similarity computation across domains. However, these simple approaches are very sensitive to factors such as popularity biases and data sparsity. Especially for

[18] http://www.amazon.com.

[19] http://music.yahoo.com.

items with very few ratings, recommendations performed in a fashion as outlined are not very reliable. However, when aiming at recommending items, it is often more desirable to directly predict user ratings instead of calculating item similarity. This is done either by estimating ratings from similar items (or users) as described above or by applying regression models, in particular by utilizing matrix factorization techniques, cf. Bell and Koren [2007]. As can be seen from the results of the KDD-Cup 2011 competition,[20] matrix factorization methods are able to substantially improve prediction accuracy in the music domain, cf. Dror et al. [2011b].

In another paper, Dror et al. [2011a] show that matrix factorization models for music recommendation can be easily extended to incorporate additional information such as temporal dynamics in listening behavior, temporal dynamics in item histories, and multi-level taxonomy information like genre. Yang et al. [2012] further address temporal aspects on a smaller and more consistent time scale. These "local preferences", as the authors call them, reflect changes in listening behavior that are strongly influenced by the listening context and occurring events rather than caused by a gradual change in general taste over time. In addition, Yang et al. [2012] propose an efficient training algorithm to find optimal parameters and explore different settings of time granularity.

Mesnage et al. [2011] strive to build a social music recommender system by investigating different strategies with a dedicated Facebook[21] application. More precisely, they compare recommendations based on friend relationships, random other users, and random recommendations.

6. OTHER SOURCES OF CONTEXTUAL MUSIC SIMILARITY

While the preceding sections cover the currently best established and most prominent sources to derive context-based music information, there also exist alternative, implicit data sources that are considered in research, for instance, as previously mentioned, *server logs*. Cohen and Fan [2000] analyze download statistics from an AT&T internal digital music repository which is monitored for three months, yielding a total of nearly 30,000 music downloads relating to about 1,000 different artists. Conceptionally, this technique is very similar to exploiting P2P network co-occurrences (cf. Section 4.4).

Fields et al. [2008] propose the usage of artist-related social network data from MySpace.[22] The authors state that similarity information based on the artists' "top friends" seems to be a promising complement to signal-based audio similarity. Fields et al. [2008] further propose to combine these similarities for the task of automatic playlist generation. MySpace is also exploited in Jacobson et al. [2008], in which the authors mine the inherent artist network to detect "communities" of similar artists. These are further shown to exhibit structures closely related to musical genre.

Music-related data also includes images such as band photographs or album artwork. Schedl et al. [2006] present methods to automatically retrieve the correct album cover for a record. Brochu et al. [2003] use color histogram representations of album covers to index music pieces. Lībeks and Turnbull [2011] calculate music similarity based on artists' promotional photographs. It is shown that the notions of similarity and genre to some extent correspond to visual appearance.

McFee and Lanckriet [2009] use different data sources, among which artist biographies stand out. These are gathered from Last.fm, stopwords removed, and a standard vector space model based on TF·IDF weighting and cosine similarity created in order to measure similarities between music artists. These TF·IDF features computed from biographies (and in addition from collaborative tags) are combined with acoustic features (MFCCs and chroma). McFee and Lanckriet [2009] then apply Partial

[20] The goal in this competition was to predict music ratings based on explicit ratings to songs, albums, artists, and genres; see http://www.sigkdd.org/kdd2011/kddcup.shtml.

[21] http://www.facebook.com.

[22] http://www.myspace.com.

Table II. Overview of Different Context-based Sources

	Tags	Web-terms	Lyrics	Co-occ.
Source	web service	web pages	lyrics portals	web (search)
Community Req.	yes	depends	no	no
Level	artist, track	artist	track (artist)	artist
Feature Dim.	moderate	very high	possibly high	item×item
Specific Bias	community	low	none	low
Potential Noise	moderate	high	low	high

	Microblogs	Playlists	P2P	Ratings
Source	API, feeds	radio, CDs, web	shared folders	users
Community Req.	yes	depends	yes	yes
Level	artist (track)	artist, track	artist, track	all
Feature Dim.	item×item	item×item	item×item	user×item
Specific Bias	community	low	community	community
Potential Noise	high	low	high	yes

Order Embedding (POE) to learn a low-dimensional embedding of the high-dimensional feature space, which in turn is tuned to match subjective human similarity judgments. The resulting similarity measure is suggested for music artist recommendation and similarity visualization.

7. DISCUSSION AND OUTLOOK

Exploiting the wealth of information provided by the "power of the crowd" is key to next-generation music services and personalized recommendations. On the other hand, the abundance of newly available and semantically meaningful information offered on Web 2.0 platforms and the like also poses new challenges, such as dealing with the large quantity and noisiness of the data, various user biases, hacking, or the cold start problem. Furthermore, with mobile devices and omnipresent web technology, music access and interaction modalities have undergone a fundamental change, requesting innovative methods to exploit contextual data for music retrieval.

In this survey article, we have given an overview of approaches to estimate music similarity which do not rely on the audio signal, but rather take into consideration various aspects of the context in which a music entity occurs. To sum up the reviewed sources of music context, we give a brief comparison of some of their key properties in Table II. To this end, we have revisited these approaches and their authors' conclusions and discussions. This allows us to contrast advantages and disadvantages with respect to important aspects such as complexity or level of contained noise.

The first row in Table II indicates the source or channel from which the respective features can be gathered. The second row shows whether a community is required to create the data set. Community here refers to a specific and dedicated subgroup of Internet users. This also includes customers of an online store, as in the case of ratings. Thus, approaches based on lyrics or general web pages and web search engines are not considered to be community dependent. In the third row, we analyze whether the source can be used to infer similarity on the level of artists or tracks. While lyrics are given on the song level, they are also used to build aggregate models on the artist level. For microblogs, good results can be achieved when modeling similarity on the artist level; however, song level experiments have also been conducted [Zangerle et al. 2012]. The forth row indicates the dimensionality of the features from which similarities are computed and hence reflects the complexity of the respective approaches. While the dimensionality of text-based approaches refers to the number of different terms selected, co-occurrence-based approaches in general operate on similarity matrices that grow quadratically with the number of music items. For methods based on page counts from search engines, this also entails the requirement of invoking the search engine for every pair of items. Rating-based strategies operate on user-rating-matrices, thus their complexity depends also on the number of users. The penultimate row in Table II indicates whether a data source is prone to a specific bias and the last row shows the sources' susceptibility to noise.

Even though the presented context-based approaches illustrate the great potential of comprehensive community data, nearly all of them suffer from similar shortcomings. First, *data sparsity*, especially for artists in the "long tail", is a problem. Second, the *popularity bias* has to be addressed, that is, that disproportionately more data is available for popular artists than for lesser known ones, which often distorts derived similarity measures. Furthermore, methods that aim at taking advantage of user-generated data are prone to include only participants of existing communities in a broad sense (from very specific services, like a certain P2P network, to the web community as a whole). It is further known that users of certain communities tend to have similar music tastes. In general, this phenomenon is known as *community* or *population bias*.

Apart from data sparsity and biases, a general challenge common to all metadata-based methods is the correct matching of identifiers to the actual entity. Reports about issues with what would initially often be considered a side task can be found with approaches of all kinds. Frequent problems are inconsistent spelling and naming, band and artist nicknames, different versions of songs, and—especially in microblogs—abbreviations. However, it is understood that user-generated data will always exhibit such artefacts and that dealing with non-standardized input is one of the central challenges when exploring contextual data.

For the future, we believe that it is crucial to transcend the idea of a generally valid notion of similarity and establish a differentiated, multi-granular concept of similarity. Such a concept should take into account regional particularities and views and further adapt to cultural areas as well as to individuals [Schedl et al. 2012]. This need becomes particularly apparent when comparing current representations of Western and non-Western music [Serra 2012]. Furthermore, we think that multi-faceted similarity measures will soon be standard in music applications. They may be defined as a mixture of content- and context-based aspects, for example, to enable retrieval systems capable of dealing with queries like "give me rhythmically similar music to the most recent chart hit in Canada, but from the 1970s".

Research in contextual music similarity and recommendation would further benefit from a systematic evaluation of the presented approaches on unified datasets. Such a comprehensive evaluation will profoundly reveal potentials and shortcomings of each method and will likely yield insights on how to combine different data sources in order to improve the quality of MIR systems.

REFERENCES

AIZENBERG, N., KOREN, Y., AND SOMEKH, O. 2012. Build your own music recommender by modeling internet radio streams. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, New York, 1–10.

AUCOUTURIER, J.-J. AND PACHET, F. 2002. Scaling up music playlist generation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02)*. 105–108.

AUCOUTURIER, J.-J., PACHET, F., ROY, P., AND BEURIVÉ, A. 2007. Signal + Context = Better Classification. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

AUCOUTURIER, J.-J., PACHET, F., AND SANDLER, M. 2005. "The Way It Sounds": Timbre models for analysis and retrieval of music signals. *IEEE Trans. Multimed. 7,* 6, 1028–1035.

BACCIGALUPO, C., PLAZA, E., AND DONALDSON, J. 2008. Uncovering affinity of artists to multiple genres from social behaviour data. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*.

BARRINGTON, L., TURNBULL, D., YAZDANI, M., AND LANCKRIET, G. 2009. Combining audio content and social context for semantic music discovery. In *Proceedings of the 32nd ACM SIGIR*.

BAUMANN, S. AND HUMMEL, O. 2003. Using cultural metadata for artist recommendation. In *Proceedings of the 3rd International Conference on Web Delivering of Music (WEDELMUSIC'03)*.

BELL, R. M. AND KOREN, Y. 2007. Lessons from the Netflix prize challenge. *SIGKDD Explorat. 9,* 2, 75–79.

BERENZWEIG, A., LOGAN, B., ELLIS, D. P., AND WHITMAN, B. 2003. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR'03)*.

BRILL, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*. 152–155.

BROCHU, E., DE FREITAS, N., AND BAO, K. 2003. The sound of an album cover: Probabilistic multimedia and IR. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*.

CANO, P. AND KOPPENBERGER, M. 2004. The emergence of complex network patterns in music artist networks. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*. 466–469.

CASEY, M. A., VELTKAMP, R., GOTO, M., LEMAN, M., RHODES, C., AND SLANEY, M. 2008. Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE 96*, 668–696.

CELMA, O. 2008. Music recommendation and discovery in the long tail. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain.

CELMA, O., CANO, P., AND HERRERA, P. 2006. SearchSounds: An audio crawler focused on weblogs. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR'06)*.

CELMA, O. AND LAMERE, P. 2007. ISMIR 2007 tutorial: Music recommendation. `http://www.dtic.upf.edu/~ocelma/MusicRecommendationTutorial-ISMIR2007/` (last accessed: April 2013).

CHARNIAK, E. 1997. Statistical techniques for natural language parsing. *AI Magazine 18*, 33–44.

CHEN, S., MOORE, J., TURNBULL, D., AND JOACHIMS, T. 2012. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 714–722.

COHEN, W. W. AND FAN, W. 2000. Web-collaborative filtering: Recommending music by crawling the web. *Computer Netw. 33,* 1–6, 685–698.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. ASIS 41*, 391–407.

DROR, G., KOENIGSTEIN, N., AND KOREN, Y. 2011a. Yahoo! music recommendations: Modeling music ratings with temporal dynamics and item taxonomy. In *Proceedings of the 5th ACM Conference on Recommender Systems*. ACM. 165–172.

DROR, G., KOENIGSTEIN, N., KOREN, Y., AND WEIMER, M. 2011b. The Yahoo! music dataset and KDD-Cup'11. *J. Mach. Learn. Res. 8*, 3–18.

ECK, D., LAMERE, P., BERTIN-MAHIEUX, T., AND GREEN, S. 2008. Automatic generation of social tags for music recommendation. In *Advances in Neural Information Processing Systems 20 (NIPS'07)*. MIT Press.

ELLIS, D. P., WHITMAN, B., BERENZWEIG, A., AND LAWRENCE, S. 2002. The quest for ground truth in musical artist similarity. In *Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR'02)*.

FIELDS, B., CASEY, M., JACOBSON, K., AND SANDLER, M. 2008. Do you sound like your friends? Exploring artist similarity via artist social network relationships and audio signal processing. In *Proceedings of the International Computer Music Conference (ICMC'08)*.

GELEIJNSE, G., SCHEDL, M., AND KNEES, P. 2007. The quest for ground truth in musical artist tagging in the social web era. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

HOFMANN, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*.

HU, X., DOWNIE, J. S., AND EHMANN, A. F. 2009. Lyric text mining in music mood classification. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR'09)*.

HU, X., DOWNIE, J. S., WEST, K., AND EHMANN, A. 2005. Mining music reviews: Promising preliminary results. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*.

JACOBSON, K., FIELDS, B., AND SANDLER, M. 2008. Using audio analysis and network structure to identify communities in on-line social networks of artists. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*.

KIM, J. H., TOMASIK, B., AND TURNBULL, D. 2009. Using artist similarity to propagate semantic information. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR'09)*.

KLEEDORFER, F., KNEES, P., AND POHLE, T. 2008. Oh Oh Oh Whoah! Towards automatic topic detection in song lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*. 287–292.

KNEES, P., PAMPALK, E., AND WIDMER, G. 2004. Artist classification with web-based data. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*. 517–524.

KNEES, P., POHLE, T., SCHEDL, M., AND WIDMER, G. 2006a. Combining audio-based similarity with web-based data to accelerate automatic music playlist generation. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'06)* (Santa Barbara, CA).

KNEES, P., POHLE, T., SCHEDL, M., AND WIDMER, G. 2007. A music search engine built upon audio-based and web-based similarity measures. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*.

KNEES, P., SCHEDL, M., AND POHLE, T. 2008. A deeper look into web-based classification of music artists. In *Proceedings of 2nd Workshop on Learning the Semantics of Audio Signals (LSAS'08)*.

KNEES, P., SCHEDL, M., POHLE, T., AND WIDMER, G. 2006b. An innovative three-dimensional user interface for exploring music collections enriched with meta-information from the web. In *Proceedings of the 14th ACM International Conference on Multimedia (MM'06)*.

LAURIER, C., GRIVOLLA, J., AND HERRERA, P. 2008. Multimodal music mood classification using audio and lyrics. In *Proceedings of the International Conference on Machine Learning and Applications*.

LAW, E., VON AHN, L., DANNENBERG, R., AND CRAWFORD, M. 2007. Tagatune: A game for music and sound annotation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

LEVY, M. AND SANDLER, M. 2007. A semantic space for music derived from social tags. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

LĪBEKS, J. AND TURNBULL, D. 2011. You can judge an artist by an album cover: Using images for music annotation. *IEEE MultiMedia, 18*, 4, 30–37.

LINDEN, G., SMITH, B., AND YORK, J. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput. 4,* 1.

LOGAN, B., ELLIS, D. P., AND BERENZWEIG, A. 2003. Toward evaluation techniques for music similarity. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03): Workshop on the Evaluation of Music Information Retrieval Systems.*

LOGAN, B., KOSITSKY, A., AND MORENO, P. 2004. Semantic analysis of song lyrics. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'04)*.

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., Statistics Series, vol. I, University of California Press, Berkeley and Los Angeles, CA, 281–297.

MAHEDERO, J. P. G., MARTÍNEZ, A., CANO, P., KOPPENBERGER, M., AND GOUYON, F. 2005. Natural language processing of lyrics. In *Proceedings of the 13th ACM International Conference on Multimedia (MM'05)*. 475–478.

MANDEL, M. I. AND ELLIS, D. P. 2007. A web-based game for collecting music metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

MAYER, R., NEUMAYER, R., AND RAUBER, A. 2008. Rhyme and style features for musical genre classification by song lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*.

MCFEE, B. AND LANCKRIET, G. 2009. Heterogeneous embedding for subjective artist similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR'09)*.

MESNAGE, C. S., RAFIQ, A., DIXON, S., AND BRIXTEL, R. P. 2011. Music discovery with social networks. In *Proceedings of the 2nd Workshop on Music Recommendation and Discovery (WOMRAD'11)*. 7–12.

NANOPOULOS, A., RAFAILIDIS, D., SYMEONIDIS, P., AND MANOLOPOULOS, Y. 2010. Musicbox: Personalized music recommendation based on cubic analysis of social tags. *IEEE Trans. Audio, Speech, Lang. Process. 18,* 2, 407–412.

PACHET, F., WESTERMAN, G., AND LAIGRE, D. 2001. Musical data mining for electronic music distribution. In *Proceedings of the 1st International Conference on Web Delivering of Music (WEDELMUSIC'01)*.

PAMPALK, E., FLEXER, A., AND WIDMER, G. 2005. Hierarchical organization and description of music collections at the artist level. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'05)*.

PAMPALK, E. AND GOTO, M. 2007. MusicSun: A new approach to artist recommendation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

POHLE, T., KNEES, P., SCHEDL, M., PAMPALK, E., AND WIDMER, G. 2007. "Reinventing the Wheel": A novel approach to music player interfaces. *IEEE Trans. Multimed. 9*, 567–575.

POHLE, T., KNEES, P., SCHEDL, M., AND WIDMER, G. 2007. Building an interactive next-generation artist recommender based on automatically derived high-level concepts. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI'07)*.

SARWAR, B., KARYPIS, G., KONSTAN, J., AND REIDL, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of 10th International Conference on World Wide Web (WWW'01)*. 285–295.

SCHEDL, M. 2008. Automatically extracting, analyzing, and visualizing information on music artists from the world wide web. Ph.D. thesis, Johannes Kepler University, Linz, Austria.

SCHEDL, M. AND HAUGER, D. 2012. Mining microblogs to infer music artist similarity and cultural listening patterns. In *Proceedings of the 21st International World Wide Web Conference (WWW'12): 4th International Workshop on Advances in Music Information Research (AdMIRe'12)*.

SCHEDL, M., HAUGER, D., AND SCHNITZER, D. 2012. A model for serendipitous music retrieval. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI'12): 2nd International Workshop on Context-awareness in Retrieval and Recommendation (CaRR'12)*.

SCHEDL, M. AND KNEES, P. 2009. Context-based music similarity estimation. In *Proceedings of the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS'09)*.

SCHEDL, M. AND KNEES, P. 2011. Personalization in multimodal music retrieval. In *Proceedings of the 9th Workshop on Adaptive Multimedia Retrieval (AMR'11)*.

SCHEDL, M., KNEES, P., POHLE, T., AND WIDMER, G. 2006. Towards automatic retrieval of album covers. In *Proceedings of the 28th European Conference on Information Retrieval (ECIR'06)*.

SCHEDL, M., KNEES, P., AND WIDMER, G. 2005. A web-based approach to assessing artist similarity using co-occurrences. In *Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI'05)*.

SCHEDL, M., POHLE, T., KNEES, P., AND WIDMER, G. 2011. Exploring the music similarity space on the web. *ACM Trans. Info. Syst. 29,* 3.

SERRA, X. 2012. Data gathering for a culture specific approach in MIR. In *Proceedings of the 21st International World Wide Web Conference (WWW'12): 4th International Workshop on Advances in Music Information Research (AdMIRe'12)*.

SHAVITT, Y. AND WEINSBERG, U. 2009. Songs clustering using peer-to-peer co-occurrences. In *Proceedings of the IEEE International Symposium on Multimedia (ISM'09): International Workshop on Advances in Music Information Research (AdMIRe'09)*.

SHEN, J., MENG, W., YAN, S., PANG, H., AND HUA, X. 2010. Effective music tagging through advanced statistical modeling. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 635–642.

SLANEY, M. 2011. Web-scale multimedia analysis: Does content matter? *IEEE MultiMedia 18*, 2, 12–15.

SLANEY, M. AND WHITE, W. 2007. Similarity based on rating data. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

SORDO, M., LAURIER, C., AND CELMA, O. 2007. Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*. 531–534.

TURNBULL, D., BARRINGTON, L., AND LANCKRIET, G. 2008. Five approaches to collecting tags for music. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*.

TURNBULL, D., LIU, R., BARRINGTON, L., AND LANCKRIET, G. 2007. A game-based approach for collecting semantic annotations of music. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*.

WANG, X., ROSENBLUM, D., AND WANG, Y. 2012. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM International Conference on Multimedia*.

WHITMAN, B. AND LAWRENCE, S. 2002. Inferring descriptions and similarity for music from community metadata. In *Proceedings of the 2002 International Computer Music Conference (ICMC'02)*. 591–598.

YANG, D., CHEN, T., ZHANG, W., LU, Q., AND YU, Y. 2012. Local implicit feedback mining for music recommendation. In *Proceedings of the 6th ACM Conference on Recommender Systems*. 91–98.

ZADEL, M. AND FUJINAGA, I. 2004. Web services for music information retrieval. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*.

ZANGERLE, E., GASSLER, W., AND SPECHT, G. 2012. Exploiting Twitter's collective knowledge for music recommendations. In *Proceedings of the 21st International World Wide Web Conference (WWW'12): Making Sense of Microposts (#MSM'12)*. 14–17.

ZHANG, B., XIANG, Q., LU, H., SHEN, J., AND WANG, Y. 2009. Comprehensive query-dependent fusion using regression-on-folksonomies: A case study of multimodal music search. In *Proceedings of the 17th ACM International Conference on Multimedia*. 213–222.

ZHAO, Z., WANG, X., XIANG, Q., SARROFF, A., LI, Z., AND WANG, Y. 2010. Large-scale music tag recommendation with explicit multiple attributes. In *Proceedings of the 18th ACM International Conference on Multimedia*. 401–410.

ZOBEL, J. AND MOFFAT, A. 1998. Exploring the Similarity Space. *ACM SIGIR Forum 32,* 1, 18–34.

Received September 2012; revised January 2013; accepted April 2013