#nowplaying Music Dataset: Extracting Listening Behavior from Twitter

Eva Zangerle, Martin Pichl, Wolfgang Gassler, Günther Specht
Databases and Information Systems
Institute of Computer Science
University of Innsbruck, Austria
firstname.lastname@uibk.ac.at

ABSTRACT

The extraction of information from online social networks has become popular in both industry and academia as these data sources allow for innovative applications. However, in the area of music recommender systems and music information retrieval, respective data is hardly exploited. In this paper, we present the #nowplaying dataset, which leverages social media for the creation of a diverse and constantly updated dataset, which describes the music listening behavior of users. For the creation of the dataset, we rely on Twitter, which is frequently facilitated for posting which music the respective user is currently listening to. From such tweets, we extract track and artist information and further metadata. The dataset currently comprises 49 million listening events, 144,011 artists, 1,346,203 tracks and 4,150,615 users which makes it considerably larger than existing datasets.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Online Information Systems—*Data Sharing*; H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Experimentation, Human Factors

Keywords

Social Media, Music Retrieval, Information Extraction

1. INTRODUCTION

Social media platforms like Facebook or Twitter gained huge popularity and have proven to be a valuable source for user-generated content as harvesting social media platforms allows for gathering huge amounts of data from a diverse set of users. The extraction of information from social media platforms has become popular not only among scientists as these data sources allow for new applications as e.g., detecting real-world incidents [1], earthquakes [15] or recommender systems aiming at recommending news [12], followees [9] or hashtags [22] based on information extracted from Twitter. However, research focused on music information retrieval (MIR) and music recommender systems hardly makes use of user-generated data gathered from online social networks (OSN). For instance, Schedl found that work leveraging social media data for music information retrieval is hardly existent except for approaches exploiting the music service Last.fm [16]. Similarly, Bertin-Mathieux et al. call for a large, publicly available dataset which can be used to evaluate scalable algorithms in the field of music information retrieval and music recommender systems [3].

To foster research in the fields of music information retrieval and music recommender systems based on user-generated data retrieved from OSNs, we present the #nowplaying dataset, which contains information about the music listening behavior of users gathered from Twitter. In particular, we leverage so-called #nowplaying tweets, i.e., tweets stating that a certain user listened to a specific track by a specific artist. An example of such a tweet is depicted in the following: “Like a Rolling Stone - Bob Dylan #nowplaying #listenlive”. In this tweet, a user states that he/she listened to the song “Like a Rolling Stone” performed by Bob Dylan. For the #nowplaying dataset, we extract information about the artist and the track title and enrich it with further metadata. The dataset currently comprises 49,921,024 listening events described by approx. 590 million triples (as of 2014/07/09). The set features 144,011 artists, 1,346,203 tracks and 4,150,615 users which makes it considerably larger than existing datasets (cf. Section 5). Additionally, the dataset is updated daily and hence, an average of 62,000 listening events are added each day.

Generally, we see our contributions as follows. Firstly, we provide a publicly available and extensive dataset of listening events which is updated daily and gathered from the Twitter platform. Most importantly, the dataset is considerably larger (more than twice the size in terms of the number of entries, users, tracks and artists) as existing publicly available datasets (an overview about existing datasets can be found in Section 5). Secondly, we interlink our dataset with the MusicBrainz database [18]. The MusicBrainz database allows for a central reference point for artist and track information which can be used to further gather information from other datasets and source. Furthermore, the multitude of datasets underlying the evaluation and comparison of different MIR and recommendation approaches relies on the

WISMM'14, November 7, 2014, Orlando, FL, USA.

ACM 978-1-4503-3157-9/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661714.2661728 .

size, quality and content of the underlying dataset. Hence, we provide a central and up-to-date reference dataset which can be leveraged for comparing and evaluating MIR and recommendation approaches. The size of the dataset also allows for evaluating the scalability of the applied algorithms.

The remainder of the paper is structured as follows. Section 2 describes the extraction framework used for crawling data from Twitter and extracting information. Section 3 describes ways to access the dataset. Section 4 subsequently presents the main characteristics of the dataset.Section 5 covers works and datasets related to our dataset and Section 6 concludes the paper and outlines future work.

2. EXTRACTION FRAMEWORK

In the following Section, we present the framework used for the creation of the #nowplaying dataset. The goal of this framework is to (i) gather raw data underlying our dataset and (ii) extract the desired information about artists and tracks from it. We rely the extraction of track and artist information from tweets on three different extractors. The basic extractor is responsible for extracting simple metadata from the data gathered via the Twitter API, whereas the Track and Artist Extractor relies on the content of the tweets and aims at extracting the artist and track mentioned in the tweet. For all tweets we could not directly resolve against MusicBrainz and were sent via the Spotify music streaming platform, we employ the Spotify extractor which exploits information from the Spotify website.

2.1 Twitter API Crawler

To gather a representative and comprehensive raw dataset, we facilitate the following data collection method. We make use of the public Twitter Streaming API which allows retrieving tweets containing given keywords [19]. In particular, we filter for tweets containing information about a user listening to certain music and hence, utilize the keywords “nowplaying”, “listento” and “listeningto” for filtering tweets. In total, we were able to gather 140 million tweets between 2011/07/11 and 2014/05/11 and the crawling is still continued. As the Twitter Filter API is subject to Twitter’s rate limiting, the number of delivered tweets matching the given keywords is capped by a rate limiting equal to the rate limiting of the public Streaming API (approximately 1% of all tweets). However, this rate limiting did not affect our crawling process as the number of tweets matching our query constantly was below this limit (maximum number of tweets crawled per day: 91,268; cf. Section 4). Hence, we were able to crawl all tweets matching the given filter keywords during the given time period.

2.2 Basic Extractor

The basic extractor’s task is to extract information from the raw tweet data gathered from the Twitter API. The information provided by the Twitter API not only features the tweet text, but also valuable metadata. We directly extract the following entries from each of the crawled documents: (i) the *date and time* when the given tweet was sent, (ii) the *service* used for publishing the tweet (e.g., the Twitter website or an online music streaming service[1] or a plugin for audio players which automatically publishes information about the song currently listened to[2]) and (iii) *username*: the user name of the user who sent the tweet. In order to anonymize this information, we provide the SHA1 hash [8] of the user name.

2.3 Track and Artist Extractor

As already laid out in the introduction, the main motivation for this work was to provide a reference dataset for research related to music streams of users. Therefore, we chose to interlink the #nowplaying dataset with the MusicBrainz dataset as it constitutes the largest publicly available and constantly updated dataset containing artist, track and release information [18]. MusicBrainz features a total of 16,551,902 tracks and 859,893 artists (as of 2014/07/09). Hence, the task of the Track and Artist Extractor module is to match artists and songs occurring within tweets with the corresponding entries in the MusicBrainz database. In the final dataset, we only include listening events which we were able to resolve against the MusicBrainz database, i.e., tweets which contain both a song and an artist which could be identified by its MusicBrainz identifiers. We chose to apply this restriction in order to firstly ensure the data quality within the dataset, as we can guarantee that the dataset features only entries which contain both valid track and artist information which furthermore is coherent. I.e., the track extracted from the tweet has to be performed by the artist extracted from the tweet according to the MusicBrainz database. Secondly, we only want to provide data which can be interlinked by using e.g., the MusicBrainz identifiers for artist and/or song.

An example input tweet for the extractor framework might look as follows: “Like a Rolling Stone - Bob Dylan #nowplaying #listenlive”. The tasks performed by the Track and Artist Extractor based on this tweet are to detect the song (“Like a Rolling Stone”) and artist (“Bob Dylan”) mentioned in the tweet, match both of these with the according Musicbrainz identifiers and store the extracted data to the #nowplaying database. As for the extraction of track and artist information we rely on patterns occurring within such #nowplaying tweets. Schedl et al. observed that tweets about music listening are similarly structured despite occasional comments [17]. Therefore, Schedl et al. proposed five different patterns for extracting artist and track information from #nowplaying tweets. We relied on these patterns for our resolution. Particularly, we made use of the delimiters used in these examples (-, :, by). Our extraction approach is depicted as pseudocode in Algorithm 1. In a first step, we clean the tweet text by removing URLs and whitespaces. In a second step, we split the text of the cleaned input tweet by the above mentioned delimiters. Subsequently, we check whether any artist contained in the MusicBrainz database is contained in the currently checked text chunk. If we can find appropriate artists, we retrieve all songs performed by these artists from MusicBrainz and check, whether one of the song titles is contained in the tweet. If we can find both artist and song within the tweet, we return this information. We have to check for a set of artists as MusicBrainz’s content is user-generated and hence, contains duplicated artists and tracks.

By facilitating this approach, we were able to resolve 30.51% of all tweets against a MusicBrainz ID for both artist and track, which is similar to the findings of Schedl et al. who

[1] e.g., `http://www.spotify.com`

[2] e.g., `http://www.nowplayingplugin.com`

Algorithm 1 Pseudo-Code for Track and Artist Extraction

```
procedure TWEETRESOLUTION(tweetText)
    $result \leftarrow \{\}$
    $tweetText \leftarrow cleanTweetText(tweetText)$
    $textChunks \leftarrow split(tweetText, \{-, :, by\})$

    // Iterate over split text and check if artist is contained
    for all $chunk \in textChunks$ do
        $artistCandidates \leftarrow getArtistsContained(chunk)$

        // if artist is matched, check songs within tweet
        if $artistCandidates.size > 0$ then
            for all $artist \in artistCandidates$ do
                $songsByArtist \leftarrow getSongsByArtist(artist)$
                for all $song \in songsByArtist$ do
                    if song.contains(textchunk) then
                        $result \cup (getMBID(artist), getMBID(song))$
    return result
end procedure
```

achieved a resolution rate of 29.70% [17]. This can be lead back to the following reasons:

- Information about track and artist names within tweets may be incorrect as e.g., in "#nowplaying Mirror - Justin Timberlake" (correct song name is "Mirrors").
- Popular hashtags are frequently abused to spread spam via popular hashtags. Trend hijacking refers to using a hashtag for a different purpose than the one originally intended [21], as e.g. in: "http://t.co/qNr8zeoQTj <– LIKE THE FACEBOOK PAGE!! #follow #twitter #instagram #nowplaying".
- The MusicBrainz database is not complete. As we only make use of artists resp. tracks with a matching MusicBrainz entry, tweets containing valid information may get discarded if they cannot be matched with the MusicBrainz database.
- We perform a rigid matching as we only consider tweets which can be fully matched to a MusicBrainz entry. By loosening the restriction to also include partial or fuzzy matches, we could increase the number of matches. However, this comes at the price of imprecise and incorrect matches.

2.4 Spotify Extractor

A second extractor facilitated to increase the quality and quantity of listening events in the #nowplaying dataset is the Spotify Extractor, which leverages tweets which were sent via the Spotify platform. Spotify is a commerical music streaming service serving 24 million users world wide.[3] Users of this platform are provided with a service which allows for posting the song the user is currently listening to on Twitter (as e.g., "#NowPlaying Peel Me A Grape by Diana Krall on #Spotify http://t.co/J6hJmHGx7s"). The shortened URL mentioned in this example tweet leads to the Spotify website where further information about the track is given. By following the identified URLs, the artist and the track can be extracted from the title tag of the according website. This information is subsequently used for matching the artist and track against the MusicBrainz database and finally the resulting information is stored in the #nowplaying database. As the data extracted from Spotify contains less noise than tweets, using this approach we are able to resolve 81.03% of all track-artist pairs extracted from the Spotify website. Adding this extractor increases the percentage of resolvable tweets from 30.51% to 31.39%. This is, as only 7.22% of all tweets contain a Spotify URL and hence, can be exploited.

2.5 RDF Format

In the following, we present the structure of RDF documents contained in the #nowplaying dataset. As for the definition of these documents, we relied on a set of popular ontologies. In particular, we relied on the SIOC ontology (Semantically-Interlinked Online Communities) for describing content related to Twitter information, as this ontology provides means to describe online community information [4]. As for the description of the music data, we made use of the Music Ontology [13]. Furthermore, we also incorporate Dublin Core elements for describing metadata of listening events [20]. Generally, we distinguish three types of resources: listening events, tracks and artists. A listening event contains metadata about the listening event (tweet) and features a link to the track mentioned in this tweet. A music track contains its title, the according MusicBrainz identifier and a link to the according artist resource. Similarly, an artist resource describes the artist's name and the according MusicBrainz identifier. Further description about the format and an example document can be found on the project's website at `http://dbis-nowplaying.uibk.ac.at`.

3. ACCESSING THE DATASET

Generally, the #nowplaying dataset can be explored and retrieved at `http://dbis-nowplaying.uibk.ac.at`. The according website also provides documentation and usage examples of the access modes (described in the following). Also, the latest statistics about the dataset are presented there (updated nightly).

(1) HTML Browser and Online Query Interface: The easiest way to access the dataset is to make use of the HTML browser at `http://dbis-nowplaying.uibk.ac.at`. This allows for directly browsing through the listening events, artists and track contained in the #nowplaying dataset. Furthermore, we also implemented an online SPARQL interface at `http://dbis-nowplaying.uibk.ac.at/snorql/`. The interface is implemented using the SNORQL SPARQL client[4].

(2) SPARQL Endpoint: We also provide a SPARQL endpoint for querying the presented dataset which can be accessed at `http://dbis-nowplaying.uibk.ac.at/sparql`. Due to performance reasons, we limit the result set of SELECT queries to 1,000 lines.

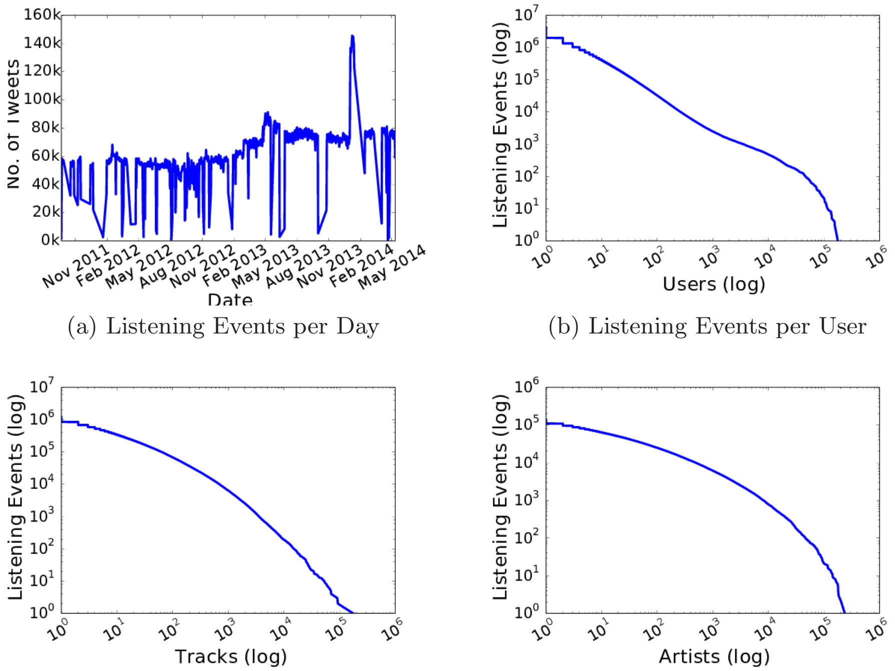
(3) RDF-Dump: In order to also make the full dataset accessible, we provide a downloadable dump of the dataset. This dump contains all RDF-documents as a compressed archive using Turtle Syntax.

4. #NOWPLAYING DATASET

In the following, we present the most important key figures of the #nowplaying dataset extracted by applying the methods described in the previous Section. In total, the dataset contains 49,921,024 listening events, 144,011 distinct artists,

[3] `https://www.spotify.com/at/2013/`

[4] `https://github.com/kurtjx/SNORQL/`



(a) Listening Events per Day

(b) Listening Events per User

(c) Listening Events per Track

(d) Listening Events per Artist

Figure 1: #nowplaying Dataset Statistics

1,346,203 distinct tracks tweets by 4,150,615 distinct users (as of 2014/07/09). The distribution of resolved tweets per day for the given period can be seen in Figure 1(a). The plunges in the chart have to be lead back to failures of either our crawling server or temporary Twitter API downtimes. The average number of new tweets per day is 61,985.67 (median=59,876, standard deviation=19,717.72). As for the artists occurring in the dataset, Table 4 features the top 10 most popular artists in our dataset. The most popular artist within the dataset is Rihanna appearing 237,108 times. The average number of listening events per artist is 333.54, however, the median value of this distribution is 8 (standard deviation=3175.49). This can also be seen in the longtail distribution featured in Figure 1(d). Similarly, also the distribution of the popularity of tracks features a long tail. For a total of 1,206,499 tracks, the mean number of occurrences of each track is 38.15, whereas the median is 3 (standard-deviation=503.57). Figure 1(b) depicts the number of listening events per user (mean number of tweets by user=11.09, median=1, standard-deviation=482.78). This distribution is heavily long-tailed and skewed to the left which is common for music datasets [6]. As for the sources used for publishing #nowplaying tweets, Securenet Systems Radio Playlist Update (a platform for providing internet radio streaming), the Spotify app, the Twitter website and the Twitter client for iPhone are among the most popular sources. The top 10 most popular sources used for sending #nowplaying tweets within the dataset can be seen in Table 4.

Description	Count
Triples	552,655,284
Listening events	46,054,607
Distinct artists	137,270
Distinct tracks	1,206,499
Distinct users	4,150,615

Table 1: #nowplaying Dataset Characteristics

To get a deeper understanding about which music genres are featured in the dataset, we queried the last.fm API [11] to gather the tags associated with the tracks featured in the #nowplaying dataset. The genres featured in the #nowplaying dataset can be seen in Figure 2, where all tags occurring in more than five percent of all tracks featured in the dataset are plotted. The most popular genres within the dataset are Rock, Pop and Alternative.

Artist	Events
Rihanna	237,108
Coldplay	204,930
Taylor Swift	183,753
Bruno Mars	180,212
One Direction	176,588
Maroon 5	165,387
Adele	160,788
Drake	142,740
Katy Perry	142,558
Eminem	136,813

Table 2: Top 10 Artists

Source	Events
Securenet Systems Radio Playlist Update	5,697,531
Spotify	5,006,898
Web	4,186,778
Twitter for iPhone	2,218,604
SAM Broadcaster Song Info	1,833,632
Twitter for Android	1,697,065
iOS	1,271,294
BigURL	1,144,326
Twitter for Blackberry	1,044,029
Now Playing	1,027,307

Table 3: Top 10 Sources

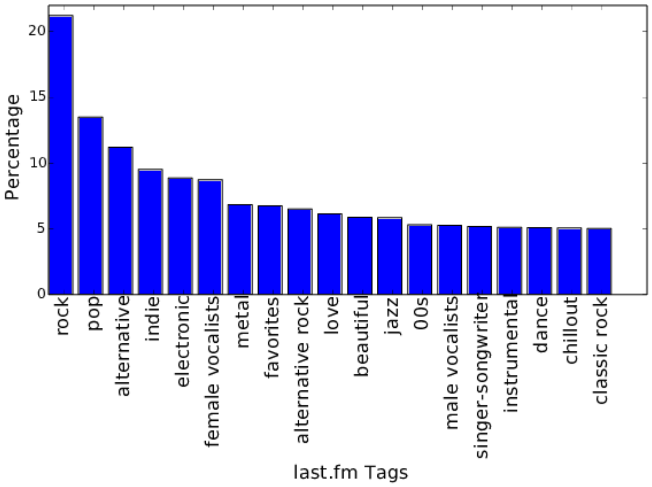


Figure 2: last.fm Tag Distribution for Tracks

5. RELATED WORK

Work related with the #nowplaying dataset can be divided into works concerned with datasets on music data and generally, data extracted from social media platforms. Table 4 features a comparison of the most comprehensive and popular music datasets. As can be seen from Table 4, the #nowplaying dataset features more than twice as many user streaming items than the currently largest dataset.

The most similar dataset available is the Million Musical Tweets Dataset (MMTD) [10]. The data contained in this dataset is also based on information extracted from #nowplaying tweets over the course of 500 days. The authors achieved a similar resolution rate as the resolution approach presented in this paper. However, the authors focus on geospatial characteristics of tweets and as approximately 3% of all tweets contain geospatial information, the dataset is smaller than the #nowplaying dataset. Nevertheless, the MMTD contains the exact geographic location at which the tweet was sent, information about the respective country and identifiers for 7digital, Amazon and MusicBrainz. Generally, the largest publicly available dataset is based on last.fm data [6]. Celma provides two versions of this dataset: a 360k users and a 1k users dataset. The 360k dataset features play-counts for artists for 359,347 unique users. However, no track information is given. The 1k dataset features the full listening history of 992 users including the tracks the users listened to. Both of these datasets include MusicBrainz identifiers for artists resp. artists and tracks. However, not all artists and track could be resolved (i.e., for the 360k dataset 60.7% of all artists feature a MusicBrainz ID). Besides the timestamp, user, song and artist information, no further information is provided. Closely related to these two datasets is a last.fm dataset which was published by the GroupLens group [5]. This dataset includes information about users listening to artists (including listening counts) for a total of 1,892 users and 17,632 artists. Furthermore, the data is enriched with network features (friend relationships between users) and 186,479 artist-tag assignments. Another publicly available datasets is the Million Song dataset (MSD) [3]. This dataset is based on data originating from The Echo Nest and contains one million songs including according metadata and audio analysis. Furthermore, the MusicBrainz and 7digital identifiers are provided. Furthermore, also Yahoo! released music datasets which contain ratings for artists and songs and additional genre information. This data was gathered between 2004 and 2006 from Yahoo's music services. The biggest dataset (R2) features 717 million ratings of 136,000 songs by 1.8 million users. However, this data is anonymized and hence, no information about songs, artists or albums can be identified. Additionally, it is important to mention that there exist further datasets which are mostly focused on providing information about audio features of songs. The MusiCLEF dataset features audio information about 200,000 songs [3]. Further data sources facilitated are internet radio streams [2] or microblogging platforms like Twitter [23]. Furthermore, also RDF sources like DBTune provide general information about music [14]. DBTune is a service platform providing interlinked access to music-related datasets such as MusicBrainz [18], Jamendo[5] or Magnatune[6].

Stemming from a different domain, however, very similar to our dataset approach, Dooms et al. released a dataset containing tweets with movie ratings, which is continuously updated [7]. As of 2014/05/11 it contains 241,215 ratings issued by 27,591 users for 16,237 unique items.

6. CONCLUSION AND FUTURE WORK

In this paper we presented the #nowplaying dataset which features listening events of users which incorporates information about users listening to certain tracks and artists over a timespan over two years. We extract this information from Twitter by gathering all tweets describing that a user listened to a certain song (so-called #nowplaying-tweets). In total, we were able to gather 49,921,024 listening events published by 4,150,615 users and containing 144,011 artists and 1,346,203 tracks. The dataset features 590 million triples and is constantly growing as it is updated daily. Future work includes an optimization of the resolution process in order to be able to resolve more tweets against MusicBrainz (or other reference databases) in respect to loosening the rigid matching approach while still obtaining correct matching results. Furthermore, the incorporation of other sources for the resolution (besides Spotify) is part of future work.

[5] `http://www.jamendo.com`

[6] `http://www.magnatune.com`

dataset	Type	Entries	#Artists	#Tracks	#Users	Updated
#nowplaying	User Streams	49,921,024	144,011	1,346,203	4,150,615	✓
Celma 1K	User Streams	19,150,819	174,090	1,084,865	992	✗
Celma 360K	User Streams	17,559,530	292,557	—	359,349	✗
MMTD	User Streams	1,086,808	25,060	133,968	15,375	✗
MSD	Audio	1,000,000	44,745	1,000,000	—	✗
MusicMicro	User Streams	594,306	19,529	71,400	136,866	✗
HetRec	Ratings	92,834	17,632	—	1,892	✗
Yahoo!	Ratings	717,872,016	9,441	136,735	1,800,000	✗

Table 4: Overview of Available Music Datasets

7. ACKNOWLEDGEMENTS

This work was partially funded by the the University of Innsbruck (Forschungsförderungsmittel der Nachwuchsförderung 2013 der Universität Innsbruck).

8. REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Semantics+Filtering+Search=twitcident. Exploring Information in Social Web Streams. In *Proc. of the 23rd ACM Conference on Hypertext and Social Media*, pages 285–294. ACM, 2012.

[2] N. Aizenberg, Y. Koren, and O. Somekh. Build Your Own Music Recommender by Modeling Internet Radio Streams. In *Proc. of the 21st Int. Conference on World Wide Web*, WWW '12, pages 1–10, New York, NY, USA, 2012. ACM.

[3] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The Million Song Dataset. In *Proc. of the 12th Intl. Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[4] J. G. Breslin, S. Decker, A. Harth, and U. Bojars. SIOC: An Approach to Connect Web-based Communities. *Int. Journal of Web Based Communities*, 2(2):133–142, 2006.

[5] I. Cantador, P. Brusilovsky, and T. Kuflik. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In *Proc. of the 5th ACM Conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

[6] O. Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010.

[7] S. Dooms, T. De Pessemier, and L. Martens. MovieTweetings: a Movie Rating Dataset Collected From Twitter. In *Workshop on Crowdsourcing and Human Computation for Recommender Systems*, 2013.

[8] D. Eastlake and P. Jones. US secure hash algorithm 1 (SHA1), 2001.

[9] J. Hannon, M. Bennett, and B. Smyth. Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. In *Proc. of the fourth ACM Conference on Recommender Systems*, pages 199–206. ACM, 2010.

[10] D. Hauger, M. Schedl, A. Kosir, and M. Tkalcic. The Million Musical Tweet Dataset - What We Can Learn From Microblogs. In *Proc. of the 14th Intl. Conference on Music Information Retrieval (ISMIR 2013)*, pages 189–194, 2013.

[11] last.fm API. `http://www.lastfm.de/api`.

[12] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to Recommend Real-time Topical News. In *Proc. of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.

[13] Y. Raimond, S. A. Abdallah, M. B. Sandler, and F. Giasson. The Music Ontology. In *Proc. of the 8th Intl. Conference on Music Information Retrieval (ISMIR 2007),*, pages 417–422. Citeseer, 2007.

[14] Y. Raimond and M. B. Sandler. A Web of Musical Information. In *Proc. of the 9th Intl. Conference on Music Information Retrieval (ISMIR 2008)*, pages 263–268, 2008.

[15] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proc. of the 19th Int. Conference on World wide web*, pages 851–860. ACM, 2010.

[16] M. Schedl. Leveraging Microblogs for Spatiotemporal Music Information Retrieval. In *Proc. of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 796–799, Berlin, Heidelberg, 2013. Springer-Verlag.

[17] M. Schedl, D. Hauger, and J. Urbano. Harvesting Microblogs for Contextual Music Similarity Estimation - A Co-occurrence-based Framework. *Multimedia Systems*, 2013.

[18] A. Swartz. Musicbrainz: A semantic web service. *Intelligent Systems, IEEE*, 17(1):76–77, 2002.

[19] Twitter API. `https://dev.twitter.com/docs/api/1.1/post/statuses/filter`.

[20] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core Metadata for Resource Discovery. *Internet Engineering Task Force RFC*, 2413(222):132, 1998.

[21] What is Hashtag Hijacking? `http://smallbiztrends.com/2013/08/what-is-hashtag-hijacking-2.html`.

[22] E. Zangerle, W. Gassler, and G. Specht. Using Tag Recommendations to Homogenize Folksonomies in Microblogging Environments. In *Proc. of the Third Intl. Conference on Social Informatics (SocInfo) 2011, Singapore.*, Lecture Notes in Computer Science, pages 113–126. Springer, 2011.

[23] E. Zangerle, W. Gassler, and G. Specht. Exploiting Twitter's Collective Knowledge for Music Recommendations. In *Proc. WWW Workshop: Making Sense of Microposts MSM*, 2012.