MusicLynx: Exploring Music Through Artist Similarity Graphs

Alo Allik
Queen Mary University of London
a.allik@qmul.ac.uk

Florian Thalmann
Queen Mary University of London
f.thalmann@qmul.ac.uk

Mark Sandler
Queen Mary University of London
mark.sandler@qmul.ac.uk

ABSTRACT

MusicLynx is a web application for music discovery that enables users to explore an artist similarity graph constructed by linking together various open public data sources. It provides a multifaceted browsing platform that strives for an alternative, graph-based representation of artist connections to the grid-like conventions of traditional recommendation systems. Bipartite graph filtering of the Linked Data cloud, content-based music information retrieval, machine learning on crowd-sourced information and Semantic Web technologies are combined to analyze existing and create new categories of music artists through which they are connected. The categories can uncover similarities between artists who otherwise may not be immediately associated: for example, they may share ethnic background or nationality, common musical style or be signed to the same record label, come from the same geographic origin, share a fate or an affliction, or have made similar lifestyle choices. They may also prefer similar musical keys, instrumentation, rhythmic attributes, or even moods their music evokes. This demonstration is primarily meant to showcase the graph-based artist discovery interface of MusicLynx: how artists are connected through various categories, how the different graph filtering methods affect the topology and geometry of linked artists graphs, and ways in which users can connect to external services for additional content and information about objects of their interest.

KEYWORDS

music discovery and recommendation, music information retrieval, machine learning, graph theory, web programming, Semantic Web, Linked Data, similarity modeling, affective computing

ACM Reference Format:
Alo Allik, Florian Thalmann, and Mark Sandler. 2018. MusicLynx: Exploring Music Through Artist Similarity Graphs. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186970

1 RECOMMENDATION AND DISCOVERY

Music recommendation systems, whether commercial or experimental, either based on collaborative filtering, content-based information retrieval or both, have a tendency to present suggestions to users in a linear list format without providing details about the nature of similarity. This is possibly due to the tendency towards efficiency and accuracy of recommendations in such services, in fear of overwhelming users with too much content. This may sound like a reasonable approach in a profit-minded music streaming or download environment, however, valuable information, which otherwise might enable a more enhanced and illuminating discovery of music, is concealed from users. MusicLynx - available as a web application at https://musiclynx.github.io - is being developed as an alternative way to browse and discover music, based on different methods of modeling artist similarity. The primary method relies on analyzing Wikipedia categories[1] found at the bottom of each artist page and deriving from these meaningful connections between artists. These connections are enhanced by creating new categories based on music information retrieval data and crowd-sourced mood tag statistics. The result is a similarity graph for each artist who has an entry in one or more of the available datasets. The goal is to create a multifaceted representation of artist similarity for enhanced music exploration and discovery. The similarity graph is displayed on each artist's page which also includes a brief biography, links to external web resources about the artist, potentially including metadata, videos, TV and radio shows, recordings of live performances and other information. The page also features a music player that streams 30 second previews of top tracks by the artist if found in the streaming service database.

2 LINKING DATA SOURCES

There are a number of openly accessible music-related knowledge bases and datasets, that crowd-source different types of information and make it available to the community. These come in many different formats and varying levels of programmatic accessibility, which makes it sometimes challenging to link them together. The linked artist dataset that was created in the process of MusicLynx development uses and makes connections between the following services:

- **MusicBrainz** (https://musicbrainz.org): an open music encyclopedia of music publishing metadata, primarily used in MusicLynx for global identifiers that it provides for musical entities
- **Dbpedia** (https://dbpedia.org): a Semantic Web triple store that contains structured information extracted from Wikipedia; in MusicLynx this enables querying and analysis of Wikipedia artist categories
- **AcousticBrainz** (https://acousticbrainz.org): a crowd-sourced publicly accessible dataset of content-based music information retrieval data, which in this context is used to build similarity models based on musical tonality, rhythm and timbre features
- **Million Song Dataset** (https://labrosa.ee.columbia.edu/millionsong/): a popular data set in the Music Information Retrieval community
- **Last.FM** (https://www.last.fm/): is a social music user site that collects information about users' listening habits and musical tastes; among other information it provides access to user tags from which a mood-based similarity model has been derived for the MusicLynx artist graph

[1] https://en.wikipedia.org/wiki/Help:Category

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186970

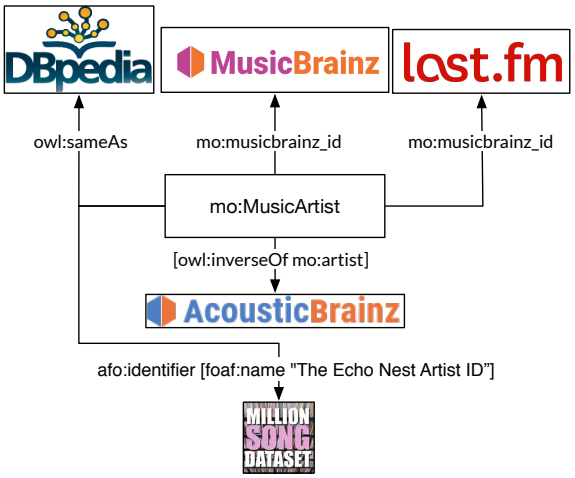

Figure 1: Linking artist identifiers from different data sources in the MusicLynx API

- **BBC Music** (https://www.bbc.co.uk/music): BBC Music provides a page for most artists in MusicBrainz; used in the process of linking an artist MusicBrainz identifier to Dbpedia URI.
- **Sameas** (http://sameas.org): a service that finds co-references between different data sets, primary linking tool between MusicBrainz and Dbpedia.
- **Wikidata** (https://www.wikidata.org): a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others. MusicLynx API uses Wikidata as a backup linking tool between MusicBrainz and Dbpedia.

Figure 1 highlights services where various types of data originate from and how artist and recording entities are linked by Semantic Web properties. MusicBrainz GUIDs and Dbpedia URIs are used as the primary identifiers of artists in the MusicLynx system, therefore making reliable links between these datasets is crucial. There are 3 methods that accomplish this: (1) using Sameas.org service and BBC Music artist URIs, since Sameas does not contain MusicBrainz URIs; (2) using Wikidata entity identifiers; (3) if the previous methods fail, we can try to convert the artist name into a Dbpedia URI, which is the least reliable way of achieving the connection. Listing 1 is an example of RDF representation of an artist in the MusicLynx dataset in Notation3 syntax. Concepts from the Music Ontology, Dbpedia Ontology and YAGO are combined to link the different sources into a Semantic Web fragment of music artists. The properties **owl:sameAs** and **mo:musicbrainz_id** are used to link the structured data from Dbpedia to information from any service that uses MusicBrainz identifiers, including, for example, Last.FM. AcousticBrainz feature data can be linked to the Semantic Web graph as well, even though the service does not support Semantic Web formats; however, it does use MusicBrainz identifiers for recordings. This can be achieved by employing an inverse of the property **dbo:artist** of class **dbo:MusicalWork** to associate recordings with artists. Mappings are also provided for some of the tracks in the Million Song Dataset using the generic identifier linking mechanism **afo:identifier** from the Audio Feature Ontology (AFO) as shown in the listing.

Links between data sources enable analyzing the existing categories of artists that can be queried from Dbpedia and adding new ones to the artist graphs using content-based audio features from AcousticBrainz and applying machine learning techniques to Last.FM mood tags in order to calculate 2-dimensional mood coordinates on the arousal-valence plane for each artist.

```
@prefix yago: <http://dbpedia.org/class/yago/> .
@prefix lynx: <https://musiclynx.github.io/> .
@prefix afo: <https://w3id.org/afo/onto/1.1#> .
@prefix mo: <http://purl.org/ontology/mo/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
<https://musiclynx.github.io/artist/> a mo:MusicArtist ;
  foaf:name "Miriam Makeba"@en ;
  mo:musicbrainz_id "bc5c2918-4aba-4ef6-a245-100563a4487f"
      ;
  owl:sameAs <http://dbpedia.org/resource/Miriam_Makeba> ;
  rdf:type yago:WikicatXhosaPeople ,
    yago:WikicatHumanRightsActivists ,
    yago:WikicatAnti-apartheidActivists ,
    yago:WikicatMusiciansWhoDiedOnStage ,
    yago:WikicatPeopleFromJohannesburg ,
    yago:WikicatSouthAfricanExiles ,
    yago:WikicatSouthAfricanPeopleOfSwaziDescent ,
    lynx:AcousticBrainzSimilarByTonality ,
    lynx:AcousticBrainzSimilarByRhythm ,
    lynx:AcousticBrainzSimilarByTimbre ,
    lynx:MoodplaySimilarArtists .
  [owl:inverseOf dbo:artist]
    [ a mo:Recording ;
    mo:musicbrainz_id
        "fed344b5-4175-4ecd-8453-cc785697e990" ;
    afo:identifier [
      afo:value "TRKQPBL128F9311048" ;
      foaf:name "Echo Nest Track ID"
    ] ;
    dc:title: "The Naughty Little Flea" . ] ,
    [ a mo:Recording ;
    mo:musicbrainz_id
        "b0a30dd6-1bc1-4423-ad51-7dd977b68dfc" ;
    dc:title: "Nyamuthla" . ]
```

Listing 1: An example of an artist RDF Notation3 representation in the MusicLynx dataset

3 GRAPH FILTERING

The Wikipedia artist categories, that can be queried from Dbpedia in the form of structured Semantic Web triples, organize artists according to style, preferred instrument, record label, geographic location or origin, ethnicity, nationality, lifestyle choices, occupation, and so on. There is a significant disparity between category degrees (i.e. number of artists that belong to a category), which implies a difference in meaningfulness of a connection between artists. For example, the connections due to the category "Living People" are arguably not as interesting, and thereby less meaningful, as ones made through "Artists Who Died On Stage". The MusicLynx API provides a number different methods of graph ranking to facilitate filtering that, on the one hand, offer a measure of "meaningfulness" of links between artists, and a balance between accuracy and diversity of similarity on the other. In the paradigm of bipartite graph similarity, which means there are only connections between 2 types of nodes (in this case artists and categories), there are a number of different ways to calculate and rank similarity, primarily using the degrees of either types of nodes as similarity

weights in order to boost the ranking of nodes with smaller degrees. The simplest measure of similarity is *global ranking* or, in this case, the number of categories shared between any pair of artists. Global ranking in itself is not a very accurate or diverse measure of similarity, but rather is used as a component in more complex calculations. The MusicLynx API supports a number of well-established graph-based similarity measures, including Jaccard similarity, the Sorensen index, collaborative filtering, Maximum Degree Weighted (MDW), and a hybrid of heat and probability spreading algorithms. For a detailed overview of different bipartite graph similarity methods, see [6]. These measures are used to rank similar artists as well as limit the number of artists displayed to the user. Since many artists are linked to potentially thousands of other artists, methods that boost selection of meaningful links and that balance between accuracy and diversity become significant. The MDW[4] method accomplishes this task more successfully than other methods, because it uses category degrees to weight global ranking which, in effect, diminishes the contribution to the similarity of large categories like "Living People". Most other measures do not take into account category degrees, with the exception of the hybrid heat-probability spreading algorithm[9]. Furthermore, the similarity measure is divided by the maximum of the degrees of the artists being compared:

$$s_i, j^u, MWD = \frac{1}{max\{k_i, k_j\}} \sum_{\alpha=1, k_\alpha>1}^{M} \frac{a_{i,\alpha} a_{j,\alpha}}{k_\alpha - 1} \quad (1)$$

The effect of this is diminished contribution to similarity of artists who belong to many categories. This combination of artist and category degrees in such a way consequently promotes artists who belong to fewer categories with smaller memberships, which arguably increases diversity. The heat-probability spreading algorithm is very similar to MDW with the exception of dividing the weighted global ranking by the degree of either the target artist (probability spreading) or each similar artist (heat spreading). That allows a choice in the system between encouraging either accuracy or diversity.

4 CONTENT-BASED RETRIEVAL

Content-based Music Information Retrieval (MIR)[2] facilitates applications that rely on perceptual, statistical, semantic or musical features derived from audio using digital signal processing and machine learning methods. These features may include statistical aggregates computed from time-frequency representations extracted over short time windows. For instance, spectral centroid is said to correlate with the perceived brightness of a sound [8], therefore it may be used in the characterization in timbral similarity between musical pieces. More complex representations include features that are extracted using perceptually modeled algorithms. Mel-Frequency Cepstral Coefficients (MFCCs) for instance are often used in speech recognition as well as in estimating music similarity [5]. Higher-level musical features include keys, chords, tempo, rhythm, as well as semantic features like genre or mood, with specific algorithms to extract this information from audio. Content-based features are increasingly used in music recommendation systems to overcome issues such as infrequent access of lesser known pieces in large music catalogues (the “long tail” problem) or the difficulty of recommending new pieces without user ratings in systems that employ collaborative filtering (“cold start” problem) [3].

MusicLynx is designed to support music discovery by encouraging users to engage in interesting journeys through the artist graph. The aim is to create links between artists based on stylistic elements of their music derived from a collection of recordings and add to the social, biographical and cultural categories from Dbpedia. It is not trivial to extract high-level stylistic descriptors directly from audio, however, correlations with lower level features can be detected such as the average tempo of a track, the frequency of musical event onsets, the most commonly occurring keys or chords, or overall spectral dynamics that indicate which instruments are playing or whether the audio contains a singing voice. Three main categories of audio descriptors are used to model similarity between artists in the domains of rhythm, tonality and timbre. Audio features in each category of interest are queried from the AcousticBrainz service using MusicBrainz recording identifiers. The features selected for the content-based similarity experiments include:

- beats-per-minute and onset rate for rhythm
- chord histograms for tonality
- MFCC-s for timbre

There are tracks by approximately 20,000 artists included in the dataset, a number arrived at after applying the constraint that requires an artist to be represented by a minimum of 10 tracks in AcousticBrainz. At the other end of the track count spectrum, there are artists who have features for well over 1,000 tracks in the database, which raises the question of how to meaningfully compare artists with such disparate representations. Two methods have been tested to solve this problem: Gaussian Mixture Models (GMM) and calculating maximum similarity between artists. The GMM method is described in [7]. While the GMM method includes all tracks available in the dataset for each artist, thus potentially comparing feature sets of significantly different sizes, the maximum similarity method has been developed specifically to alleviate this disparity. The significant variance in the number of tracks from artist to artist, of course, reflects a general truth: some artists are more prolific than others. However, in the case of AcousticBrainz, this phenomenon is made even more complicated, since artist representation depends on the unpredictability of crowd-sourced data, particularly since the domain of music information retrieval is relatively unknown to the general public. The solution we propose here is that only the most similar tracks between two artists are included in the comparison; the number of tracks considered for each pair of artists is simply determined by which artist has fewer tracks in the dataset. The maximum similarity method stores track features for each category of each track in a separate track vector of fixed dimension M (i.e. number of features). Then, we define a proximity function as the squared Euclidean distance that will determine if two track vectors are close and therefore similarity between tracks.

5 CROWD-SOURCED TAG STATISTICS

While automatic feature extraction has significantly enhanced organization and categorization of large music collections, it is still rather challenging to derive high level semantic information relating to mood or genre. Complementing signal processing and

machine learning methods with crowd-sourced social tagging data from platforms like Last.fm can enrich and inform understanding of general listening habits and connections between artists. Mood-based similarity is another experimental enhancement to MusicLynx. This method involves using the Semantic Web version of ILM10K music mood dataset originally created for a mood-based interactive music player - Moodplay - that consists of over 4,000 unique artists. The dataset is based on crowd-sourced mood tag statistics from Last.fm users, which have been transformed to numerical coordinates in a Cartesian space. A more detailed account of this machine learning process is outlined in [1]. Every track in the collection is associated to 2-dimensional coordinates reflecting energy and pleasantness respectively. The similarity between artists in this case is measured by first calculating the location of the target artist in the mood space by averaging the coordinates of all the associated tracks. The same procedure is repeated for all other artists which then enables computing Manhattan distances between the target from the rest and using the ranking as similarity metric. Each artist in the mood dataset can thereby be included in the new category "Moodplay similar artists" by linking to the top N closest artists by mood coordinates distance.

6 BUILDING A SIMILARITY GRAPH

The similarity graph that is displayed to users is built utilizing the different graph filtering methods described above. The first step in the process of constructing a similarity graph for an artist involves sending a query to the Dbpedia SPARQL endpoint which retrieves all artists linked to the target artist, their degrees, i.e. how many categories does each artist belong to, and the global ranking or count of shared categories between the target and linked artists. In order to ensure that all categories that a given artist belongs to are presented in the interface, while at the same time curbing links to artists in larger categories, the system uses one of the filtering methods - the above-mentioned MDW by default - to boost artists and categories of smaller degrees for ranking. Even though artists typically belong to more than one category, each can only belong to one in this context, due to constraints of the graph structure. This means that artists that belong to small categories get assigned with priority and larger categories end up with smaller number of artists as a result. The preference here is given to uniqueness and diversity over accuracy. The maximum number of artists is fixed by a system parameter as it is not feasible to display graphs that exceed a certain number of nodes, which means that there is a cutoff point which is also taken into account when the graph is being calculated.

7 THE APPLICATION

The MusicLynx application consists of two components: the Angular2 front-end that serves content to users and the MusicLynx API implemented in ExpressJS which consists of modules that query the different services for data and process it. Figure 2 shows a screenshot of an artist page. The application is deployed on publicly available and free infrastructure with consideration to sustainability and legacy. The front-end application is compiled into a static assembly and deployed to Github pages, which also provides a free URI for the application. The API makes use of

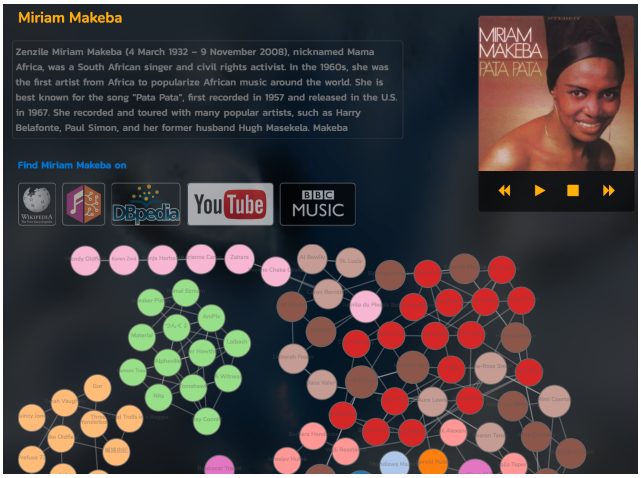

Figure 2: Screenshot of an artist page on MusicLynx with the interactive artist similarity graph at the bottom.

the Heroku free server hosting plan and is also publicly accessible for querying: https://musiclynx-api.herokuapp.com/. MusicLynx is developed as an open source project. The source code is available on Github for both the front-end (https://github.com/darkjazz/musiclynx/tree/static) and the server component (https://github.com/darkjazz/musiclynx-server) The API modules include SPARQL query interfaces for Dbpedia, Wikidata, and Sameas services, MusicBrainz, Youtube, and Deezer search functionality for external content, as well as SPARQL query builder, graph constructing, and graph filtering modules. AcousticBrainz and Moodplay data is provided as static datasets integrated into the API component.

REFERENCES

[1] Mathieu Barthet, György Fazekas, Alo Allik, and Mark B. Sandler. 2015. Moodplay: an interactive mood-based musical experience. In *Proceedings of the Audio Mostly 2015 on Interaction With Sound, AM '15, Thessaloniki, Greece, October 7-9, 2015.* 3:1–3:8. https://doi.org/10.1145/2814895.2814922

[2] Michael A. Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. 2008. Content-Based Music Information Retrieval: Current Directions and Future Challenges, Vol. 96. IEEE Proceedings.

[3] Ò. Celma. 2010. *Music Recommendation and Discovery:The Long Tail, Long Fail, and Long Play in the Digital Music Space.* Springer Verlag.

[4] A. Fiasconaro, M. Tumminello, V. Nicosia, V. Latora, and R. N. Mantegna. 2015. Hybrid recommendation methods in complex networks. *Phys. Rev. E* 92 (Jul 2015), 012811. Issue 1. https://doi.org/10.1103/PhysRevE.92.012811

[5] Beth Logan. 2000. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proc. Int. Symp. of Music Information Retrieval (ISMIR).*

[6] Linyuan Lü, Matus Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender Systems. *CoRR* abs/1202.1112 (2012). http://dblp.uni-trier.de/db/journals/corr/corr1202.html#abs-1202-1112

[7] Mariano Mora-Mcginity, Alo Allik, GyÃűrgy Fazekas, and Mark B. Sandler. 2016. MusicWeb: Music Discovery with Open Linked Semantic Metadata.. In *MTSR (Communications in Computer and Information Science)*, Emmanouel Garoufallou, Imma Subirats Coll, Armando Stellato, and Jane Greenberg (Eds.), Vol. 672. 291–296. http://dblp.uni-trier.de/db/conf/mtsr/mtsr2016.html#Mora-McginityAF16

[8] Emery Schubert and Joe Wolfe. 2006. Does Timbral Brightness Scale with Frequency and Spectral Centroid? *Acta Acustica united with Acustica* 92, 5 (2006), 820–825.

[9] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515. https://doi.org/10.1073/pnas.1000488107 arXiv:http://www.pnas.org/content/107/10/4511.full.pdf