Interactive Multimedia Visualization for Exploring and Fixing a Multi-Dimensional Metadata Base of Popular Musics

Maroua Tikat, Marco Winckler, and Michel Buffa

University Côte d'Azur, SPARKS/wimmics team, Inria, CNRS, I3S, France
[maroua.tikat, winckler, michel.buffa]@univ-cotedazur.fr

Abstract. In this position paper we discuss the use of information visualization techniques as a mean to find and characterize inconsistencies in music datasets. This idea is supported by empirical findings in a previous work dedicated to the visualization of a large dataset of music metadata called WASABI. During the development process of a visualization technique for exploring the multitude of multimedia attributes in the WASABI dataset (which includes lyrics, chords, audio, graphics describing sound analysis, etc.), we found visual patterns suggesting data inconsistencies (ex. ambiguities, inaccuracies, missing data, conflicts, etc.), which might have occurred during the integration from diverse sources. Traditionally, information visualization techniques are used to understand the data corpus and identify causal relationships, trends, patterns of data concentrations. Nevertheless, our findings suggest that information visualization techniques can be used to inspect data quality and highlight the parts of the datasets that need to be corrected/improved. Furthermore, we suggest that information visualization could be used as an entry point for repairing the dataset. More specifically, our aim is to use information visualization techniques to: communicate data quality issues to users, compare the outcomes of methods (such as crowdsourcing, matrix vectorization, graph reasoning, among other) used to fix the dataset, and observe the evolution of problem solving during the maintenance of the dataset.

Keywords: music dataset, multimedia data, visualization techniques, data quality, multivariate data

1 Introduction

This position paper presents our findings and ideas raised whilst creating a knowledge graph for music datasets. Working on a knowledge graph makes it possible to interlink resources with rich semantic relationships, which helps increase knowledge. If we take for example a music dataset, it would offer a large scope of metadata describing music, including multimedia dimensions that count textual (e.g. title, lyrics), graphical (e.g. curves), and audio (e.g., the sound produced), as well as metadata qualifying the works (e.g., date and time of recording, authors, performers...) and uses (e.g. song covers, classification for a specific use,...). All these resources are given URIs that can be referenced to documents describing them and can be shared across datasets. In order to

explore music datasets, we have proposed in a previous work [6] the use of information visualization techniques. Nonetheless, during the visualization process we found some patterns that suggest inconsistencies in the dataset. As a matter of fact, the related work section shows that all the music datasets containing a large volume of data have been built from several data sources, either from public sources on the web, or from audio analysis or lyrics analysis, which will inevitably generate data incoherence. These observations led us to consider a new perspective for using information visualization techniques as a tool for assessing data quality, comparing the results of fixing methods and communicating data inconsistencies to the user. In the rest of this position paper we discuss the underlying background for characterizing the problem, especially on the field of music datasets, and we envisage a method aimed at helping to fix data quality problems.

2 Background and related work

2.1 Music datasets

Musical contents can be described in many multimedia dimensions (ex. lyrics, chords, sounds, metadata...). Nonetheless, most of music datasets are specialized on a few types of attributes or on a specific musical genre. In order to have a more complete description of songs, some new datasets have been created by aggregating multiple data sources:

- *The WASABI dataset [6]*: available at https://github.com/micbuffa/WasabiDataset, it contains over 2 MM commercial songs issued from multiple sources. It features a rich set of cultural metadata on songs, albums and artists, and also contains metadata extracted from NLP analysis of lyrics and MIR (Music Information Retrieval) analysis of song audio contents. An Ontology describing the WASABI dataset is also publicly available [1].
- *The Million Song Dataset [5]*: this one includes over 1 MM contemporary popular songs, with metadata extracted mainly from audio content analysis and lyric-related metadata in the form of "bag of words".
- *MusicBrainz*: a free online music database, available under an open licence, which collects music metadata (ex. artits, albums, labels, etc) and makes it available to the public. There is no information about lyrics but a companion web site (i.e. AcousticBrainz) offers MIR audio analysis of a subset of the songs (crowd sourced).
- *MusicWeb [2] and its latest version MusicLynx [3]*: allow the exploration of a graph of artist similarities, built by linking several free public data sources.

2.2 Visualization techniques used in the musical field

Music datasets are often huge, containing hundreds of thousands, or millions of songs. In this context, information visualization techniques might be a suitable alternative for exploring music contents, identify patterns, trends, and correlations. According to a recent survey done on visualizations for music related data [8], musical collections

[1] https://github.com/micbuffa/WasabiDataset/blob/master/ontology/wsb-1.0.ttl

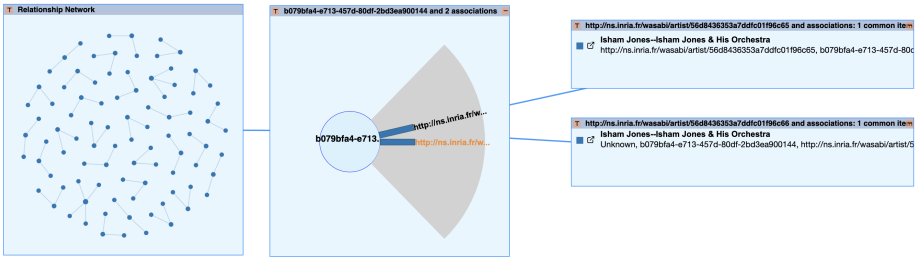

Fig. 1: Visualization of artists by their MusicBrainz id

(albums, playlists, music archives) are mostly visualized through the means of map visualizations, on the other hand musical works are represented by different and special techniques such as glyph, some of which cannot be grouped explicitly. Musicians are viewed most often through graphs and timelines, while 3D rendering is used for instruments. Charts can be used for all data types listed. Sunbursts, Node-Link Trees, Bubble Charts and Treemaps are used to explore music rankings using interactive visualization [9]. They all allow to represent music data content (artist's name, tracks' position,...) and music genre. As far as audio analysis is concerned, several approaches have been considered to determine the structure of a song [14] [11] and classify them (intro, theme, verse, chorus, solo and outro) based on different musical dimensions such as melody, harmony, rhythm, and timbre. We cite the following visualization techniques: recurrence plots, arc diagrams, chroma features and rythmograms.

2.3 Types of data quality defects found using visualizations

Whilst inconsistencies are not frequent in the WASABI dataset, we were able to spot some of them during the exploration process using a visualization tool called MGExplorer. Figure 1 illustrates an example of duplicated entry for artists found during the exploration process. Hereafter we present a short list of the types of data quality defects we have found. We classify the problems in two categories, as follows:

1) Intrinsic defects encompass cases of duplicated data, missing data, disambiguation of artists with similar names, wrong formats for dates, conflicts between values of the same property collected from different data sources, broken links, etc.

2) Scenario problems related are specific to the application domain, such as: i) the same song with more than one producer or performer, according to the source; ii) a big difference between the number of songs produced in one single year, compared to others years (is it normal - related to some real life events? Is it an error during data collection?); iii) songs classified in multiples genres, which might be the expression of errors in the classification of divergence of opinions among how songs could be classified.

2.4 Methods to assess data quality, to complete and to fix the dataset

Some of the methods we found in the literature allows both the detection and the correction of dataset inconsistencies, as follows:

- *Sourcing knowledge [1]*: a solution that takes advantage of the development of the web and online communities to help fill blanks in a dataset. It can also be used to detect and fix outliers in a dataset.
- *Support Vector Machines [13] and matrix factorization*: methods that have been used to predict entities' types.
- *association rule mining [7] and neural network* : can be used to predict relations based on chains of other relations,
- *Retrospective evaluation*: an approach that allows people to assess a value by saying whether it is true or false.
- *Reasoning*: inference techniques are highly used in the semantic web community to check for errors in knowledge graphs, by adding rules and restrictions on the ontology, provided that the latter is rich.
- *Interquartile range and kernel density estimation*: these methods are used to correct numerical outliers in a dataset.
- *DeFacto [10]*: one of the few methods to correct a dataset, it assigns confidence scores to statements, based on their occurences in different web pages.
- *Fact validation through consensus measurement [12]*: an automatic approach using knowledge graph interlinks to detect erroneous numerical values. It exploits the links between identical resources and apply different matching functions between the properties of individual sources. Facts in a knowledge graph are assumed to be wrong if several other sources have a consensus for a contradictory fact.
- *Graph embedding TransE [4]*: a method that allows link prediction on large databases by interpreting relationships as translations operating on the low-dimensional embeddings of the entities.

Whilst several methods for detecting problems with data quality exist, our research questions are focused on how to communicate these problems to the users. For example, it is a duplicated record and not a homonyms artists. In the example given by Figure 1, the disambiguation can be made by allowing users to explore other attributes associated to the artists.

3 Our approach

As we shall see, detecting and solving data quality problems might require a decision-making processing involving users. For example, different algorithms might provide divergent analysis for music genre and the correct classification will depend on the users' needs. For that, we propose to put users in the loop for detecting data quality problems and for deciding how the problems must be solved. Moreover, rather than propose completion/repairing methods (e.g., link prediction, duplication detection, missing URL, etc.), our approach allows users to apply existing methods and visualize the outcomes in terms of repairs proposed; ultimately users should be able to check if the corrections

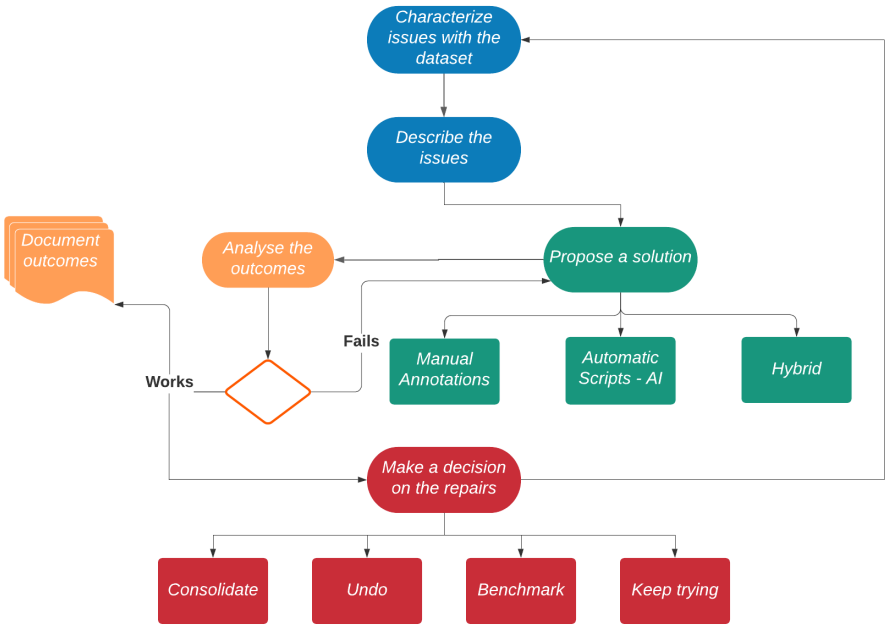

Fig. 2: Methodology

fulfill their needs (or not). Figure 2 presents a view at glance of our approach for detecting and solving quality problems whilst exploring music dataset using visualization techniques. Our approach encompasses the following premises:

- Allow continuous integration of data from several sources, including results of audio and song lyrics analysis; In addition to standard music data, we will be adding emotion analysis, structure detection (verse, chorus, etc.), topic analysis, etc.;
- Use information visualization techniques for multivariate data (such as MGExplorer and Parallel Coordinates) to represent all sort of data;
- Identify the problems with data quality during the exploration process using visualization techniques; Figure 1 shows an example of a problem detected in the WASABI dataset using MGExplorer, which is a duplicate artist.
- Solutions for detecting problems can be: manual, automatic or both;
- Visualization techniques can accommodate the results of data correction adding an explanation to the context of use;
- Compare the results of methods for detecting problems, allowing the users to decide about the outcomes of the corrections proposed. Validation of data correction/completion can be done quantitatively, by running against some measures (i.e number of missing properties, number of artists that require disambiguation) or qualitatively (i.e present a before/after visualization, and do user testing with real users to evaluate the refined data compared to what we had previously and evaluate the tools themselves).

4 Discussion and future work

This paper presented a preliminary work towards the visualization of music datasets. It includes a discussion about problems and potential solutions for integration methods for fixing data quality problems along the visualization process. As we have discussed, these data quality problems arose by continuous data integration (ex. new songs, new analysis of old songs, conflicts detected over time...). So that, we claim that data quality of music datasets is part of a continuous curation process rather than a definite state of the dataset. For that, we propose a methodology aiming at visualizing music datasets, highlighting problems found during the exploration process, as well as allowing to compare solutions to fix those problems. The next steps should be to deal with integration of corrective methods for fixing problems and support the versioning of data (a main requirement for comparison of the methods' outcomes). For this purpose, MGExplorer will be used as a proof of concept to illustrate the feasibility of the methodology proposed. We are currently working on running prototype that could be used to test our hypothesis with end users. Finally, despite the fact that our work is focused on the music dataset WABASI, we suggest that the solution might be suitable to other types of datasets to be explored in the future.

References

1. M. Acosta and al. Crowdsourcing linked data quality assessment. In *International semantic web conference*, pages 260–276, 2013.
2. A. Allik, M. Mora-Mcginity, G. Fazekas, and M. Sandler. Musicweb: an open linked semantic platform for music metadata. In *Proc. 15th International Semantic Web Conf.*, 2016.
3. A. Allik, F. Thalmann, and M. Sandler. Musiclynx: Exploring music through artist similarity graphs. In *Companion Proceedings of the The Web Conference*, 2018.
4. N. U. "Antoine Bordes and al". Translating embeddings for modeling multi-relational data. pages 2787–2795, 2013.
5. T. "Bertin-Mahieux and al". The million song dataset. pages 591–596, 2011.
6. M. BUFFA and al. The wasabi dataset: cultural, lyrics and audio analysis metadata about 2 million popular commercially released songs. 2021.
7. U. G. "Jochen Hipp and G. Nakhaeizadeh". Algorithms for association rule mining – a general survey and comparison. In *ACM SIGKDD Explorations Newsletter*, 2000.
8. R. KHULUSI, J. KUSNICK, C. MEINECKE, and al. A survey on visualizations for musical data. In *Computer Graphics Forum*, pages 82–110, 2020.
9. C. F. Leandro Guedes. Exploring music rankings with interactive visualization leandro. In *Proceedings of the Symposium on Applied Computing*, pages 214–219, 2017.
10. J. Lehmann and al. Defacto – deep fact validation. In *International semantic web conference. Springer, Berlin, Heidelberg*, pages 312–327, 2012.
11. J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *Proc. of the 11th International Society for Music Information Retrieval Conf.*, pages 625–636, 2010.
12. E. M. Shuangyan Liu, Mathieu d'Aquin. Towards linked data fact validation through measuring consensus. In *CEUR Workshop Proceedings*, 2015.
13. J. Sleeman and T. Finin. Type prediction for efficient coreference resolution in heterogeneous semantic graphs. In *IEEE 7th International Conference on Semantic Computing*, pages 78–85, 2013.
14. H. H. Wu and J. P. Bello. Audio-based music visualization for music structure analysis. In *Proceedings of the 7th Sound and Music Computing Conference*, 2010.