Tools for building and evaluating new MIR datasets

Roman Tsukanov
Supervisor: Dmitry Bogdanov

Sound and Music Computing
Universitat Pompeu Fabra

Motivation

Datasets are very important part of research in Music Information Retrieval. Unfortunately, their quality is not good enough for “real-world” tasks.

AcousticBrainz project showed this problem. Users have much more extensive collections of music compared to what is available within the MIR community.

Most datasets contain no more than a 1,000 recordings and their structure has problems (for example, not enough genre labels).

Goals

Find different ways to allow people to improve quality of datasets.

Create an open framework with tools for creation. Evaluate improvements by measuring accuracy of models produced from these datasets or using some other metric.

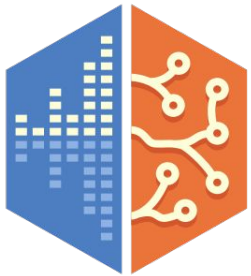
We already have a basic version of dataset creation and evaluation tool *(introduced at ISMIR 2015)*. Provides a way to create collections of recordings and generate models for extracting high-level information like genre or mood.

Methodology

We can try several things:

- Streamline process of dataset creation
- Let people collaborate and improve existing datasets
- Make ready-to-use labeled lists of recordings
- Add a way for creators to gather feedback about quality of their datasets
- Organize “dataset challenges”

All parts of the project are meant to be open and freely available. This includes implementation and data.



AcousticBrainz

http://acousticbrainz.org/

https://github.com/metabrainz/acousticbrainz-server
https://github.com/MTG/acousticbrainz-client
https://github.com/MTG/acousticbrainz-gui