Overview of MusiCLEF 2011

Nicola Orio[1] and David Rizo[2]

[1] Department of Information Engineering
University of Padova, Italy
orio@dei.unipd.it
[2] Department of Software and Computing Systems
University of Alicante, Spain
drizo@dlsi.ua.es

Abstract. MusiCLEF is a novel benchmarking activity that aims at promoting the development of new methodologies for music access and retrieval on real public music collections. A major focus is given to multimodal retrieval that, in the case of music collections, can be obtained by combining content-based information, automatically extracted from music files, with contextual information, provided by users via tags, comments, or reviews. Moreover, MusiCLEF aims at maintaining a tight connection with real application scenarios, focusing on issues on music access and retrieval that are faced by professional users. To this end, the benchmark activity of the first year of MusiCLEF focused on two main tasks: automatic categorization of music to be used as soundtrack of TV shows and automatic identification of the digitized material of a music digital library.

1 Introduction

The increasing availability of digital music accessible by end users is boosting the development of Music Information Retrieval (MIR), a research area devoted to the study of methodologies for content- and context-based music access. As it appears from the scientific production of the last decade, research on MIR encompasses a wide variety of different subjects that go beyond pure retrieval: the definition of novel content descriptors and multidimensional similarity measures to generate playlists; the extraction of high level descriptors – e.g. melody, harmony, rhythm, structure – from audio; the automatic identification of artist and genre. As it is well known, the possibility to evaluate the different research results using a shared dataset has always played a central role in the development of information retrieval methodologies, as it is witnessed by the success of initiatives such as TREC and CLEF, which focus on textual documents.

The same need has been perceived in MIR, motivating the development of an important evaluation campaign, the Music Information Retrieval Evaluation eXchange (MIREX). MIREX campaigns[3] are organized since 2005 [1] by the International Music Information Retrieval Systems Evaluation Laboratory

[3] http://www.music-ir.org/mirex

(IMIRSEL) at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Due to the many limitations posed by the music industry, the organizers of the MIREX chose to distribute only publicly available test collections. Participants are in charge to create their own collection and after local experimentation submit their software that is run by the organizers. This approach has two drawbacks, which have already been debated by the MIR research community: the results of previous campaigns cannot be easily replicated and the results depend on the individual training sets and not only on the submitted algorithms.

A recent relevant initiative, that aims at overcoming the limitations imposed by not sharing the datasets between researchers, is the Million Songs Dataset (MSD). Thanks to MSD[4], researchers can access a number of features from a very large collection of songs [2]. Unfortunately, the algorithms used to extract these features are not public, limiting the possibility to carry out research on content description techniques. Another ongoing initiative related to the evaluation of MIR approaches is the Networked Environment for Music Analysis (NEMA), that aims at providing a web-based architecture for the integration of music data and analytic/evaluative tools[5]. NEMA builds upon the achievements of MIREX campaigns regarding the evaluation of MIR approaches, with the additional goal of providing tools for resource discovery and sharing.

Within this scenario, MusiCLEF is an additional benchmarking initiative, that has been proposed in 2011 as part of the activities of the Cross-Language Evaluation Forum (CLEF). CLEF focuses on multilingual and multimodal retrieval[6] and gathers researchers in different aspect of information retrieval, ranging from plagiarism and intellectual property rights to image retrieval.

The goal of MusiCLEF is to promote the development of novel methodologies for music access and retrieval, which can combine content-based information, automatically extracted from music files, with contextual information, provided by users through tags, comments, or reviews. The combination of these two sources of information is still under-investigated in MIR, although it is well known that content-based information alone is not able to capture all the relevant features of a given music piece (for instance, its usage as a soundtrack or the year of release), while contextual information suffers from the typical limitations for new items and new users (also known as cold start).

Aiming at investigating and promoting research on the combination of textual and music information, MusiCLEF has a strong focus on multimodality that, together with multilingualism, is the main objective of the CLEF evaluation forum. Moreover, the tasks proposed for MusiCLEF 2011 are motivated by real scenarios, discussed with private and public bodies involved in music access and dissemination. In particular, MIR techniques can be exploited for helping music professionals to describe music collections and for managing a music digital library of digitized analogue recordings. To this end, the organizers

[4] http://labrosa.ee.columbia.edu/millionsong/
[5] http://www.music-ir.org/?q=nema/overview
[6] http://clef-campaign.org/

of MusiCLEF exploited the ongoing collaborations with both a company for music broadcasting services (LaCosa s.r.l.) and a public music library (University of Alicante's Fonoteca).

Two tasks are proposed within MusiCLEF 2011, and both are based on a test collection of thousands of songs in MP3 format. To completely overcome copyright issues, only low-level descriptors will be distributed to participants. Figure 1 depicts the tasks workflow of MusiCLEF, which is described in more detail in the following sections.

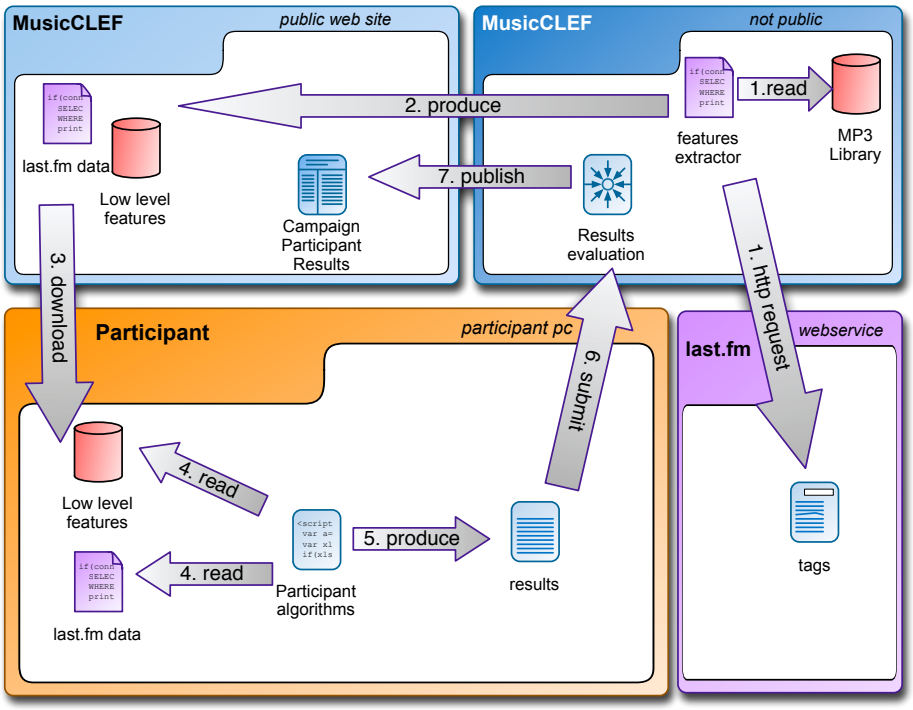

Fig. 1: Task workflow in MusiCLEF.

2 Professional Partners

A major goal of MusiCLEF is to maintain a tight connection with real application scenarios, in order to promote the development of techniques that can be applied to solve issues in music accessing and retrieval that are faced by professional users. The choice of focusing on professional users is motivated by the fact that they need to address a number of real-life issues that are usually not taken into

account by music accessing systems aimed at the general public. At the same time, the evaluation of the effectiveness of the proposed automatic solution is easier to assess, because professional users have a clear idea of what are their information needs.

2.1 LaCosa s.r.l.

LaCosa was founded as a service provider of the major TV broadcasting – public and private – companies in Italy with the goal of managing and describing a large music collection of songs to be used for TV programs, including jingles, background and incidental music, and music themes for TV shows. LaCosa has a strong cooperation with RTI, a company that, apart from buying and storing songs issued by the major record companies, produces its own music catalogue. At present, RTI library contains about 320,000 songs of pop-rock, jazz, and classical music. Besides playing the role of music consultant, being one of the biggest private music repositories in Italy, RTI offers a number of services to external companies of music consultants, who can browse remotely the repository. Audio features distributed to the participants are thus extracted remotely, without downloading the audio files.

The typical job of a music consultant is to select a list of songs that are suitable for a particular application, for instance a TV commercial, the "promo" of a new program, the background music for a documentary, and so on. The availability of large online collections, such as Last.fm and YouTube, is representing an alternative to the services of a music consultant. For instance, journalists are increasingly selecting by themselves the music for their news stories, instead of asking to music consultants. The goal of LaCosa is then to provide high quality descriptions, that are tailored to the particular application domain, in order to represent still a more interesting alternative to free recommendations.

Given these considerations, the requirements of LaCosa can be summarized as follows: How to improve the acquisition process, extracting the maximum amount of information about music recordings from external resources? How to provide good suggestion about possible usages of music material, minimizing the amount of manual work?

Because of the interest on the development of automatic systems for addressing these two requirements, LaCosa decided to provide at its own expenses a number of assessors to create the ground truth for evaluation. The involvement of professional users included also the definition of a vocabulary of 167 terms describing music genre (terms are organized in two levels, genre and subgenre), and of 188 terms describing the music mood. It is important to note that, in this case, the concept of mood is related to the usage of a particular song within a video production. As explained in more detail in Section 3.1, only a subset of the mood terms have been used in the evaluation campaign.

2.2 Fonoteca de la Universidad de Alicante

Some years ago, the local radio broadcast station *Radio Alicante Cadena Ser* transferred its collection of vinyls to the Library of the University of Alicante. This collection contains approximately 40,000 vinyls of an important cultural value, containing a wide range of genres. The library decided to digitize the vinyls, sound and covers, to overcome the preservation problems when allowing library users to access the discs and to enable its reproduction embedded in the library's Online Public Access Catalog (OPAC) with the name *Fonoteca*[7].

The process was carried out following library cataloguing techniques to make the inventory of the collection. Vinyls were catalogued using Universal Decimal Classification, and classified into subjects based on the Library of Congress subject headings. Digitized covers and audio were linked to the corresponding records. The cataloguing data consists of the album's title, the name of the discographic company, the release year, its physic description, several entries for genres classified manually by the cataloguers, and finally notes about the content. Regarding the sound content, each vinyl was digitized in two files, one for each side. For 45 rpm discs each side usually contains only one song, while for 33 rpm LPs, which are more common in the collection, each side contains several tracks.

Having catalogued and digitized the material, some drawbacks emerge that strongly limit the browsing capabilities in the OPAC. The separation of tracks from a continuous stream could be easily solved in most cases just by finding silences between tracks. However, this may not be the case for live recordings or classical music tracks, where the music itself contains long rests. A related problem is the correct entitling of the tracks. Although some catalogued albums contain details of the contained tracks, there are many others, mainly operas, where the track names are not present. Another common situation is that of finding two different recordings of the same work whose tracks have been labeled using two different languages or naming schemes, e.g., "Symphony No. 9" knowns as "Novena Sinfonía" as well as "Choral Symphony". Audio fingerprinting techniques can hardly be applied to solve this task because of disc age, besides the fact that some of the discs may not have been reissued on CD and thus may not have been included in any audio fingerprint dataset.

Besides these drawbacks, the staff of the library demands some features that cannot be implemented given the current structure of the data. For example, given an album, find it in music sites like *Last.fm* or *Grooveshark*. Similarly, find a given song/track and its different recordings in those music sites and inside the library regardless of language or naming schemes. In order to locate music, they want the users to be able to query the library given metadata not contained in the catalog, like the lyrics of the songs.

[7] `http://www.ua.es/en/bibliotecas/SIBID/fonoteca`

3 The Benchmark Activities

Each of the two professional partners motivated a particular task that has been organized within MusiCLEF 2011. The tasks can be summarized as follows:

- *Automatic categorization* of pop and rock music, for its use in TV broadcasts, made in cooperation with LaCosa.
- *Automatic identification* of classical music, for the creation of computer assisted bibliographic records, made in cooperation with the Fonoteca.

Both tasks that are typical of the MIR research area. On the one hand, the categorization of music documents is related to *automatic tagging* (or autotagging). Tags are short free-text descriptions of multimedia items, which are shared through web-based systems. These descriptions are usually provided by end-users and have been often exploited for automatic recommendation systems. One the other hand, the identification of unknown recordings is related to *cover identification*. A cover is usually an alternative version of a previously released song, and the term is particularly used in pop and rock genres. The automatic identification of alternative versions of a given song can be used for intellectual property management or simply to provide the end-user with additional information about the music he is accessing to.

3.1 Automatic Categorization

The goal of the first task is to exploit both automatically extracted information about the content and user generated information about the context to carry out categorization. The application scenario that motivates this task is the following:

> Songs of a "commercial music library" need to be categorized according to their possible usage in TV and radio broadcasts, for instance as commercials, soundtracks and jingles.

For professional applications it is common practice to use different sources of information to assess the relevance of a given song to a particular usage. At first candidate songs are selected depending on the result of Web searches and on the analysis of user-generated tags. Since these sources of information are usually very noisy, experts make the final choice depending on the actual music content.

The dataset made available to participants of MusiCLEF 2011 is composed of 1355 different songs, played by 218 different artists; each song has a duration between 2 and 6 minutes. In order to simulate this scenario, participants of MusiCLEF were provided with three different sources of information for each of the songs in the dataset:

- Content descriptors, extracted directly from audio;
- User tags, downloaded from the Web-service of an Internet radio;
- Web pages, related to the artist that performed the song.

Since CLEF campaigns aim at promoting multilingualism, tags could be in any language (although most of the music terms for pop and rock genres are in English) while Web pages have been downloaded using queries in five different languages (English, Italian, German, French, and Swedish).

As regards the ground truth to evaluate the different systems, we exploited the collaboration with LaCosa (see Section 2.1). Each song was manually categorized by music professionals, using at least one term for genre and five terms for mood. The vocabulary of tags defined by the experts was initially composed of 355 tags divided in two categories – genre (167) and mood (288) – loosely inspired by the Music Genome Project[8]. At the end, we discarded all the tags that were assigned to less than twenty songs; this led to the final released vocabulary of 94 tags.

Songs were grouped in two sets: a training set of 975 song, available to participants, and test set of the remaining 380 songs, for the final evaluation.

3.2 Identification of Classical Music

The second task gives an higher importance to content descriptors, which are used to automatically provide contextual information. The real-life scenario that has been considered for this task is the following:

> A "music digital library" contains a set of loosely labeled digital acquisitions of old analogue recordings of classical music; recordings should be automatically annotated with metadata, such as composer, title, movement, excerpt.

It has to be noted that systems for automatic music identification already give good results, at least in terms of effectiveness (efficiency being still an issue). Yet the combination of segmentation and identification of continuous recordings is not well investigated yet.

To this end participants were provided by a dataset divided in two parts: 6680 music files containing single music pieces (e.g., a Sonata in four movements is contained in four different files) and 22 music files each one containing a side of LP. Hence this task was split in two subtasks: to identify the pieces that are alternative recordings of the same work (for single files) and to match the content of the LP recordings with the corresponding songs (for digital acquisitions).

Also for this task, participants of MusiCLEF were provided with different sources of information for each of the songs in the dataset

- Content descriptors, extracted directly from audio;
- Metadata regarding title, movement and composer for the single files;
- Information reported in the LP covers for the digital acquisitions.

It is interesting to note that metadata descriptors are mostly in Italian, while the information on the LP covers is mostly in Spanish.

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

The ground truth was already available for the single recordings, because all the files have been catalogued by music experts, while the correspondence of the cover information for the LPs and the actual content has been manually inspected by the organizers.

The training set contained: 661 individual files organized in groups, considering that two songs were in the same group if they were alternative recordings of the same music work and 3 files with LP acquisitions. The test set used to evaluate the different systems contained 600 additional single files and the remaining 19 digital acquisitions.

4 Conclusions

MusiCLEF is a new benchmarking activity that aims at fostering content- and context-based analysis techniques to improve music information retrieval tasks, with a special focus on multimodal approaches. We reported on the motivation and on the organization of tasks. At the time of writing evaluation is still ongoing, thus results will be available directly at the CLEF conference.

5 Acknowledgments

The authors are grateful for the support of the staff of LaCosa s.r.l. and the University of Alicante's Fonoteca. MusiCLEF has been partially supported by Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191. CLEF is an activity of PROMISE. This research is also supported by the Spanish Ministry projects DRIMS (TIN2009-14247-C02-02) and Consolider Ingenio MIPRCV (CSD2007-00018), both partially supported by EU ERDF.

References

1. Downie, J.S., West, K., Ehmann, A.F., Vincent, E.: The 2005 Music Information retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. In: Proc. of ISMIR. (2005)
2. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: Proc. of ISMIR. (2011)