Towards an Integrated Approach to Music Retrieval

Emanuele Di Buccio[1], Ivano Masiero[1], Yosi Mass[2], Massimo Melucci[1], Riccardo Miotto[1], Nicola Orio[1], and Benjamin Sznajder[2]

[1] Department of Information Engineering – University of Padova
Padova, Italy
{dibuccio,masieroi,melo,miottori,orio}@dei.unipd.it
[2] IBM Research Lab
Haifa, Israel
{yosimass,benjams}@il.ibm.com

Abstract. This paper describes a research work on peer-to-peer music search based on the combination of content based descriptors and textual metadata. The envisaged scenario is the one of a user who searches for documents according either to their melodic and rhythmic content, and to additional information about the title, the instrumentation, or tempo in the form of textual metadata. Two different overlay networks are used to deal with music content and text. A user interface has been developed, which allows the user to perform a query, to merge the results using alternative approaches, and to listen to the retrieved music documents.

1 Introduction

The main goal of SAPIR[3] (Search in Audio-visual content using Peer-to-peer Information Retrieval), a project funded by the EU, is the development of a large-scale, distributed peer-to-peer (P2P) infrastructure to allow users searching in audio-visual content using a *Query By Example* paradigm. According to this approach, the user queries the system using an example of what he is looking for, possibly with additional metadata in order to better describe his information need. SAPIR aims at providing large scale search capabilities in P2P network for different types of media. The SAPIR vision is illustrated in Fig. 1 and it includes a media analysis frameworks for different media – images, video, speech, music, and text – together with scalable and distributed P2P index structures supporting similarity search and support for multiple devices embedding social networking in a trusted environment. In this paper we focus on a particular component of the SAPIR architecture, that is a music search based on content descriptors and textual metadata.

Music search is increasingly gaining interest because of the wide diffusion of music files in P2P networks. Yet, few P2P architectures allow for a content based search of music files. One major problem in music searching over a P2P network

[3] http://www.sapir.eu

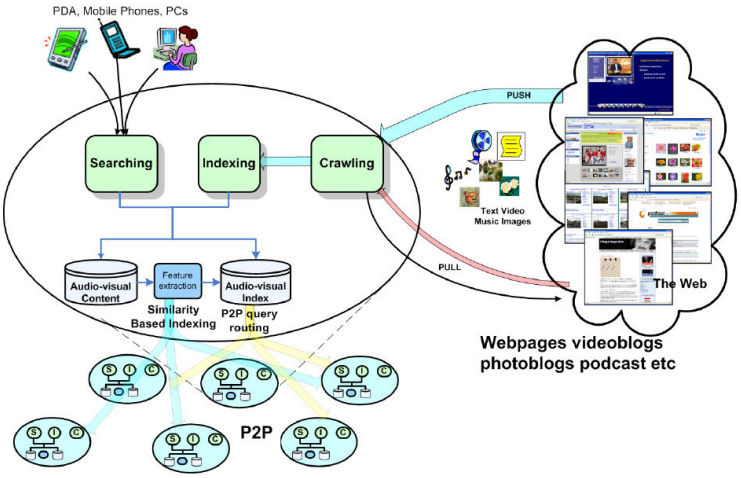

Fig. 1. SAPIR Components and Functions

is, as in the general case, the lack of knowledge of the content location. To overcome this problem, the work reported in [1] proposes a DHT-based system applied to music retrieval. The system exploits both manually specified tag-like information – e.g. artist, album, title – and automatic feature extraction techniques to search by content.

Even though structured networks allow for efficient query routing, they require an high degree of collaboration among peers. The latter may not be a suitable solution for networks that are highly dynamic, heterogeneous, or protective of intellectual property. This kind of networks is well-matched by unstructured overlays. To this end, two P2P music retrieval systems for unstructured networks have been proposed, in [2] and in [3] respectively. Yet, both approaches lack in terms of P2P searching algorithms: the former routes the query to all the peers, whereas the latter exploits a breadth-first search algorithm. The approach proposed in [4] utilizes a centralized coordinator that stores the identifiers of all the peers in the network, and for each peer also its PC-feature, that is a feature which describes the music content of the peer and that is used to select the most promising peers to answer to a given query. A more efficient solution in terms of network traffic can be obtained by decreasing decentralization.

In this paper we present an approach to P2P music search based on the combined use of descriptive metadata and content features. The approach has been developed as part of the SAPIR project, according to one of the envisaged scenarios – called *music and text* – where a user interacts with the system using a query by example paradigm but is also able to provide additional information about his information need in the form of textual metadata.

2 Description of Music Documents

We propose to retrieve music documents through the integration of two alternative descriptions: textual metadata and content features. Both descriptions are automatically extracted from the documents themselves, without using external sources. The approach has been developed and tested on a collection of about 60,000 music files in MIDI format [5].

MIDI files have been downloaded through a focused crawling on a number of Web sites that granted free access for non commercial usages. Files are typically provided by end users and thus there is no control on their quality. Moreover, there could be a replication of the same songs, sequenced and provided by different users, often with slightly different names. Exact copies may also be present in different Web sites, although this happened for less than 5% of the files. Given all these characteristics, the music collection can be considered a good approximation of the content that can be found in typical P2P network.

2.1 Textual Metadata

Two kinds of metadata can be found in MIDI files: textual metadata and coded information useful for the representation of the music score. Textual metadata is directly embedded in the file format by the user who sequenced the file, in order to provide additional information about the file content. It is important to notice that this information is not displayed on screen by most of the available music players, and thus is used freely by the users and sometimes is automatically added by the sequencing software. Textual information is structured in a number of fields that are defined by the MIDI standard, which includes: the name of the different music tracks that may be present in the file, the name of the instruments associated to each track, copyright information, and the markers that are added to a given position in the music score – which are often used to insert the lyrics.

Unfortunately, most of the MIDI files available on the Web are provided by end users on a voluntary basis, and thus there is no guarantee that the different fields for structuring metadata are used consistently. For instance, an analysis on some samples of our music collection showed that the copyright field may contain either the name of the artists, or the name of the software used to generate the file, or even the URL of the Web site where the file has been made available. Analogously, the information about the tracks can be used either to represent the name of the performer in the original recording, or the name of the played instrument, or other information unrelated to the track.

The information about artist and title, which is usually not included in the MIDI format – there is actually no field to represent information about title and artist in the MIDI standard – has been extracted by analyzing the structure of the Web pages containing the links to the MIDI files and the text surrounding the links to the music files. In about 90% of the cases we were able to identify either the title alone or the title and artist. Being provided by the Web sites, the information about title and artist is likely to be more reliable than textual metadata added by the user.

The second kind of metadata that can be extracted from MIDI files regards information strictly related to the music language. It can be computed through the automatic analysis of the music content of the files and thus is somehow more reliable than textual metadata. We choose to extract some general information about the main features of a music piece, including: tonality (e.g., C major or E flat minor), time division (e.g., 4/4 or 3/4) and tempo in beats per minutes. The latter information has been represented using a simple perceptual scale from "very slow" to "very fast". Moreover, the information about the sound samples to be used during playback – it is worth noting that all MIDI players synthesize the music by using a bank of internal sound samples – has been used as an alternative description of the music instruments that are associated to each track (in general, this information can be different from the one manually provided by the user as textual metadata).

The use of music terms may not be particularly easy for a user without a music education. The fact that a song is in a given tonality or has a particular time division may not give additional information about the relevance to a given information need. Nevertheless, there are applications where such information can be relevant. For instance, in case that professional user is interested to create a playlist of songs or to mix different songs to create a new music product, or when a musician is interested in retrieving songs that are more suitable for the tonality of his instrument.

2.2 Content-based Descriptors

The approach to content description of music documents is based on the use of high level features, which should be perceptually relevant for the final users. According to the approaches presented in the literature on music information retrieval, the most relevant content descriptors are the rhythm and the melody of the leading voice. This information is readily available in the case of MIDI files, while it can be computed from audio documents using signal processing techniques achieving an accuracy of more than 85% in onset detection and note description. In this paper, we focus on the use of MIDI format.

The first step regards the automatic identification of the track containing the main melody, which has to be analyzed in order to extract the relevant melodic information. It has to be noted that a MIDI file contains a number of tracks, related to the different instruments that are employed in the overall music score. The approach to identify the main track, which achieves an accuracy of 97.7%, is described in [6]. In case the selected track was polyphonic, it has been transformed to monophonic using the approach described in [7].

The melodic information has been quantized, in order to take into account local variations of pitch and tempo. To this end, the fundamental frequency values has been quantized to the 12 semitones of the chromatic scale used in Western music, while rhythm has been normalized using state of the art techniques for music transcription. After quantization and normalization, melodic information was described by two main parameters:

– Pitch intervals (PIT): that is the distance between two subsequent tones.
– Interonset interval log-ratio (IOI): that is the ratio, in logarithmic scale between the duration of two subsequent tones.

The second step regards the segmentation of the extracted melody in musical lexical units, which are used as content descriptors. To this end, pattern analysis techniques have been applied to the sequence of symbols forming either the melody or the rhythm, highlighting all different patterns with a length from 3 to 6 notes, as described in [8]. These thresholds have been experimentally evaluated with the test collection available for the Music Information REtrieval Evaluation eXchange campaign [9].

The final step in the representation regards a suitable coding of the patterns, which differs according to the features to be represented. Given that all features undergo quantization, a textual representation is used to describe the patterns, by assigning a different symbol, taken from a small alphabet, to all the elements in melodic or rhythmic music patterns. The size of the alphabet depends on the choice of the quantization step. Commonly used values are 15 symbols for PIT and 9 symbols for IOI. The relative frequency with which a pattern occurs in a documents has been computed as well for each feature.

3 Music Retrieval in SAPIR

The information extracted from music documents is indexed using two different SAPIR overlays: a text overlay developed by IBM, and a content based music overlay, called SPINA, developed at the University of Padova [10]. The two overlays index the same collection of music documents in MIDI format. The file URI is used as unique identifier for the music documents.

Document content is represented using a common framework. The SAPIR approach aims at making use of standards concerning metadata representation as well as content analysis methods: metadata are expressed in XML derived from MPEG-7 (ISO/IEC 15938), while the UIMA framework [11] has been chosen to analyze content and extract its relevant features. Additional information about content analysis and representation can be found in [12]. Music information is represented as a set of melodic and rhythmic patterns, as described in Sect. 2.2. The textual metadata extracted from the music files is indexed using the structured fields as described in Sect. 2.1 and can be retrieved either by using the structure or as free text. Given the considerations made in Sect. 2.1 the UI is more oriented towards queries with unstructured text.

In order to use the textual information as real metadata, in the present implementation the user has to format the query using a XML-like representation, which is a simplified version of MusicXML [13]. For instance, if a user is interested in searching music documents played with a fast tempo and where the electric guitar is one of the instruments, he can either explicitly express his information need with a structured query like *<tempo>fast</tempo> <instrument>electric guitar</instrument>* or simply using a free text query such as *fast electric guitar*.

A stand-alone Java application implementing a SAPIR peer has been developed, providing a user interface (UI) for creating complex queries that combine music content and textual metadata. The UI allows the user:

- To upload a music file to be used as an example of his information need.
- To insert a textual query, possibly using the musical terms that describe the music content (e.g., tonality and tempo)
- To combine the content based description with the textual metadata.

The peer is connected to a SPINA super peer that acts as the entry point to the SPINA overlay and to a text overlay. Once a user creates a query using the interface of the peer, the query is parsed by a component called *query analyzer*, which parses the XML representation of the combined query – music content plus textual metadata (in the terminology proposed in SAPIR these are two *FeatureGroups*) and sends the music content features to the SPINA overlay and the textual part to the text overlay. It is important to note that SAPIR architecture allows for more complex combinations of features. This characteristics could be exploited in case the music overlay were combined with other media, such as e.g. video.

3.1 Merging of the Results

Each overlay processes independently the textual and music components of the query. The individual results can be merged using different strategies. For the particular application of music retrieval, one possible scenario is of a user who is mainly interested in retrieving music documents that are similar to the one provided as a query, while the textual information can be used only to refine the results. Alternatively, the user could be interested in obtaining two different lists of potentially relevant documents, one for the music features and one for the textual part, and a third list with the results merged together. A screenshot of the music and text peer is shown in Fig. 2.

Given these considerations, at the moment the SPINA peer presents the retrieval results through four different lists of documents:

1. A list of music documents retrieved by the SPINA music overlay alone, with additional information about the peer that actually contains the document; the user can download and listen to the music file by simply clicking on an icon, because the SPINA peer can directly contact the peer containing the file.
2. A list of music documents retrieved by the text overlay; in this case it is not possible to listen to the music files, which are stored as XML documents containing the metadata; a possible extension of the functionalities may regard the possibility to access to the textual metadata, in order to have a better description of the file content.
3. A merged list of music documents, retrieved either by the SPINA and the text overlays. The list is obtained by applying data fusion techniques. Given that in this list there could be documents retrieved only by the text overlay, the user can listen only to a limited number of documents in this list.

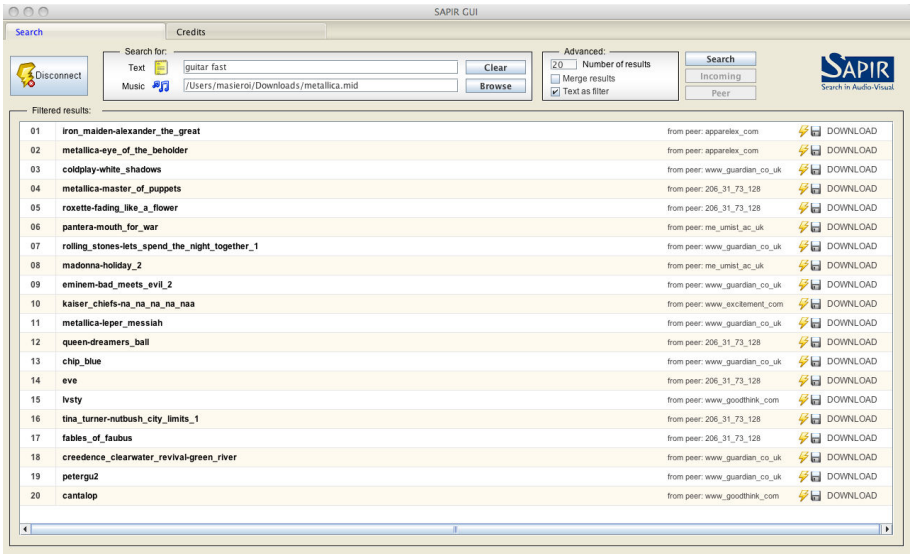

Fig. 2. Screenshot of the user interface for the music and text peer

4. An alternative merged list, where the documents retrieved by the SPINA overlay are reranked according to the results of the text overlay. In this case, which is probably the most useful for a user who needs to evaluate the results by listening to the retrieved documents, the user can listen to all the documents in the list.

The fact that a user is presented with a single rank list, including the query results on the two overlays as for case 3 and case 4, is an important step towards an integrated approach to retrieve music documents. For case 3, it can be noted that, being based on different retrieval schemes, the two overlays may retrieve different music documents and the data fusion approach allows for giving a high rank to music documents that are both similar to the music query and relevant for the textual query. At the same time, also music documents that are particularly relevant either for their content or for their metadata can have a high final rank. For case 4, the reranking approach of documents retrieved by the SPINA overlay allows the user for a content based search where additional textual information can be used to give an higher rank to documents with the required metadata.

4 Conclusions

The approach has been tested using a collection of about 60,000 MIDI files. Some files were replicas of existing ones, in order to simulate a P2P network. The approach showed to be scalable in the number of documents. Further analysis

need to be carried out on the effectiveness of the content based description of music documents. Preliminary results using a centralized system and a small test collection of about 600 music documents in MIDI format showed that the method for music retrieval can achieve a mean average precision of 0.54. Future work will address the effects of a distributed collection on retrieval effectiveness.

Acknowledgments

This work was partially supported by the SAPIR project, funded by the European Commission under IST FP6 (Sixth Framework Programme, Contract no. 45128). The authors thank Maristella Agosti, Nicola Ferro and Giorgio Di Nunzio for their valuable support in the development of the proposed approach.

References

1. Tzanetakis, G., Gao, J., Steenkiste, P.: A scalable peer-to-peer system for music content and information retrieval. In: Proceedings of the International Conference on Music Information Retrieval. (2003)
2. Yang, C.: Peer-to-peer architecture for content-based music retrieval on acoustic data. In: Proceedings of the International Conference on World Wide Web. (2003) 376–383
3. Karydis, I., Nanopoulos, A., Papadopoulos, A., Manolopoulos, Y.: Musical retrieval in p2p networks under the warping distance. (2006)
4. Wang, C., Li, J., Shi, S.: A kind of content-based music information retrieval method in peer-to-peer environment. In: Proceedings of the 3rd International Conference on Music Information Retrieval. (2002)
5. Rothstein, J.: MIDI: A comprehensive introduction. A-R Editions, Madison, WI (1991)
6. Orio, N., Zen, C.: Song identification through hmm-based modeling of the main melody. In: Proceedings of International Computer Music Conference. (2007) 248–251
7. Uitdenbogerd, A., Zobel, J.: Manipulation of music for melody matching. In: Proceedings of the ACM Conference on Multimedia. (1998) 235–240
8. Neve, G., Orio, N.: Indexing and retrieval of music documents through pattern analysis and data fusion techniques. In: Proceedings of the International Conference on Music Information Retrieval. (2004) 216–223
9. Mirex-2006: Second annual Music Information Retrieval Evaluation eXchange (July 2006) `http://www.music-ir.org/mirex2006/`.
10. Di Buccio, E., Ferro, N., Melucci, M.: Content-based information retrieval in SPINA. In: Proceedings of the Italian Research Conference on Digital Library Systems. (2008)
11. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering **10**(3-4) (2004) 327–348
12. Allasia, W., Falchi, F., Gallo, F., Kacimi, M., Kaplan, A., Mamou, J., Mass, Y., Orio, N.: Audio-visual content analysis in p2p: the sapir approach. In: Proceedings of the AEIMPro08 Workshop. (2009)
13. MusicXML: Recordare: Internet music publishing and software (December 2008) `http://www.musicxml.org/`.