Design of an Information Retrieval System Based on the Peer-to-Peer Paradigm: An Application to Music Retrieval

Emanuele Di Buccio, Nicola Ferro, Massimo Melucci, Riccardo Miotto and Nicola Orio

Department of Information Engineering, University of Padua, Italy
{dibuccio,ferro,melo,miottori,orio}@dei.unipd.it

Abstract. The Peer-To-Peer (P2P) paradigm is a good choice for providing federated search capabilities to collections of documents spread across Internet or to Digital Libraries. In traditional P2P networks, search operations focus on properly labeled files, and the search is often limited to textual metadata. The explosive growth of available multimedia documents in recent years called for more flexible search capabilities, namely search by content. In this paper we present a novel P2P architecture to provide a distributed content-based multimedia search engine where the search operations exploit some features related to the content of the documents rather than their metadata. The proposed system aims at retrieving music and text documents.

1 Introduction

Peer-To-Peer (P2P) networks integrate autonomous computing resources without requiring a central authority: the basic rationale is that peers are entities able to work as both client and server. P2P is a good choice to provide federated search capabilities to collections of documents spread across Internet or to Digital Libraries. In the beginning, the popularity of the P2P paradigm was due to the diffusion of applications for distributing and sharing digital documents, in particular music files. Among all the proposed systems in late 90s', Napster can be certainly considered the most famous one. In general, the weaknesses of this kind of systems concerned both the scalability and the search capabilities and during the last decade the problem of Information Retrieval (IR) across P2P networks was widely investigated. Although, a considerable part of the usage performed in P2P networks concerned music files sharing, the great part of the proposed solutions allowed to search music information by metadata only. In different scenarios, metadata could be either not suitable, or unreliable or even missing. Moreover, as it is well known, providing metadata for large collections is an extremely time consuming task and could also entail the problem of multilingual access to the documents.

The design of a system for Music Information Retrieval (MIR) across P2P networks should be performed both at modeling and architectural level. In [1] a

weighing framework for addressing the design of a P2P-IR system at modeling level was proposed. The following experimental evaluation reported in [2] showed that the scheme helps the retrieving of a significant proportion of relevant data after traversing only a small portion of a P2P hierarchical network in a depth-first manner. In [3], then, the same problem was addressed at an architectural level by proposing the Superimposed Peer Infrastructure for iNformation Access (SPINA) software architecture which allows to index and retrieve unstructured documents distributed across a P2P network.

MIR approaches [4] provide different methodologies of music processing and retrieval by exploiting some features related to the music content rather than the documents metadata. These approaches are very challenging and aimed at satisfying the need of the users which, for instance, could prefer to retrieve music documents by humming a little part of the melody or by submitting as query a fragment of an audio file.

In this paper we describe an ongoing capability of the SPINA architecture in order to automatically index and retrieve music documents by content. The automatic content-based retrieval of music documents is gaining increasing interest because it can provide new tools for music accessing and distribution. These can be exploited in several contexts such as recommendation systems, digital libraries population, Web searches, detection of copyright infringement and so on. In the last years, different music identification approaches have been proposed and, in particular, more recently in [5], [6] and [7]. In our work, in particular, the general ideas proposed in [7] has been extended to index and retrieve music files by content in a P2P network.

In the following sections, we describe the components of the system, ranging from the architecture designing to the music content indexing.

2 Design of a P2P-IR System

In the last decade the problem of text retrieval across P2P networks was widely investigated [8, 9]. On the contrary, there are only few solutions for the problem of P2P-MIR.

A Distributed Hash Table (DHT)-based system to retrieve music documents was proposed in [10]. 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 — e.g. DHT-based — enable for efficient query routing, they require an high degree of collaboration among peers, requirement which is not suitable for networks that are highly dynamic, heterogeneous, or protective of intellectual property. This kind of networks is well-matched by *unstructured* overlays. P2P-MIR systems for unstructured networks were proposed in [11] and in [12], but both approaches lack in terms of network traffic. A more efficient solution may be achieved by decreasing decentralization. The approach proposed in [13] exploits a centralized coordinator that stores the identifiers of all the peers in the network together with its own *PC feature*, which is a feature describing the music content of the peer, used to

select the most promising peers to answer to the formulated query. As suggested by the authors, the architecture might be improved in terms of efficiency and robustness by increasing the numbers of coordinators.

The type of overlays adopted by the software architecture described in this paper does not require the presence of fixed central entities. Indeed, some peers are *dynamically* elected as ultra-peers or super-peers. Because of the presence of ultra-peers this kind of overlays is named as hybrid. Moreover, the adopted overlay is a particular type of unstructured hybrid network because it is also hierarchical: indeed each peer refers to one and only one ultra-peer. The presence of ultra-peers, which act as hubs, enables to decrease the number of messages during query routing. The adoption of unstructured networks might be a suitable solution also because, as pointed out in [12], in such overlay topologies each peer can share just its own resources without keeping information about the resources of the other peers. This approach may be also exploited to localize peers that share illegal content or files which violate copyright.

The system designed aims at being independent not only from the underlying network infrastructure, but also from the media of the documents stored in the network. An API called SPINA has been designed and implemented [3] in order to achieve a flexible software architecture whose functionalities enable to search text and music documents by content. In the system design, each peer is provided with a local search engine that supplies all the functionalities required to perform the indexing and retrieval operations on the local collection of documents. Each peer locally stores two indexes, one for information about the text collection and the other one for the music documents.

Besides providing these functionalities, the system provides a retrieval process across the P2P network. The rest of the section describes the rationale of the search process and how to represent and select resources at higher levels starting from the statistics locally extracted by the peers about their content.

2.1 Search and Query Routing

Resource selection in P2P systems is related to the task of query routing because of the topology of the network. For this reason, in this section the adopted overlay topology and the query routing mechanism will be briefly described before focusing on the considered solution to the resource selection problem. Figure 1 depicts an instance of the logical layers that SPINA superimposes on top of the underlying network infrastructure. In the figure three layers are considered: starting from the lower one, we can distinguish the document, the peer and the ultra-peer layer. Another characteristic shown in Figure 1 is that, because of the hierarchical nature of the underlying overlay, the network is subdivided in groups – no clustering is adopted – each of which refers to an ultra-peer. For instance, p_a, p_b and p_c belong to the group of peers that refers to the ultra-peer up_d. The notion of **group** is crucial to explain the first part of the retrieval process.

When interacting with the peer, the end user can perform a music or textual search by submitting a query as a bag of features. At present time SPINA supports free text search queries and the possibility to submit a MIDI file in order

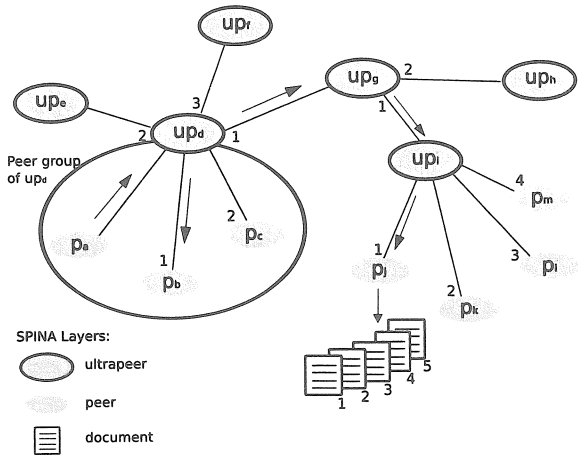

Fig. 1. P2P topology adopted in SPINA and query routing mechanism.

to identify it or to find similar songs. If the query is formulated in a peer, the request is forwarded to the ultra-peer which the peer refers to. But the request can be formulated also in an ultra-peer – remember that also an ultra-peer is a peer and so it is provided by a local engine. Whatever the starting resource is, the resources at higher levels – peers and ultra-peers – are ranked and selected by ultra-peers according to a certain criterion. Then the selected peers are contacted according to the order they appear in the ranked peer list. After ranking its local resources each peer returns to the referring ultra-peer a ranked list of objects in response to the formulated query. For instance, in Figure 1, the query is formulated in peer p_a, then it is forwarded to the referring ultra-peer up_d, and finally to the other peers of the group, that is p_b and p_c. The order in which the peers are contacted, that is first p_b and then p_c, is denoted by the number nearby the edge connecting the peer with its ultra-peer.

Retrieval is not limited to the group a peer belongs to. The ultra-peers communicate among them to form the backbone of an hybrid decentralized network. Before explaining how the search process continues, the notion of **neighbor** has to be clarified. Each ultra-peer stores some information about a subset of the other ultra-peers in the network: this subset of ultra-peers constitutes the set of neighbors of the ultra-peer. According to the information about its neighbors, an ultra-peer can extend the search to the peers of the other groups. In particular, the ultra-peer ranks its neighboring ultra-peers and forwards the query to the most promising neighbors which in their turn search in the groups that serve. For instance, in Figure 1, the ultra-peer up_d contacts first up_g, then up_e and finally up_f. The order in which the ultra-peers are contacted is denoted by the number nearby the edge connecting the ultra-peer which forwards the query and its neighbors. The search process can be extended by propagating the query to

the neighbors of the latest contacted ultra-peers, for a number of hops bounded by the Time-To-Live (TTL) of the formulated query.

Each ultra-peer is able to merge the results obtained by the peers of the group that serves together with the merged list of results returned by its neighbors. The final merged result list is then returned to the requesting peer by following the inverse path of the query.

2.2 Ranking Criterion and High Granularity Indexes

In Section 2.1 we referred to a ranking criterion to rank resources at the different levels of the hierarchy. In particular, an ultra-peer has to rank the peers in its group in order that the query is forwarded to the most promising peers. At the higher level, ultra-peers have to be ranked so that the query can be forwarded to the most promising neighbors of the ultra-peer. Indeed, the resource selection problem in the considered network topology seems to be characterized by a recursive nature exploited in the design of the system, but even before in the weighing framework proposed in [1], that is the Term Weighted Frequency (TWF) Inverse Resource Frequency (IRF) weighing scheme. In this scheme the weight assigned to a resource — peer or ultra-peer — is computed by the statistics about the features extracted from the full content of the documents in the peer's collections, aggregated according to the level hierarchy. The rationale of this choice is to use the features that occurs in the collection stored in the peer, or in the set of collections stored in a group of peers, and the statistics derived from their frequency of occurrence, to describe respectively a peer or an ultra-peer. The methodology described in Section 3 allows to use the same description of resources at higher levels also in terms of musical features. Since only the exchange of little information – features with their weights – is required to rank resources at different levels, the adopted weighing scheme is a suitable solution in terms of network load.

Each ultra-peer locally stores two indexes, one for each resource level, where the information required by the weighing scheme is stored.
The first index has a "peer-granularity" and it is achieved by aggregating the information about the content of the peers in the group the ultra-peer serves. The aggregation takes place in each single peer that communicates this summary information. The ultra-peer collects and stores this information in a local index, which basically provides the list of peers in the ultra-peer group associated to each feature, as well as the total weight of every feature occurring in every peer. According to this information the peers in the group are ranked by the adopted weighing scheme and then the query is routed to the most promising peers.
The information required to rank and select the neighbors is stored in the second typology of index, that is the "ultra-peer-granularity" index. In this index the list of neighboring ultra-peers is associated to each feature, as well as the total weight of every feature occurring in every neighboring ultra-peer. Each ultra-peer computes its contribution locally by aggregating the information about the peers in its group thus achieving a list of all the features in the peer-granularity

index and the weight such feature has in the ultra-peer itself. Thus each ultra-peer has two indexes – peer and ultra-peer granularity – for each medium – text and audio.

3 Music Content Representation

A general way to represent and index music content is required to integrate a content-based music search component in the proposed architecture. In an IR system, the document indexing is a crucial step because it enables a compact representation of the content of a collection, aimed at an efficient and scalable access and retrieval. General indexing techniques can be extended also to music, providing that significant descriptors are computed from music documents. These descriptors can be defined as the *lexical units* of music, and depend on the dimension that are taken into account – melody, harmony, rhythm, timbre – and are related to the way listeners perceive music. With the aim of creating a common retrieval strategy for different media, the index for music documents has to be consistent with the index for all the other media, following a similar scheme. The basic idea underlying the approach is that a music document can be effectively described by excerpts of its melodic features [14]. The main goal then becomes the automatic extraction of relevant excerpts from an unstructured flow of notes, stored in different digital formats. With this aim we can introduce some general terminology. In particular, according to [7], we can define as **features** the characteristic that describes subsequent notes in a music documents. Features can be of different types, such as the pitch, the pitch interval with the previous note, the duration, the chroma features [15] and so on. In our approach they are mostly related to pitch and rhythm which can also be treated independently and can be considered valid descriptors for our purposes. A sequence of features is defined as **string**. For example, any sub-sequence of notes in a melody can be considered as a string. A string repeated at least twice in a music flow is defined as a **pattern**. The repetition can be due to the presence of different choruses in the flow or by the use of the same melodic material. Patterns can be considered as the descriptor of the music documents and can carry different information about their content. The structure of the patterns can be related to the textual case where patterns play the same role of words of a document. Thus the most common indexing textual techniques can be exploited.

Following the terminology introduced previously, each document has a number of patterns of different length and with different multiplicity. Among all the computed patterns, some of them could have little or no musical meaning. For instance, a pattern that is repeated only two or three times in a document is likely to be computed by chance just because the combination of features is repeated in some notes combination. Moreover, some patterns related are likely to appear in almost all documents and hence to be poor discriminant among documents. In general, the degree by which a pattern is a good index may vary depending on the pattern and on the document. This is a typical scenario of textual information retrieval where words may describe a document to a differ-

ent extent. For this reason, a weighing scheme based on the *TF·IDF* measure might be a viable solution for MIR too as reported in [14]. Thus, at the end of the process, the music documents will be described by a sequence of patterns together with their weight (which represents the number of times a pattern is repeated in the document) and their positions (in *ms*) along the music flow. In the designed system each peer stores this information in its local audio index.

As stated previously, each pattern is composed by a sequence of features which describes the music content. Features computation is a difficult task which strictly depends on the digital format in which music documents are stored. Indeed, according to [16], some digital formats can represent music scores, whereas others can represent music performances. At the current state of the art, the most common formats are MIDI for music scores and both WAVE and MP3 for music performances.

According to the digital formats, thus, the algorithms of content extraction, even if with the same objective, might be completely different.

4 Conclusions

In this paper the current status of the design of a P2P content-based search engine is reported. The architecture has been thought to be flexible enough to handle different types of media, both in terms of representation and retrieval.

At the present time, the network infrastructure liable to the communication among peers has been implemented together with the functionalities for the search engine. A weighing scheme to rank resources at different level hierarchy and the indexes which stored the information used to compute the weights have been already developed. Each peer of the network stores music and textual documents which can be retrieved by the engine. The implemented functionalities enable to perform content-based retrieval in a P2P networks, where a generic user can submit as query both a sequence of words and a musical file.

Some questions are still open and under investigation. Concerning the P2P architecture, the major issues pertain to the dynamism of both the peers and the statistics. A first point concerns the set of the policies which handle the composition of the peers groups. It is important to define some rules in order to manage the behavior of the network, especially when an ultra-peer is shut down and when a new peer joins the network and has to be associated with an ultra-peer. Other studies will be aimed at the definition of some policies to manage the dynamics due to the change of the documents stored in a peer.

Concerning the documents representation, the major issue is aimed at handling the music recordings, in particular stored in the MP3 format. The music content representation, both of scores and recordings, then has to be deeply tested in order to formally evaluate the retrieval efficiency of the system. In particular, some standard music collections should be exploited in order to make the achieved results comparable with other systems.

Finally, a valid strategy which enables an ultra-peer to merge the results returned by the peers during a search process will be investigated.

Acknowledgments

The authors are grateful to Maristella Agosti and Giorgio Maria Di Nunzio for the fruitful discussions on the topic of this paper. The work reported in this paper has been partially supported by the SAPIR project, as a part of the Information Society Technologies (IST) Program of the European Commission (Contract IST-045128).

References

1. Castiglion, R., Melucci, M.: An evaluation of a recursive weighing scheme for information retrieval in peer-to-peer networks. In: Proceedings of P2PIR 2005, Bremen, Germany (2005) 9–16
2. Melucci, M., Poggiani, A.: A study of a weighting scheme for information retrieval in hierarchical peer-to-peer networks. In: Proceedings of ECIR 2007, Rome, Italy (2007) 136–147
3. Di Buccio, E., Ferro, N., Melucci, M.: Content-based Information Retrieval in SPINA. In: Proceedings of IRCDL 2008, Padua, Italy (2008) 89–92
4. Orio, N.: Music retrieval: a tutorial and review. Foundations and Trends in Information Retrieval 1(1) (2006) 1–96
5. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A Review of Audio Fingerprinting. Journal of VLSI Signal Processing Systems 41(3) (2005) 271–284
6. Miotto, R., Orio, N.: A methodology for the segmentation and identification of music works. In: Proceedings of ISMIR 2007, Vienna, Austria (2007) 271–284
7. Neve, G., Orio, N.: Indexing and retrieval of music documents through pattern analysis and data fusion techniques. In: Proceedings of ISMIR 2004, Barcelona, Spain (2004) 216–223
8. Lu, J., Callan, J.: Full-text federated search of text-based digital libraries in peer-to-peer networks. Information Retrieval 9(4) (2006) 477–498
9. Nottelmann, H., Fuhr, N.: Comparing Different Architectures for Query Routing in Peer-to-Peer Networks. In: Proceedings of the ECIR 2006, London, UK (2006) 253–264
10. Tzanetakis, G., Gao, J., Steenkiste, P.: A scalable peer-to-peer system for music information retrieval. Computer Music Journal 28(2) (2004) 24–33
11. Yang, C.: Peer-to-peer architecture for content-based music retrieval on acoustic data. In: Proceedings of WWW2003, Budapest, Hungary (2003) 376–383
12. Karydis, I., Nanopoulos, A., Papadopoulos, A.N., Manolopoulos, Y.: Musical retrieval in p2p networks under the warping distance. In: Proceedings of ICEIS 2005, Miami, USA (2005) 100–107
13. Wang, C., Li, J., Shi, S.: A Kind of Content-Based Music Information Retrieval Method in Peer-to-peer Environment. In: Proceedings of ISMIR 2002, Paris, France (2002) 178–186
14. Melucci, M., Orio, N.: Musical information retrieval using melodic surface. In: Proceedings of DL'99, Berkeley, California, United States (1999) 152–160
15. Peeters, G.: Chroma-based estimation of musical key from audio-signal analysis. In: Proceedings of ISMIR 2007, Victoria, Canada (2006) 115–120
16. Orio, N.: Alignment of performances with scores aimed at content-based music access and retrieval. In: Proceedings of ECDL 2002, Rome, Italy (2002) 479–492