A Digital Rights Aware Similarity Measure for Multimedia Documents

Walter Allasia
EURIXGROUP
Via Carcano, 26
Torino, Italy
allasia@eurixgroup.com

Fabrizio Falchi
ISTI-CNR
Via Moruzzi, 1
Pisa, Italy
fabrizio.falchi@isti.cnr.it

Francesco Gallo
EURIXGROUP
Via Carcano, 26
Torino, Italy
gallo@eurixgroup.com

Nicola Orio
University of Padova
Via Gradenigo, 6/a
Padova, Italy
orio@dei.unipd.it

ABSTRACT

This paper presents a novel approach to the retrieval of multimedia documents that considers Intellectual Property Rights (IPR) metadata as a multidimensional feature in a metric space. The approach allows us to perform similarity searches on the IPR attributes of digital items and to integrate these searches in a common query-by-example paradigm. We aim at managing the metadata related to the IPR in both centralized and Peer-to-Peer systems with metric indexing capabilities. Together with content-based similarity search, IPR similarity search can help the end user to deal with a huge amount of similar items with different licenses. Moreover, content providers may be able to detect fake copies or illegal uses. Two use cases, related to the retrieval of music and images respectively, are presented to describe the possible applications of the approach.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*; H.5.1 [**Information Systems**]: Information Interfaces and Presentation—*Multimedia Information Systems*; K.5.1 [**Computing Milieux**]: Legal Aspects of Computing—*Hardware/Software Protection*

General Terms

Management, Legal Aspects

Keywords

digital rights, information retrieval, multimedia information systems, metric spaces

MS'07, September 28, 2007, Augsburg, Bavaria, Germany.

1. INTRODUCTION

The amount of digital items that are produced every day is increasingly growing. Even non professional end users can access several devices that allow them to create their own digital content, in the form of images, videos, music, and text. Moreover, end users may benefit from a variety of means to distribute and share digital content: electronic mail, personal Web sites, chats, forums, multimedia messaging services. Systems for sharing digital content are becoming very popular, either with a centralized architecture where the users upload their material or with a distributed architecture.

The sources for the digital material range from the result of a personal production by professionals, which can be used for self-promotion and possibly sold, to just the recording of an event that the user wants to digitally store and eventually share with his friends. In the former case, users create the content with high quality standards, because the goal is to exploit channels for disseminating digital content that are alternative to the regular distribution. In the latter case, the digital content is often created through portable devices, which normally does not produce high quality multimedia objects.

A significant side effect is that our cultural heritage is no longer made up only of videos, images and text documents provided by *institutional* public or private bodies but also of the digital contents provided by every connected digital device as well. The increasingly relevance of *blogs* in modern societies with a high technological development is just an example of how digital material can become part of the contemporary culture without passing through the common editorial steps. In order to be able to guarantee the preservation of and access to these digital items, we have to take into account their Digital Rights Management (DRM) during the creation phase and especially during the search of something provided by someone else.

Clearly, the digital content created by end users is mixed with the digital content provided by private, and sometimes public, institutions. Thus, it is of paramount importance for the content provider that the IPR is correctly dealt with. Needless to say, the same channels that are easing the distribution of digital contents created by the private users for the

purpose of being shared, are massively exploited to illegally distribute copyrighted material. A user who is searching for digital content, either for leisure or for his profession, has to deal with items that may differ dramatically on the kind of applied license.

The effectiveness of a retrieval task depends both on the relevance of the retrieved objects to the query and on the relevance of the type of license for their intended usage. Thus DRM has to be considered as part of the user's information needs, and dealt in a similar fashion as other objets features. Although several approaches have been proposed so far for managing Digital Rights and many standards are available for representing them (for a review the interested reader can refer to [24]), usually open as well as trusted systems provide a simple attribute search on a single specific type of license.

In this paper, we propose a novel approach for indexing and retrieval of the information related to the licenses of the digital items, where licenses are considered as features in a metric space. The approach is meant to be part of a flexible and open network infrastructure. The paper is organized as follows. Section 2 introduces the context and some concepts related to the proposed approach. Section 3 describes the way Digital Rights can be mapped on a metric space allowing for similarity search on license attributes. Two possible applications of the approach are presented in Section 4. Finally, some conclusions and ideas for future work are given in Section 5.

2. BACKGROUND

Before introducing the novel approach for dealing with DRM at query time, a number of related concepts are discussed in the following sections.

2.1 Digital Rights Management

The technologies created for dealing with IPR of digital contents are still debated. During the 6^{th} Framework Programme, the European Community has sponsored a project named Networked Audiovisual Systems and Home Platforms, headed by the best experts at dealing with digital content environments. Their effort produced at the end of 2005 an important report [24], describing a set of requirements that reasonably have to be satisfied by whatever DRM system. Some of these requirements are harshly criticized, mainly concerning the analogy with the contract laws and the development and use of free and open source content [8]. At the same time there are many initiatives that are trying to provide the basis for a DRM infrastructure, such as DMP [10], Chillout [6], and MediaLive [21].

The considerations reported above prove that it is very difficult today to deal with DRM systems in such an heterogeneous environment as the Web and at the moment we are very far from having a common agreement on the adoption of a standard DRM system.

The expression of the kind of the license is evolving in a more agreed way, and some standards on the expression of the license are arising and are commonly used. However it should be noted that, from the copyright law point of view, there are still some differences between the license and the actions that a DRM system is able to provide according to that license. It is important to point out the difference between the "license" itself on a digital item and the "control" that the item is really used according to its license. This article is dealing with the definition of the license and in particular the search capabilities of a digital item by its license. We do not want to provide any guideline nor implementation of the software for controlling the respect of use of the license as a DRM system has to guarantee. Our goal is to provide users with an innovative approach for managing the license attributes in order to be able to search for digital objects that have a notion of *metric distance* from a query object. It is interesting to note that the existence of a metric distance between IPR characteristics may allow also for querying objects that are dissimilar on IPR, while being similar on their visual or music content.

Up to now, many solutions have been adopted so far in the Web for expressing a license which are widely used, such as Adobe Content Manager [3], Creative Commons [9], MPEG-21 Right Expression Language (REL) [17], Open Digital Right Language (ODRL) [16] and Publishing Requirements for Industry Standard Metadata (PRISM) [28].

As discussed above the digital items available on the Web are produced by the personal use of digital devices and are mainly audio/video files. Hence we focus on the language for expressing the license that are widely adopted for defining this kind of items. In this paper we consider three different license formats: CreativeCommons, MPEG-21 REL and ODRL. CreativeCommons is widely adopted for photos, music and educational contents; MPEG-21 REL is commonly used for music and videos while ODRL has been adopted for example for the ringing tones in the Mobile Phone because it has been used in the OMA [26] environment.

Unfortunately the metadata for expressing the license in the three formats described above are quite different from each other and a mapping is required in order to process them in the same query. Table 1, adopted from [15], is representing an example of the mapping for the group of metadata referring to the different Rights managed by the three formats in four classes. As it can be seen, CreativeCommons has a reduced number of data elements compared to MPEG-21 REL and ODRL, while the last two formats have different focuses, the former on reuse and the latter on transfer and management.

2.2 Distributed and centralized systems

Many network infrastructures are arising in order to provide the bases for Web sharing and searching functionalities on digital items. Most of them are peer-oriented networks, such as eMule [2] or BitTorrent [1] for images and audio/video files and Joost [19] for video streaming. Furthermore, several multimedia platforms enable the automatic audio/video processing for the cataloging and the indexing of digital items [22] and in combination with the network infrastructure will provide powerful solutions for digital content management.

Most of these networks are useless in a business context, where service providers and users want to be aware of the license related to the searched content and want to know what are the actions that can be done on the downloaded digital items. Moreover, in a distributed network, it is likely that very different approaches to DRM are present at the same time, with additional problems in the mapping of the types of licenses.

Also centralized systems are very popular for distributing content created by end users. A centralized system, providing access to hundreds of thousands, if not millions, of digital items, represent a sort of *entry point* where users searching

	Use-Type Rights
CC	reproduction
MP21	execute, play, print
ODRL	display, execute, play, print
	Transfer-Type Rights
CC	distribution
MP21	transferControl
ODRL	send, lend, give, lease
	Manage-Type Rights
CC	
MP21	delete, move, install, uninstall
ODRL	delete, move, install, uninstall
	duplicate, backup, verify, save
	Reuse-Type Rights
CC	derivativeWorks
MP21	adapt, diminish, embed, enhance
	enlarge, modify
ODRL	modify, excerpt, annotate, aggregate

Table 1: Mapping of the data elements defined by CreativeCommons (CC), MPEG-21 REL (MP21), and ODRL in four main classes related to Rights

for digital content are reasonably sure to find something that is relevant to their information needs.

Sometimes the information need is very generic, because users may browse and search the system just because they know there is a large amount of multimedia items available. It could be argued that this is the *added value* of a centralized system that convinces so many end users to upload their material, and to loose the control of it. In many cases, a centralize system chooses a particular typology of license – e.g. CreativeCommons – and allows the users providing the content to choose within a number of variants about use, distribution, modification and so on. Also in this case, retrieval should be carried out also taking into account IPR metadata.

2.3 The metric space approach

As previously mentioned, we propose to deal with IPR metadata as a multidimensional feature in a metric space. Although many similarity search approaches have been proposed, the most generic one considers the mathematical metric space as a suitable abstraction of similarity [31]. The simple but powerful concept of the metric space consists of a domain of objects and a distance function that measures the proximity of pairs of objects.

In the metric space $M = (\mathcal{D}, d)$ defined over a domain of objects $\mathcal{D}$ with a total (distance) function $d : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$, the following properties hold $\forall x, y \in \mathcal{D}$:

$$
\begin{array}{ll}
d(x,y) \geq 0 & (\textit{non-negativity}) \\
d(x,y) = 0 \text{ iff } x = y & (\textit{identity}) \\
d(x,y) = d(y,x) & (\textit{symmetry}) \\
d(x,z) \leq d(x,y) + d(y,z) & (\textit{triangle inequality})
\end{array}
$$

The metric space approach has been proved to be very important for building efficient indexes for similarity searching. A survey of existing approaches for centralized structures can be found in [29] and [31]. Two well known examples of them are M-tree [7] and D-Index [11].

Very recently scalable and distributed index structures based on Peer-to-Peer networks have also been proposed for similarity searching in metric spaces, i.e. GHT* [4], VPT*, MCAN [13] and M-Chord [25] (see [5] for a comparison of their performances).

Currently many research projects are investigating these fields, such as SAPIR [30], a project funded by European Research Area in the 6^{th} Framework Program, that aims to develop cutting-edge technology that will break the barriers and enable search engines to look for large scale audio-visual information by content, using the query by example paradigm. SAPIR intends to propose new solutions for an innovative technological infrastructure for next-generation Multimedia Search Engines. This research effort should lead towards a distributed, P2P based, search engine architecture, as opposed to today parallel search engines within a centralized Web data warehouse.

3. METRIC DISTANCE EXAMPLE FOR LICENSES

We now illustrate an example of a metric distance defined over the IPR metadata. The main common groups of the expression languages of the licenses can be classified, as proposed by [15], in a number of categories:

- *Agent Data Element.*
- *Rights*: including Manage-type, Reuse-type, Transfer-type, and Use-type Rights.
- *Constraint*: including User, Device, Limits, Temporal, Aspect, Target, and Payment Constraints.
- *Usage Conditions.*

Considering the scheme representations of the three licenses mentioned above, we noticed that there are more than 50 candidate attributes to be compared in the ODRL. Moreover some of the attributes are complex types, like for example, the *price* for the MPEG-21 REL, where the *FeeMetered* of the *Conditions* group is represented by the use of a mathematical function.

3.1 IPR-based distance functions

We propose different approaches to compute a measure of the distance (i.e. dissimilarity) between IPR metadata. We consider distance for convenience but everything could be easly adapted to evaluate similarity. Since metadata may have different domains, the distance function has to be tailored to the particular domain in order to be meaningful.

Let $\mathcal{D}$ be the domain of metadata related to the license of any given object. For any $x \in \mathcal{D}$ we define $x_1, x_2, ..., x_n$ as the n main groups and $x_{i,1}, x_{i,2}, ..., x_{i,n}$ as the n_i attributes for the i-th main group. The global distance is defined as the weighted sum of the distances between main groups, i.e.

$$d(x,y) = \sum_{i=1}^{n} w_i \cdot d_i(x_i, y_i) \qquad (1)$$

The distance between the same groups of two distinct licenses can be defined as:

$$d_i(x_i, y_i) = \sum_{j=1}^{n_i} w_{i,j} \cdot d_{i,j}(x_{i,j}, y_{i,j}) \qquad (2)$$

Once a general definition of a distance function is given, the actual computation of the distance $d_i(x_i, y_i)$ between two values $x_{i,j}$ and $y_{i,j}$ of the j-th attribute of the group i must be defined considering the specific attribute type. We consider three different types.

3.1.1 Binary attributes

In many cases attributes are binary values, describing whether or not a given action is permitted by the license. Thus, in case $x_{i,j}, y_{i,j} \in \{0, 1\}$ we can use the L_1 norm, which assumes binary values as well. Hence the distance is directly computed as:

$$d_{i,j}(x_{i,j}, y_{i,j}) = |x_{i,j} - y_{i,j}| = x_{i,j} \otimes y_{i,j} \quad (3)$$

Please note that a specific weight to this distance can be given by setting $w_{i,j}$, thus the overall distance $d(x, y)$ may be any positive value.

3.1.2 Numeric attributes

In the case of numeric attributes with $x_{i,j}, y_{i,j} \in \mathbb{R}$, which for example can represent a fee for accessing the digital content, it is possible to apply the L_1 norm as well, or the euclidian distance. However, more sophisticated metric distances could be used for specific numerical attributes. For instance, we suggest to define the distance between fess as:

$$d_{i,j}(x_{i,j}, y_{i,j}) = |\log(x_{i,j}) - \log(y_{i,j})| = \left|\log\left(\frac{x_{i,j}}{y_{i,j}}\right)\right| \quad (4)$$

According to Equation 4, the distance between \$5 and \$10 is the same as the distance between two fees of \$50 and \$100 respectively. We believe that given a fee as query, the user is much more interested on the proportion between its query and a given fee. Unfortunately in this case any non \$0 fees would be at infinite distance from \$0 objects. Incidentally, in the case of fees this characteristics may reflect a common attitude, because users that are looking for free digital content are not interested in non free items. For instance, they are not willing (or able) to purchase items on the Web, or they are browsing for material just because it is free.

Anyway, to avoid this problem we suggest that whenever the fee value is smaller than a given threshold, say \$0.01, the value used for evaluating the distance is automatically set to 0.01. In this way, the distance between \$0 and \$1 becomes the same of the distance between \$1 and \$100, which is a reasonable assumption. It is worth noting that the distance is still a metric.

It should also be noted that there are other possible approaches for measuring differences in prices, for example the *gap-ratio* computed by the following equation:

$$d_{i,j}(x_{i,j}, y_{i,j}) = \frac{x_{i,j} - y_{i,j}}{y_{i,j}} \quad (5)$$

Yet, the gap-ratio is not symmetric and thus it cannot be used for the proposed metric approach.

3.1.3 Textual attributes

For an attribute whose value can be a term in a given vocabulary, we propose an alternative approach. If the j-th attribute of the i-th group is a term taken from a specific vocabulary of m terms, we can define the distance $d_{i,j}(x_{i,j}, y_{i,j})$ between the two values according to what reported in Table 2.

	$term_1$	$term_2$	$term_3$	...	$term_m$
$term_1$	0	$\alpha^{i,j}_{2,1}$	$\alpha^{i,j}_{3,1}$	...	$\alpha^{i,j}_{m,1}$
$term_2$	$\alpha^{i,j}_{2,1}$	0	$\alpha^{i,j}_{3,2}$	...	$\alpha^{i,j}_{m,2}$
$term_3$	$\alpha^{i,j}_{3,1}$	$\alpha^{i,j}_{3,2}$	0	...	$\alpha^{i,j}_{m,3}$
...	...	...	...	...	...
$term_m$	$\alpha^{i,j}_{m,1}$	$\alpha^{i,j}_{m,2}$	$\alpha^{i,j}_{m,3}$	...	0

Table 2: Distance values for attributes taken from terms in a given dictionary

It is assumed that the values of α are manually chosen according to the semantic of the given terms. Clearly, this is a critical step that needs to be further investigated. For instance, the actual values may reflect the similarity given by either the content producers or the end users. It should be noted that for $d_{i,j}(x_{i,j}, y_{i,j})$ to be a metric the matrix must be symmetric, thus $\alpha^{i,j}_{a,b} = \alpha^{i,j}_{b,a}$, all the diagonal values must be 0 and

$$\forall l \ , \ \alpha^{i,j}_{a,b} \leq \alpha^{i,j}_{a,l} + \alpha^{i,j}_{l,b} \quad (6)$$

A user interface should help the administrator setting the α values according to these requirements. A trivial solution is to set all $\alpha^{i,j}_{a,b} = 1$ when $a \neq b$, and in this case textual attributes are considered as binary attributes.

As an example, let x and y be the metadata about the license attributes. Taking into account the CreativeCommons schema for the *Requirements*[1] (the restrictions imposed by the license), we can assign a set of values to the $\alpha^{i,j}_{m,n}$ terms as shown in Table 3, where:

- **Notice** requires that the copyright and license notices must be kept intact
- **Attr** stands for Attribution and requires that credit must be given to copyright holder and/or author
- **SA** stands for ShareAlike and requires that derivative works must be licensed under the same terms as the original work
- **SC** stands for SourceCode and requires that source code (the preferred form for making modifications) must be provided for all derivative works.

	Notice	*Attr*	*SA*	*SC*
Notice	0	0.3	0.6	1
Attr	0.3	0	0.3	0.7
SA	0.6	0.3	0	0.4
SC	1	0.7	0.4	0

Table 3: Proposed distance values for Creative Commons terms for the Requirements

3.2 Metric distances

We now prove that if all the distances used to compare the values of each attribute are metric, the proposed distance between two licenses is still a metric. In other words, a weighted sum of metric distances is a metric distance too.

[1]http://creativecommons.org/technology/metadata/implement

In fact, $\forall x, y \in \mathcal{D}$ and for any given group $i = 1, ..., n$ and $\forall j = 1, ...n_i$ we get:

$$d_{i,j}(x_{i,j}, y_{i,j}) \leq d_{i,j}(x_{i,j}, z_{i,j}) + d_{i,j}(z_{i,j}, y_{i,j}) \Rightarrow$$

$$\sum_{j=1}^{n_i} w_{i,j} \cdot d_{i,j}(x_{i,j}, y_{i,j}) \leq$$

$$\leq \sum_{j=1}^{n_i} w_{i,j} \cdot [d_{i,j}(x_{i,j}, z_{i,j}) + d_{i,j}(z_{i,j}, y_{i,j})] \Rightarrow$$

$$d(x, y) \leq d(x, z) + d(z, y).$$

Thus, if all the distances defined for the attributes in a given group are metric, the proposed distance for the given group is still a metric. Defining the global distance between two licenses as the weighted sum of distances between the main groups, this global distance is still a metric.

3.3 Combining different IPR-based distance functions

In order to index the licenses metadata using the global distance d in a single index for similarity searching in metric spaces, all the weights w should be fixed in advance. However, if we want to specify at query time the weights w_i for the single groups to be used for searching, we can use distinct indexes for each d_i and then combine the results coming from the various indexes using optimal aggregation algorithms as the ones described in [12]. Moreover, in this case we do not need the global distance function to be metric, but just all the d_i. In this case the aggregation must be monotone. Thus, using separate indexes for each d_i and then combining them using the algorithms described in [12], more aggregation functions could be used and they could even be specified at search time. Obviously there is a price to be paid for that: efficiency. A single global metric distance function can be more efficiently indexed using a single index structure for metric spaces. An extension of the proposed global distance which is still metric is a sort of Minkowski Distance combination:

$$d(x, y) = \sqrt[k]{\sum_{i=1}^{n} w_i \cdot |d_i(x_i, y_i)|^k} \ . \tag{7}$$

The same approach could be used to combine the distance values among the attributes of the same group. Settin k the importance of groups with higher distance can be tuned. In fact, the greater k, the more important far away groups or attributes values are.

3.4 Merging content-based and IPR-based distances

An important issue that has to be taken into account is that IPR-based distances should be merged with common content-based distance functions. A straightforward approach would be to simply include both classes of distances either in Equation 1 or in Equation 7, and use the weights to balance the effect of content features, which are related to the way the digital object is perceived, with the IPR features, which are related to the way the object can be used.

A linear, or more complex, combination of content-based and IPR-based distances puts at the same level the intrinsic qualities of the digital objects – the result of a creation process – with the external properties – the result of an agreement on how, when and for which purpose the object could be used. The basic idea is that, for instance, a user is more interested to a digital music recording that is less similar to the one he used to query the system, but that costs less (e.g. a free low quality live recording of a song instead of the $0.99 standard quality studio version).

The risk of this approach is that, if individual weights are not fine tuned, the overall distance of a totally irrelevant digital item might be smaller than the one of a relevant item that has different IPR properties. In this way the perceived recall and precision, which are likely to be based on the content similarity of the two objects, will both decrease. It is worth recalling that the goal of our approach is to provide users with items that may have a positive distance on the IPR metric space, and not just filtering out all the items that have a different kind of license. The approach can be considered as a way to sort, or re-rank, potentially relevant items according to their IPR-based distance.

For this purpose, it is possible to perform the retrieval in two steps:

1. An initial rank list of potentially relevant objects is computed from content-based features only. Assuming that the system has good levels of precision at N retrieved documents, only the first N items are kept, and the others are discarded.

2. The list of documents is reranked according to IPR-based distances. Different strategies to rerank can be envisaged, for example taken from the literature on data fusion techniques for meta search engines [20].

The two step approach could also be exploited for building an IPR aware search engine on the top of an existing retrieval system.

Another approach could be to have two distinct indexes for IPR and content-based retrieval. In this case, the same considerations reported in Section 3.3 to combine the distance obtained for each IPR main group are valid. Making use of well know algorithms as the one described in [12], it is possible to combine the results coming from the two indexes. In this case the only requirement for the aggregation function (used to obtain the score given the two distinct scores obtained for IPR and content-based) is to be monotone. Moreover, the aggregation function could be different for different types of users and also the same users could decide to use a specific one even considering more relevant (e.g. using weights) the score obtained for one of the two aspects (IPR, content-based). Obviously, it is also possible to have more than one content-based similarity search (e.g. color, edges, shape and texture for images). In such a system the user could be free to decide the features to use for content-based and how to combine them together and, eventually, with the IPR similarity search.

4. SIGNIFICANT USE CASES

It is worthwhile to underline that in order to be able to handle the Rights of digital contents, they should be defined during the creation phase of the items. Hence the GUI that a sharing system should provide has to take into

account the multiple choice of available licenses. For instance, Jamendo [18] that distributes music and Flickr [14] that distributes photos, are currently providing at least one type of license definition (CreativeCommons). So far the Flickr search engine for the Rights is nothing but an attribute search, looking for the same value of a specific attribute of the expressed license, while Jamendo does not provide an IPR-based search.

In the following, we will give two use cases of our approach of considering IPR metadata as special features in a given space, instead of simple attributes. We assume that searching for a digital item can be done by similarity search and/or by textual attributes. For example we can look for an image or a sound similar to what someone provided as query to the search engine. Moreover we can add to our query some keywords that the search engine will take into account as specific attributes. Yet, nowadays most of the search engines available on the web are providing nothing but the "full text" and/or "attribute" search capabilities. However, many research projects are developing music, image, and video similarity search.

4.1 Searching for digital music

Content-based access to music is not popular yet, even if there is a number of systems that are already available [23]. We can envisage that, in the short term, the effectiveness of music information retrieval systems will be improved enough to allow users to search for music using a query by example paradigm. It is important to note that music information retrieval deals with different forms in which music can be instantiated, namely the audio recordings of (live or studio) performances and the music score [27].

Historically, digital music has been associated to problems in DRM. The music industry and the main distributors, have different ideas on the way IPR has to be managed for music, and sometimes the systems for DRM have been very invasive for the music consumer. At the moment, all the publishers gave up to include DRM on audio CDs, and there is a similar tendency also in online stores. It is not a secret that hundreds of thousands of users simply ignore any IPR and download hours of music. Nevertheless, there are many users that are interested in being fair with DRM, and need new tools for retrieving music according to the content and on the kind of license. Among these, there are professional users: composers, performers, deejays, music critics, audio engineers, responsible of music programs, music assistants for tv shows, and so on.

In order to exemplify the applications of the proposed approach, three possible scenarios of music information needs that encompass both content-based and IPR-based features are presented:

1. A professional musician is looking for new interesting scores to be part of a new album, he is interested on songs with a similar mood, which have also similar licenses.

2. The responsible of a place open to the public (e.g. a bar, a restaurant, a shop) looks for background music to be played for his customers.

3. A user listens to a short excerpt of a song, which he likes, and queries the system for similar versions of the same song, which he wants to listen many times.

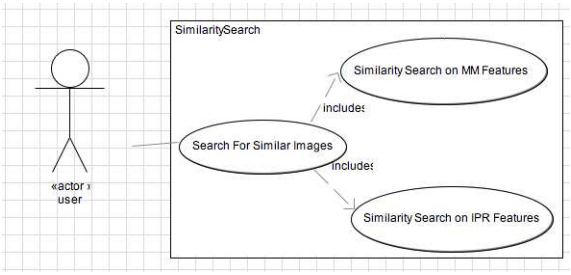

Figure 1: Use case diagram for similarity search

The first scenario could be based on an incremental search. Starting from a seed song and directions on the kind of license that should be granted by the composer, the user finds new music scores. He can then choose which kind of repertoire should be part of the new album, according both on the quality of the scores, and on the kind of permissions to perform and, if possible, adapt the scores with a transcription. For instance, composers may choose to permit only the integral reproduction of their work. It could be noted that, in this case, both content and license similarity are important, because the musician may want to have coherent music and coherent licenses for his new album.

Being the second scenario a part of business, and not for personal use, the user is probably equally interested in music quality – which might increase the number of customers – and in the kind of usage – which may cost. Based on the proposed approach, the rank list presents the results sorted both in terms of similarity to a particular style and in terms of adherence to the kind of license the user is interested in.

In the third scenario, the user has a precise information need regarding the content, but at the time he is querying the system does not know the IPR status of the song and of alternate versions. He can provide the system with an example of license that he usually uses. The system then ranks the songs according to their similarity, yet weighted by the IPR features, allowing the user to choose either the original studio version with restricted use, or a less expensive live bootleg with unlimited use, or the remake of a less known band which is totally free. Incidentally, this approach to rerank will help the user discover new performers and new songs, which may not be the ones that he uses for his query – because they are not famous – but are closer to the query in terms of the kind of license.

Additional scenarios can be envisaged, for instance looking for songs that have IPR features that differ from the song used in the query by example paradigm, in order to retrieve copies that are illegally distributed.

4.2 Searching for digital photos

A user may need to search for an image similar to the one provided considering both the multimedia content, which defines the kind of images he is interested in, and the related license provided by the user as well, which defines the kind of usage he has in mind. Thus the user can apply for searching similar images regarding the multimedia content and a specific kind of license defined by means of attributes.

Since the user can search for content-based similarity and license similarity independently, we are now focusing on scenarios where they are combined.

In Figure 1 we are pointing out that the *similarity search use case* is made up of the inclusion of two distinct use cases:

1. *Co-occurrence of perceptual and usage similarity*: the user is searching for images similar to a given one both considering its visual appearance and license "file", for instance because he has to choose a particular image to illustrate an article.

2. *Opposite needs regarding the perceived similarity and the intended use*: the user is searching for images similar to a given one but with a different license, whose definition can be taken from another image, for instance for discovering illegal copies of an image.

In the first case, the user is interested in images similar to a given one both considering its content and its license. This is the typical case where the user has an image which satisfied his needs both in terms of content and license. He made a picture of a painting, he does not remember the painter and the picture covers only a small piece of the paint. Also he is providing a license defined in ODRL as query, the one he is adopting for his pictures. The search engine will display as result the ranking list of images similar to the provided one according to the content and to the license.

For example a result could be a picture of the same painting made by a tourist and released according to a CreativeCommons license and another result, probably distant from the previous in the result list, could be a picture of the painting made by a professional photographer and the license defined according to MPEG-21 REL, expressing for example a price for downloading the high resolution picture. The user is able to decide according to both the quality of the image in the rank list and the kind of license, having immediately the feeling of how far is the result from the provided object query.

In the second case the user has an image which he does like, but that has a license which does not satisfy his needs. The user can search for an image similar to the given one but with a different license. In this case the license can be either taken from another image or specified using a form.

As already mentioned, a special case of this second scenario is searching for copyright violation. Imagine a professional photographers agency that wants to be sure that nobody is making a fake use of their own pictures and/or non authorized use of the associated copyrights. The agency can query the system providing the picture to be searched and can provide the attributes for an open license or something "similar" to an open one. If the system will find a result, it means either that someone has made the same picture or that someone is sharing a non authorized copy of the picture. This use case is innovative because the current search engines are focused on the content sharing and are not addressed to the "control" of the contents themselves, delegating this feature entirely to the DRM systems.

5. CONCLUSIONS AND FUTURE WORKS

We have proposed an innovative approach for managing the attributes and metadata referred to the expression language adopted for describing a license for Digital Rights. The metadata shown are taken as examples and should be changed to fit the needs of the software infrastructure that the users have to deal with. This approach considers the IPR attributes as *special features* which a specific distance function that can be applied to. For efficiently indexing the data it is important for this distance to be a metric.

The Right Management warrantee has been deeply studied in the last few years and lots of solutions are available. However not so much has been done so far concerning the "retrieval" of the license associated to the digital items. Since many standards are available, we will reasonably have many types of license and once we have to deal with thousands of items, the attribute search on the license files could be not enough to handle the problem. We are proposing here the adoption of the *Similarity Search* for the IPR attributes. In this way the kind of license we are looking for can be easily provided, instead of all the attributes of a specific license format in a complex GUI. On the other hand we can also have a ranking list of results according to the metric function, by defining the distance between the license attributes and by a mapping between license types that the users have to manage.

The next step of our research is the development of a framework to test the approaches in the two application domains described in Section 4. For this purpose we are in the process of collecting available material, images and music, with different IPR features. A major problem is given by the novelty of the proposed approach, because obviously no common benchmark is available for measuring the effectiveness of similarity search applied to Digital Rights.

6. ACKNOWLEDGMENTS

This work was partially supported by the SAPIR (Search In Audio Visual Content Using Peer-to-Peer IR) project, funded by the European Commission under IST FP6 (Sixth Framework Programme, Contract no. 45128).

7. REFERENCES

[1] BitTorrent. `http://www.bittorrent.com/`. Last visited on May 31st 2007.

[2] eMule. `http://www.emule.org/`. Last visited on May 31st 2007.

[3] Adobe. Content server 3. `http://www.adobe.com/products/contentserver/`. Last visited on May 31st 2007.

[4] M. Batko, C. Gennaro, and P. Zezula. Similarity grid for searching in metric spaces. In *Peer-to-Peer, Grid, and Service-Orientation in Digital Library Architectures6th Thematic Workshop of the EU Network of Excellence DELOS, Cagliari, Italy, June 24-25, 2004, Revised Selected Papers*, volume 3664 of *LNCS*, pages 25–44, 2004.

[5] M. Batko, D. Novak, F. Falchi, and P. Zezula. On scalability of the similarity search in the world of peers. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*, page 20, 2006.

[6] Chillout. The interoperable DRM platform. `http://chillout.dmpf.org/`. Last visited on May 31st 2007.

[7] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 426–435, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[8] E. A. C. Cory Doctorow. Critique of NAVSHP (FP6) DRM requirements report. Technical report, Electronic Frontier Foundation, 2006.

[9] CreativeCommons. Share, reuse, and remix – legally. `http://creativecommons.org/`. Last visited on May 31st 2007.

[10] DMP. The Digital Media Project Homepage. `http://www.dmpf.org/`. Last visited on May 31st 2007.

[11] V. Dohnal, C. Gennaro, P. Savino, and P. Zezula. D-index: Distance searching index for metric data sets. *Multimedia Tools Appl.*, 21(1):9–33, 2003.

[12] R. Fagin. Combining Fuzzy Information from Multiple Systems. In *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, 1996, Montreal, Canada*, pages 216–226, 1996.

[13] F. Falchi, C. Gennaro, and P. Zezula. A content-addressable network for similarity search in metric spaces. In *DBISP2P '05: Proceedings of the the 2nd International Workshop on Databases, Information Systems and Peer-to-Peer Computing, Trondheim, Norway*, volume 4125 of *LNCS*, pages 98–110. Springer, 2005.

[14] Flickr. The best way to store, search, sort and share your photos. `http://www.flickr.com/`. Last visited on May 31st 2007.

[15] R. González. *A Semantic Web Approach to Digital Rights Management*. PhD thesis, Department of Technologies, Universitat Pompeu Fabra, Barcelona, Spain, Apr 2006.

[16] R. Iannella. Open digital rights language (odrl), version 1.1, 2002. World Wide Web Consortium, W3C Note.

[17] ISO/IEC. Information technology - Multimedia framework (mpeg-21), 2005. 21000-5.

[18] Jamendo. Open your ears. `http://www.jamendo.com/`. Last visited on May 31st 2007.

[19] Joost. The new way of watching TV. `http://www.joost.com/`. Last visited on May 31st 2007.

[20] J. Lee. Analysis of multiple evidence combination. In *Proceedings of the ACM-SIGIR Conference*, pages 267–275, Philadelphia, USA, 1997. ACM Press.

[21] Medialive. Enabling secured content distribution. `http://www.medialive.com/`. Last visited on May 31st 2007.

[22] A. Messina, L. Boch, G. Dimino, W. Bailer, P. Schallauer, W. Allasia, M. Groppo, M. Vigilante, and R. Basili. Creating rich metadata in the tv broadcast archives environment: The prestospace project. In *Proc. AXMEDIS 2006, Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*, pages 193–200, Dec. 2006.

[23] MiDoMi. Search with your voice. `http://www.midomi.com/`. Last visited on May 31st 2007.

[24] Networked Audiovisual Systems and Home Platforms Group. Navshp (fp6) drm requirements report. Technical report, European Community 6^{th} Framework Programme, Sep 2005.

[25] D. Novak and P. Zezula. M-Chord: a scalable distributed similarity search structure. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*, page 19, 2006.

[26] OMA. The open mobile alliance. `http://www.openmobilealliance.org/`. Last visited on May 31st 2007.

[27] N. Orio. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 21(1):1–90, 2006.

[28] PRISM. Developing a stnadard XML metadata vocabulary or the publishing industry. `http://www.prismstandard.org/`. Last visited on May 31st 2007.

[29] H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Computer Graphics and Geometric Modeling. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

[30] SAPIR. Search in audio visual content using peer-to-peer ir. `http://sysrun.haifa.il.ibm.com/sapir/index.html`. Last visited on May 31st 2007.

[31] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search. The Metric Space Approach*, volume 32 of *Advances in Database Systems*. 233 Spring Street, New York, NY 10013, USA, 2006.