Proceedings of I-MEDIA '07 and I-SEMANTICS '07
Graz, Austria, September 5-7, 2007

A Similarity Approach on Searching for Digital Rights

Walter Allasia
(EURIX, Torino - Italy
allasia@eurixgroup.com)

Fabrizio Falchi
(ISTI-CNR, Pisa - Italy
fabrizio.falchi@isti.cnr.it)

Francesco Gallo
(EURIX, Torino - Italy
gallo@eurixgroup.com)

Abstract: We present an innovative approach that treats the right management metadata as metric objects, enabling similarity search on IPR attributes between digital items. We show how the content base similarity search can help both the user to deal with a huge amount of similar items with different licenses and the content providers to detect fake copies or illegal uses. Our aim is the management of the metadata related to the Digital Rights in centralized systems or networks with indexing capabilities for both text and similarity searches, providing the basic infrastructure enabling the private use and the commercial exploitation as well.
Key Words: rights, information retrieval, multimedia information systems
Category: H.3.3, H.5.1, K.5.1

1 Introduction

Nowadays we are dealing with several devices able to consume and produce lots of digital items, exponentially growing. People from all over the world are creating their own digital contents like images and audio/video files, sharing them by means of electronic mail, Web sites, chats, multimedia messaging services and several distributed systems. The digital contents are mostly provided by the personal use of high tech devices. 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.

In order to be able to guarantee the preservation and access of these digital items, we have to take into account their Digital Rights management during the creation phase and especially during the search of something provided by someone else. Several approaches have been proposed so far for managing Digital Rights and many standards are available for representing them, but 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 different approach for indexing and searching the information related to the licenses of the digital items, towards a more flexible and open network infrastructure.

2 Backgrounds

2.1 Digital Rights Management

The European project *Networked Audiovisual Systems and Home Platforms* produced an important report [NAVSHP 2005], describing a set of requirements to be satisfied by whatever DRM system. Some of these requirements are harshly criticized, mainly those concerning the analogy with the contract laws and the development and use of free and open source content [Doctorow 2006]. At the same time many initiatives, such as DMP[1], Chillout[2] and MediaLive[3] are trying to provide the basis for a DRM infrastructure. Recently some standards on the expression of the license have been proposed and started to be commonly used, but it is still very difficult today to deal with DRM systems in such an heterogeneous environment as the Web, since at the moment we are very far from having a common agreement on the adoption of a standard DRM system. In this paper we are dealing with the definition of the license and in particular with the search capability 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. We want to provide an innovative approach for managing the license attributes in order to be able to search for a similar digital object as query with its related digital license. Many solutions have been adopted so far in the Web for expressing a license which are widely used, such as AdobeContentManager[4], CreativeCommons[5], MPEG-21 REL [MPEG-21 REL 2005], ODRL [Iannella 2002] and PRISM[6].

Since the digital items available on the Web are mainly audio/video files produced by the personal use of digital devices, we focus on the language for expressing the license that are widely adopted for defining this kind of items. We consider as an example three different license formats: CreativeCommons, MPEG-21 REL and ODRL (see Table 1 from [Roberto García González 2006]). 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. An example of the mapping for the group of metadata referring to the *Use-type Rights* group is shown in Table 1.

[1] http://www.dmpf.org/
[2] http://chillout.dmpf.org/
[3] http://www.medialive.com/
[4] http://www.adobe.com/products/contentserver/
[5] http://creativecommons.org/
[6] http://www.prismstandard.org/

Use-Type Rights		
CreativeCommons	**MPEG-21 REL**	**ODRL**
reproduction		
		display
	execute	execute
	play	play
	print	print

Table 1: Use-Type Rights group mapping

2.2 Distributed 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[7] or BitTorrent[8] for images and audio/video files and Joost[9] for video streaming. Furthermore several multimedia platforms enable the automatic audio/video processing for the cataloging and the indexing of digital items [Messina et al. 2006] and in combination with the network infrastructure will provide powerful solutions for digital content management.

2.3 The Metric Space Approach

Although many similarity search approaches have been proposed, the most generic one considers the mathematical metric space as a suitable abstraction of similarity [Zezula et al. 2006]. 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 $\forall x, y \in \mathcal{D}$:

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

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 [Zezula et al. 2006] and [Samet 2006]. Two examples of well known centralized structure for indexing metric objects are M-tree [Ciaccia et al. 1997] and D-Index [Dohnal et al. 2003].

[7] http://www.emule.org
[8] http://www.bittorrent.com
[9] http://www.joost.com

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* [Batko et al. 2004], VPT*, MCAN [Falchi et al. 2005] and M-Chord [Novak and Zezula 2006] (see [Batko et al. 2006] for a comparison of their performances). Currently many research projects are investigating these fields, such as SAPIR[10], 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 information. The main common groups of the expression languages of the licenses can be identified as [Roberto García González 2006]: *Agent Data Element*, *Manage-type Rights*, *Reuse-type Rights*, *Transfer-type Rights*, *Use-type Rights*, *User Constraint*, *Device Constraint*, *Limits Constraint*, *Temporal Constraint*, *Aspect Constraint*, *Target Constraint*, *Payment Constraint*, *Usage Conditions*.

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 the main groups, i.e.

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

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

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

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. In case $x_{i,j}, y_{i,j} \in 0, 1$ we can use the distance $|x_{i,j} - y_{i,j}|$. Please note that a specific weight to this distance can be given by setting $w_{i,j}$.

The same distance can be used whenever $x_{i,j}, y_{i,j} \in \mathbb{R}$. However, more sophisticated metric distances could be used for specific numerical attributes. As

[10] http://sysrun.haifa.il.ibm.com/sapir/index.html

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

	$Notice$	$Attr$	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 2: Distance values for attributes taken from terms in a given dictionary (*left*) and proposed values for *CreativeCommons* terms for the *Requirements* (*right*)

an example, for fees we suggest to define the distance as:

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

since given a fee as query, the user would be much more interested on the proportion between its query and a given fee. Unfortunately any non 0 fees would be at infinite distance from 0 objects. To avoid this problem we suggest that whenever the fee is 0 the value used for evaluating the distance is 0.01. Please note that the well known distance gap ratio $((x - y)/y)$ would avoid the 0 fee problem but it is not a metric.

For an attribute whose value can be a term in a given vocabulary, we propose a specific 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 (left).

It is assumed that the values of α are manually chosen according to the semantic of the given terms. In particular, if all $\alpha^{i,j}_{a,b} = 1$ when $a \neq b$, textual attributes are considered as binary attributes. For $d_{i,j}(x_{i,j}, y_{i,j})$ to be a metric the matrix must be symmetric and all the diagonal values must be 0 and

$$\forall l \ , \ \alpha^{i,j}_{a,b} \leq \alpha^{i,j}_{a,l} + \alpha^{i,j}_{l,b} = \alpha^{i,j}_{l,a} + \alpha^{i,j}_{l,b} \ .$$

Let x and y be the metadata about the license attributes. Considering for example the CreativeCommons schema for the *Requirements* [11] (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 2 (right), 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.

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

Using the *triangle inequality* reported above, it can be shown that 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 license as the weighted sum between the main groups, this global distance is still a metric one. For indexing 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 [Fagin 1996]. 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 [Fagin 1996], 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:

$$d(x, y) = \sqrt[k]{\sum_{i=1}^{n} w_i \cdot |d_i(x_i, y_i)|^k} \tag{3}$$

The same approach could be used for combining the distance values among the attributes of the same group.

4 Significant Use Cases for Photo Search

Most of the search engines available on the Web provide nothing but the "full text" and/or "attribute" search capabilities. However, many research projects are developing audio and image "similarity" search. According to our proposal, a user will be able to search for an image similar to the one provided considering both the multimedia content (content base) and the related license (provided by the user as well). Furthermore the user can apply for searching similar images regarding the multimedia content and a specific kind of license defined by mean 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. Two important combination scenarios are:

1. The user is searching for images similar to a given one both considering its visual appearance and license "file"

2. The user is searching for images similar to a given one but with a license similar to a different one (informations from other images can be provided)

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 in which the user has an image which satisfied his needs both in terms of content and license. The search engine will display as result the ranked list of images similar to that provided according to the content and to the license. In the second case the user has an image which he does like, but that has a license which does not satisfy its needs. The user can search for an image similar to the given one but with a different license. In this case the license the results are requested to be similar to a license that can be either taken from another image or specified using a form.

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

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 the user has to deal with. This approach considers the IPR attributes as *special features* which a specific distance function can be applied to. For efficiently indexing the data it is important that this distance is a metric.

The Right Management warrantee has been deeply studied in the last few years and lots of solutions are available. However not much has been done 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 over the licenses could be not enough to handle the problem. We propose the adoption of the *Similarity Search* for the IPR attributes. In this way the license we are looking for can be easily provided, instead of all the attributes of a specific license format in a complex GUI. Moreover, we can also have a ranked list of results, according to the metric function, by defining the distance between the licenses.

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).

References

[Messina et al. 2006] A. Messina, L. Boch, G. Dimino, W. Bailer, P. Schallauer, W. Allasia, M. Groppo and M. Vigilante: "Creating Rich Metadata in the TV Broadcast Archives Environment: The PrestoSpace Project"; Proc. Axmedis 2006, IEEE Computer Society (2006), 193-200.

[NAVSHP 2005] Networked Audiovisual Systems and Home Platforms Group: "NAVSHP (FP6) DRM Requirements Report". Technical Report, European Community 6^{th} Framework Programme (Sep 2005).

[Doctorow 2006] Cory Doctorow, European Affairs Coordinator: "Critique of NAVSHP (FP6) DRM Requirements Report"; Technical report, Electronic Frontier Foundation (2006).

[MPEG-21 REL 2005] ISO/IEC - Information Technology - Multimedia Framework (MPEG-21), 21000-5 (2005).

[Iannella 2002] R. Iannella: "Open digital rights language (ODRL)"; Version 1.1, World Wide Web Consortium, W3C Note (2002).

[Roberto García González 2006] Roberto García González: "A Semantic Web Approach To Digital Rights Management"; PhD Thesis, Department of Technologies, Universitat Pompeu Fabra, Barcelona, Spain (Apr 2004).

[Zezula et al. 2006] P. Zezula, G. Amato, V. Dohnal and M. Batko: "Similarity Search. The Metric Space Approach"; Volume 32 of Advances in Database Systems. Springer, Heidelberg / New York (2006).

[Ciaccia et al. 1997] P. Ciaccia, M. Patella and P. Zezula: "M-tree: an efficient method for similarity search in metric spaces"; Proc. VLDB '97: 23^{rd}, Morgan Kaufmann, Publishers Inc. (1997), 426-435.

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

[Samet 2006] H. Samet: "Foundations of Multidimensional and Metric Data Structures"; Computer Graphics and Geometric Modeling. Morgan Kaufman Publishers Inc. (2006).

[Batko et al. 2004] 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 Architecture. 6^{th} Thematic Workshop of the EU Network of Excellence DELOS. LNCS Springer, 3664 (2004), 25-44 .

[Falchi et al. 2005] F. Falchi, C. Gennaro and P. Zezula: "A content-addressable network for similarity search in metric spaces"; Proc. DBISP2P '05, LNCS Springer, 4125 (2005), 98-110.

[Novak and Zezula 2006] D. Novak and P. Zezula: "M-chord: a scalable distributed similarity search structure". Proc. Infoscale'06: 1^{st}, ACM Press (2006), 19.

[Batko et al. 2006] M. Batko, D. Novak, F. Falchi and P. Zezula: "On scalability of the similarity search in world of peers"; Proc. Infoscale '06, ACM Press (2006), 20.

[Fagin 1996] R. Fagin: "Combining Fuzzy Information from Multiple Systems"; Proc. Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM Press (1996), 216-226.