Publication Framework for Content-Based Search in Heterogeneous Distributed Multimedia Databases

François Boisson, Michel Crucianu, Dan Vodislav
Vertigo - CEDRIC & Wisdom, CNAM
292 rue St. Martin, 75141 Paris cedex 03, France

10 September 2008

Abstract

Search for multimedia content by the large public is expected to experience at least two major evolutions in the near future: addition of content-based criteria to those issued from textual metadata and significant increase in the volume of digitized content available at each provider. In this context, most providers will need a local index in order to search in their own database. We argue that the local indexes are unlikely to be all of the same type and that many providers want full control over their local index. To support these requirements in a flexible and scalable way, we outline a framework based on the publication by every provider, in a consensual format, of *representatives* associated to those parts of the local index where the provider intends to make content "findable" by others. The representatives published by all the providers serve to build a global distributed index employed for answering content-based similarity queries.

1 Introduction

The evolution of user requirements and of enabling technologies will have a significant impact on how online search for multimedia content is performed. A major demand of the users is to avoid the "keyword bottleneck", in part by the inclusion of *content-based* search criteria [21]. But indexing the content of images, videos or music is very different from indexing hypertext, both because of the nature and volume of the data and because of specific rights issues. To clarify the implications of large scale content-based search in distributed repositories and highlight some research challenges that we deem important in this context, we start by taking a closer look at two major types of existing or potential multimedia content providers: the general public and large institutional archives.

With widespread digital imaging and cheap high-capacity storage, end users became creators and potential providers of multimedia content. According to the prevailing paradigm for access to user-generated content (UGC), producers upload their content to the central servers of a provider that makes it available to a large public. Consequently, content creators have difficulties in keeping control over their content. This approach also has scalability problems because of the reliance on central servers. One solution could consist in the use by the provider of a distributed system instead of central servers and in the setting up of comprehensive contractual relations between the creators and the provider. Another solution, that is likely to coexist with the previous one, would be the establishment of a peer-to-peer (P2P) network where every content producer is also a provider. To make search possible in such a network, a global index has to be created and then distributed among the peers. In Section 2.2 we mention some recent proposals that focus on content-based multimedia retrieval (CBMR) and follow this paradigm. These proposals assume that all the peers employ the same content descriptors and agree with their dissemination over the network, according to the distributed indexing scheme. However, with the recent availability of inexpensive digital video cameras, a significant growth can be expected for the volume of multimedia content owned by each peer. An index can then become necessary even for the content of a single peer and, to remain autonomous, the peer may prefer to have a local copy of his index. It can be expected that end-user software for content-based indexing of images and video will be released in the near future by several editors, using various image or video descriptors and multidimensional index structures.

After significant efforts dedicated to the digitization of multimedia content (e.g. cultural heritage, scientific documents), many institutional archives now provide facilities for the online exploration of all or part of their collections. We can mention, for example, the Institut National de l'Audiovisuel (`http://www.ina.fr/archivespourtous/`), the Photo Agency of the Réunion des Musées Nationaux (`http://www.photo.rmn.fr/`) or the European Archive (`http://www.europarchive.org`). Access to the digitized content is usually controlled and only synoptic representations (thumbnail images, low resolution video fragments) are freely available. To extend the capabilities of the current keyword-based search tools, several large institutional archives attempt to develop or acquire proprietary tools that can provide content-based indexing for their own collections of multimedia content. To optimize its use of resources, every institution would naturally fine-tune its indexing tool to the specific characteristics of the content it provides and of the users it is familiar with. Different archives will probably adopt different content descriptors and multidimensional index structures, depending on the specific content of the archive and on the selected technology provider. Also, an archive may independently decide to replace this supporting technology. An important issue for most institutional archives is keeping full control over the index of the content they own. First, many archives do not accept to make public a major part of their index because they consider this index to be a *core asset* for their activity. Second, in order to be able to control what part of the content is searchable ("visible") from the outside, an archive should not make public its full index. The concern for "visibility control" (or, more generally, for provider-side privacy) may then also become significant for user-generated content.

The above discussion highlights the fact that content is likely to be widely distributed among many autonomous providers that employ *heterogeneous* local index structures (and maybe heterogeneous content descriptions) and are *concerned by the privacy* of their index. How can then global, uniform, efficient and effective search capabilities be proposed to the users? Indeed, a user can hardly accept to search a high number of archives independently; having to deal with a specific search tool for every archive would be a further hindrance. We suggest here a general publication framework that aims to make the requirements of the providers compatible with the demands of the users. We believe that a provider would accept to pay a reasonable price in order to make its content visible through such a global and uniform search network. We only focus on the *search* stage; *access* to the content that was found is an independent issue, not addressed here.

After a brief description in Section 2 of content-based multimedia retrieval and a review of existing work regarding distributed multidimensional or metric indexing, a more thorough description of the problems under consideration is provided in Section 3. Section 4 outlines a general framework for taking into account the above mentioned heterogeneity and privacy requirements. It is based on the publication by every provider, in a consensual format, of *representatives* associated to those parts of the local index where the provider intends to let others find the content it owns. Some potential problems related to this publication framework for distributed search are then mentioned in Section 5 and possible preventive measures are discussed.

2 Index structures for CBMR

Before studying the issues raised by heterogeneity and provider-side privacy in a widely distributed environment, the basics of content-based multimedia retrieval are reminded and existing work on corresponding multidimensional or metric index structures is briefly revised.

2.1 Content-based search for multimedia

Keyword-based search is still the leading paradigm for querying multimedia databases, mainly because it is so easy to use and can be so readily provided when the content is annotated. But it also has important drawbacks: it is language-dependent, annotating is expensive and inherently incomplete, the relation between words and concepts is complex (e.g. synonymy, homonymy) and many search criteria just can't be well described in a few words. These problems boosted research activities in the field of *content-based* multimedia retrieval, where various descriptors are automatically extracted from the content of multimedia documents and serve to evaluate the similarity between documents (see

[9], [30], [14], [4]).

A descriptor is either computed for an entire document (e.g. edge orientation histogram for an image) or for a part of a document (e.g. differential description for an *interest point* in an image). The evaluation of similarity (or dissimilarity) can focus on a single characteristic or combine several complementary characteristics; for example, when comparing images one can consider color alone or combine color, texture and shape, etc. Also, a document can be represented by a single vector associated to one or several complementary descriptors of the entire document, by a set of vectors associated to different parts of the document (e.g. to the interest points found in an image), by a graph, etc. It follows that in some cases the dissimilarity can be evaluated using a standard metric in a vector space, but in many relevant situations a more complex measure has to be employed (e.g. a metric defined between sets of vectors or between graphs).

Content-based retrieval typically relies on such low-level descriptions automatically extracted from the multimedia content. CBMR in general also has limitations, mainly as a consequence of the *semantic gap* between this low-level description and the higher level descriptions that users may find meaningful for evaluating the similarity between documents. Among the various proposals that aim at a reduction of the semantic gap [21], can be mentioned here the use of *intermediate-level* descriptors (e.g. exploiting, for images, spatial relationships between interest points) or of search paradigms relying more on supervised learning.

The content-based retrieval paradigm that was applied to the highest diversity of documents is Query by Example (QBE). It consists in using a document (or a part of it) as query, to obtain a list of documents ordered by decreasing similarity between their descriptions and the description of the query. One can distinguish *within distance* search, where the system must return all the documents whose distance to the query is smaller than a given bound, from *k-nearest-neighbor* (kNN) search, where the k documents that are closest to the query (for a given k) are returned.

A related paradigm can be employed when no relevant document is available as initial query but the user can create a simplified document and employ it as a query: query by sketch (e.g. [2]) for image retrieval and query by humming (e.g. [19]) for music retrieval. Another solution when no relevant document is initially available is to start from a *visual summary* of the database, consisting of the prototypes of document clusters. Search with hierarchical summaries can also be seen as a full-grown CBMR paradigm and not just as an initialization stage. Mapping the data to a low-dimensional space supports the visualization of summaries (e.g. [20] for images or [26] for music). Another retrieval paradigm exploiting visual summaries is the logical composition of image region categories [13], which is an image retrieval counterpart of Boolean queries using keywords.

Since they usually rely on a predefined metric in the space of the low-level descriptions, QBE and the creation of database summaries suffer most from the consequences of the semantic gap. To bridge the gap while still using low-level descriptions, some supervision from the user is needed. A CBMR paradigm following this idea is iterative search with *relevance feedback*: during consecutive rounds, the user qualifies the documents returned by the system as either relevant or irrelevant, allowing the system to progressively identify the characteristics of user's target (see e.g. [31], [8]). Relevance feedback was mainly studied for image databases because users can immediately evaluate the relevance of an image.

Our focus here is on query by example because it can valuably complement keyword-based search and a more extensive use of intermediate-level descriptors should make it less prone to the semantic gap.

2.2 Data structures for content-based search

Query by example relies on the evaluation of the similarity between the description of the query and the descriptions of potential candidates. Multidimensional or metric index structures (with associated retrieval methods) can be employed in order to make the complexity of search sub-linear with respect to the size of the database. Most of them can address both within distance search and kNN search. Comprehensive reviews of centralized index structures can be found in [29], or [3] and [5].

Some recent work also concerns index structures for speeding up retrieval with relevance feedback, where the problem cannot be easily reduced to standard queries in the space of document descriptions. Indeed, in RF a query may correspond to a complex space region or to a frontier between relevant and irrelevant images. In [23], such a complex query is translated into a set of classical kNN queries and an existing index structure is used. It was shown that different RF-related queries could be simplified by transposing search in the feature space associated to a positive definite kernel function; a VA-file index was adapted for processing such queries in [16], a specific index structure for approximate search was put forward in [24] and an M-tree was employed with hyperplane queries in [7].

Search for distributed multimedia content was first developed in a P2P context, relying exclusively on the textual tags associated to the multimedia files in order to process *exact* queries. The P2P architectures employed were quite diverse [1]. For example, while Napster used a centralized P2P approach, another popular system, Gnutella, was a more resilient hybrid and unstructured P2P network [27]. Recent research focuses more on *similarity*-based search using descriptors automatically extracted from the multimedia content. Content itself is left at the originating peers and a global distributed index of all the available content is created for search purposes. The aim is to make all the data searchable in an uniform way, while minimizing the response time (and communication overhead) of the search operations and of the construction and update of the index at the arrival or departure of data or peers.

Given the large number of existing centralized index structures for similarity-based retrieval (usually using data partitioning, space partitioning or spatial approximation), many proposals focus on adapting these structures to a distributed setting. Every data or space partition, containing descriptions of content that may be situated on many different peers, is allocated to one or a few peers who will answer similarity queries regarding this partition. This allocation either relies on distributed hash tables (DHT, [10]) or uses specific solutions; queries are routed accordingly. A rather general solution, applicable to different tree-based data partitioning index structures, is the VBI-tree put forward in [17]. Some methods attempt to make distributed indexing schemes initially defined for unidimensional range queries work with multidimensional similarity-based retrieval, e.g. by using DHT with locality sensitive hashing [15] or with a unidimensional mapping of metric relations [22]. Finally, some indexing schemes were developed from the start for multidimensional distributed data, such as the content-addressable network [25] (extended to metric data in [12]) or the search based on reference vectors [28].

There is a rather wide consensus among recent proposals to use a *structured* P2P approach, where either the content or the index structure is distributed among the peers following some straightforward computation (usually hashing). Structured P2P systems provide efficient retrieval, with good load balancing and low communication overhead. However, this approach based on homogeneous sharing of the content or index is not appropriate if peers restrict the access to their local content and index or if they already provide heterogeneous local indexing and retrieval services.

3 Extending the P2P framework

Our focus in the following is on these new conditions—privacy and heterogeneity of the local indexes—and on their impact on the content-based search for multimedia in distributed repositories.

3.1 Provider-side privacy

Content providers think of both their local content and their local index as being private, do not intend to distribute a large share of them and want to fully control the access to them following archive-specific rules. Clearly, the standard structured P2P approach cannot be directly applied in this context.

While index privacy and access control are seen as prerequisites by many content providers, they also want part of their content to be “visible” for global, uniform searches that can query the entire network of archives. Therefore, they must agree to feed a global index with a minimum of information regarding their own content.

We propose that each archive should publish to the global index a *coarse view* of its local index. This coarse view is composed of a set of *representatives* for the local data, building up a controlled summary of the multimedia content owned by the archive. An archive manages a local index for its data using the descriptions extracted from every content unit; this index is used to define the representatives that aggregate sets of indexed descriptions to form the coarse view. The global index gathers all the representatives published by each archive in a common, distributed structure.

A global content-based search uses the global index to detect the nearest coarse view representatives, then the query is directed to the corresponding archives to get references to local results. This approach guarantees provider-side privacy, since the local index is kept private and only a coarse view of it is published at the global level. Each archive also keeps full control over the access to its own data, because queries are only allowed to return references to the content.

3.2 Levels of heterogeneity

Another important difference with respect to standard structured P2P management for multimedia data comes from the *heterogeneity* of the archives, which is a natural consequence of their autonomy. We consider here two heterogeneity parameters, (i) the *type of descriptor*, and (ii) the *index structure* used by each content provider.

One extreme situation is full homogeneity, i.e. every archive uses the same descriptor and the same type of local index. In this case, the global index can be of the same type as the local indexes. A query in the global index will return a set of representatives coming from different archives; for each representative, a query to the corresponding archive will return a set of results. An important issue is that the global index not only allows to select the archives that should be queried, but may also *accelerate* the search in each local archive by keeping a backward pointer from each representative in the global index to its source component in the local index. If every archive would publish its entire local index (including the descriptions), the global index would be sufficient for answering queries. This is the standard structured P2P approach using a distributed index.

Another extreme situation is when different archives employ different multimedia content descriptors. To help performing global searches, a common description scheme has to be found and every archive must identify a mapping between the local and the common description scheme (e.g. approximate mapping between histograms with a different number of bins in the same feature space). During search, the description of every query must be directly computed using the common scheme or translated by the inquirer into this scheme; eventually, every archive receiving the query should translate it into its local scheme, perform the retrieval operations locally and then translate the results into the common scheme before sending them as an answer in order to let the inquirer make the final selection for a kNN query. Such a solution is conceivable for specific description schemes, but not in a general case. Moreover, it generates a significant overhead and loss in precision. In the worst case, no global index can be created and each query, packed with its full multimedia content, should be directly sent to every archive in the network; each archive would extract its own descriptions from the query and rely on its own local index to answer the query. The inquirer would simply put all the results together, since it can not compare the different similarity values returned by the different archives.

Since the use of heterogeneous descriptors appears to raise insurmountable difficulties for global searches, our focus in the following is on *homogeneous* *descriptors and associated metrics*, but *heterogeneous index structures*. All the content providers must employ the same descriptors for their content, but their local indexes can be different. The main problem that has to be solved is the definition of a global indexing scheme and of a mapping between this scheme and various local indexes.

4 Publication-based framework

Our proposal is based on the idea that each archive publishes a *summary* (a coarse view) of its indexed content. The local index of each archive is composed of the descriptions (*signatures* in the following) of all the content elements, organized into a structure (usually hierarchical) whose nodes correspond to regions of the description space that are obtained by various approaches [29] such as data partitioning, space partitioning or space approximation. We consider *homogeneous descriptions*: all the archives use the same descriptor type, so they all share the same description space.

We suggest that the coarse view published by an archive consists in a *set of representatives* of the content. Each representative is an aggregation of the set of signatures corresponding to a region in the local index of the archive. The construction of representatives is *driven by the local index*, but subject to constraints defined at a global level. Fig. 1 illustrates a possible mapping between the local index of an archive and the set of representatives published by the archive. A representative corresponds here to a node of the local index, more precisely to the set of signatures in that subtree and the corresponding range. Furthermore, each representative stores a backward pointer toward the source node of the local index, to be used at query processing.

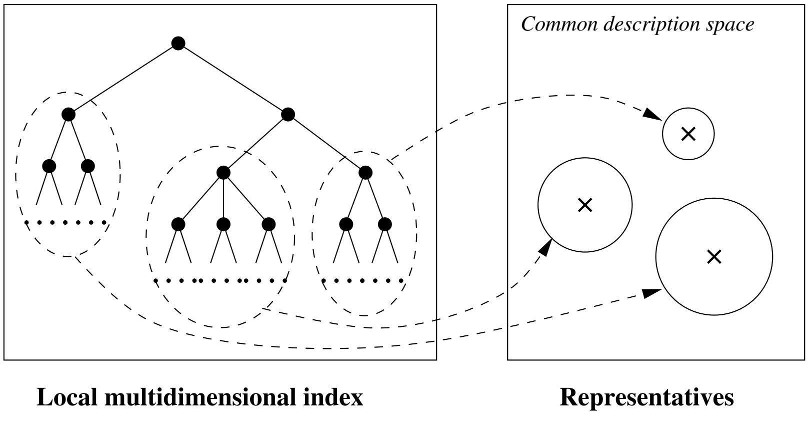

Figure 1: Local index and representatives

The selection of the index nodes (or regions) that produce representatives depends on the index type and on criteria such as the number of signatures in the region or the size of the region, further described below. An archive may decide to publish only *part* of its indexed content and may choose different “levels of granularity” for representatives coming from different parts of its index.

The representatives published by all the archives share the same description space and can be used to build the global index necessary for query processing. The global index is distributed among the content providers.

4.1 Representatives: definition, publication

A representative aggregates a set of content signatures from an archive, corresponding to a part of the common description space; this part is determined by the local index structure, for example as in Fig. 1.

Individual archives can use multidimensional (vector) or metric index structures. In multidimensional indexes, regions associated to individual nodes can have different shapes (depending on index type): multidimensional rectangles, spheres, pyramids, etc. In the more general case of a metric index, a node is usually associated to a “sphere” containing all the data points within some distance to a central point. For generality, we retain spheric regions (defined by a center and a radius) that can be employed for both vector and metric spaces. All the archives share a unique description space, which is either vectorial or just metric. Both local vectorial or metric indexes are supported in the former case, while only metric indexes can be used in the latter one.

We will consider that a representative is defined by:

- A *spheric region*, characterized by a covering radius and containing all the signatures of the local index that are aggregated by the representative.
- A *central point*, i.e. a signature in the description space that is the center of the spheric region. Given the metric space requirements, this signature must be one of the points from the local index region aggregated to produce the representative.
- The *number of signatures* aggregated by the representative, that can be used to determine the *density* of signatures (or a comparable measure in a metric space) within the region of the representative.
- A *backward pointer* to the local index node that produced the representative (e.g. the network ID of the archive site, plus path and filename of its local index, plus address of the page on mass storage).

The *granularity* of the local index regions that would produce a representative may be decided on the basis of various criteria, such as minimal and maximal thresholds for the size of the region, for the number of signatures within the region, for the density of signatures, etc. Each archive may have its own criteria for defining the thresholds employed locally, depending on the granularity level intended for publication, but also on the type of local index structure and on the distribution of its signatures in the description space.

The choice must take into account two conflicting goals of each content provider: (i) the need for *privacy*, requiring regions that are large enough to hide the details of the local index, and (ii) the need for *precision*, requiring representatives that are only selected by a query if they have good chances to provide relevant results and suggesting to consider dense and "not-too-large" regions. Other criteria, expressed from the point of view of the network, highlight a similar conflict: (i) larger regions, to minimize the number of representatives (so the size of the global index), and (ii) smaller dense regions, to improve precision and avoid useless communication. Even if all these criteria can be expressed by minimum and/or maximum thresholds for the number of signatures, the radius and the density, the publishing algorithm may have to handle exceptions, such as representatives for isolated points.

As an example, a simple algorithm for defining the local index regions to be published as representatives could follow a top-down traversal of the local index tree. Suppose that the granularity criteria provide an upper bound for the radius and a lower bound for the density. Then, the publication of representatives starts with the root node and applies the granularity test: the main idea is that if the region associated with the node is too large or not dense enough, the algorithm continues by examining the children of the current node (if they exist), otherwise the node's region is used to generate one or several representatives.

Algorithm 1 presents this generic, recursive algorithm, called ***publishRepresentatives***. The inputs are *a node* of a local index tree and the *publication predicate*, expressed using granularity conditions (e.g. an upper bound for the radius and a lower bound for the density) and specific conditions for "visibility control" (e.g. excluding parts of the local index). If the node satisfies the predicate, it is used to publish a representative through a call to the ***publishNode*** method. Note that in some cases ***publishNode*** may produce several representatives, depending on the type of the local index, as illustrated below.

If the node does not satisfy the publication test, it cannot produce a representative; its underlying signatures should be further divided in subsets that might satisfy the test. If the node is a leaf, this is performed by the ***publishLeaf*** method, specific to the type of the local index. Note that ***publishLeaf*** may produce a *singleton*, i.e. a representative composed of only one signature. If the node is not a leaf, the ***publishRepresentatives*** algorithm is recursively called for every child of the node.

To deal with heterogeneous local indexes, appropriate solutions are required for mapping different types of index structures to spheric regions for defining representatives. For instance, in the case of an R-tree index (only efficient in low-dimensional description spaces) using multidimensional rectangular regions, one could determine the minimal bounding sphere of all the signatures in the rectangle. For a pyramid tree (more appropriate for higher-dimensional spaces), the indexed regions ("slices" of a multidimensional pyramid) are less appropriate for being approximated with a sphere because of their shape; it is possible to cover the signatures in a slice of the pyramid by a set of spheres, thus producing several representatives. We aim at devising an inventory of precise translation methods from various index structures to such spherical representatives, taking

Algorithm 1 A simple generic algorithm ***publishRepresentatives*** for the recursive identification and publication of representatives in a tree-like local index

```
Require: $n$ = node of the tree-like local index,
    $P$ = predicate for the publication test on an index node
Ensure: the result is a set of representatives to publish
if $P(n)$ then
    return publishNode($n$)
else
    if leaf($n$) then
        return publishLeaf($n$)
    else
        $result \leftarrow \emptyset$
        for all $c \in children(n)$ do
            $result \leftarrow result \cup$ publishRepresentatives($c$, $P$)
        end for
        return $result$
    end if
end if
```

into account the impact of the high dimension of typical vectorial descriptions of multimedia content.

The recent proposal in [11] also defines a kind of representatives in a metric space in order to allow the construction of a global distributed metric index, relying on clustering operations performed by every peer rather than on heterogeneous local indexes.

4.2 Global index: construction, distribution

The global index is built from the representatives published by all the archives. As explained above, our intention is to be as general as possible in handling local index heterogeneity. Consequently, we make the choice of using a metric index structure (like the M-tree [6]) for the global index, compatible with both vector and metric local indexes. A new feature is the use of representatives (instead of ordinary data points) with specific characteristics such as range and density, which has an impact on the construction of the index and on search, thus requiring an extension of the existing index structure.

The size of the global index may grow very fast, depending of the number of providers (archives), the sizes of their local indexes and the granularity of the published representatives. For scalability and robustness reasons, the global index is distributed among the providers. We intend to explore several distribution techniques for metric indexes, focusing on *structured P2P overlays* that guarantee good retrieval performance, and compare them with distributed metric indexes such as M-Chord [22] or M-CAN [12].

Figure 2 illustrates the construction and distribution of the global index.

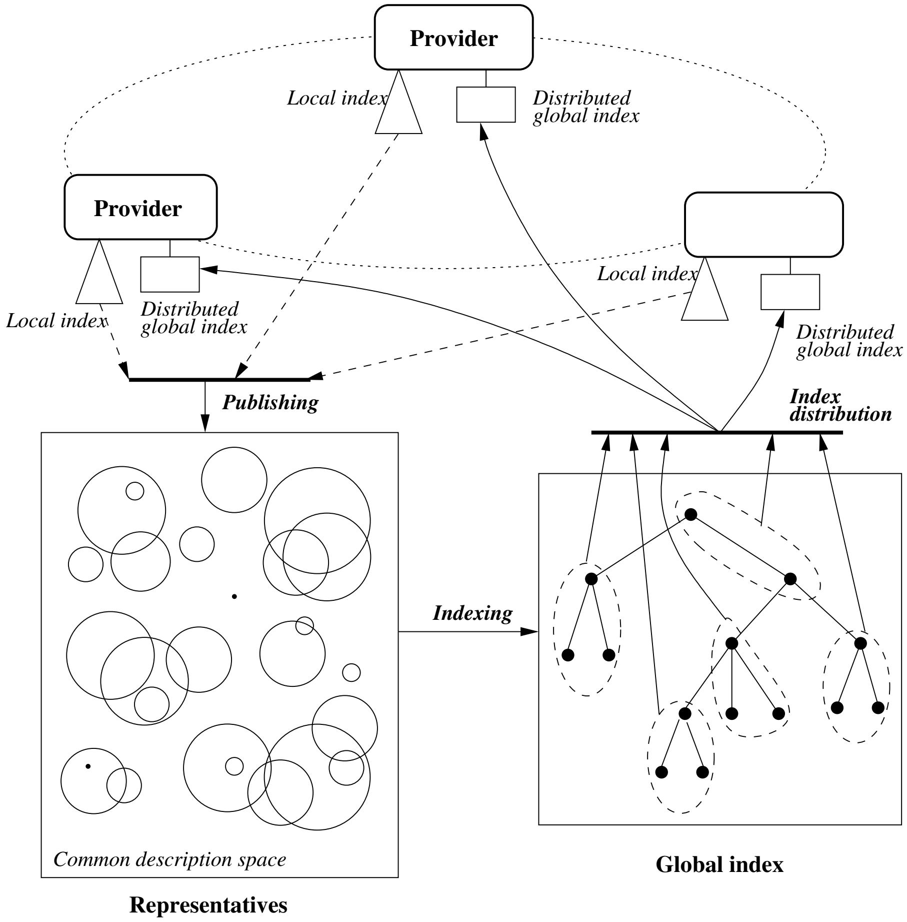

Figure 2: Construction and distribution of the global index

Each provider publishes a set of representatives based on its local index, by using a variant of the ***publishRepresentatives*** algorithm. Representatives published by all the providers share a common description space (and associated metric), as explained above. This global set of representatives is indexed following a metric indexing schema, in order to produce the global index. The distribution of the global index is performed by partitioning the index tree into small segments (nodes or subtrees) that are stored by the provider sites, following a specific distribution algorithm. Searching the global index, necessary for both querying and index update (when new representatives are published) is based on the same distribution algorithm.

4.3 Uniform search with the distributed index

The processing of similarity queries (within-distance or kNN queries) over the whole set of archives is driven by the site that receives the initial user query. We distinguish the following query processing phases:

- *Source identification*: the signature extracted from the query object is used for finding, with the help of the global distributed index, the set of representatives that should provide the relevant answers. A within-distance query defines a query sphere in the description space; only the representatives whose regions intersect the query sphere are selected by using the distributed search in the global index. For processing kNN queries, existing variable-range algorithms can be adapted to take advantage of the information provided by every representative regarding the number of signatures in the region it covers. This phase selects a set of representatives that cover the complete answer with the signatures they aggregate. Only the sources that published the representatives retained during this first phase are queried in order to produce the final result.

- *Local query processing*: every archive selected in the first phase receives the query parameters (query point and radius or k) and the set of selected representatives from that archive. These archives perform in parallel the local queries and return their results, sorted by increasing distance from the query, to the site that launched the search. For each representative, the backward pointer is employed to directly access the local index node that produced it. Search starts from these nodes rather than from scratch, resulting in a significant speedup of local query processing. For within-distance queries all the signatures within the range are returned, while for kNN queries the number of local answers is limited to the k best ones (or less, if additional information is available about results form other sources).

- *Merging local results*: the sorted lists of results coming from the selected archives are merged to produce the global list, sorted by decreasing similarity with the query. Each result contains a reference to the actual content (or to an associated synoptic representation), that can be used to retrieve the content from the corresponding archive, following the access rules defined by this archive. For within-distance queries all the local results are part of the global answer, while for kNN queries only the best k are retained in the final result.

5 Preventing abnormal behavior

The design of a distributed network dedicated to content-based search for multimedia documents should also address (and propose solutions to) some problems that are likely to occur in this environment, due to an unfair use of the network.

We will only consider here two such problems: spam and implicit transfer of load.

Spam is a well-known phenomenon in information retrieval and can also be expected in the context of content-based search for multimedia documents in distributed autonomous repositories. Since search is performed through the signatures and not directly through the content itself, a provider may pretend to have the sought content by returning high similarity scores (and even highly relevant signatures, obtained by querying the network) and then send the user a different content, such as an unsolicited advertisement.

Among the potentially many answers to this kind of problem we only highlight two: certification and community-based reputation. Certification can be rather restrictive and consists in providing the ability to check whether the signatures returned by a provider were indeed obtained by that provider on content it legitimately owns and with certified software. Since certification concerns the generation of signatures that should all be obtained in the same way and belong to the same description space, we consider that the constraints it introduces are reasonable. Community-based reputation can be progressively built with the help of user feedback (see e.g. [18]). Every provider should have an overall reputation score but also specific scores for different parts of the content description space.

The implicit transfer of load is what can happen when a provider does not want to pay the cost of maintaining a local index for the signatures of its content and perform local searches; instead, it publishes all its signatures as representatives and, thanks to the distributed index, relies on the other providers in the network to do the work. In this case, certification may introduce undesired constraints because it would have to verify the way representatives are generated by the *heterogeneous* local index structures. But a distributed mechanism based on generating queries can be defined for experimentally checking whether the published representatives have indeed the characteristics they pretend to have (range and density). Those representatives that do not respect the thresholds imposed by the network would progressively be removed from the global index (e.g. those that are found to be too “hollow”: too small covering radius or too low density). The additional use of network-defined reputation can speed up this process.

6 Conclusion

The number of multimedia content providers and the volume of content they can make available online are expected to increase fast in the near future. To reduce the “keyword bottleneck”, the retrieval of multimedia documents is likely to include content-based criteria to a higher extent. Given the volume of content they own and the need for content-based search, the providers need local index structures to support fast retrieval. Autonomous providers would prefer to keep their index private since they may consider it a core asset for their activity and want to control what part of their content is “visible”. Because they can have specific content and can adopt indexing solutions from various software editors, providers may end up employing different local index structures. In this context, we argue that the current peer-to-peer model for content-based retrieval of multimedia content, based on a complete global index and on the open distribution of all content signatures, may need to be revised.

In order to make their content visible through a global and uniform search network, the providers may nevertheless accept a few compromises, such as using a consensual content description scheme and publishing a small part of their index. To support both the privacy and the heterogeneity of the local indexes, we outlined a general framework based on the publication by every provider, in a common format, of representatives associated to those parts of the local index where the provider intends to make content findable. This framework is general and should accommodate several different local indexes, as well as various solutions for the global index. Our on-going work aims to prove this concept by proposing a complete algorithmic solution and an associated software implementation.

References

[1] K. Aberer and M. Hauswirth. An overview of peer-to-peer information systems. In W. Litwin and G. Lévy, editors, *WDAS*, volume 14 of *Proceedings in Informatics*, pages 171–188. Carleton Scientific, 2002.

[2] A. D. Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(2):121–132, 1997.

[3] C. Böhm, S. Berchtold, and D. A. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322–373, 2001.

[4] N. Boujemaa, J. Fauqueur, and V. Gouet. What's beyond query by example? In R. V. L. Shapiro, H.P. Kriegel, editor, *Trends and Advances in Content-Based Image and Video Retrieval*. Springer Verlag, 2004.

[5] E. Chavez, G. Navarro, R. A. Baeza-Yates, and J. L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[6] P. Ciaccia, M. Patella, and P. Zezula. M-tree: an efficient access method for similarity search in metric spaces. In *Proceedings of the 23rd IEEE International Conference on Very Large Data Bases (VLDB'97)*, pages 426–435, Athens, Greece, August 1997.

[7] M. Crucianu, D. Estevez, V. Oria, and J.-P. Tarel. Speeding up active relevance feedback with approximate kNN retrieval for hyperplane queries. *International Journal of Imaging Systems and Technology, Special issue on Multimedia Information Retrieval, in press*, 2008.

[8] M. Crucianu, M. Ferecatu, and N. Boujemaa. Relevance feedback for image retrieval: a short survey. In *State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction, Including Datamodels and Languages*. DELOS Network of Excellence, June 2004.

[9] A. del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann, 1999.

[10] R. Devine. Design and implementation of DDH: A distributed dynamic hashing algorithm. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO'93)*, pages 101–114, London, UK, 1993. Springer-Verlag.

[11] C. Doulkeridis, A. Vlachou, Y. Kotidis, and M. Vazirgiannis. Peer-to-peer similarity search in metric spaces. In *Proceedings of the 33rd international conference on Very large data bases (VLDB'07)*, pages 986–997. VLDB Endowment, 2007.

[12] F. Falchi, C. Gennaro, and P. Zezula. Nearest neighbor search in metric spaces through content-addressable networks. *Information Processing and Management*, 43(3):665–683, 2007.

[13] J. Fauqueur and N. Boujemaa. New image retrieval paradigm: Logical composition of region categories. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'2003)*, Barcelona, Spain, September 2003.

[14] T. Gevers and A. W. M. Smeulders. Content-based image retrieval: An overview. In G. Medioni and S. B. Kang, editors, *Emerging Topics in Computer Vision*, chapter 8. Prentice Hall, 2004.

[15] A. Gupta, D. Agrawal, and A. E. Abbadi. Approximate range selection queries in peer-to-peer systems. In *Proceedings of the First Biennial Conference on Innovative Data Systems Research*, Asilomar, California, United States, January 2003.

[16] D. R. Heisterkamp and J. Peng. Kernel VA-files for relevance feedback retrieval. In *Proceedings of the first ACM international workshop on Multimedia databases*, pages 48–54, New Orleans, LA, USA, 2003. ACM Press.

[17] H. V. Jagadish, B. C. Ooi, Q. H. Vu, R. Zhang, and A. Zhou. VBI-tree: A peer-to-peer framework for supporting multi-dimensional indexing schemes. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, pages 34–43, Washington, DC, USA, 2006. IEEE Computer Society.

[18] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 640–651, New York, NY, USA, 2003. ACM.

[19] N. Kosugi, Y. Sakurai, and M. Morimoto. Soundcompass: a practical query-by-humming system; normalization of scalable and shiftable time-series data and effective subsequence generation. In *Proceedings of the ACM SIGMOD international conference on Management of data (SIGMOD'04)*, pages 881–886, New York, NY, USA, 2004. ACM.

[20] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja. PicSOM—content-based image retrieval with self-organizing maps. *Pattern Recognition Letters*, 21(13-14):1199–1207, 2000.

[21] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transaction on Multimedia Computing, Communications and Applications*, 2(1):1–19, February 2006.

[22] D. Novak and P. Zezula. M-Chord: a scalable distributed similarity search structure. In *Proceedings of the 1st international conference on Scalable information systems (InfoScale'06)*, pages 19–28, New York, NY, USA, 2006. ACM Press.

[23] N. Panda and E. Y. Chang. Efficient top-k hyperplane query processing for multimedia information retrieval. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 317–326, New York, NY, USA, 2006. ACM Press.

[24] N. Panda, K.-S. Goh, and E. Y. Chang. Active learning in very large databases. *Multimedia Tools and Applications*, 31(3):249–267, 2006.

[25] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM'01)*, pages 161–172, New York, NY, USA, 2001. ACM.

[26] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003.

[27] M. Ripeanu. Peer-to-peer architecture case study: Gnutella network. In *Proceedings of the First International Conference on Peer-to-Peer Computing (P2P'01)*, pages 99–106, Washington, DC, USA, 2001. IEEE Computer Society.

[28] O. D. Sahin, A. Gulbeden, F. Emekci, D. Agrawal, and A. E. Abbadi. PRISM: indexing multi-dimensional data in p2p networks using reference vectors. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 946–955, New York, NY, USA, 2005. ACM Press.

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

[30] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[31] X. S. Zhou and T. S. Huang. Relevance feedback for image retrieval: a comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.