Multimedia Information Retrieval Model

Carlo Meghini[1], Fabrizio Sebastiani and Umberto Straccia[1]
[1] *The Italian National Research Council, Pisa, Italy*
[2] *Qatar Computing Research Institute, Doha, Qatar*

1 Synonyms

Content-based retrieval, semantic-based retrieval, multimedia information discovery.

2 Definition

Given a collection of multimedia documents, the goal of multimedia information retrieval (MIR) is to find the documents that are relevant to a user information need. A multimedia document is a complex information object, with components of different kinds, such as text, images, video and sound, all in digital form.

3 Historical Background

The vast body of knowledge nowadays labelled as MIR, is the product of several streams of research, which have arisen independently of each others and proceeded largely in an autonomous way, until the beginning of 2000, when the difficulty of the problem and the lack of effective results made it evident that success could be achieved only through integration of methods. These streams can be grouped into three main areas:

The first area is that of *information retrieval* (IR) proper. The notion of IR attracted significant scientific interest from the late 50's in the context of textual document retrieval. Early characterizations of IR simply relied on an "objective" notion of topic-relatedness (of a document to a query). Later, the essentially subjective concept of relevance gained ground, and eventually became the cornerstone of IR. Nowadays, IR is synonymous with "determination of relevance" [3].

Around the beginning of the 80's, the area of multimedia documents came into existence and demanded an IR functionality that no classical method was able to answer, due to the *medium mismatch problem* (in the image database field, this is often called the *medium clash problem*). This problem refers to the fact that, when documents and queries are expressed in different media, matching is difficult, as there is an inherent intermediate mapping process that needs to reformulate the concepts expressed in the medium used for queries (*e.g.* text) in terms of the other medium (*e.g.* images). In response to this demand, a wide range of methods for achieving IR on multimedia documents has been produced, mostly based on techniques developed in the areas of *signal processing* and *pattern matching,* initially foreign to the IR field. These methods are nowadays known as *similarity-based* methods, due to the fact that they use as queries an object of the same kind of the sought ones (*e.g.* a piece of text or an image) [2]. Originally, the term *content-based* was used to denote these methods, where the content in question was not the content of the multimedia object under study (*e.g.* the image) but that of the file that hosts it.

The last area is that of *semantic information processing* (SIP) which has developed across the information system and the artificial intelligence communities starting from the 60's. The basic goal of SIP was the definition of artificial languages that could represent relevant aspects of a reality of interest (whence the appellation *semantic*), and of suitable operations on the ensuing representations that could support knowledge-intensive activities. Since the inception of the field,

SIP methods are rooted in first-order mathematical logic, which offers the philosophically well-understood and computationally well-studied notions of syntax, semantics and inference as bases on which to build. Nowadays, SIP techniques are mostly employed in the context of Knowledge Organization Systems. In MIR, SIP methods have been used to develop sophisticate representations of the contents (in the sense of "semantics") of multimedia documents, in order to support the retrieval of these documents based on a logical model. According to this model, user's information needs are predicates expressed in the same language as that used for documents representations, and a document is retrieved if its representation logically implies the query. A wide range of logical models for IR have been proposed, corresponding to different ways of capturing the uncertainty inherent in IR, of expressing document contents, of achieving efficiency and effectiveness of retrieval [7].

To a lesser extent, the database area has also contributed to MIR, by providing indexing techniques for fast access to large collections of documents. Initially, typical structures such as inverted files and B-trees were employed. When similarity-based retrieval methods started to appear, novel structures, such as R- or M-trees were developed in order to support efficient processing of range and k nearest neighbors queries [15].

4 Scientific Fundamentals

MIR is a scientific discipline, endowed with many different approaches, each stemming from a different branch of the MIR history. All these approaches can be understood as addressing the same problem through a different aspect of multimedia documents.

Documents can be broadly divided from a user perspective into two main categories: simple and complex.

A document is *simple* if it cannot be further decomposed into other documents. Images and pieces of text are typically simple documents. A simple document is an arrangement of symbols that carry information via meaning, thus concurring in forming what is called the *content* of the document. In the case of text, the symbols are words (or their semantically significant fractions, such as stems, prefixes or suffixes), whereas for images the symbols are colors and textures. Simple documents can thus be characterized as having two parallel dimensions: that of *form* (or *syntax*, or *symbol*) and that of *content* (or *semantics*, or *meaning*). The form of a simple document is dependent on the medium that carries the document. On the contrary, the meaning of a simple document is the set of states of affairs (or "worlds") in which it is true, and is therefore medium-independent. For instance, the meaning of a piece of text is the set of (spatio-temporally determined) states of affairs in which the assertions made are true, and the meaning of an image is the set of such states of affairs in which the scene portrayed in the image indeed occurs.

Complex documents (or simply documents) are structured sets of simple documents. This leads to the identification of *structure* as the third dimension of documents. Document structure is typically a binary relation, whose graph is a tree rooted at the document and having the component simple documents as leaves. More complex structures may exist, for instance those requiring an ordering between the children of the same parent (such as between the chapters of a book), or those having an arity greater than 2 (such as synchronization amongst different streams of an audio-visual document).

Finally, documents, whether simple or complex, exist as independent entities characterized by (meta-)attributes (often called *metadata* in the digital libraries literature), which describe the relevant properties of such entities. The set of such attributes is usually called the *profile* of a document, and constitutes the fourth and last document dimension.

Corresponding to the four dimensions of documents just introduced, there can be four categories of retrieval, each one being a projection of the general problem of MIR onto a specific dimension. In addition, it is possible, and in some cases desirable, to combine different kinds of retrieval within the same operation.

Retrieval based on document structure does not really lead to a genuine discovery, since the user must have already seen (or be otherwise aware of) the sought document(s) in order to be able to state a predicate on their structure. Retrieval based on document profile, from a purely logical point of view, is not different from content-based retrieval and in fact many metadata schema used for document description (notable, the Dublin Core Metadata Set) include attributes of both kinds.

4.1 Form-based multimedia information retrieval

The retrieval of information based on form addresses the syntactic properties of documents. In particular, form-based retrieval methods automatically create the document representations to be used in retrieval by extracting low-level features from documents, such as the number of occurrences of a certain word in a text, or the energy level in a certain region of an image. The resulting representations are abstractions which retain that part of the information originally present in the document that is considered sufficient to characterize the document for retrieval purposes. User queries to form-based retrieval engines may be documents themselves (this is especially true in the non-textual case, as this allows to overcome the medium mismatch problem), from which the system builds abstractions analogous to those of documents. Document and query abstractions are then compared by an appropriate function, aiming at assessing their degree of similarity. A document ranking results from these comparisons, in which the documents with the highest scores occur first.

In the case of text, form-based retrieval includes most of the traditional IR methods, ranging from simple string matching (as used in popular Web search engines) to the classical *tf-idf* term weighting method, to the most sophisticated algorithms for similarity measurement. Some of these methods make use of information structures, such as thesauri, for increasing retrieval effectiveness; however, what makes them form-based retrieval methods is their relying on a form-based document representation. Two categories of queries addressing text can be distinguished:

1. *full-text* queries, each consisting of a *text pattern*, which denotes, in a deterministic way, a set of texts; when used as a query, the text pattern is supposed to retrieve any text layout belonging to its denotation;

2. *similarity* queries, each consisting of a text, and aimed at retrieving those text layouts which are similar to the given text.

In a full-text query, the text pattern can be specified in many different ways, *e.g.* by enumeration, via a regular expression, or via *ad hoc* operators specific to text structure such as proximity, positional and inclusion operators [3].

Queries referring to the form dimension of images are called *visual* queries, and can be partitioned as follows:

1. *concrete visual queries:* these consist of full-fledged images that are submitted to the system as a way to indicate a request to retrieve "similar" images; the addressed aspect of similarity may concern color [5,9], texture [10,14], appearance [12] or combination thereof [13];

2. *abstract visual queries:* these are artificially constructed image elements (hence, "abstractions" of image layouts) that address specific aspects of image similarity; they can be further categorized into:

(a) *color queries:* specifications of color patches, used to indicate a request to retrieve those images in which a similar color patch occurs [2,9];

(b) *shape queries:* specifications of one or more shapes (closed simple curves in the 2D space), used to indicate a request to retrieve those images in which the specified shapes occur as contours of significant objects [2,11];

(c) combinations of the above [5].

Visual queries are processed by matching a vector of features extracted from the query image, with each of the homologous vectors extracted from the images candidate for retrieval. For concrete visual queries, the features are computed on the whole image, while for abstract visual queries only the features indicated in the query (such as shape or color) are represented in the vectors involved. For each of the above categories of visual queries, a number of different techniques have been proposed for performing image matching, depending on the features used to capture the aspect addressed by the category, or the method used to compute such features, or the function used to assess similarity.

4.2 Semantic content-based multimedia information retrieval

On the contrary, semantic-based retrieval methods rely on symbolic representations of the meaning of documents, that is descriptions formulated in some suitable knowledge representation language, spelling out the truth conditions of the involved document. Various languages have been employed to this end, ranging from net-based to logical. Description Logics [4], or their Semantic Web syntactic forms such as OWL, are contractions of the Predicate Calculus that are most suitable candidates for this role, thanks to their being focussed on the representation of concepts and to their computational amenability. Typically, meaning representations are constructed manually, perhaps with the assistance of some automatic tool; as a consequence, their usage on collections of remarkable size (text collections can reach nowadays up to millions of documents) is not viable. The social networking on which Web 2.0 is based may overcome this problem, as groups of up to thousands of users may get involved in the collaborative indexing process (flicker).

While semantic-based methods explicitly apply when a connection in meaning between documents and queries is sought, the status of form-based methods is, in this sense, ambiguous. On one hand, these methods may be viewed as pattern recognition tools that assist an information seeker by providing associative access to a collection of signals. On the other hand, form-based methods may be viewed as an alternative way to approach the same problem addressed by semantic-based methods, that is deciding relevance, in the sense of connection in meaning, between documents and queries. This latter, much more ambitious view, can be justified only by relying on the assumption that there be a systematic correlation between "sameness" in low-level signal features and "sameness" in meaning. Establishing the systematic correlation between the expressions of a language and their meaning is precisely the goal of a *theory of meaning* (see, *e.g.*[8]), a subject of the philosophy of language that is still controversial, at least as far as the meaning of natural languages is concerned. So, pushed to its extreme consequences, the ambitious view of form-based retrieval leads to viewing a MIR system *as an algorithmic simulation of a theory of meaning*, in force of the fact that the sameness assumption is relied upon in every circumstance, not just in the few, happy cases in which everybody's intuition would bet on its truth. At present, this assumption seems more warranted in the case of text than in the case of non-textual media, as the representations employed by form-based textual retrieval methods (*i.e.* vectors of weighted words) come much closer to a semantic representation than the feature vectors employed by similarity-based image retrieval methods. Anyway, irrespectively of the tenability of the sameness assumption, the

identification of the alleged syntactic-semantic correlation is at the moment a remote possibility, so the weaker view of form-based retrieval seems the only reasonable option.

4.3 Mixed multimedia information retrieval

Suppose a user of a digital library is interested in retrieving all documents produced after January 2007, containing a critical review on a successful representation of a Mozart's opera, and with a picture showing Kiri in a blue-ish dress. This need addresses all dimensions of a document: it addresses structure because it states conditions on several parts of the desired documents; it addresses profile because it places a restriction on the production date; it addresses form- (in particular color-) and semantic-based image retrieval on a specific region of the involved image (the region must be blue and represent the singer Kiri) as well as on the whole image (must be a scene of a Mozart's opera); it addresses from-based text retrieval by requiring that the document contains a piece of text of a certain type and content. This is an example of mixed MIR, allowing the combination of different types of MIR in the context of the same query [1].

Emerging standards in multimedia document representation (notably, the ISO standard MPEG21) address all of the dimensions of a document. Consequently, their query languages support more and more mixed MIR.

5 Key applications

Nowadays, MIR finds its natural context in *digital libraries,* a novel generation of information systems [6], born in the middle of the 90's as a result of the First Digital Library Initiative. Digital Libraries are large collections of multimedia documents which are made on-line available on global infrastructures for discovery and access. MIR is a core service of any DL, addressing the discovery of multimedia documents.

6 Cross references

Information Retrieval: a. IR retrieval models, d. Text retrieval, h. Digital Libraries
IX.c Multimedia database

7 Recommended reading

[1] Carlo Meghini, Fabrizio Sebastiani, and Umberto Straccia. A model of multimedia information retrieval. *Journal of the ACM*, 48(5):909–970, 2001.

[2] Alberto Del Bimbo. *Visual Information Retrieval.* Morgan Kaufmann Publishers, Inc., 1999.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval.* Cambridge University Press, 2007.

[4] Franz Baader, Diego Calvanese, Deborah McGuiness, Daniele Nardi, and Peter Patel-Scheneider, editors. *The description logic handbook.* Cambridge University Press, 2003.

[5] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu. The Virage image search engine: An open framework for image management. In

Proceedings of SPIE-96, 4th SPIE Conference on Storage and Retrieval for Still Images and Video Databases, pages 76–87, San Jose, US, 1996.

[6] Leonardo Candela, Donatella Castelli, Pasquale Pagano, Constantino Thanos, Yannis Ioannidis, Georgia Koutrika, Seamus Ross, Hans-Jörg Schek, and Heiko Schuldt. Setting the foundations of digital libraries. The DELOS manifesto. *D-Lib Magazine*, 13(3/4), March/April 2007.

[7] F. Crestani, M. Lalmas, and C.J. van Rijsbergen, editors. *Logic and Uncertainty in Information Retrieval: Advanced models for the representation and retrieval of information*, volume 4 of *The Kluwer International Series On Information Retrieval*. Kluwer Academic Publishers, Boston, MA, October 1998.

[8] Donald Davidson. Truth and meaning. In *Inquiries into truth and interpretation*, pages 17–36. Clarendon Press, Oxford, UK, 1991.

[9] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, and W. Niblack. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994.

[10] F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modelling and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733, 1996.

[11] Euripides G. Petrakis and Christos Faloutsos. Similarity searching in medical image databases. *IEEE Transactions on Data and Knowledge Engineering*, 9(3):435–447, 1997.

[12] S. Ravela and R. Manmatha. Image retrieval by appearance. In *Proceedings of SIGIR-97, 20th ACM Conference on Research and Development in Information Retrieval*, pages 278–285, Philadelphia, US, 1997.

[13] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998.

[14] John R. Smith and Shih-Fu Chang. Transform features for texture classification and discrimination in large image databases. In *Proceedings of the 1st IEEE International Conference on Image Processing*, pages 407–411, Austin, US, 1994.

[15] P. Zezula, G. Amato, Dohnal, V., and M. Batko. *Similarity Search: The Metric Approach*. Springer-Verlag, 2006.