
Using textual and structural context for searching Multimedia Elements

Mouna Torjmen*, Karen Pinel-Sauvagnat
and Mohand Boughanem

IRIT-SIG,
University of Toulouse,
118 route de Narbonne, 31062 Toulouse Cedex 4, France
E-mail: torjmen.mouna@gmail.com
E-mail: Karen.Sauvagnat@irit.fr
E-mail: Mohand.Boughanem@irit.fr
*Corresponding author

Abstract: We investigate in this paper the use of XML structure in multimedia retrieval, particularly in context-based image retrieval. We propose two methods to represent multimedia objects: the first one is based on an implicit use of textual and structural context of multimedia objects, whereas the second one is based on an explicit use of both sources. Experimental evaluation is carried out using the *INEX MultimediaFragments* Task 2006 and 2007. We show that there is a strong vocabulary relation between the query and the multimedia object representation, and that using XML structure improves significantly the effectiveness of multimedia retrieval.

Keywords: XML documents; context-based multimedia retrieval; hierarchical structure; multimedia element; textual and structural context.

Reference to this paper should be made as follows: Torjmen, M., Pinel-Sauvagnat, K. and Boughanem, M. (2010) 'Using textual and structural context for searching Multimedia Elements', *Int. J. Business Intelligence and Data Mining*, Vol. 5, No. 4, pp.323–352.

Biographical notes: Mouna Torjmen is currently Teaching Assistant (ATER) at the Mirail University of Toulouse and at the IRIT laboratory (France). She has received her Ph.D in computer science from the University of Toulouse (France) in 2009. She is interested in semi-structured and multimedia retrieval.

Karen Pinel-Sauvagnat is Assistant Professor at the Paul Sabatier University of Toulouse and a member of the SIG-RFI team of the IRIT laboratory (France) since 2006. She is currently working on Information retrieval on semi-structured documents (relevance feedback, multimedia retrieval, use of graph theory in retrieval). She is also interested in the evaluation of Information retrieval (protocol, metrics, . . .) and aggregated search.

Mohand Boughanem is a Professor at the University of Toulouse III and the leader of the Information Retrieval and filtering (IRf) group

at IRIT/SIG team. His current research focuses on IR models, XML information retrieval, Contextual/personalised/social IR, aggregated search. He has served as programme committee member of the major IR conferences (e.g., SIGIR, ECIR, CIKM), he also served as programme committee chair of ECIR'2009, CORIA'2009 and and co-chair of WI'2009 Tutorial. In 2007, he was appointed Editor-in-Chief of the I3 journal (<http://www.revue-i3.org>). He has been invited as reviewers by several related-IR journals (IR journal, IP&M, JASIST). He is one of the founders of ARIA (the French Association of Information Retrieval) and CORIA (the annual French conference in information retrieval), both were founded in 2004. He published more than 100 papers in international conferences and journals.

1 Introduction

The joint evolution of user needs and electronic documents constantly raises new challenges in the Information Retrieval (IR) field. On the one hand, plain-text documents give way to semi-structured or structured documents (in HTML or XML format for example), and on the other hand multimedia contents (photos, music, videos, ...) are now largely included in those documents.

In the literature, there are two main approaches in Multimedia Information Retrieval (MIR): content-based approaches and context-based approaches.

Content-based MIR approaches use low level features extracted from multimedia objects themselves. These features are specific for each media type. For example, content-based Image Retrieval uses visual features as texture and colour to retrieve images. Such techniques showed their effectiveness in some specific applications (let us for instance cite the medical image retrieval field), but are still limited in other generic domains, like for example image retrieval on the web.

Context-based MIR approaches use the multimedia object context (as for example the surrounding text) to evaluate its relevance. One of the main advantages of these retrieval approaches is that they can be applied on any type of multimedia objects (e.g., image, video, audio), because the semantic representation of the multimedia object is defined using its context and not its content.

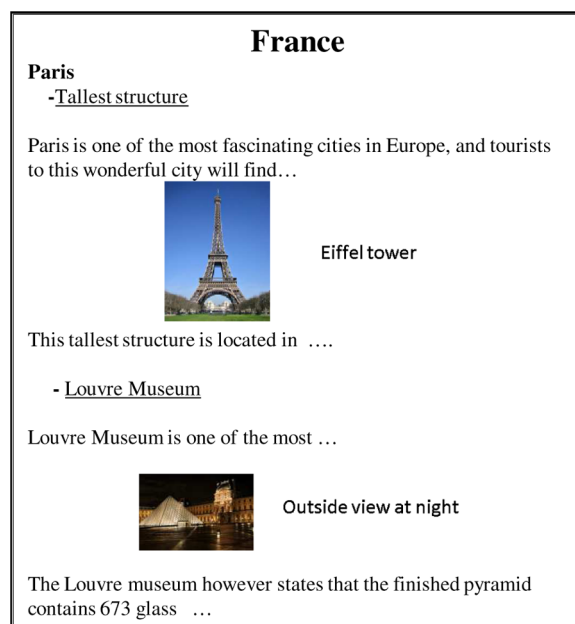
Existing research on multimedia information retrieval techniques (*Content-based MIR* or *Context-based MIR*) has already shown that both approaches are far from trivial. The main problem of *Content-based MIR* techniques is the lack of semantic representation of the multimedia content, while most of the *Context-based MIR* techniques only focus on the textual context surrounding the multimedia objects.

In this work, we are interested in *Context-based MIR* techniques, and more precisely in *Context-based MIR* techniques for *image retrieval*. Nonetheless, proposed approaches and discussions are applicable to any other media. As in context-based approaches users generally express their needs with queries composed of keywords terms, we will limit our work to those type of queries.

Image context is composed of all information surrounding the image and allowing a good description of it or giving a good interpretation of its meaning. For instance, to retrieve images of the document presented in Figure 1, many

contextual factors can be used: the document title, the images' names, the images' captions, text surrounding images, etc. In our example, caption seems to be the best information to use to retrieve the image concerning the 'Eiffel Tower' while the title and the surrounding text seem to better describe the image about 'Louvre'.

Figure 1 Example of a multimedia object context (see online version for colours)



Although most of the existing works in image retrieval uses text surrounding images (Chen et al., 2005), other sources of evidence have recently been explored. Among them one can cite hyperlinks of documents containing images (Hliaoutakis et al., 2006) or documents structure (van Zwol et al., 2005). The main challenging task in context-based image retrieval is thus to identify the most suitable context for representing multimedia objects (i.e., images) and to define a way to exploit it in the multimedia retrieval model.

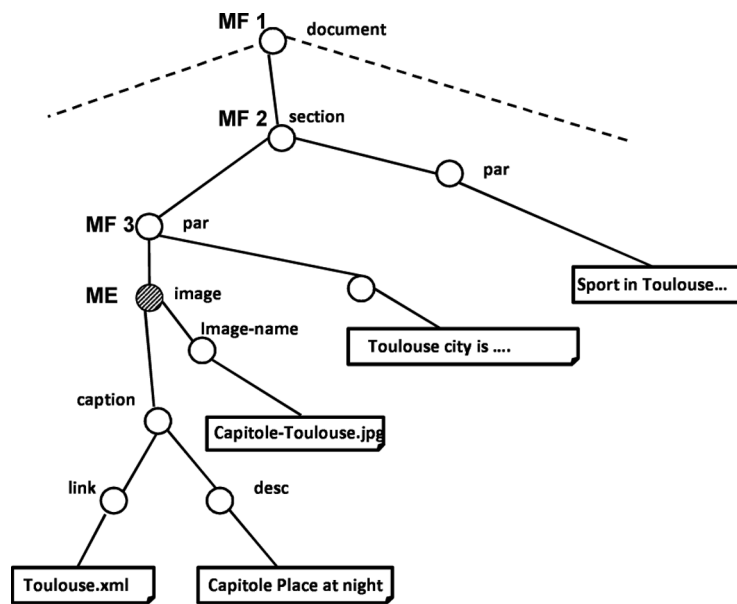
As most of multimedia documents contain structural information and as the use of structure in textual retrieval has already showed its benefit (Fuhr et al., 2004, 2005, 2006, 2007), this latter is obviously being considered as an interesting contextual factor to retrieve multimedia objects. The XML language is nowadays the most common language used to structure documents. A wide range of works use the XML syntax to annotate and describe multimedia objects (one can cite the *MPEG*, *SVG*, or *SMIL* formats). All documents following these formats share the same standard structure defined by the format specification. At the same time, XML structure is also very used to semantically and hierarchically organise document contents (text+images, videos, ...). In this case, XML documents do not share a standard structure (structure is heterogeneous) and may not have the same organisation. We focus on this latter type of structured documents in this work.

In XML-based multimedia retrieval, two types of results can be defined according to the documents structure:

- The multimedia object itself, that is an XML element¹ containing the reference entity to the multimedia object content (file name) and possibly associated information, as caption for example. We talk in this case about ‘*multimedia element*’.
- The multimedia object possibly associated with textual information. We talk here about ‘*multimedia fragment*’.

As XML documents can be considered as trees, both types of results (multimedia elements or multimedia fragments) are nodes of the document tree. Let us consider the example in Figure 2 and the query ‘*Toulouse city*’.

Figure 2 Example of a Multimedia Element/multimedia fragment



In this document, the image node is a Multimedia Element (ME). Multimedia fragments that are also related to the query are the following: MF1, MF2, MF3 and ME, which is also a multimedia fragment.

Retrieving each result type is far from trivial. The main challenge in retrieval is the evaluation of the relevance of MEs using contextual information composed of text and structure. In multimedia fragment retrieval, the challenge is to specify the most relevant multimedia fragments to be returned by the system. They should have an appropriate granularity, as they can be composed of the media itself, or a mixture of text and media elements.

In this paper, we are interested in ME retrieval, and especially in the impact of structure as a contextual factor for MEs retrieval. Our main intuition is to use the structure to select ‘relevant’ textual information, that is textual information that gives the best possible description of MEs. For this purpose, we will consider that textual information in the document tree which is close to the MEs is more likely to be relevant than information far located from it. We propose two methods for MEs retrieval:

- The first one (called *CBA*) (Torjmen et al., 2009) is based on an implicit use of textual and structural context of MEs. Relevance scores of children, brothers and ancestors nodes of MEs are evaluated with a traditional XML retrieval system, and the impact of each of these sources of evidence on the ME relevance is then studied.
- The second one (called *OntologyLike*) (Torjmen et al., 2008a) is based on an explicit use of textual and structural contexts thanks to an analogy between XML documents and ontologies. More precisely, we proposed to use a concept-based similarity measure to evaluate the participation degree of each textual node of documents in the relevance score of the MEs they contain.

Our aim in this paper is to present new and detailed experiments concerning these methods:

- concerning the *CBA* method, we study the relationship between the query type (specific or generic) and the best sources of evidence (children, brothers or ancestors nodes) to use to retrieve relevant MEs
- concerning the *OntologyLike* method, we study new structural factors in the similarity measure, and we compare the effectiveness of using textual context, structural context and the combination of both
- a comparison of both methods concludes these experiments.

The rest of the paper is organised as follows. Section 2 presents a brief overview of existing approaches in XML Multimedia Retrieval, and a reminder on similarity measures used in the semantic resources domain. Sections 3 and 4 describe our two methods for retrieving MEs. These sections aim at making the paper self-contained, allowing the reader to know our methods before introduce, in Section 5, the related experiments and the comparison of both methods. The last section provides concluding remarks and promising future directions.

2 Related works

We present in the first part of this section some related works in XML multimedia retrieval. In the second part, we present some similarity measures used in the semantic resources domain as our method ‘*OntologyLike*’ is based on an analogy between XML documents and ontologies.

2.1 Multimedia retrieval in semi-structured documents

Existing work in XML-multimedia retrieval can be classified into two categories:

- existing IR approaches (approaches for traditional XML retrieval or content-based MIR approaches) that are adapted to XML-Multimedia retrieval
- specific approaches for XML-Multimedia retrieval, that only aim at returning MEs or fragments to users.

Queries processed by state-of-the-art systems are generally textual queries (i.e., queries composed of keywords terms), but they can possibly be associated with image queries (i.e., with examples images).

2.1.1 Existing IR approaches adapted to XML-multimedia retrieval

Some XML-related approaches use a linear combination of evidences to merge the results of content-based image retrieval and text-based XML retrieval. Others filter adhoc results into multimedia ones.

For example, combining textual and image retrieval in XML multimedia retrieval is studied in Tjondronegoro et al. (2005), Mihajlovic et al. (2005) and Lau et al. (2006). Results show that the use of visual features degrades the system accuracy. Authors of Iskandar et al. (2006) proposed to use the content-based system GIFT on one hand and the textual retrieval system *Zettair* on the other hand. Results in the INEX campaign did not show the effectiveness of the method.

A method proposed by CWI/UTwente team (Tsikrika et al., 2007) consists of using a traditional retrieval method based on language models and of using different length priors. Retrieved results are limited to fragments that contain at least one image, and no further multimedia processing is used. This method shows its effectiveness when retrieved fragments are documents.

2.1.2 Specific approaches for XML-Multimedia retrieval

Authors of Kong and Lalmas (2005, 2007b) proposed a method to represent and retrieve MEs (images). It consists of dividing textual content into *Region Knowledge*² (RKs): self level RK (RK of ME); sibling level RK (RK of the sibling elements of the multimedia node); 1st ancestor level RK (RK of the first ancestor of ME excluding nodes already used); 2nd ancestor level RK; ...; *N*th ancestor level RK. Then, authors used the vector space model to evaluate each Region Knowledge. The final score of the image is evaluated by combining the different regions which participate in the image representation with different degrees. Even though this method exploits the document structure, it does not take into account the element distribution in the same Region Knowledge.

Another approach in Kong and Lalmas (2007a) uses a bayesian network incorporating element-based language models for the retrieval of a mixture of text and image (i.e., a multimedia fragment). The approach was evaluated with a small collection (*Lonely Planet* of INEX Multimedia 2005) and showed its efficiency. However this method has never been validated with a larger collection (as the *Wikipedia* collection of INEX Multimedia Fragment task 2006–2007).

To our knowledge, until 2005 when the INEX evaluation campaign introduced a new task called Multimedia Task (van Zwol et al., 2005), only a few studies were interested in multimedia retrieval in semi-structured documents. This is why most of the works presented here were proposed in this framework. The INEX Multimedia track moved to the imageCLEF WikipediaMM Task in 2008. The used collection is now composed of images annotated in XML format, and structure is only used for annotation purpose: all nodes containing useful information have the same depth in documents, and the same information can be found for all images in all documents: author, date, caption, format, etc. Some approaches dealing with this types of XML documents can be found in Torjmen et al. (2008b), Tsikrika and Vries (2009), or

Moulin et al. (2009). This track is however not of high interest for our work, since structure cannot really be used as a contextual factor to improve multimedia retrieval.

As a conclusion, state-of-the-art approaches use either a combination of adhoc XML and *Content-based MIR* retrieval, or a filtering of XML adhoc results by keeping only fragments having at least one ME. Only a few approaches offer a real study of the impact of the XML structure in Multimedia Retrieval.

2.2 Similarity measures on semantic resources

Ontology is a formal and explicit representation of knowledge which consists of concept lexicon, concept properties, and relations among concepts (Bertini, 2005) while a concept is a cognitive unit of meaning.

Many semantic measures on ontologies have been proposed in literature to evaluate the strength of the semantic link between two concepts or two groups of concepts inside an ontology. Authors in Slimani et al. (2006) classify semantic measures into three categories:

- *Edge counting methods*: These methods measure the similarity between two concepts as a function of the path length linking the concepts and on the position of the concepts in the ontology (Hirst and St-Onge, 1998; Rada et al., 1989; Wu and Palmer, 1994).
- *Information content methods*: They measure the difference in information content between two concepts as a function of their probability of occurrence in a corpus. Moreover, the similarity between two concepts is obtained by the degree of shared information (Miller and Charles, 1991; Resnik, 1999; Jiang and Conrath, 1997).
- *Hybrid methods*: They combine the above methods (Rodriguez and Egenhofer, 2003; Leacock and Chodorow, 1998; Li et al., 2003).

In what follows, we present some measures related to the first type of similarity measures. Indeed, in our work, we are interested in this type of measures as textual content of a ME is generally small (a few terms) or can even be absent.

Rada measure (Rada et al., 1989): It is the first measure proposed to compute the semantic similarity between concepts through ‘*is-a*’ relations. This measure is based on ‘*specific/generic*’ hierarchical relations. It computes the distance $Dist(C_1, C_2)$ between the two concepts C_1 and C_2 , which is the shortest path length between the two concepts in terms of edges count. The similarity between two concepts is then defined as follows:

$$Sim_{Rada}(C_1, C_2) = \frac{1}{Dist(C_1, C_2)}. \quad (1)$$

HiOn measure (Hirst and St-Onge, 1998): In this measure, the similarity between two concepts depends on the number of edges between them ($Dist(C_1, C_2)$) and on how many times direction changes in the path between the concepts. The measure is defined by the following formula:

$$Sim_{HiOn}(C_1, C_2) = C - Dist(C_1, C_2) - k \times NbDir \quad (2)$$

where C and k are two constant parameters and $NbDir$ is the number of changed directions between C_1 and C_2 .

WP measure (Wu and Palmer, 1994): Also called the *Wu-Palmer* measure, this measure takes into account the concepts position with the ontology root. It is defined by the following expression:

$$Sim_{WP}(C_1, C_2) = \frac{2 \times N}{N_1 + N_2 + 2 \times N} \quad (3)$$

where N_1 and N_2 are the distances which separate concepts C_1 and C_2 from their most specific common ancestor CS , and N is the distance which separates CS from the root R . The WP measure has the advantage of being simple to implement and achieves good performances compared to the other similarity measures based on edge counting (Lin, 1998).

3 CBA method

In this section, we presented our CBA method (Torjmen et al., 2009), which implicitly exploits the XML content and structure for ME retrieval. This method consists in:

- 1 evaluating a relevance score for each XML node of documents using a traditional XML retrieval system
- 2 evaluating a relevance score for MEs using their closest nodes having a positive relevance score in the document tree according to Step 1.

In this method, the XML structure is mainly used in Step 1, where we use a traditional XML retrieval system (like for example the state-of-the-art systems *GPX* (Geva, 2005) or *XFIRM* (Sauvagnat, 2005) to evaluate a relevance score to the query for each non-textual node (also called inner nodes). The relevance score of MEs is then evaluated using surrounding inner nodes having a positive score according to the XML retrieval system. XML structure is thus exploited implicitly.

In our method, three sources of evidences are used to represent ME:

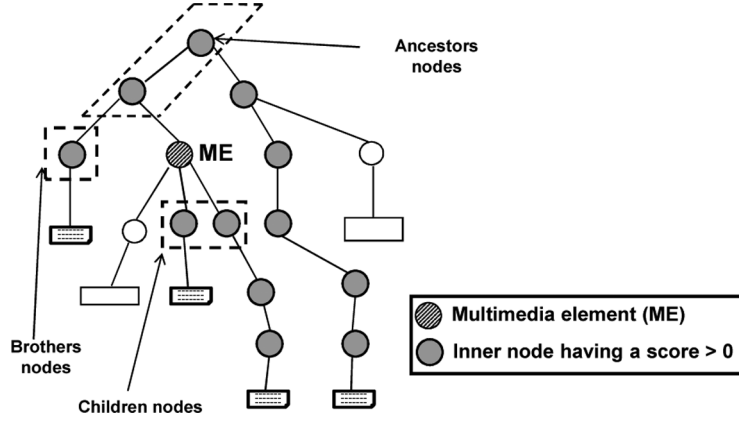
- child nodes because they may contain specific information
- brother nodes because they have more chance than others to share the same topic than ME
- ancestor nodes in order to take into account the document context.

Figure 3 illustrates the definition of the three sources of evidence.

For a ME, a relevance score ($S^{E_k}(\text{ME})$) is evaluated for each source of evidence E_k with $k = \{\text{Children, Brothers, Ancestors}\}$ using the following formula:

$$S^{E_k}(\text{ME}) = \left(\sum_{i=1}^{|IN_k|} Score_{trad-XML}(IN_{i,k}) \right) \times \frac{|Rel - IN_k| + 1}{|IN_k| + 1} \quad (4)$$

Figure 3 Definition of the three sources of evidence: children nodes, brothers nodes and ancestors nodes



where

E_k is one of the three sources of evidences, i.e., children node, brothers nodes or ancestors nodes

$IN_{i,k}$ is an inner node (i.e., a non textual node) with $IN_{i,k} \in E_k$

$Score_{trad-XML}(IN_{i,k})$ is the score of an inner node $IN_{i,k} \in E_k$, evaluated with a traditional XML retrieval system

$|Rel - IN_k|$ is the number of inner nodes $\in E_k$ having a score > 0

$|IN_k|$ is the total number of inner nodes $\in E_k$.

The final score of a ME is evaluated as follows:

$$S(\text{ME}) = p_1 \times S^{E_{Children}} + p_2 \times S^{E_{Brothers}} + p_3 \times S^{E_{Ancestors}} \quad (5)$$

where p_1, p_2, p_3 are three parameters to emphasise respectively the score of children, brothers and ancestors of ME, with $p_1 + p_2 + p_3 = 1$.

The representation of MEs using one, two and three sources of evidence is evaluated in Section 5.3. As the CBA method uses in a first step a traditional XML retrieval system to evaluate inner nodes relevance values, its performance thus strongly depends on the traditional XML retrieval system performance. In our experiments, we used the XFIRM system presented in Section 5.2.

4 Ontologylike method

We present in this section our OntologyLike method (Torjmen et al., 2008a), where MEs can be represented using only textual context, only structural context or using both contexts. This way, the XML structure is explicitly exploited in multimedia retrieval.

4.1 Multimedia Element representation by textual context

We mean by textual context all textual nodes of the document containing ME to represent. The XML structure is ignored here.

For each textual node (also called leaf node), we choose to use a $tf \times idf \times ief$ weighting formula to evaluate its relevance score, since this formula has been shown to be effective in Pinel-Sauvagnat and Boughanem (2004). However, any other weighting model can be used to estimate scores of textual nodes. The relevance score of a textual node TN_i is thus evaluated as follows:

$$S^{tf \times idf \times ief}(TN_i) = \sum_{j=1}^{|Q|} tf_j \times idf_j \times ief_j \tag{6}$$

where

j is a query term and $|Q|$ is the number of query terms

tf_j is the frequency of j in the textual node

$idf_j = \log(|D|/(|d_j| + 1)) + 1$, with $|D|$ the total number of documents in the collection

$|d_j|$ the number of documents containing j , and ief_j is the inverse element frequency of term j , i.e., $\log(|TN|/|TN_j| + 1) + 1$, where

$|TN_j|$ is the number of textual nodes containing j and $|TN|$ is the total number of textual nodes in the collection.

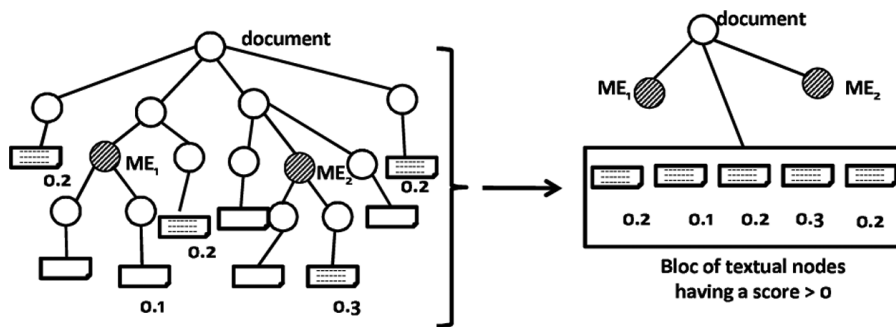
The score of each ME is computed as the sum of the relevance scores of its associated textual nodes, as described in the following formula:

$$S_{Text.Cont.}(ME) = \sum_{i=1}^{|TN|} S^{tf \times idf \times ief}(TN_i) \tag{7}$$

where $TN_i \in TN$ is a textual node in the document containing ME.

Figure 4 illustrates the textual context definition.

Figure 4 Multimedia elements representation through textual context



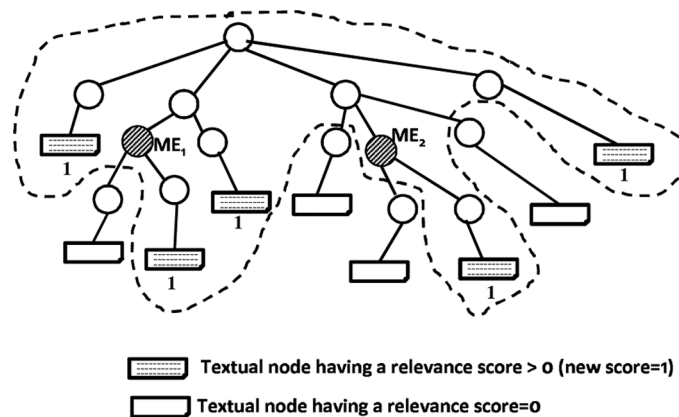
We notice here that when using only textual context, a document is considered as a composition of multimedia objects and a bloc of textual information. Using the textual context will assign the same score for all MEs in the same document (e.g., ME_1 and ME_2 in Figure 4 are assigned the same relevance score).

4.2 Multimedia Element representation by structural context

As there is no direct relation between the query composed of keywords terms and the XML structure, we propose to only consider in a first step if each textual node has a positive relevance score using our weighting formula or not, i.e., all relevance scores will be either 0 (if the relevance score according to equation (6) equals 0) or 1 (if the relevance score according to equation (6) is > 0). A textual node will thus have a relevance score = 1 if it contains at least one query term.

Figure 5 shows an example of structural context definition to represent MEs. All textual nodes having a relevance score > 0 according to equation (6) have a new score equal to 1, while others still have a relevance score = 0.

Figure 5 Multimedia elements representation through structural context



The challenge is now how to define the degree of participation of each textual element to evaluate the relevance score of ME. This definition must take into account the structure and our following intuition: textual descendants of ME must participate more to the relevance score of ME than textual descendants of brothers of ME, and the latter nodes must participate more than textual descendants of the ancestors of ME. Indeed, we think that textual descendants are the most specific nodes to represent MEs, that textual descendants of brothers nodes have a high probability of sharing the same information than MEs and that descendants of the root node should less participate since they are far from ME in the document tree.

As explained before, an XML document can be represented as a hierarchical tree, composed of a root (document), simple nodes (elements and/or attributes) and leaf nodes (textual nodes). An inner node is any node of the tree that has children nodes (i.e. a non-leaf node). Thanks to this representation, an XML document can be considered as a simple ontology : nodes are considered as concepts linked with the 'IsPartOf' relationship. For example, 'Section IsPartOf Article', 'Paragraph IsPartOf Section'. The main idea of our approach is to use a semantic similarity measure between ontological concepts to evaluate how much each inner node in the document should participate in the representation of ME.

We consider the multimedia node as a concept C_1 , and each other node of the document as another concept C_2 . Table 1 shows the analogy between an XML document and a simple ontology.

Table 1 Analogy between an XML document and a simple ontology

<i>Simple ontology</i>	<i>XML document tree representation</i>
Ontology concepts	XML document nodes
' <i>IsPartOf</i> ' relationship between concepts	Structural relationship between nodes
Similarity measure between two concepts C_1 and C_2	Measure of participation of a textual node TN in the multimedia element node ME representation
$\text{Sim}(C_1, C_2)$	$\text{Rep}(\text{ME}, \text{TN})$

Based on this analogy between the XML document tree and a simple ontology, we propose to use concept similarity measures to calculate a participation degree of each textual node in the relevance score of ME.

In Zargayouna (2004), the WP measure is used in a semantic indexing scheme for XML documents. Nevertheless, with this measure, brothers can be ranked before descendants of a concept whereas authors of Zargayouna (2004) want to select all descendants of a concept before its brothers.

They thus proposed to penalise brothers scores by adding a spec (C_1, C_2) function to the WP measure, which evaluates the specificity of the two concepts (C_1 and C_2) comparing to the lowest concept (bottoms³). The spec function is defined as follows:

$$\text{spec}(C_1, C_2) = \text{depth}(CS) \times N_1 \times N_2 \quad (8)$$

where N_1 and N_2 are the distances which separate concepts C_1 and C_2 from their most specific common ancestor CS and $\text{depth}(CS)$ is the maximum number of edges between CS and bottom (Figure 6).

As seen in Figure 6, the depth factor reflects the vertical hierarchical structure to differentiate the participation degree of the textual nodes of ancestors of each ME.

Let us consider the example of Figure 6, the children node F of ME participates more in ME representation than node N , a child of node B ($\text{Depth}(\text{ME}) < \text{Depth}(\text{B})$). This factor appears thus valid to reflect our intuition concerning the participation degree of each textual node according to its hierarchical position in the document tree. Based on the three factors of the spec function, several experiments were done and are presented in Section 5.4.2. They allow us to define the best measure in our case which is defined as follows:

$$\text{Rep}_{\text{spec}}(\text{ME}, \text{TN}_i) = \frac{1}{(N_1 + 1) \times \text{Depth}(CS) \times N_2} \quad (9)$$

where TN_i is a textual node having a non-zero score according to equation (6) and belonging to the same document than ME. N_1 and N_2 are respectively the distance between the multimedia element ME and textual element TN_i and their most specific common ancestor.

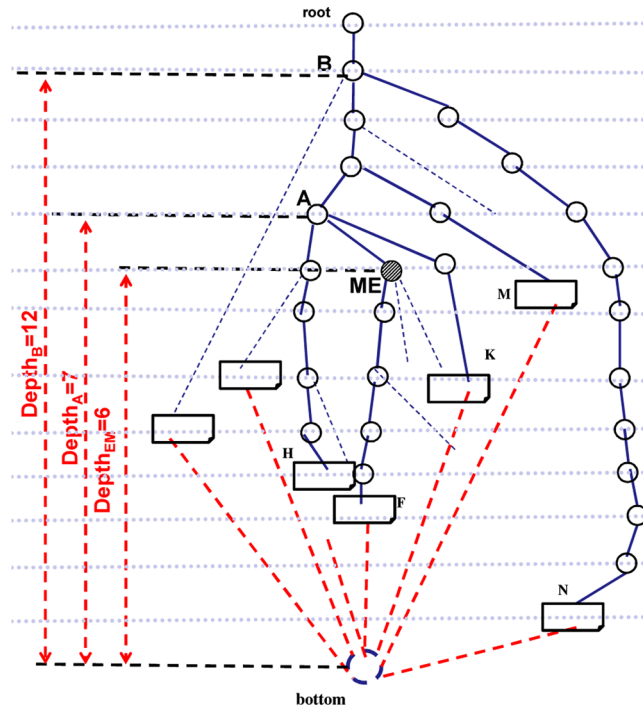
We add 1 to the N_1 factor because it can be equal to zero: when the used textual node is a ME child, the common specific ancestor is the ME itself. Each of these factors is evaluated separately in Section 5.4.2. The final score for each ME

is evaluated in function of $\text{Rep}_{\text{spec}}(\text{ME}, \text{TN}_i)$ of all textual nodes of the document. We use the following formula:

$$S_{\text{Struct. Cont}}(\text{ME}) = \sum_{i=1}^{|\text{TN}|} \text{Rep}_{\text{spec}}(\text{ME}, \text{TN}_i) \quad (10)$$

where $|\text{TN}|$ is the number of textual nodes in the document containing the multimedia element ME.

Figure 6 Definition of the depth factor (see online version for colours)



4.3 Multimedia Element representation by textual and structural context

We propose in this section to combine the textual and the structural context to semantically represent ME. The participation degree of each textual-node to evaluate ME relevance is thus defined by the following Rep_{Ont} formula.

$$\text{Rep}_{\text{Ont}}(\text{ME}, \text{TN}_i) = \frac{S^{tf \times idf \times ief}(\text{TN}_i)}{(N_1 + 1) \times \text{Depth}(\text{CS}) \times N_2} \quad (11)$$

where ME is the multimedia element, TN_i is a textual node participating in its representation and the $S^{tf \times idf \times ief}(\text{TN}_i)$ is the score of the textual node TN_i , evaluated with equation (6).

In our experimental evaluation (Section 5.4.3), by comparing the *RepOnt* measure with the *Rada* measure, we observed that results are not always statistically different. Indeed, the *Rada* measure uses another structural factor, which is the distance between the two nodes (image node and textual node) and that favours textual nodes of ME compared to other textual nodes in the representation.

Based on this observation, we evaluated a similar structural factor which is the number of changed directions between the multimedia node and the textual node (Hirst and St-Onge, 1998). This factor favours textual nodes of ME because the number of changed directions between a ME and a textual child node is equal to 1, while it is equal to 2 between ME and any other non-children textual element.

Experiments related to this factor are done in Section 5.4.4 and show its effectiveness. Our formula thus becomes:

$$\text{Rep}_{\text{OntNbDir}}(\text{ME}, \text{TN}_i) = \frac{S^{tf \times idf \times ief}(\text{TN}_i)}{(N_1 + 1) \times \text{Depth}(\text{CS}) \times N_2 \times \text{NbDir}} \quad (12)$$

where NbDir is the number of changed directions between the textual node and ME. The final score of each ME is:

$$S(\text{ME}) = \sum_{i=1}^{|\text{TN}|} \text{Rep}_{\text{OntNbDir}}(\text{ME}, \text{TN}_i) \quad (13)$$

where TN_i is a textual node of the document and $|\text{TN}|$ is the number of all textual nodes in the document that contains the multimedia element ME.

5 Evaluation

We evaluated our two methods using the Multimedia task of the INEX evaluation campaign (Westerveld and van Zwol, 2006; Tsikrika and Westerveld, 2008).

More precisely, concerning the *CBA* method, our aim is to identify the best source of evidence to represent MEs. We thus study the use of one, two or three sources of evidence and discuss the benefit of each.

Concerning the *OntologyLike* method, we evaluate the use of the textual context only, the structural context only and the combination of both. We then study the impact of our structural factors separately and together. Our measure is also compared to other semantic concept-based measures as *Rada* and *Wu-Palmer*.

We finally conclude this section with a comparison and discussion of both approaches.

5.1 Evaluation protocol

5.1.1 Collection

Our experiments are based on the INEX Multimedia Fragment task. The aim of this task is to find relevant XML fragments (that must contain at least one image) given a multimedia information need. More details about can be found in Westerveld and van Zwol (2006), and Tsikrika and Westerveld (2008). The core collection of the task is the English version of the Wikipedia XML collection which is composed

of XML documents that can contain or not images. This collection contains about 660,000 documents (4.6 Giga-Bytes without images), 30 millions elements, and more than 300,000 images. On average, an article contains 161.35 XML nodes, where the average depth of an element is 6.72. Details of this collection are given in Denoyer and Gallinari (2006).

The 2006 and 2007 Multimedia Fragment query sets are respectively composed of 9 and 19 topics. We only use the title field (keywords terms) to process queries.

In the official campaign, assessments of both queries sets 2006 and 2007 are done on multimedia fragments and not on MEs. In order to properly evaluate our methods, we thus constructed a new base of assessments composed of relevant MEs (i.e., image elements) extracted from the original assessments provided by organisers.

5.1.2 Metrics

Effectiveness of our approaches is evaluated with the Mean Average Precision (MAP) which is traditionally used in the evaluation of information retrieval approaches. MAP is the mean of the Average Precision scores for a group of queries and is defined as follows:

$$\text{MAP} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \text{NonInterpolatedAveragePrecision}(j) \quad (14)$$

where $|Q|$ is the number of queries and $\text{NonInterpolatedAveragePrecision}$ is defined for each query as $\frac{1}{|R|} \sum_{k=1}^{|R|} \text{Precision}_k$, with $|R|$ the number of relevant MEs and Precision_k is the precision at recall level $k/|R|$.

To statistically validate our results, we used the signed-rank test of Wilcoxon test (Wilcoxon, 1945) which is the non-parametric equivalent of the paired samples t -test. This test consists in evaluating a value of significance $p \in [0, 1]$ which estimates the probability that the difference between two methods is due to chance. We can thereby conclude that two methods are statistically different when $p < \alpha$, where $\alpha < 0.05$ is commonly used (Hull, 1993). More precisely, the more $p \rightarrow 0$, the more two methods are supposed to be different.

In our experiments, we consider that the difference between two methods is significant when $p < 0.1$, and it is very significant when $p < 0.05$.

5.2 XFIRM system

As explained in Section 3, our CBA method needs a traditional XML retrieval system to evaluate the relevance value of nodes surrounding MEs. In our experiments, we choose to use the XFIRM system (Sauvagnat et al., 2006), which is based on a relevance propagation method.

During query processing, relevance values are assigned to leaf nodes (i.e., textual nodes), and relevance scores of inner nodes are then computed dynamically thanks to the propagation and aggregation of leaf nodes scores.

5.2.1 Queries processing

Let $q = t_1, \dots, t_n$ be a query composed of keywords terms. Relevance values of textual (i.e., leaf) nodes are computed thanks to a similarity function $RSV(q, TN)$.

$$\text{RSV}(q, \text{TN}) = \sum_{i=1}^n w_i^q \times w_i^{\text{TN}}, \quad (15)$$

with $w_i^q = tf_i^q$ and $w_i^{\text{TN}} = tf_i^{\text{TN}} \times idf_i \times ief_i$

w_i^q and w_i^{TN} are the weights of term i in query q and leaf node TN respectively

tf_i^q and tf_i^{TN} are the frequency of i in q and TN respectively

$idf_i = \log(|D|/(|d_i| + 1)) + 1$, with $|D|$ the total number of documents in the collection, and $|d_i|$ the number of documents containing i

ief_i is the inverse element frequency of term i , i.e., $\log(|\text{TN}|/|\text{TN}_i| + 1) + 1$, where $|\text{TN}_i|$ is the number of leaf nodes containing i and $|\text{TN}|$ is the total number of leaf nodes in the collection.

Each inner node in the document tree is then assigned a relevance score which is function of the relevance scores of the leaf nodes it contains and of the relevance value of the whole document.

$$r_n = \rho \times |L_n^r| \cdot \sum_{\text{TN}_k \in L_n} \alpha^{\text{dist}(n, \text{TN}_k)-1} \times \text{RSV}(q, \text{TN}_k) + (1 - \rho) \times r_{\text{root}} \quad (16)$$

$\text{dist}(n, \text{TN}_k)$ is the distance between node n and leaf node TN_k in the document tree, and $\alpha \in (0) \dots [1]$ allows to adapt the importance of the dist parameter. $|L_n^r|$ is the number of leaf nodes being descendant of n and having a non-zero relevance value (according to equation (15)). $\rho \in]0..1]$, inspired from work presented in Mass and Mandelbrod (2005), allows the introduction of document relevance in inner nodes relevance evaluation, and r_{root} is the relevance score of the root element, i.e., the relevance score of the whole document.

5.2.2 Multimedia retrieval settings

In INEX 2006 adhoc task, the best values of α and ρ of the XFIRM system are respectively 0.1 and 0.9 according to the *MAeP* measure and the thorough strategy.⁴ $\alpha = 0.1$ implies that the distance between textual nodes and their ancestors is strongly taken into account. Consequently, element specificity is an important factor to determine XML elements evaluation. On the other hand, $\rho = 0.9$ implies that the root element score slightly contributes in the inner score evaluation.

In INEX 2007 adhoc task, best values of α and ρ of the XFIRM system are respectively 0.3 and 1 according to *MAiP* measure and the focused strategy.⁵ $\alpha = 0.3$ has a similar impact than for the 2006 campaign. $\rho = 1$ means that the root score does not contribute in the inner nodes evaluation.

5.3 Evaluation of the implicit use of XML structure in Multimedia Retrieval (CBA method)

In this section, we present the experiments we performed with the implicit use of text and structure in multimedia retrieval. The aim of these experiments is to answer the following questions: which is the best source of evidence to use in order to represent ME? Should we combine all sources?

5.3.1 Multimedia elements representation using one source of evidence

Table 2 presents the results when ME is represented by only one source of evidence, i.e., either child nodes (*C*), brother nodes (*B*) or ancestor nodes (*A*).

Table 2 Results of representing multimedia elements by one source of evidence (*E*)

<i>E</i>	p_1	p_2	p_3	<i>MAP</i>	
				2006	2007
<i>C</i>	1	0	0	0.3969	0.1654
<i>B</i>	0	1	0	0.2649	0.1950
<i>A</i>	0	0	1	0.2872	0.2682

We notice here that for the INEX 2006 test set, the best representation of images is obtained by using only child nodes, while for INEX 2007 test set, the best representation is obtained through ancestor nodes.

This contradiction between the two test sets can be explained by the specific/generic notion of the vocabulary used in queries and images representation. More precisely, the difference between the query vocabulary and the image representation vocabulary can have a great impact on results. Indeed, if the query vocabulary is specific, child nodes are sufficient to represent images as they are the most specific for the image. In a similar way, if the query vocabulary is generic, the specific information of the image (i.e., child nodes) is not sufficient to represent the image elements while generic information allow to improve performance. In this case, the document context (i.e., ancestor nodes) is the most appropriate source to represent images and to answer to the query.

Figure 7 presents an example of the vocabulary relation between the query and the image representation (specific or generic).

According to this example, if the query is expressed with a generic vocabulary (as ‘*animal*’), the image representation through generic information (ancestor nodes) allow to retrieve the image ‘*dog.jpg*’, whereas the image representation through specific information is not sufficient to return this image to the user. On the other hand, if the query vocabulary is specific (as ‘*herding dog*’), the image representation through specific information (i.e., children nodes) allows to better judge its relevance comparing to generic information (i.e., ancestor nodes). In this case, representing the image with ancestor nodes will lead to the following cases: the image will be returned with a low relevance score or will not be retrieved.

To better understand this problem, we made a query by query results analysis. Let us consider for example, query 22 of the INEX 2006 test set which is a specific query ‘*London bridge*’, and query 528 of the INEX 2007 test set which is a generic query ‘*skyscraper building tall towers*’. Comparing the two queries, relevant images to query 22 all represent the same monument, whereas relevant images to query 528 can represent many monuments like the ‘*Eiffel Tower*’ and ‘*London Tower*’. The user need is thus not specific.

Table 3 shows the results of the two queries when images are represented only by children nodes, only by brothers nodes and finally only by ancestors nodes.

It can be easily seen from Table 3 that the best representation for query 22 is obtained when only children nodes are used (an improvement of 47% can

be observed comparatively to the image representation through ancestor nodes), whereas the best representation of the query 528 is obtained when only ancestor nodes are used (an improvement of 236% is observed comparatively to the image representation through child nodes).

Figure 7 Vocabulary relation between the query and sources of evidence representing the multimedia element

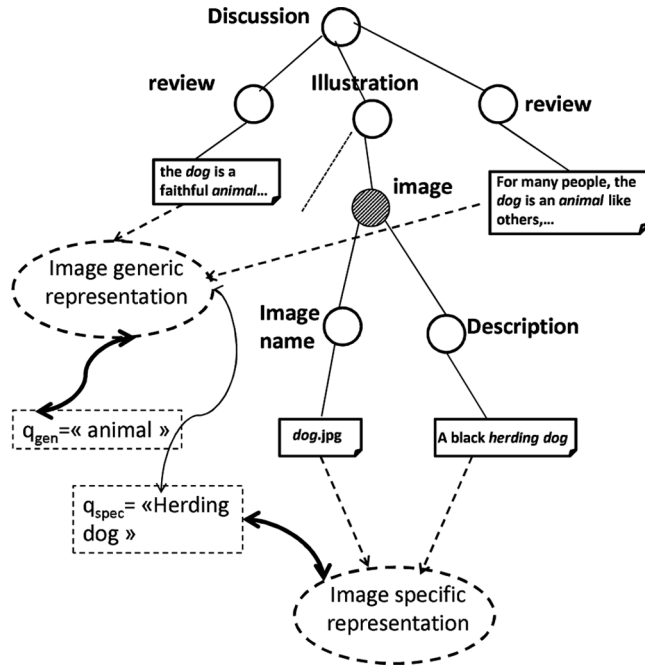


Table 3 Results comparison of a specific and a generic query

<i>E</i>	p_1	p_2	p_3	<i>MAP</i>	
				<i>q22-2006 (spec)</i>	<i>q528-2007 (gen)</i>
<i>C</i>	1	0	0	0.1668	0.0670
<i>B</i>	0	1	0	0.0951	0.1742
<i>A</i>	0	0	1	0.1137	0.2251

Table 4 presents the percentages (determined manually) of generic and specific queries in the whole test sets.

As there is more specific queries than generic ones in 2006, children nodes will be more useful than ancestors nodes for this test set. The inverse situation can be observed for 2007: there is more generic queries than specific ones, and ancestors nodes will probably be more useful than children nodes in this case.

We then calculated separately the MAPs of all specific and generic queries, using only children nodes or only ancestor nodes. Results of Table 5 show confirms our hypothesis that it is better to use specific information (children nodes of MEs) for

specific queries and to use general information (ancestors nodes of MEs) for generic queries.

Table 4 Percentage of specific and generic queries in 2006 and 2007 test sets

	<i>% of specific queries</i>	<i>% of generic queries</i>
2006	56	44
2007	22	78

Table 5 MAPs of specific and generic queries of INEX 2006 and 2007

	<i>Specific queries</i>	<i>Generic queries</i>
2006		
<i>C</i>	0.6055	0.1227
<i>A</i>	0.3969	0.1500
2007		
<i>C</i>	0.3140	0.1258
<i>A</i>	0.2508	0.2728

Furthermore, based on this query classification (specific/generic), we calculated the global MAP using children nodes for specific queries and ancestor nodes for generic queries. We obtained for INEX 2006 a MAP equals to 0.4030 (instead of 0.3965) and for INEX 2007 a MAP equals to 0.2815 (instead of 0.2682). These results show that if we can correctly classify queries into specific and generic, this classification can be used to select the appropriate source of evidence and improve results.

5.3.2 Multimedia Elements representation using two sources of evidence

In this section, we present results when combining two sources of evidence to represent image elements. Table 6 shows our best results.

Table 6 Results using two sources of evidence (*E*) in INEX 2006 and 2007

<i>E</i>	p_1	p_2	p_3	MAP-2006
<i>C + B</i>	0.9	0.1	0	0.4166
<i>C + A</i>	0.9	0	0.1	0.4249
<i>B + A</i>	0	0.6	0.4	0.3182
<i>E</i>	p_1	p_2	p_3	MAP-2007
<i>C + B</i>	0.3	0.7	0	0.2277
<i>C + A</i>	0.2	0	0.8	0.2871
<i>B + A</i>	0	0.3	0.7	0.2843

Looking at these results for 2006, we notice that combining the children nodes with another source of evidence gives the best performances, especially when $p_1 = 0.9$,

i.e., when the score of the children nodes is strongly used. This observation was expected as most of the 2006 queries are specific.

Concerning the INEX 2007 test set, best results are obtained when the score of ancestors is strongly used, which is also expected as most of the queries in this set are generic.

According to Tables 2 and 6, we notice that representing MEs by two sources of evidence (child and brother nodes for specific queries vs. ancestor and brother nodes for generic queries) slightly improves the results (+12 % for the 2006 test set, and +7% for the 2007 test set).

5.3.3 Multimedia Elements representation using three sources of evidence

Runs were carried out with various possible combination of evidences within both 2006 and 2007 test sets. Table 7 summarises the most significant results obtained when the image representation is constructed through the combination of the three sources of evidences.

Table 7 Results of representing a Multimedia Element with three sources of evidence (E)

E	p_1	p_2	p_3	MAP	
				2006	2007
$C + B + A$	0.33	0.33	0.33	0.3969	0.2778
$C + B + A$	0.8	0.1	0.1	0.4257	0.2423
$C + B + A$	0.7	0.2	0.1	0.4162	0.2414
$C + B + A$	0.7	0.1	0.2	0.4224	0.2492
$C + B + A$	0.1	0.8	0.1	0.3717	0.2494
$C + B + A$	0.1	0.7	0.2	0.3717	0.2603
$C + B + A$	0.2	0.7	0.1	0.3886	0.2507
$C + B + A$	0.1	0.1	0.8	0.3288	0.2963
$C + B + A$	0.1	0.2	0.7	0.3344	0.3022
$C + B + A$	0.2	0.1	0.7	0.3538	0.2974
$C + B + A$	0.1	0.3	0.6	0.3581	0.3008
$C + B + A$	0.3	0.1	0.6	0.3629	0.2902

Comparing the contribution of each source of evidence with the others, we notice that best results are obtained when $p_1 = 0.8$, $p_2 = 0.1$ and $p_3 = 0.1$ for the 2006 test set, and $p_1 = 0.1$, $p_2 = 0.2$ and $p_3 = 0.7$ for the 2007 test set. These results are expected as we demonstrated above that there is a strong vocabulary relation between the query and the sources of evidence used to represent ME.

Another observation which seems very important is that according to Table 1, we conclude that using the three sources of evidence with the appropriate parameters improves results comparing to using one source of evidence to represent MEs. In fact, an improvement of 7% and 13% is respectively observed for the 2006 and 2007 test sets. The same conclusion can be drawn with Table 6 and the use of two sources of evidence: an improvement of 0.1% and 5% is respectively observed on the 2006 and 2007 test sets.

5.3.4 Conclusion

Evaluation of our CBA method showed that there is a strong relationship between the query vocabulary (specific/generic) and the best sources of evidence we should use to represent MEs. On the one hand, when the query is specific, children nodes allows a better representation of MEs. On the other hand, when the query is generic, ancestors nodes seem to be more useful. Finally, we also showed that using three sources of evidence is significantly better than using one or two sources to evaluate MEs relevance.

Even if this method allowed us to draw some conclusions, it was not possible to study separately the impact of the structural and textual context. This is done in the following section with our second method, the *OntologyLike* method.

5.4 Evaluation of the explicit use of XML structure in multimedia retrieval (*OntologyLike* method)

The aim of the experiments in this section is to show the effectiveness of XML structure in MEs retrieval. For this purpose, we evaluated separately the use of textual context only, structural context only, as well as the combination of the two.

5.4.1 Evaluation of images representation by textual context

We evaluate here the image representation using only textual context (see Section 4.1). The XML structure is not taken account. For INEX 2006 and INEX 2007 test set, we respectively obtain the following MAP values: 0.3119 and 0.2145.

5.4.2 Evaluation of images representation by structural context

Experiments carried out in this section are based on a binary evaluation of textual nodes. i.e., score of textual nodes is either 0 or 1. Here, the XML structure will determine the image relevance score and will differentiate between images.

Structural factors can be classified into two types:

- Some factors are used to differentiate between hierarchical levels of MEs (as N_1 and $Depth^6$). These factors allow us to consider the textual nodes of ME or of an ancestor as a textual bloc, i.e., all textual nodes having the same most specific common ancestor with the image node have the same score.
- The second type of structural factors aims at differentiating between textual nodes of ME or an ancestor (N_2)⁷ having the same hierarchical position, i.e., we are interested here to define the participation degree of each textual node belonging to a common ancestor.

We believe that the first type of structural factors is more important to define the representation measure than the second one as it represents the vertical hierarchical structure. We thus firstly evaluate the impact of $Depth$ and N_1 separately, and then combined. Results are presented in Figure 8.

As we can see, according to the MAP measure on both test sets, and comparing $1/Depth$, $1/(N_1 + 1)$ and $1/(Depth \times (N_1 + 1))$, best results are obtained when the two factors $Depth$ and N_1 are used. To differentiate between textual elements related

to the same image ancestor, we added the N_2 factor in our experiments. Table 8 presents results without and with the use of this factor.

Figure 8 Results of image representation through structural context: use of the N_1 and Depth factor (see online version for colours)

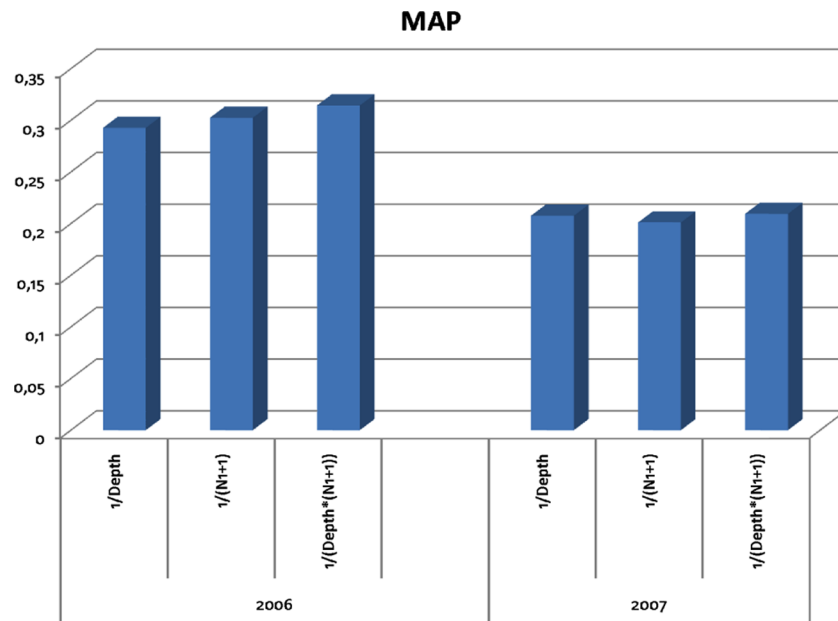


Table 8 Results of image representation through structural context: impact of the N_2 factor

	$1/(depth \times (N_1+1))$	$1/(depth \times (N_1 + 1) \times N_2)$	Gain
2006	0.3144	0.3072	-2%
2007	0.2095	0.2196	+7% *

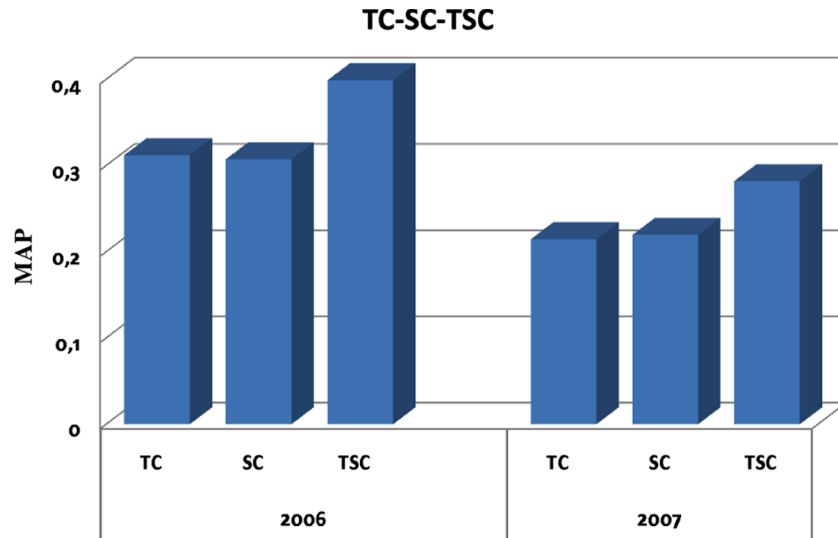
Symbol * after the gain indicates statistical significance using the Wilcoxon test at $p < 0.1$. Even if results are contradictory between the 2006 and 2007 test sets, the N_2 factor seems to be useful as the improvement obtained on INEX 2007 test set is statistically significant ($p \leq 0.1$). Thus, even if the N_2 factor did not demonstrate its effectiveness on the INEX 2006 test set, it will be kept in our formulas, because we believe it can have a positive impact in other collections.

5.4.3 Evaluation of images representation by combining textual and structural context

In this section, we discuss the images representation by combining the textual and structural contexts.

First, we compare between the use of Textual Context (TC), Structural Context (SC) and the combination of both (TSC) with the measure RepOnt (equation (11)). Figure 9 shows results for both test sets INEX 2006 and INEX 2007.

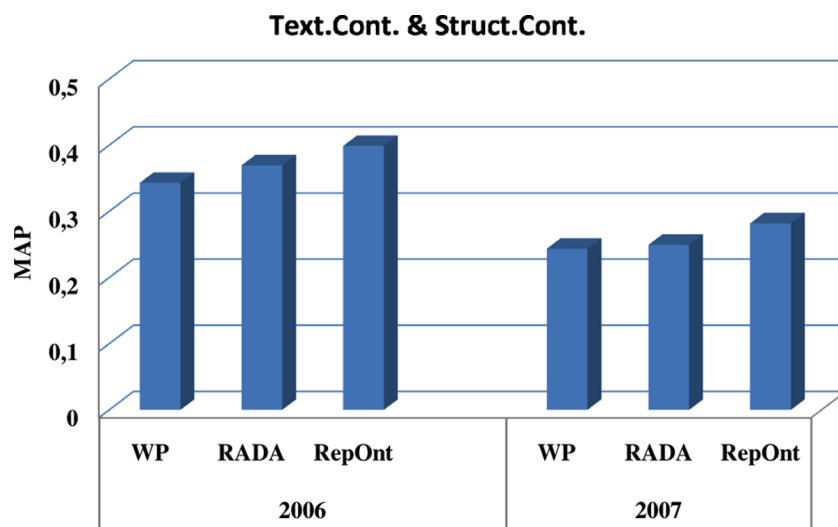
Figure 9 Comparison of the use of textual, structural contexts and combination of the two (see online version for colours)



As we can see, textual and structural contexts give almost the same results. However, the combination of both contexts improves significantly the results. We can thus conclude that the XML structure improves the multimedia retrieval effectiveness.

Results presented in Figure 10 compares between the use of *WP* measure (equation (3)), *Rada* measure (equation (1)) and our proposed measure (*RepOnt*). In this evaluation, both textual and structural contexts are used.

Figure 10 Comparison between WP, Rada and *RepOnt* measures (see online version for colours)



Best results are obtained by our measure. Gains in MAP measure are presented in Table 9.

Table 9 Gain between the RepOnt metric and the WP and Rada measures

	<i>RepOnt/WP</i>	<i>RepOnt/Rada</i>
2006	+16% **	+8%
2007	+16% **	+13% **

Symbol ** after the gain indicates statistical significance using the Wilcoxon test at $p < 0.05$.

For all the comparisons, we observe more than 5% of improvement. Moreover, for both data sets, RepOnt is statistically more effective than the WP measure. By comparing our proposed measure (RepOnt) and the Rada measure, we observed that the difference of results between the two measures is very significant in the INEX 2007 test set, whereas it is not the case for the INEX 2006 set. This observation could be explained by the fact that, as explained above (Section 5.1.3), the vocabulary relation between ME representation and the query has an impact on results. More precisely, the Rada measure depends on a structural factor, which is the distance between the two nodes (image one and textual one), that favours queries having a specific vocabulary. Using this factor, we may obtain a most suitable representation of ME. The distance between image node and textual node is evaluated using N_1 and N_2 . $N_1 = 0$ when we represent images by descendant nodes. Consequently, image descendant nodes participate more than other nodes in the image score. Inversely, *Rada* measure is not a suitable measure for answering generic queries as it favours specific information (descendant nodes).

Based on the previous observations, we looked for another factor that favours specific information and consequently improves results obtained by our proposed measure RepOnt comparing to ones obtained by the Rada measure. The *HiOn* measure uses a factor that can be used to increase the participation degree of the image descendants compared to other nodes. This factor is the number of changed directions between the image and the textual node: this number equals 1 if the textual node is a child of the image node, and equals 2 otherwise. The evaluation of this factor is done in the following section.

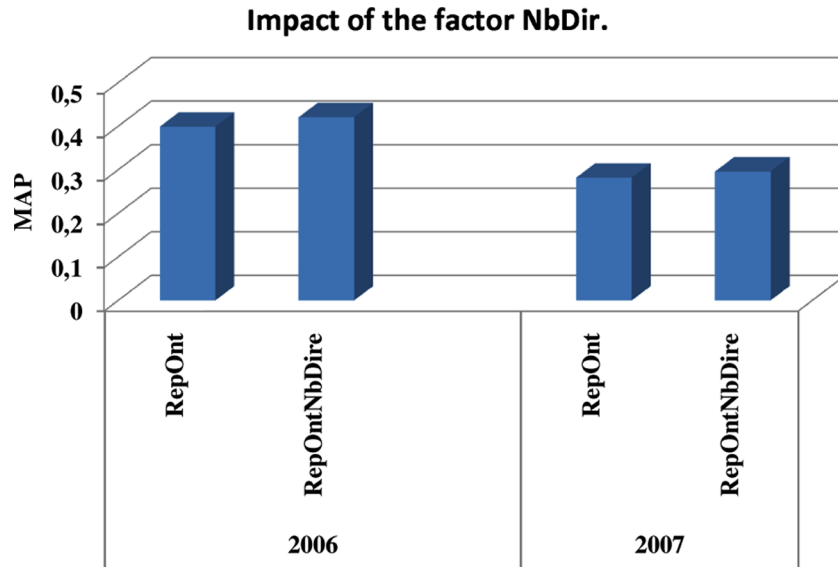
5.4.4 Impact of the number of changed directions

We recall that the measure including the NbDir factor is the following:

$$Rep_{OntNbDir}(ME, TN_i) = \frac{S^{tf} \times idf \times ief(TN_i)}{(N_1 + 1) \times Depth(CS) \times N_2 \times NbDir} \quad (17)$$

To evaluate the impact of this new measure, we compare it with the RepOnt measure. Figure 11 presents the results according to the MAP measure. Gains are reported in Table 10.

We recall that symbol ** after the gain indicates statistical significance using the Wilcoxon test at $p < 0.05$. For both test sets, we observe a significant improvement $\geq 5\%$: this implies that this structural factor is very useful in ME representation.

Figure 11 Impact of the NbDir factor in the image representation(see online version for colours)**Table 10** Gain between the RepOntNbDir metric and the RepOnt and Rada measures

	<i>RepOntNbDir/RepOnt</i>	<i>RepOntNbDir/Rada</i>
2006	+5% **	+14% **
2007	+5% **	+19% **

In addition, using the measure *RepOntNbDir* allows now a significant improvement compared to the *Rada* metric ($p(\text{RepOntNbDir}/\text{Rada}) \leq 0.05$).

5.4.5 Discussion: Comparison between implicit (CBA) and explicit use (OntologyLike) of XML structure in multimedia retrieval

We aim in this section at comparing the implicit and explicit use of textual and structural context to represent MEs.

In our first method, using implicitly textual and structural context means that these contextual factors are used indirectly to evaluate MEs scores. Indeed, inner nodes of XML documents are assigned a relevance score using these two contextual factors, and then these inner nodes are used to represent ME. The disadvantage of this method is that it strongly depends of the used system to evaluate inner node scores. In addition, this method has many parameters, some related to the XML textual system and some others related to the method itself. Consequently, to obtain the best results showed here, many experiments are needed which influences on the stability of the method. Nevertheless, this method allowed us to evaluate the impact of the children nodes, brothers nodes and ancestors nodes in ME representation, and to discover the vocabulary relation between the query and the image representation. More precisely, we concluded that representing images through specific information

(i.e., children nodes) gives the best image representation if the query vocabulary is specific, whereas representing images through generic information (i.e., ancestors nodes) gives the best image representation if the query vocabulary is generic.

In our second method, using explicitly textual and structural context means that these contextual factors can be used either separately or combined to represent MEs. This way, we can compare between text and structure in MEs representation. The main advantage of this method is that it shows its effectiveness without taking into account query type (specific or generic vocabulary). In addition, it is not dependent from another system or from parameters. The *OntologieLike* method is thus more reliable to use for ME representation than the CBA method. The evaluation of this method shows that the textual and the structural context are both useful to represent MEs. Combining the two contextual factors improves results compared to using only one context.

Finally, comparing the two methods experimentally, we notice that using implicitly both contexts (CBA) slightly improves results compared to the explicit use of the contexts (*OntologieLike*). However, according to the *Wilcoxon* test, this difference between the two methods is not significant ($p > 0.1$).

6 Conclusion

We have presented in this paper a study about the impact of XML structure on Multimedia element retrieval through two methods: the CBA method which uses implicitly the textual and structural context, and the *OntologieLike* method which uses explicitly the textual and structural context. The basic idea of both CBA and *OntologieLike* methods was presented respectively in Torjmen et al. (2008a, 2009), but we presented in this paper a more detailed description and further experimental evaluation in addition to a discussion about their effectiveness.

We have conducted experiments in order to show which sources of evidence (children nodes, brothers nodes or ancestors nodes) offers the best ME representation in the CBA method. Results show that there is a strong vocabulary relation between the query and ME representation. Concerning the *OntologieLike* method, new structural factors are proposed and evaluated in the measure of the participation degree of each textual node in ME representation. Moreover, a complete comparison between the textual context, the structural context and their combination is done. Best results are obtained with the use of both contexts.

By comparing explicit use (*OntologieLike* method) vs. implicit use (CBA method) of XML structure in multimedia retrieval, we have shown that there is no significant difference. However, we believe that the *OntologieLike* method is more effective than the CBA one as it is vocabulary-independent and it is not based on the performance of other techniques.

In future work, we plan to study the impact of textual and structural context on multimedia *fragment* retrieval, where the user need can be a multimedia element (e.g., an image) or a mixture of text and multimedia element (e.g., text + image). In our preliminary experiments presented in Torjmen et al. (2009), our model would have been ranked first in the official INEX 2007 Multimedia Focused Fragment

Task. Results are improved of respectively 17% and 53% compared to the best adhoc run and the best Multimedia Fragment run, according to the iP[0.01] measure (Kamps et al., 2007).

Moreover, we are currently studying the impact of another source of evidence to evaluate MEs: hyperlinks between XML elements in the multimedia retrieval context can also be useful.

References

- Bertini, M., Bimbo, A.D. and Torniai, C. (2005) 'Automatic video annotation using ontologies extended with visual information', *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*, Hilton, Singapore, pp.395–398.
- Chen, Z., Joyce, C. and Rong, J. (2005) 'User term feedback in interactive text-based image retrieval', *The 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'05*, Salvador, Brazil, pp.51–58.
- Denoyer, L. and Gallinari, P. (2006) 'The Wikipedia XML corpus', *SIGIR Forum*, Vol. 40, No. 1, pp.64–69.
- Fuhr, N., Lalmas, M., Malik, S. and Szlávik, Z. (Eds.) (2004) 'Advances in XML information retrieval', *Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, Lecture Notes in Computer Science, 6–8 December, Springer, Dagstuhl Castle, Germany, Vol. 3493.
- Fuhr, N., Lalmas, M., Malik, S. and Kazai, G. (Eds.) (2005) 'Advances in XML information retrieval and evaluation', *4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, 28–30 November, Springer, Dagstuhl Castle, Germany.
- Fuhr, N., Lalmas, M. and Trotman, A. (Eds.) (2006) 'Comparative evaluation of XML information retrieval systems', *5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'06*, Lecture Notes in Computer Science, 17–20 December, Springer, Dagstuhl Castle, Germany, Vol. 4518.
- Fuhr, N., Kamps, J., Lalmas, M. and Trotman, A. (Eds.) (2007) 'Advances in XML information retrieval', *6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'07*, Lecture Notes in Computer Science, 17–19 December, Springer, Dagstuhl Castle, Germany, Vol. 4862.
- Geva, S. (2005) 'GPX – gardens point XML IR at INEX', *INEX'05*, pp.240–253.
- Hirst, G. and St-Onge, D. (1998) 'Lexical chains as representation of context for the detection and correction malapropisms', *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, The MIT Press, pp.305–332.
- Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M. and Milios, E.E. (2006) 'Information retrieval by semantic similarity', *Int. J. Semantic Web Inf. Syst.*, Vol. 2, No. 3, pp.55–73.
- Hull, D. (1993) 'Using statistical testing in the evaluation of retrieval experiments', *The 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, Pennsylvania, USA, pp.329–338.
- Iskandar, A., Pehcevski, J., Thom, J. and Tahaghoghi, S. (2006) 'Social media retrieval using image features and structured text', *Proceedings of INEX 2006 Workshop*, Dagstuhl, Germany, pp.358–372.

- Jiang, J.J. and Conrath, D.W. (1997) 'Semantic similarity based on corpus statistics and lexical taxonomy', *International Conference Research on Computational Linguistics (ROCLING X)*, September, pages 9008+.
- Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M. and Robertson, S. (2007) 'INEX 2007 evaluation measures', *INEX'07*, pp.24–33.
- Kong, Z. and Lalmas, M. (2005) 'XML multimedia retrieval', *The 9th International Symposium on String Processing and Information Retrieval, SPIRE'05*, Buenos Aires, Argentina, pp.218–223.
- Kong, Z. and Lalmas, M. (2007a) 'Combining multiple sources of evidence in XML multimedia documents: an inference network incorporating element language models', *29th European Conference on Information Retrieval (Poster), ECIR'07*, pp.716–719.
- Kong, Z. and Lalmas, M. (2007b) 'Using XML logical structure to retrieve (multimedia) objects', *European Conference on Digital Libraries, ECDL'07*, Budapest, Hungary, pp.100, 111.
- Lau, C., Tjondronegoro, D., Zhang, J., Geva, S. and Liu, Y. (2006) 'Fusing visual and textual retrieval techniques to effectively search large collections of Wikipedia images', *INEX'06*, pp.345–357.
- Leacock, C. and Chodorow, M. (1998) 'Combining local context and wordnet similarity for word sense identification', *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, The MIT Press, pp.265–283.
- Li, Y., Bandar, Z.A. and McLean, D. (2003) 'An approach for measuring semantic similarity between words using multiple information sources', *IEEE Transactions on Knowledge and Data Engineering*, pp.871–882.
- Lin, D. (1998) 'An information-theoretic definition of similarity', *The 15th International Conference on Machine Learning*, pp.296–304.
- Mass, Y. and Mandelbrod, M. (2005) 'Using the inex environment as a test bed for various user models for xml retrieval', *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'05*, Dagstuhl Castle, Germany, pp.187–195.
- Mihajlovic, V., Ramírez, G., Westerveld, T., Hiemstra, D., Blok, H. and Vries, A. (2005) 'TIJAH scratches INEX 2005: Vague element selection, image search, overlap, and relevance feedback', *Proceedings of INEX 2005*, Dagstuhl, Allemagne, Dagstuhl Castle, Germany, pp.72–87.
- Miller, G.A. and Charles, W.G. (1991) 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, Psychology Press, Vol. 6, No. 1, pp.1–28.
- Moulin, C., Barat, C., Lematre, C., Géry, M., Ducottet, C. and Largeton, C. (2009) 'Combining text/image in WikipediaMM task 2009', *ECDL 2009 – Workshop CLEF*, Corfu, Greece, pp.1–6.
- Pinel-Sauvagnat, K. and Boughanem, M. (2004) 'The impact of leaf nodes relevance values evaluation in a propagation method for XML retrieval', in Baeza-Tates, R., Marek, Y., Roelleke, T. and de Vries, A.P. (Eds.): *XML and Information Retrieval Workshop – SIGIR 2004, Sheffield, Angleterre, 29/07/2004*, University of Sheffield, pp.19–22.
- Rada, R., Mili, H., Bicknell, E. and Blettner, M. (1989) 'Development and application of a metric on semantic nets', *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 19, No. 1, pp.17–30.
- Resnik, P. (1999) 'Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research*, Vol. 11, pp.95–130.

- Rodriguez, M.A. and Egenhofer, M.J. (2003) 'Determining semantic similarity among entity classes from different ontologies', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 4, pp.442–456.
- Sauvagnat, K. (2005) *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*, Thèse de doctorat, Université Paul Sabatier, Toulouse, France.
- Sauvagnat, K., Boughanem, M. and Chrisment, C. (2006) 'Answering content-and-structure-based queries on XML documents using relevance propagation', *Information Systems, Special Issue SPIRE 2004*, Vol. 31, pp. 621–635.
- Slimani, T., Ben-Yaghlane, B. and Mellouli, K. (2006) 'A new similarity measure based on edge counting', *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 17.
- Tjondronegoro, D., Zhang, J., Gu, J., Nguyen, A. and Geva, S. (2005) 'Integrating text retrieval and image retrieval in XML document searching', *5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'05*, pp.511–524.
- Torjmen, M., Sauvagnat, P. and Boughanem, M. (2008a) 'Towards a structure-based multimedia retrieval model', *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR'08*, Vancouver, British Columbia, Canada, pp.350–357.
- Torjmen, M., Pinel-Sauvagnat, K. and Boughanem, M. (2008b) 'Evaluating the impact of image names in context-based image retrieval', *Advances in Multilingual and Multimodal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF'08*, Aarhus, Denmark, pp.756–762.
- Torjmen, M., Pinel-Sauvagnat, K. and Boughanem, M. (2009) 'XML multimedia retrieval: from relevant textual information to relevant multimedia fragments', *31th European Conference on Information Retrieval, ECIR'09*, Toulouse, France, pp.150–161.
- Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly, R., Hiemstra, D. and Vries, A. (2007) 'Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah', *INEX*, pp.273–286.
- Tsikrika, T. and Vries, A. (2009) 'CWI at the photo retrieval task of ImageCLEF 2009', *Working Notes of the 10th Workshop of the Cross-Language Evaluation Forum, CLEF-Campaign*.
- Tsikrika, T. and Westerveld, T. (2008) 'The inex 2007 multimedia track', *Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'07*, Dagstuhl Castle, Germany, pp.440–453.
- van Zwol, R., Kazai, G. and Lalmas, M. (2005) 'INEX 2005 multimedia track', *INEX*, pp.497–510.
- Westerveld, T. and van Zwol, R. (2006) 'The INEX 2006 multimedia track', *INEX'06*, pp.331–344.
- Wilcoxon, F. (1945) 'Individual comparisons by ranking methods', *Biometrics Bulletin*, Vol. 1, No. 6, pp.80–83.
- Wu, Z. and Palmer, M. (1994) 'Verb semantics and lexical selection', *Proceedings of the 23rd Annual Meetings of the Associations for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, pp.133–138.
- Zargayouna, H. (2004) 'Contexte et sémantique pour une indexation de documents semi-structurés', *Conference en Recherche d'Information et Applications*, Toulouse, France, pp.571–581.

Notes

- ¹In the following, we will indifferently use element or node to design an XML document part beginning by an opening tag and ending with a closing tag.
- ²A Region Knowledge is the textual content of the multimedia object and elements hierarchically surrounding it.
- ³Bottom is a virtual node at the end of the ontology that links all leaf nodes.
- ⁴The *MAeP* metric (uninterpolated Mean Average effort-Precision) and the thorough task (which consists in simply asking systems to return elements ranked by their relevance on the topic) were used for the official Multimedia Fragment task in 2006.
- ⁵*MAiP* (Mean Average interpolated Precision) and the focused strategy (in which systems should not return overlaped elements) were used for the official Multimedia Fragment task in 2007.
- ⁶We remind that the N_1 factor is the distance between multimedia element ME and the most specific common ancestor of itself and the textual node participating to its representation, while *Depth* is the maximum number of edges between the most specific common ancestor and bottom.
- ⁷We remind that the N_2 factor is the distance between the textual node participating in ME representation and the most specific common ancestor of the multimedia element ME and this textual node.