Investigating the document structure as a source of evidence for multimedia fragment retrieval

Mouna Torjmen-Khemakhem, Karen Pinel-Sauvagnat, Mohand Boughanem

Investigating the document structure as a source of evidence for multimedia fragment retrieval

Mouna Torjmen-Khemakhem [a], Karen Pinel-Sauvagnat [b,*], Mohand Boughanem [b]

[a] *ReDCAD, DGIMA, National School of Engineering of Sfax, Tunisia*
[b] *SIG, IRIT, University of Toulouse, France*

A B S T R A C T

Multimedia objects can be retrieved using their context that can be for instance the text surrounding them in documents. This text may be either near or far from the searched objects. Our goal in this paper is to study the impact, in term of effectiveness, of text position relatively to searched objects. The multimedia objects we consider are described in structured documents such as XML ones. The document structure is therefore exploited to provide this text position in documents. Although structural information has been shown to be an effective source of evidence in textual information retrieval, only a few works investigated its interest in multimedia retrieval. More precisely, the task we are interested in this paper is to retrieve multimedia fragments (i.e. XML elements having at least one multimedia object). Our general approach is built on two steps: we first retrieve XML *elements* containing multimedia objects, and we then explore the surrounding information to retrieve relevant *multimedia fragments*. In both cases, we study the impact of the surrounding information using the documents structure.

Our work is carried out on images, but it can be extended to any other media, since the physical content of multimedia objects is not used. We conducted several experiments in the context of the Multimedia track of the INEX evaluation campaign. Results showed that structural evidences are of high interest to tune the importance of textual context for multimedia retrieval. Moreover, the proposed approach outperforms state of the art approaches.

1. Introduction

Multimedia Information Retrieval (MIR) aims at retrieving multimedia contents such as images, videos or audio objects, in response to a user information need. Two classes of approaches were developed in literature. *Content-based approaches* exploit the physical content of multimedia objects such as the color and the texture in image retrieval, or the pitch and the timbre in audio retrieval. The second class of approaches, called *context-based*, extract information around multimedia objects, which is then used to represent the objects. In this case, the physical content of multimedia objects is not exploited at all, and objects can be retrieved independently of the media type. Contextual information can be for example the text surrounding the multimedia object or the associated document title (one can for instance cite approaches of Gong, Hou, & Cheang (2006) or Noah, Azilawati, Sembok, & Meriam (2008) for image retrieval Müller, Kurth, Damm, Fremerey, & Clausen (2007) for audio retrieval or Volkmer & Natsev (2006) for video retrieval). Other contextual information such as

* Corresponding author. Tel.: +33 5 61 55 63 22.
E-mail addresses: torjmen.mouna@redcad.org (M. Torjmen-Khemakhem), Karen.Sauvagnat@irit.fr (K. Pinel-Sauvagnat), Mohand.Boughanem@irit.fr (M. Boughanem).

hyperlinks or semantic resources is also considered in Dunlop and Rijsbergen (1993) and Popescu, Grefenstette, and Moëllic (2008).

The basic assumption of approaches exploiting the text surrounding the multimedia object is that this text is included to describe the multimedia objects. Therefore it may contribute to evaluate the relevance of these objects with respect to a query. Our aim in this paper is to study the impact of text proximity in the relevance of search objects. We exploit structural information as a contextual source for multimedia retrieval. Indeed, although structural information is now extensively used in documents, only a few studies exploited the document structure to tune the importance of the different textual parts surrounding the multimedia objects.

XML documents are natural candidates for our study. Indeed, XML (*eXtended Markup Language*) is the most common language used to structure documents. This encoding standard can be used either to annotate and describe multimedia objects (as for MPEG, SVG, or SMIL formats), or to hierarchically organize documents content (text and images, videos, etc.). In the first case, all documents share the same standard structure defined by the format specification whereas in the second one, structure is heterogeneous across the different collections of documents. In this paper, we focus on this latter type of structured documents, where textual content can be easily understood by human readers.

In the particular context of XML multimedia retrieval and as defined in Westerveld and Zwol (2006) and Tsikrika and Westerveld (2008), two types of results can be returned to users queries: *multimedia elements*, i.e. the multimedia objects themselves (images for example) or *multimedia fragments*, which are composed of multimedia objects and associated text. They can be considered as document parts containing at least one multimedia object.

The main issue in multimedia element retrieval is the evaluation of the relevance of multimedia objects using contextual information composed of structure and associated text. In multimedia fragment retrieval, in addition to the object relevance, the challenge is to identify and select the most relevant multimedia fragments to be returned to the user. The resulting fragments should have an appropriate granularity, they can be composed either of the multimedia object itself, or of both text and multimedia objects.

In this paper, we focus on multimedia fragment retrieval. The approach we propose is based on two steps, first we retrieve multimedia elements and then we explore the surrounding information to retrieve relevant multimedia fragments. For both steps, we will study the impact of text proximity for relevance evaluation thanks to the underlying structural information.

Although our multimedia retrieval approach is applicable to any media type as it is only based on the multimedia object context and not on its content, we chose to illustrate and evaluate it on images for two reasons: first, the image is the the most used and easiest media (other than text) to integrate into digital documents, and secondly, to the best of our knowledge, existing collections to evaluate the use of document structure in multimedia retrieval only contain images (e.g. INEX Multimedia[1] and CLEFImage[2]).

The rest of the paper is structured as follows. We first discuss related work in Section 2 and describe our approach in Section 3. Section 4 presents evaluation and results, and our approach is compared to the state-of-the-art approaches in Section 5. Results and future works are discussed in Section 6.

2. Related work

We review in this section existing approaches for context-based multimedia retrieval, more precisely for context-based image retrieval, where queries are expressed using keywords (text) and the images annotated (indexed) by keywords provided manually or built automatically. We then focus in the second part of the section on approaches using also structural context to index images.

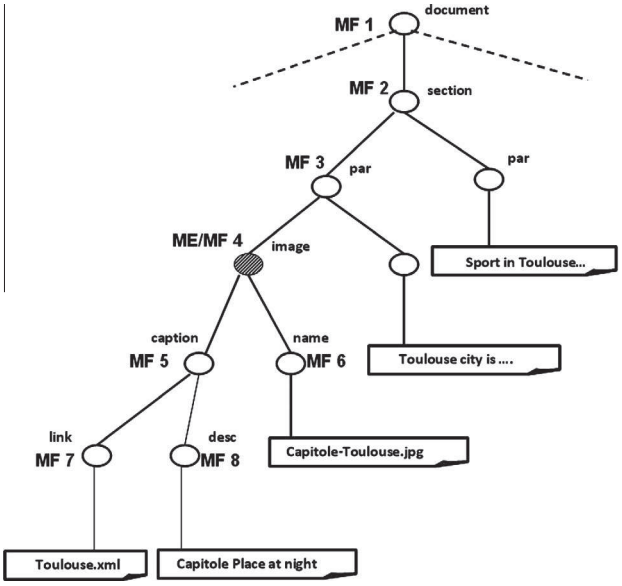
2.1. Using textual context

A first way to index images is to manually or automatically annotate them by concepts provided by the user and/or derived from semantic resources (Akbas & Yarman-Vural, 2007; Fan & Li, 2006; Hliaoutakis, Varelas, Voutsakis, Petrakis, & Milios, 2006; Piotrowski, 2009; Popescu et al., 2008).

Other approaches state that there is a strong correlation between an image and its surrounding text in the document. Therefore images search is often carried out using the textual content of the image name (and sometimes its extension) or using the associated text of the image, extracted from the document. In web collections for instance (HTML pages), the associated text of the image is generally extracted from `src` and `alt` tags (Shen, Ooi, & Tan, 2000), title of the web page, other particular tags (Noah et al., 2008), or also from text close to the image (Chen, Liu, Zhang, Li, & Zhang, 2001; Guglielmo & Rowe, 1996; LaCascia, Sethi, & Sclaroff, 1998; Gong et al., 2006; Srihari, Zhang, Rao, Baird, & Chen, 2000). Many images search engines on the Web (as Google[3] and Lycos[4]) use such methods.

The context of images can also be enlarged to other documents, thanks for example to (hyper) links (Chakrabarti et al., 1998; Chakrabarti, Punera, & Subramanyam, 2002; Dunlop, 1991; Dunlop & Rijsbergen, 1993; Haveliwala, Gionis, Klein, &

[1] INEX: Initiative for the Evaluation of XML Retrieval, multimedia track.
[2] CLEFImage: Cross-Language Evaluation Forum, Image Track.
[3] http://www.google.com.
[4] http://www.lycos.fr.




```
<?xml version="1.0"?>
<document>
…
<section>
  <par>
    <image>
      <caption>
        <link>Toulouse.xml</link>
        <desc>Capitole Place at night</desc>
      </caption>
      <name>Capitole-Toulouse.jpg</name>
    </image>
    <spar>
     Toulouse city is…
    </spar>
  </par>
  <par>
   Sport in Toulouse…
  </par>
</section>
…
</document>
```

Fig. 1. Example of multimedia elements and multimedia fragments.

Indyk, 2002; LaCascia et al., 1998). Dunlop (1991) and Dunlop and Rijsbergen (1993) proposed a link-based method where textual documents (called *textual nodes*) are linked to multimedia elements (called *multimedia nodes*). The two types of nodes are represented in a graph, and links between nodes are then used to retrieve images. More precisely, textual documents linked to the same image form a class that will be used as a textual representation of this image.

2.2. Using structural context

Although structure showed its interest in textual retrieval (Fuhr, Lalmas, Malik, & Kazai, 2006; Fuhr, Lalmas, & Trotman, 2007), only a few approaches have exploited it as a contextual source for multimedia retrieval. Structural information of documents may indeed help to evaluate images relevance, by giving information on the interest of their surrounding textual parts.

In particular, in XML-based multimedia retrieval, two types of results can be returned according to the document structure (Tsikrika & Westerveld, 2007, 2008; Westerveld & Zwol, 2006):

- the multimedia objects themselves (called "*multimedia elements*" in the INEX[5] terminology), that is XML elements containing the reference entity to the multimedia object content (*file name*) and possibly associated information, as caption for example,
- *multimedia fragments*, composed of multimedia objects possibly associated with textual information. The returned objects might be the multimedia objects themselves or their ancestors or their descendants.

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 Fig. 1 and the query "*Toulouse city*". 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, ME (it is also a multimedia fragment: MF4), MF5, MF6, MF7 and MF8.

To our knowledge, only a few approaches for multimedia element retrieval using document structure were proposed in the literature. Kong and Lalmas (2005, 2007) approach consists of dividing textual content into *Region Knowledge*[6] (RKs): self level RK (RK of the multimedia element); sibling level RK (RK of the sibling elements of the multimedia node); first ancestor level RK (RK of the first ancestor of the multimedia element excluding nodes already used); second ancestor level RK; . . . ; *N*th ancestor level RK. Then, authors used the vector space model to evaluate the relevance of each Region Knowledge w.r.t. a query. The final score of the image is evaluated by combining the different regions participating in the image representation with different degrees. Even though this method exploits the document structure, it does not take into account the element position in the same Region Knowledge.

[5] INEX (INitiative for the Evaluation of XML retrieval) is the reference evaluation campaign for structured retrieval.

[6] A Region Knowledge is the textual content of the multimedia object and elements hierarchically surrounding it.

In our previous work, we proposed two methods to retrieve multimedia elements (Torjmen, Pinel-Sauvagnat, & Boughanem, 2010). The first one, called *Children, Brothers and Ancestors* (CBA), consists of evaluating a score for each multimedia element through the scores of its children, brothers and ancestors, already evaluated by an XML retrieval model based on relevance score propagation (XFIRM). The second one, called *OntologyLike*, consists of first computing content scores of leaf nodes using a scoring formula based on tf-idf (*tf: term frequency; idf: inverse document frequency*), and then evaluating scores of the multimedia elements using the content scores and the hierarchical structure of XML documents. As this approach is used in our multimedia fragment retrieval method, we detail it in Section 3.1.

Concerning multimedia fragment retrieval approaches, most of the proposals in literature combine textual XML retrieval with image content-based approaches (Iskandar, Pehcevski, Thom, & Tahaghoghi, 2006; Lau, Tjondronegoro, Zhang, Geva, & Liu, 2006; Mihajlovic et al., 2005; Tjondronegoro, Zhang, Gu, Nguyen, & Geva, 2006). Content-based algorithms are used to score images similar to the one in the query (queries are often composed of text and an example image), and image scores are then combined with the ones obtained on text with traditional XML retrieval systems. These approaches can also be applied to retrieve multimedia elements.[7] Several limits can however be outlined: (1) results show that in most of the cases, the use of visual features decreases the system accuracy, (2) to use a combination of evidences, queries should always contain multimedia hints, and (3) document structure effect in multimedia retrieval cannot be really studied as it is used on a classical XML retrieval framework.

Another approach consists of using an XML retrieval system to assign a score to each XML element and then filtering retrieved results by keeping only fragments having a multimedia object (i.e. fragments having at least one image). For example, the method proposed by Tsikrika et al. (2007) uses a traditional retrieval method based on language models and on different length priors, and then retrieved results are limited to fragments that contain at least one image. No further multimedia processing is used. This method shows its effectiveness when the retrieved fragments are the whole documents.

Kong et al. in Kong and Lalmas (2007) use 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 effectiveness compared to official participants. Their results need however to be confirmed on a larger collection, such 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, Kazai, & Lalmas, 2005), only 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 collection associated to the task 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 type of XML documents can be found in Torjmen, Pinel-Sauvagnat, and Boughanem (2008), Tsikrika and Vries (2009), or Moulin et al. (2010). 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 for fragments selection use either a combination of classical XML and content-based multimedia retrieval, or a filtering of classical XML results by keeping only fragments having at least one multimedia element. Only a few approaches offer a real study of the impact of the XML structure (and therefore text position) in Multimedia Retrieval, and this is the purpose of this paper. Structural context will be used as a clue for evaluating the importance of textual content surrounding multimedia elements and fragments.

3. From multimedia elements to multimedia fragments

The approach we propose in this paper retrieves multimedia fragments from multimedia elements. The relevance scores of both elements and fragments are evaluated thanks to their textual context, whose importance is estimated using structural information, i.e. text position in the document. As defined in Tsikrika and Westerveld (2008) and Westerveld and Zwol (2006) and as aforementioned, a multimedia fragment can be a multimedia element, a sub-tree containing at least one multimedia element or also a descendant of a multimedia element. All ancestors and descendants of a multimedia element and the multimedia element itself are consequently "good" answers (i.e. fragments to be returned).

Our approach follows two main steps: (1) we first retrieve candidate multimedia elements (Section 3.1) and (2) we then explore the surrounding information to retrieve relevant multimedia fragments. More precisely, the score of those fragments is evaluated as follows: we first compute an initial score using a classical XML retrieval system (Section 3.2), that is then combined with the corresponding element score using the hierarchical relation between the multimedia element and the considered multimedia fragment (Section 3.3).

3.1. Multimedia element retrieval: Ontologylike approach

To retrieve relevant multimedia elements, we have already proposed in previous work two methods based on textual and structural contexts (Torjmen et al., 2010): *CBA* and *Ontologylike*. The two methods obtain similar performances, but as the

[7] A multimedia element can also be considered as a multimedia fragment.

OntologyLike method is independent of any other system (contrary to CBA that uses the XFIRM system), it will be preferred here to retrieve multimedia elements. In order to make the paper self-explanatory, this method is described in the following.

An XML document can be represented as a hierarchical tree, composed of a root (document), simple nodes (element and/or attributes) and leaf nodes (values as text and images). An inner node is any node of the tree that has child nodes (i.e. a non-leaf node). The relevance score of a multimedia element *me* according to query $q = t_1, \ldots, t_n$ composed of n keywords terms can be evaluated thanks to the following general formula:

$$S(me, q) = \sum_{ln_i \in L_{doc(me)}} \phi(me, ln_i) \cdot RSV(ln_i, q) \tag{1}$$

where

- $L_{doc(me)}$ is the set of textual (leaf) nodes of the document containing the multimedia element *me*.
- $RSV(ln_i, q)$ is the content relevance score of a textual (leaf) node ln_i belonging to the same document than *me*, evaluated as follows:

$$RSV(ln_i, q) = \sum_{j=1}^{n} w_i^q * w_j^{ln_i},$$
$$\text{where } w_j^q = tf_i^q \quad \text{and} \quad w_j^{ln_i} = tf_i^{ln_i} * idf_j * ief_j \tag{2}$$

  w_j^q and $w_j^{ln_i}$ are the weights of term j in query q and leaf node ln_i respectively. tf_j^q and $tf_j^{ln_i}$ are the frequency of term j in q and ln_i respectively, $idf_j = log(|D|/(|D_j| + 1)) + 1$, with $|D|$ the total number of documents in the collection, and $|D_j|$ the number of documents containing j, and ief_j is the inverse element frequency of term j, i.e. $log(|LN|/|LN_j| + 1) + 1$, where $|LN_j|$ is the number of leaf nodes containing j and $|LN|$ is the total number of leaf nodes in the collection.
- $\phi(me, ln_i)$ is a factor allowing to weight the importance of leaf node ln_i in the relevance evaluation of *me*.

Factor ϕ reflects the proximity of the leaf nodes to the considered multimedia element. We use structural information to evaluate this proximity and to differentiate the impact of each leaf node on the multimedia relevance. Intuitively, factor ϕ should convey the following insights: textual descendants of the multimedia element should more participate to its relevance score than textual descendants of its brothers, and the latter nodes should more participate than textual descendants of its ancestors. Indeed, textual descendants of the multimedia element can be considered as the most specific nodes to represent multimedia elements; textual descendants of brothers nodes have a high probability of sharing the same information than the multimedia elements; and descendants of the root node should less participate since they are far from the multimedia element in the document tree. For instance, if we consider the XML document of Fig. 2, we argue that the relevance

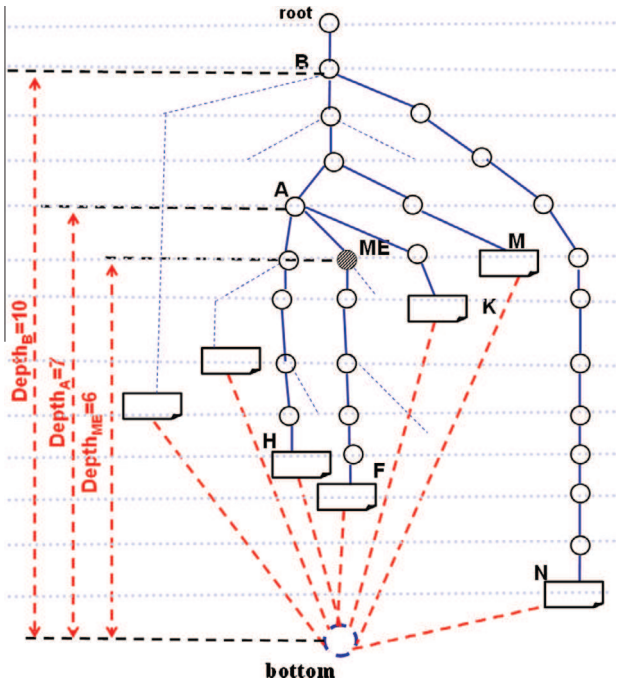


Fig. 2. Definition of bottom and the depth factor.

score of element *ME* should be first influenced by the relevance of *F*, then by the relevance of nodes *H* and *K*, and finally by the relevance of nodes *M* and *N*.

To evaluate factor ϕ, we make the following assumption: thanks to its tree representation, an XML document can be seen as a simple ontology: nodes can be considered as concepts linked with the "*IsPartOf*" relationship. For example, "*Section IsPartOf Article*", "*Paragraph IsPartOf Section*", etc. The main idea of the *OntologyLike* approach is to exploit a semantic similarity measure between ontological concepts to estimate the participation degree of each textual node to the relevance of the multimedia element.

A first way to evaluate ϕ is to directly use semantic similarity measures that can be found in the literature. They are divided into edge counting measures and information content measures. In this work, we are interested in the first type of measures as textual content of a multimedia element is generally small or can even be absent. Among these metrics, one can cite Rada, Mili, Bicknell, and Blettner (1989) and Wu and Palmer (1994) that can be simply adapted as follows:

- Rada:

$$\phi_{Rada}(me, ln_i) = \frac{1}{dist(me, ln_i)} \tag{3}$$

where $dist(me, ln_i)$ is the distance (number of edges) between node *me* and node ln_i in the document tree;

- Wu–Palmer:

$$\phi_{WP}(me, ln_i) = \frac{2 * N}{N_1 + N_2 + 2 * N} \tag{4}$$

with $N_1 = dist(me, CS)$ and $N_2 = dist(ln_i, CS)$ are the distances which separate *me* and ln_i from their most specific common ancestor *CS* and $N = dist(CS, root)$ is the distance between CS and the root.

Even if these metrics allow to distinguish between leaf nodes, some brother nodes of a multimedia element may more influence the relevance score than its descendants, although we intuitively assume the contrary. This is illustrated in Fig. 2. Using ϕ_{Rada} or ϕ_{WP} factors will give more importance to node K than node F in the relevance evaluation of *ME*. To overcome this problem, we introduced a *depth* factor as follows (Zargayouna, 2004):

$$\phi_{OntLike}(me, ln_i) = \frac{1}{(N_1 + w) * N_2 * depth(CS(me, ln_i))} \tag{5}$$

$Depth(CS(me, ln_i))$ is maximum number of edges between $CS(me, ln_i)$ and *bottom* node which is a virtual concept linking all leaf nodes (see Fig. 2). This factor reflects the vertical hierarchical structure of the document and allows to differentiate the participation degree of the textual nodes of the considered multimedia element's ancestors. w (with $w > 0$) is a factor added to $N1$ to avoid zero division when the multimedia element is itself the common ancestor between the textual node and the multimedia element.

To illustrate all these factors, let us consider the example in Fig. 2. Considering the multimedia element *ME* and the three leaf nodes *F*, *K* and *N*, we first determine the most specific ancestor between *ME* and the leaf nodes: $CS(ME, N) = B$, $CS(ME, K) = A$ and $CS(ME, F) = ME$. The *Depth* factor of these common ancestors is then: $Depth(B) = 12$, $Depth(A) = 7$ and $Depth(ME) = 6$. We then have for all three leaf nodes:

$$\phi_{OntLike}(ME, N) = \frac{1}{(4 + w) * 15 * 12} \tag{6}$$

$$\phi_{OntLike}(ME, K) = \frac{1}{(2 + w) * 2 * 7} \tag{7}$$

$$\phi_{OntLike}(ME, F) = \frac{1}{(0 + w) * 5 * 6} \tag{8}$$

The descendant node *F* of the multimedia element *ME* participates thus more in the *ME* representation than nodes *K* and *N*: $\phi_{OntLike}(ME, F) > \phi_{OntLike}(ME, K) > \phi_{OntLike}(ME, N)$ for small values of w ($w < 1.75$ in our example).

3.2. Initial score of multimedia fragment

Once each multimedia element *me* is assigned a score according to the *OntologyLike* approach, we then compute an initial score for each associated multimedia fragment (i.e. to each ancestor and each descendant of the multimedia element, given that a document may contain several multimedia elements). This score does not take into account the multimedia information of the fragments, it is only based on textual and structural information. It is evaluated thanks to the XFIRM XML retrieval model (Sauvagnat, 2005). This model is based on a relevance propagation method. For each query, relevance scores are assigned to textual nodes (which are content bearer), and relevance scores of inner nodes (i.e. here multimedia fragments) are then computed. The relevance score $S_{XFIRM}(mf, q)$ of a multimedia fragment *mf* is evaluated according to formula (9). The first part of the formula uses the scores of the descendant leaf nodes of *mf*, while the second part takes into account the whole

document score: a multimedia fragment contained in a relevant document is more likely to be relevant than another contained in a non-relevant document.

$$S_{XFIRM}(mf, q) = \rho * |L^r_{mf}| \cdot \sum_{ln_k \in L_{mf}} \alpha^{dist(mf, ln_k)-1} * RSV(q, ln_k) + (1-\rho) * S_{XFIRM}(root, q) \quad (9)$$

where

- mf is a descendant or an ancestor of a relevant multimedia element (i.e. e multimedia element having a relevance score >0 according to Eq. (5)),
- L_{mf} is the set of leaf nodes being descendant of mf,
- $dist(mf, ln_k)$ is the distance between node mf and leaf node ln_k in the document tree, and $\alpha \in]0 \cdots 1]$ allows to adapt the importance of the $dist$ parameter,
- $|L^r_{mf}|$ is the number of leaf nodes being descendant of mf and having a non-zero relevance value (according to Eq. (2)),
- $\rho \in]0 \cdots 1]$, inspired from work presented in Mass and Mandelbrod (2005), allows the introduction of document relevance in inner nodes relevance evaluation,
- and $S_{XFIRM}(root, q)$ is the relevance score of the $root$ element, i.e. the relevance score of the whole document.

3.3. Multimedia fragments retrieval

To compute the final score of each multimedia fragment, we then combine the score of their associated multimedia element (obtained in Section 3.1) with their initial score (obtained in Section 3.2). A simple combination was proposed in Torjmen, Pinel-Sauvagnat, and Boughanem (2009), which is considered in this paper as a baseline for comparison with two others combination methods, described in Section 3.3.2.

3.3.1. Simple combination

As proposed in Torjmen et al. (2009), a simple way to evaluate the score $S(mf, q)$ of each ancestor and descendant of one or more multimedia elements is to linearly combine its initial score obtained by the XFIRM system ($S_{XFIRM}(mf, q)$) and the scores of the associated multimedia elements:

$$S(mf, q) = \lambda * S_{XFIRM}(mf, q) + (1-\lambda) * \sum_{i=1}^{|me|} S(me_i, q) \quad (10)$$

with me_i a multimedia element which is ancestor or descendant of mf, $\lambda \in [0 \cdots 1]$ and $|me|$ the number of multimedia elements contained in mf when mf is ancestor of me_i or 1 otherwise (mf is descendant of me_i).

If two multimedia fragments associated to the same multimedia elements have the same relevance score, we first rank the one having the highest hierarchical level, the latter is supposed to be more exhaustive.

This simple way to evaluate scores of multimedia fragments raises two problems: first this may favor root elements (i.e. whole documents) when λ is small and second multimedia fragments containing many multimedia elements may also be favored compared to those having only one multimedia element.

3.3.2. Adding structural information in the relevance score of multimedia fragments

To overcome the aforementioned limits, we propose to correlate the importance of multimedia elements in the relevance of fragments with the distance that separates them from the fragment nodes: the larger the distance between a multimedia element node and a multimedia fragment node, the less the multimedia element participates to the relevance of the fragment. This allows to promote specificity in the one hand: the first ancestor (resp. descendant) of the multimedia element will have a higher score than its second ancestor (resp. descendant), etc. On the other hand, the use of this factor will solve the problem of fragments containing more than one multimedia element. In order to control the participation of multimedia elements in the relevance of multimedia fragments, we integrate in formula (10) the θ factor as follows:

$$S(mf, q) = \lambda * S_{XFIRM}(mf, q) + (1-\lambda) * \sum_{i=1}^{|me|} \theta * S(me_i, q) \quad (11)$$

We evaluated two different values of θ, which both take into account the distance between nodes me_i and mf:

$$\theta = \frac{1}{Dist(me_i, mf) + 1} \quad (12)$$

Or

$$\theta = K^{(Dist(me_i, mf)+1)} \quad (13)$$

where $Dist(me_i, mf)$ is the distance (number of edges) between the multimedia element me_i and the associated multimedia fragment mf (ancestor or descendant). Adding 1 to the Dist factor is done to avoid the zero value.

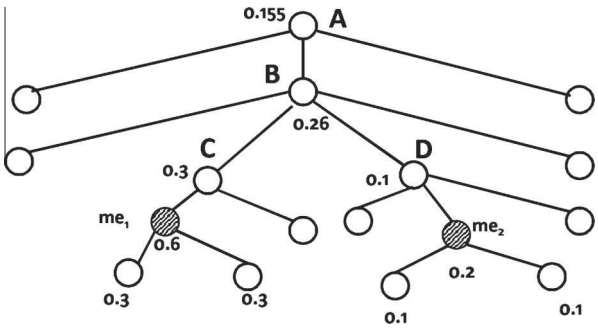


Fig. 3. Using the factor $Dist(me_i, mf)$ to compute ancestors scores.

In formulas (11) and (12), the structure score of the multimedia element contained in *mf* is simply divided by the distance separating the two nodes. As this may lead the *Dist* factor to impact too strongly in the evaluation of *mf* relevance, we propose in formula (13) to use the K parameter to tune its importance, with $K \in [0.1, 1]$.

To illustrate the impact of the *Dist* factor in our formulas, let us consider the example of Fig. 3. The document contains two multimedia elements, me_1 and me_2. If we apply formula (10) with $\lambda = 0$ (only the structural score is taken into account), fragments *A* and *B* would have the highest score in the document (0.8) since they contain the two multimedia elements. If we now apply Eqs. (11) and (12) with $\lambda = 0$, fragment *C*, which has only one multimedia element will be returned before fragments *B* and *A*, which contain two multimedia elements but more irrelevant information.

In the following section, we describe the evaluation of our approach.

4. Experimental evaluation

Our aim in these experiments is to study the impact of structural factors to first evaluate the relevance of multimedia elements and then the relevance of multimedia fragments.

For this purpose, we used the INEX (*INitiative for the Evaluation of XML Retrieval*) Multimedia tracks 2006 and 2007 (Tsikrika & Westerveld, 2008; Westerveld & Zwol, 2006)[8] framework: a test collection, a set of queries, the associated relevance assessments and appropriate evaluation metrics.

Concerning multimedia element retrieval, formula (5) was validated in previous work (Torjmen et al., 2010). We focus here on the impact of its different factors compared to content only as a source of evidence, and compared to traditional semantic measures (Section 4.2). Then, we discuss the benefit of our approach to retrieve multimedia fragments according to two retrieval strategies: *Thorough* and *Focused* (Sections 4.3 and 4.4). Our multimedia fragment approach is also compared to official participants to INEX (Section 5).

4.1. Experimental setup

The core collection of the Multimedia track is the English version of the Wikipedia XML collection, composed of about 660,000 XML documents (4.6 Giga-Bytes without images). This collection contains 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). Multimedia elements in the XML corpus are images. Topic sets in 2006 and 2007 are respectively composed of 9 and 19 topics.[9] We only use the *title* field (keywords terms) of topics for our experiments.

4.1.1. Evaluation of multimedia element retrieval

The effectiveness of our approach of multimedia element retrieval is evaluated with the *Mean Average Precision* (MAP), commonly used in IR.

In the official INEX campaign, assessments on both 2006 and 2007 test sets are done on multimedia fragments and not on multimedia elements. In order to properly evaluate our method, we constructed a new assessments base composed only of relevant multimedia elements (i.e. images) extracted from the original assessments provided by organizers.

[8] The multimedia track at INEX was launched in 2006 and 2007. The task was then transfered to the CLEF (Cross-Language Evaluation Forum) campaign, but with materials that cannot be used to evaluate our approach. The proposed collections are composed of documents having exactly the same structure used to annotate multimedia objects and not to semantically organize the document data.

[9] Building a test set for the INEX campaign is a collaborative effort: participants submit topics that are then selected by the organizers, and also make relevance assessments. This can explain the relatively small number of queries.

Table 1
Results for some representative values of w in Eq. (5).

w	INEX-MM2006	INEX-MM2007
0.01	0.4348	0.2281
0.1	**0.4496**	0.2914
0.2	0.4257	**0.2969**
0.3	0.4201	0.2909
0.4	0.4160	0.2907
0.5	0.4146	0.2932
0.6	0.4081	0.2869
0.7	0.4104	0.2844
0.8	0.4099	0.2832
0.9	0.3997	0.2827
1	0.3989	0.2819
2	0.3926	0.2730
3	0.3919	0.2707

Best results are in bold.

4.1.2. Evaluation of multimedia fragment retrieval

Our approach for fragment retrieval is evaluated according to two strategies:

- in the first one, fragments can be returned with overlap (multimedia elements and/or descendants and/or ancestors can be returned). The challenge here is to correctly rank these fragments. This task is called *Thorough* (Westerveld & Zwol, 2006) in the INEX terminology;
- in the second one, the returned fragments cannot overlap (i.e. in our approach, we should decide whether we should return the multimedia element itself, an ancestor or a descendant). The aim here is to focus on the user need and to select the best fragment to be returned. This task is called *Focused* (Tsikrika & Westerveld, 2008) in the INEX terminology.

As each strategy is evaluated with a different metric, we detail them in Sections 4.3 and 4.4. In addition, all results presented in the following were tested for statistical significance using the *signed-rank test of Wilcoxon* test (Wilcoxon, 1945) which is the non-parametric equivalent of the paired samples t-test.

In our experiments, we consider that the difference between two methods is very significant when $p < 0.05$ (results marked by $*$).

4.1.3. Relevance assessments of multimedia fragments

The relevance assessments of multimedia fragments provided by INEX organizers contain some pure textual fragments (without any image) that were judged as relevant by INEX assessors. Although the specificity of the Multimedia tack was clearly defined before doing the assessments (multimedia fragments must be multimedia elements or must contain or be contained in at least one multimedia element (Tsikrika & Westerveld, 2007)), assessments provided to participants do not respect this restriction: by analyzing them, we found that 84.71% (for INEX 2007) and 70.80% (for INEX 2006) of the relevance assessments are "pure" textual fragments.

These official relevance assessments can therefore not be directly used to evaluate our approach. To be coherent with the task definition, we decided to filter the assessments by keeping only fragments associated with at least one image.

Jointly to this assessment filtering, we filtered official submissions[10] of INEX 2006 and 2007 participants (as some participants also returned purely textual fragments) in order to compare approaches.

4.2. Evaluation of multimedia element retrieval

Our aim in this section is to compare the influence of the structural factors introduced in factor ϕ of Eq. (1). As ϕ might depend on parameter w in Eq. (5), some preliminary runs are necessary to calibrate our model.

4.2.1. System calibration

Results with some representative values of w in Eq. (5) are shown in Table 1.

Best results are obtained with relatively small values of w (0.1 and 0.2 for the 2006 and 2007 test sets). We will keep these values in the rest of the paper.

4.2.2. Evaluating factor ϕ

We evaluated different runs, corresponding to different values of factor ϕ:

[10] These submissions are called runs in INEX terminology.

Table 2
Results of multimedia elements retrieval with different values of ϕ.

Models			INEX-MM2006				INEX-MM2007			
	ϕ	Structural factors	MAP	% Change over TC	% Change over Rada	% Change over WP	MAP	% Change over TC	% Change over Rada	% Change over WP
TC	1	–	0.3116	–	–	–	0.2145	–	–	–
Rada	Eq. (3)	Dist (me, ln_i)	0.3695	+18%*	–	–	0.2493	+16%*	–	–
WP	Eq. (4)	Dist (me, CS) = N1, Dist (ln_i, CS) = N2, Dist (CS,root)	0.3436	+10%*	−8%	–	0.2434	+13%*	+2%*	–
OntLike_w = 0.1	Eq. (5)	Dist (me, CS) = N1, Dist (ln_i, CS) = N2, Depth (CS)	0.4496	+44%*	+22%*	+31%*	0.2914	+36%*	+17%*	+20%*
OntLike_w = 0.2	Eq. (5)	Dist (me, CS) = N1, Dist (ln_i, CS) = N2, Depth (CS)	0.4257	+37%*	+15%*	+24%*	0.2969	+38%*	+19%*	+22%*

* Statistical significance with the Wilcoxon test at $p < 0.05$.

- a first run is done by taking $\phi = 1$. This allows to evaluate the impact of textual context independently of any structural information.
- two runs (Rada and WP) respectively correspond to Eqs. (3) and (4).
- the other runs (OntLike_w = *X*) correspond to Eq. (5) with $w = 0.1$ or $w = 0.2$.

Results are shown in Table 2. The first conclusion that can be drawn is that the use of structural context (WP, Rada and OntLike runs) improves significantly the results compared to the simple use of textual context (TC). *Considering text position positively influences the relevance.*

If we now consider in detail runs using structural factors, we notice that the OntLike_w = X runs perform better than the Rada and WP ones (improvements are statistically significant):

- The Rada measure only uses the distance between the image and each textual node: this does not ensure that textual nodes included in the multimedia element participate more in its relevance than textual nodes included in its brother nodes, etc. The use of the *Depth* factor in OntoLike_w = *X* run solves this issue.
- The main difference between the WP and OntLike runs is that factor *Dist*(*CS*, *root*) in WP is replaced by *Depth*(*CS*) in Ontlike. *Depth*(*CS*) seems to be more effective than the *Dist*(*CS*, *root*). This can be explained as follows. We proposed formula (3)–(5) in order to allow each leaf node of a considered document to participate (at a certain degree) to the relevance of the multimedia elements. However, when considering leaf nodes that are not descendants of an ancestor (other than the root) of the multimedia element, the root element and the CS element will coincide (leading thus to *Dist*(*CS*, *root*) = 0). As a consequence, in WP run (formula (4)), we will have $\phi = \frac{2 * Dist(CS,root)}{N1+N2+2 * Dist(CS,root)} = 0$ for those leaf nodes, and consequently they will not participate to the evaluation of the relevance score. To solve this problem, a solution would have been to add a constant to the *Dist*(*CS*, *root*) factor to avoid 0 values for those leaf nodes. This will however not allow us to distinguish the contribution of each leaf node.
  This is illustrated in Fig. 4. Although we want that all leaf nodes participate to the relevance score of the *image* element, this will not be the case for the leaf nodes of section[2] and section[3]: for those leaf nodes, CS and the root element coincide (and thus *Dist*(*CS*, *root*) = 0) and they not participate to the image relevance.

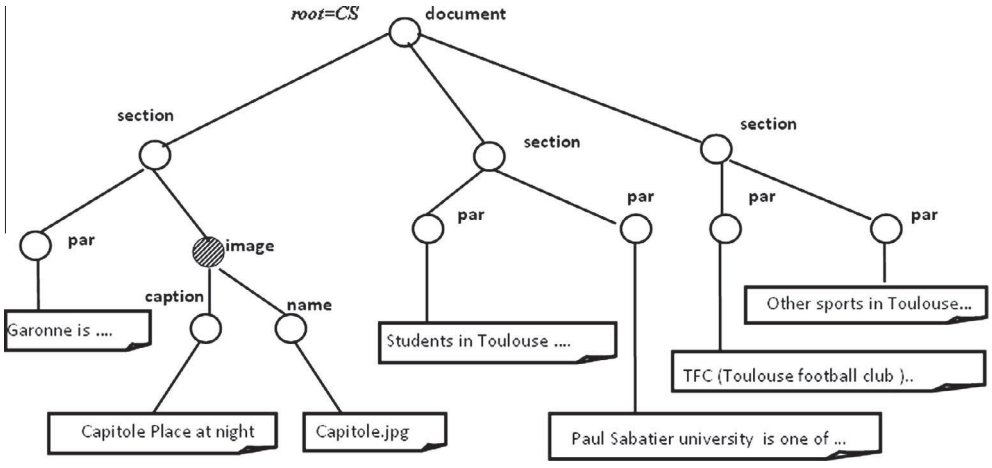


Fig. 4. For leaf nodes of section[2] and section[3], the CS element coincide with the root element.

As Eq. (5) (OntLike runs) allows to obtain the best results, it will be used to retrieve multimedia elements in our multimedia fragment retrieval approach. Even if results for the w parameter are very similar between the two collections, its best value is different for the two test sets (0.1 for 2006 and 0.2 for 2007). $w = 0.1$ is chosen as a common value for the rest of the experiments (best compromise). A 2-fold cross validation with the two different values of w will be done when comparing results with official INEX participants (parameters are learned from INEX 2006 ane evaluated on INEX 2007 and vice versa).

4.3. Evaluation of Multimedia fragment retrieval according to the Thorough strategy

The challenge of this strategy is to select and rank all relevant multimedia fragments, even if the set of retrieved results contain fragments that overlap (*Thorough Retrieval*) (Westerveld & Zwol, 2006). Before showing and discussing results, we briefly describe in the following section the metric used to evaluate systems.

4.3.1. Evaluation metrics

The INEX official metric for the *Thorough* strategy is MAeP (*Mean Average Effort Precision*) (Lalmas et al., 2006), which is based on "effort-precision/gain-recall".

Effort-precision (ep) is calculated, at a given cumulated gain value (r), as follows:

$$ep[r] = \frac{i_{ideal}}{i_{run}} \tag{14}$$

where i_{ideal} is the rank position at which the cumulated gain of r is reached by the ideal curve and i_{run} is the rank position at which the cumulated gain of r is reached by the system run.

Gain-recall (gr) is calculated as follows:

$$gr[i] = \frac{\sum_{j=1}^{i} specS(e_j)}{\sum_{j=1}^{n} specI(e_j)} \tag{15}$$

where i is the ith element in the result list. n is the total number of relevant elements in the full recall-base of the given topic. $specS(e_j)$ is the specificity of the jth element in the system ranking and $specI(e_j)$ is the specificity of the jth element in the ideal ranking.

The non-interpolated mean average effort-precision, denoted *MAeP*, is evaluated by averaging the ep values obtained for each rank where a relevant document is returned.

4.3.2. Evaluation of the simple combination between scores

In order to find the best multimedia fragment types to return to users (the multimedia element itself, ancestors or descendants), we evaluated the following cases on our simple combination of scores (Eq. (10)):

- Only images (i.e. multimedia elements) (I) are returned.
- Only image descendants (D) are returned.
- Only image ancestors (A) are returned.
- Only images and image descendants (ID) are returned.
- Only images and image ancestors (IA) are returned.
- Only image descendants and ancestors (DA) are returned.
- Multimedia fragments composed of images, or image ancestors or descendants (IDA) are returned.

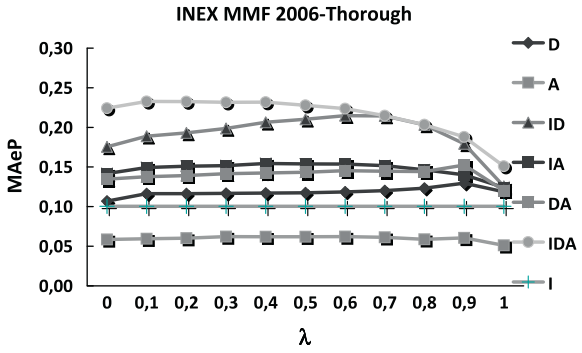


Fig. 5. *MAeP* variation against λ, 2006 test set.

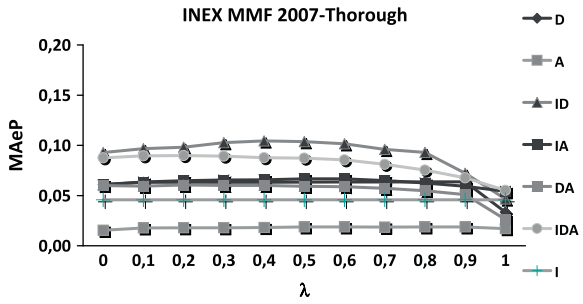


Fig. 6. *MAeP* variation against λ, 2007 test set.

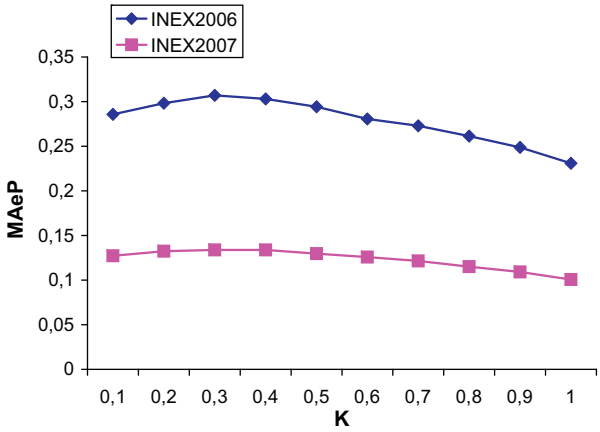


Fig. 7. Impact of K when $\lambda = 0$ with the possibility to return images, ancestors and descendants.

Results presented here are based on filtered relevance assessments, as explained in Section 4.1. Figs. 5 and 6 show the *MAeP* evolution against λ for all the cases mentioned above, and for both 2006 and 2007 test sets.

The curve denoted by *I* on both figures show results when only image elements are returned. This curve does not depend of λ, but it was plotted in order to directly compare all the types of returned elements cited above.

We observe for most of curves that using only the image score ($\lambda = 0$) leads to obtain better results than using only the initial scores of fragments evaluated by the XFIRM system ($\lambda = 1$).

In INEX 2006, best results are obtained by returning images, image ancestors and image descendants (*IDA*). On this run, combining results by a classical linear function provides slightly better results ($0.1 < \lambda < 0.4$ in Eq. (10)) than using only image scores.

In INEX 2007, best results are obtained by returning only images and image descendants (*ID*). These results are however comparable to those obtained by returning image, image ancestors and image descendants (*IDA*). The slight improvement of the *ID* run comparatively with *IDA* in INEX 2007 can be explained by the nature of relevance assessments: the percentage of results in these assessments where images are more relevant[11] than their ancestors is 96.52%. This question is discussed in details in Section 4.4.4.

At last, we also observe that combining both scores (image and XFIRM score) leads to a slight improvement of results (with $0.1 < \lambda < 0.6$ on the *ID* run).

To summarize, we conclude that returning images and/or image descendants and/or image ancestors (*IDA*) is generally the best strategy for the Thorough task. We will thus keep this strategy for the rest of experiments.

4.3.3. Is structure useful to improve results?

As mentioned in Section 3.3.2, structural information can be taken into account in our multimedia fragment retrieval method by adding factor $Dist(me_i, mf)$ (Eq. (11)). We evaluated two ways to take into account this factor (Eqs. (12) and (13)).

Before comparing both formulas, we first varied parameter K (Eq. (13)) in order to identify its best values. Fig. 7 shows results obtained by varying K when $\lambda = 0$, i.e. when using only the image score to evaluate scores of multimedia fragments.

[11] The relevance of an element is evaluated according to the fraction of relevant content in contains against its overall size.

Table 3
Improvements using $K = 0.3$ against $K = 1$ when $\lambda = 0$ for Thorough strategy.

Test set	$K = 0.3/K = 1$
2006	+31%*
2007	+33%*

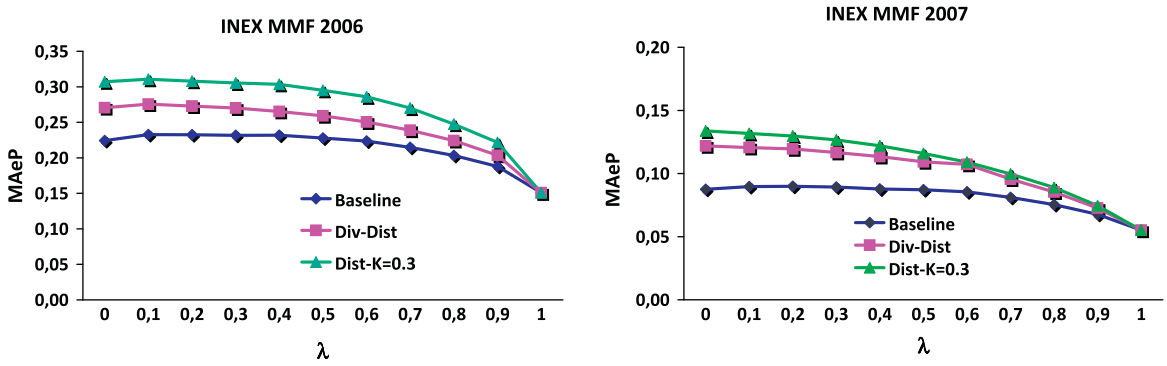


Fig. 8. Comparison between combination with and without the factor $Dist(im_i, Anc|Desc)$.

Table 4
Improvements using Eq. (13) ($Dist - K$) against Eq. (12) ($Div - Dist$) with $\lambda = 0.1$ for Thorough strategy.

Test set	$Dist - K/Div - Dist$
2006	+39%*
2007	+9%*

Best results are obtained with low values of K (between 0.2 and 0.4) for both test sets, which means that the factor *Dist* plays an important role in evaluating the relevance scores of multimedia fragments. To confirm statistically the importance of the factor *Dist*, we compared in Table 3 the results obtained with $K = 0.3$ and $K = 1$, using the Wilcoxon test. We recall that $K = 1$ means that the factor *Dist* is not taken into account in the equation (i.e. it is equivalent to use Eq. (10) with $\lambda = 0$). For both test sets 2006 and 2007, $p < 0.05$, which shows the importance of the distance between the image and its descendants/ancestors in our formula.

We now vary λ and compare results when using the classical combination of Eq. (10) (run *Baseline*), and results when using factor $Dist(me_i, mf)$ according to Eqs. (12) (run $Div - Dist$) and (13) with $K = 0.3$ (run $Dist - K$). Results are showed in Fig. 8. We notice that using *Dist* improves results compared to the classical combination (*Baseline*), and that best results are obtained with Eq. (13): the decay factor K thus allows us to better take into account the distance between elements and fragments. Table 4 details the gains obtained by the $Dist - K$ equation compared to the ones obtained by the $Div - Dist$ equation, when λ equals 0.1 (best combination is obtained when λ is between 0 and 0.2).

For both test sets 2006 and 2007, $Dist - K$ outperforms $Div - Dist$.

4.4. Evaluation of multimedia fragment retrieval according to the Focused strategy

In Focused retrieval strategy (Kamps, Pehcevski, Kazai, Lalmas, & Robertson, 2007; Tsikrika & Westerveld, 2008), overlapping elements are not allowed. The challenge for an information retrieval system is to decide which are the more exhaustive and specific elements of the documents to be returned.

As our approach is evaluated in the INEX context, we decided to not return fragments composed of image descendants in order to respect the focused task definition (Tsikrika & Westerveld, 2008): "*... topics have a clear multimedia character would only judge elements relevant if they contain at least one image*". In the experiments presented in this section a multimedia fragment can be a multimedia element (image) or a multimedia element ancestor.

4.4.1. Evaluation measure

We evaluated our method in the focused strategy using the official measure of INEX (Kamps et al., 2007). The used metric is the interpolated precision at four selected recall level: $iP[jR]$, $j \in [0.00, 0.01, 0.05, 0.1]$. Precision at rank r is defined as follows:

$$P[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)} \tag{16}$$

where p_r is the document part assigned to rank r in the ranked list L_q of document parts returned by a retrieval system for a topic q.

$rsize(p_r)$ is the length of relevant text contained by p_r in characters and $size(p_r)$ is the total number of characters contained by p_r.

Recall at rank r is defined as follow:

$$R[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{Trel(q)} \tag{17}$$

where $Trel(q)$ is the total amount of relevant text for topic q.

The interpolated precision measure $iP[x]$ is as follows:

$$iP[x] = \begin{cases} max_{1 \leqslant r \leqslant |L_q|}(P[r] \wedge R[r] \geqslant x) & \text{if } x \leqslant R[|L_q|] \\ 0 & \text{if } x \succ R[|L_q|] \end{cases} \tag{18}$$

where $R[|L_q|]$ is the recall over all documents retrieval.

The INEX official metric was $iP[0.01]$ (Kamps et al., 2007).

4.4.2. Evaluation of the classical combination between scores

We study in this section the classical combination between the image scores and initial scores of multimedia fragments obtained by the XFIRM system (Eq. (10)). Fig. 9 shows the $iP[0.01]$ evolution against parameter λ.

Using only scores obtained by the XFIRM system ($\lambda = 1$) provides better results in terms of $iP[0.01]$ than those obtained using only images scores or by combining both scores. This shows the limit of using a classical combination, where images contained in the same ancestor participate in its relevance without taking into account its hierarchical proximity. In fact, each image element will contribute with the same relevance score to evaluate the score of its ancestors, and therefore the ancestor having the largest number of images will be top ranked. This leads to always return elements having the highest hierarchical level in the XML document tree (i.e. the *root* element).

4.4.3. Is structure useful to improve results?

Before comparing the interest of both formulas (Eqs. (12) and (13)), and as for the Thorough strategy, we evaluated the effect of parameter K in Eq. (13). Fig. 10 shows results when varying parameter K with $\lambda = 0$, i.e. using only image scores to compute relevance scores of ancestors. These results are listed according to the $iP[0.01]$ metric, with the possibility to return images and/or ancestors.

Best results are obtained with low values of K (K between 0.1 and 0.4 for both test sets), which means that the factor $Dist(em_i, mf)$ has a noticeable influence on the relevance scores of image ancestors. These conclusions are confirmed by the significance tests presented in Table 5.

By varying the parameter K, we also studied the percentage of returned image elements. Fig. 11 presents this percentage in function of K.

We observe that when giving a large importance to the distance between the image element and its ancestors to calculate ancestor scores (K between 0.1 and 0.4), image elements are almost always returned. On the contrary, when giving a low importance to the distance between images and their ancestors (K between 0.6 and 0.9), the percentage of image elements varies between 50% and 90%. If the $Dist(me_i, Anc)$ factor is not used ($K = 1$), image elements are never returned. Considering this curve and the curve of Fig. 10, we note that best results are obtained with low values of K, which means that best results

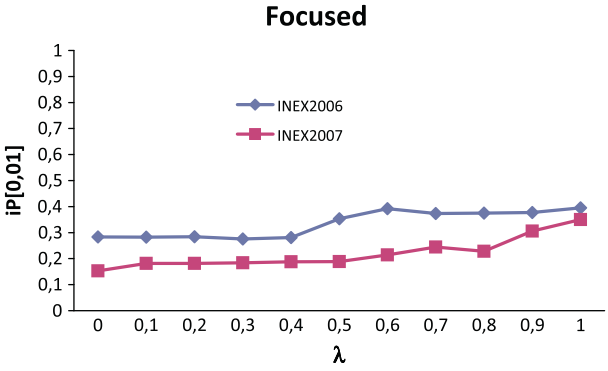


Fig. 9. Evolution of $iP[0.01]$ according to λ for both INEX 2006 and 2007 test sets.

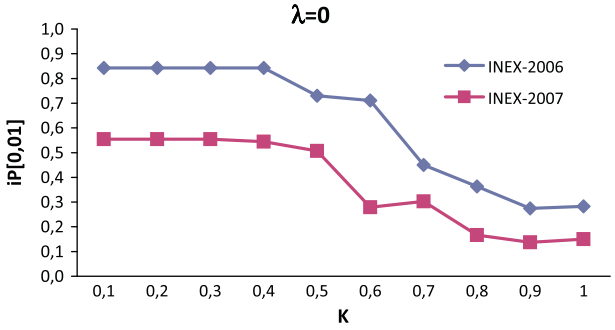


Fig. 10. Impact of the K factor when $\lambda = 0$ with the possibility to return images and/or image ancestors.

Table 5
Improvements using $K = 0.1$ against $K = 1$ when $\lambda = 0$ for Focused strategy.

Test set	$K = 0.1/K = 1$
2006	>100%*
2007	>100%*

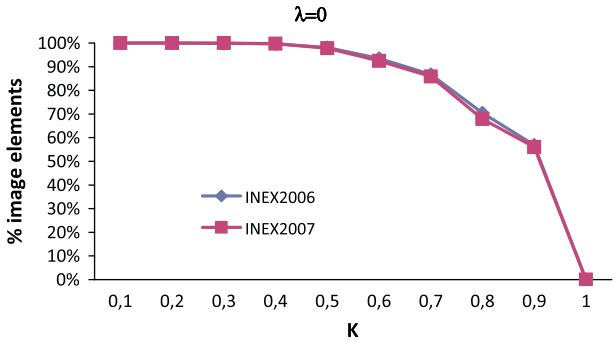


Fig. 11. Percentage of image elements when varying K for both test sets INEX 2006 and 2007.

are obtained by returning only image elements. This observation leads to the following question: is it interesting, in terms of effectiveness, to return images and ancestors? We discuss this question in Section 4.4.4 (see Table 6).

Let us now compare results obtained with the classical combination of Eq. (10) (run *Baseline*) and with the $Dist(me_i, mf)$ factor according to Eq. (12) (run $DIV - Dist$) and Eq. (13) (run $Dist - K = 0.1$). Results are shown in Fig. 12.

The impact of the structure is very clear. Using the factor *Dist* outperforms results comparing to those obtained by the classical combination (*Baseline*) for both test sets (2006 and 2007). Results also clearly show that for both test sets, the λ value seems not to be of importance (results are comparable with λ varying from 0 to 0.8). At last, using Eq. (13) leads to

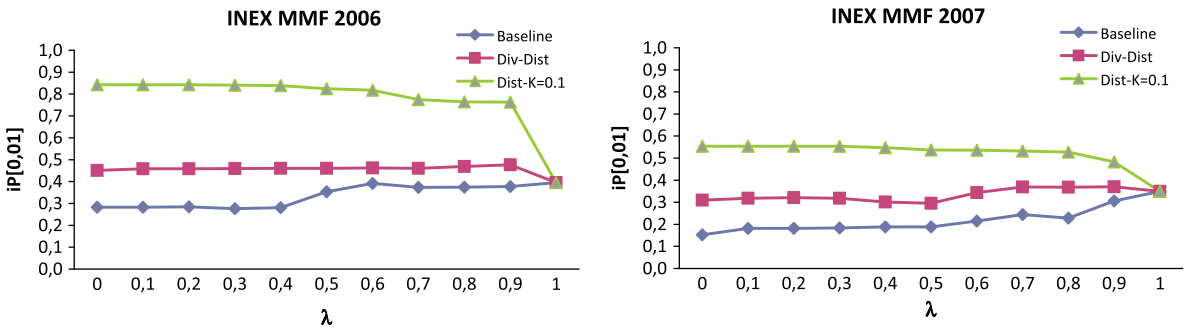


Fig. 12. Impact of the factor $Dist(em_i, Anc)$ according to the $iP[0.01]$ metric.

Table 6
Evaluation of the types of returned elements

	Run	*iP*[0.01]	%images
2006	Images	**0.8428**	100
	Images − Ancestors − Dist − $K = 0.1$	**0.8828**	100
	Ancestors − Dist − $K = 0.1$	0.3394	0
2007	Images	**0.5543**	100
	Images − Ancestors − Dist − $K = 0.1$	**0.5543**	100
	Ancestors − Dist − $K = 0.1$	0.2698	0

Best results are in bold.

obtain the best results (with $K = 0.1$). These improvements are confirmed by a Wilcoxon test ($p < 0.05$ for results using Eqs. (12) and (13), on both test sets).

4.4.4. Effect of multimedia fragment type on search performance

In the *Focused* strategy, returned results should be multimedia fragments composed of either multimedia elements (objects themselves) or their ancestors. Consequently, our approach could return three types of results: (1) images and image ancestors, (2) only image ancestors, or (3) only images.

Table 6 compares results obtained with the three types of multimedia fragments for INEX 2006 and INEX 2007. The comparison is done between:

- results obtained when only images are returned (*Images*),
- best results when images and/or image ancestors are returned (*Images* − *Ancestors* − *Dist* − $K = 0.1$ with $\lambda = 0$), and
- best results when only image ancestors are returned (*Ancestors* − *Dist* − $K = 0.1$ with $\lambda = 0$).

The percentage of returned image elements is also given for each result.

Returning images and/or image ancestors leads in fact to always return image elements (100% of results are image elements for both INEX 2006 et 2007 test sets). This can be explained by the use of low values of K in the Eq. (13). In the case where always image ancestors are returned (*Ancestors* − *Dist* − $K = 0.1$), results significantly decrease.

To better understand this result, we computed the percentage of multimedia fragments which are more relevant than the images they contained in the relevance assessments. For INEX 2006 test set, we found a mean percentage of 34%, while this percentage is 3.48% for INEX 2007.

As a conclusion, returning a text containing images (i.e. returning image ancestors) has shown its interest for some queries only. This means that users prefer in most cases an image with a very small description (images elements can contain small texts) than some text and an image (most of the more relevant elements in the relevance judgments are images).

Of course, an image can be considered as self-explanatory and looking at an image requires less user's effort than reading text. However, we must note that an image can be ambiguous (i.e. can have multiple interpretations for example) which may lead the user to read associated text. It is thus difficult to draw a definitive conclusion, results presented here are strongly dependent on the collection and associated relevance judgments.

5. Comparison of our approach with INEX official approaches

To show the effectiveness of our proposed method, we compared our results to those obtained by official participants to INEX 2006 and 2007. In INEX 2006, the official submissions concern the *Thorough* strategy, and they are classified according to the official measure *MAeP*. In INEX 2007, the official submissions concern the *Focused* strategy, and they are classified according to the official measure *iP*[0.01].

The different values used for the comparison are shown in Table 7. Parameter values are fixed using a 2-fold cross-validation (parameters are learned on the 2006 collection and evaluated on the 2007 one, and vice versa).

5.1. Thorough strategy: INEX MMF 2006

Table 8 ranks the results of all participants of INEX 2006, multimedia task, and our results, using the official relevance assessments. Column *%Text. Frag.* indicates the percentage of elements in runs that do not contain any images.

Table 7
Parameter values used for comparison of our approach with official participants.

Parameter	Best value learned from INEX 2006 (used for evaluation on the 2007 one) – Focused task	Best value learned from INEX 2007 (used for evaluation on the 2006 one) – Thorough task
ϕ in Eq. (1)	Eq. (5) with $w = 0.1$	Eq. (5) with $w = 0.2$
Type of returned elements	Images − Ancestors	Images − Descendants
θ in Eq. (11)	Eq. (13) with $K = 0.1$	Eq. (13) with $K = 0.1$
λ in Eq. (11)	0	0

Table 8
Ranking of our approach compared to INEX 2006 official participants, multimedia task, according to the official relevance assessments and using the *MAeP* metric.

Rank	MAeP	Organisation	Run Id	%Text. Frag.
1	0.1592	Qutau	MMfragmentstitlePSname	54
2	0.1564	Qutau	MMfragmentstitlePS	48
3	0.1544	Qutau	MMfragmentstitle	49
4	0.1536	Qutau	MMfragmentstitleName	54
5	0.1168	Qutau	MMfragmentsCAStitle	27
6	0.1147	Qutau	MMfragmentscastitlePS	37
	0.0744		**Our approach**	**0**
7	0.0656	RMIT	zet-Gift-MMF-Mix-10	38
8	0.0093	RMIT	zet-Gift-MMF-Title-10	47
9	0.0030	Utwente	frag-art-title	0

Best results are in bold.

According to these official relevance assessments, our system would be ranked second after the Queensland University of Technology *Qutau*, but 7th compared to all runs. The system of the university *Qutau* is the GPX system *GPX* (Geva, 2005, 2006). The GPX system rewards elements having the greatest number of unique query terms. Moreover, if an element has only one relevant child, it will be ranked after it, on the other side, if an element has more than one relevant child, it will be ranked before all its descendants. We however note here that at almost half of the results of this approach are purely textual fragments which do not meet the aim of the task.

Let us now compare results using our filtered assessments (only fragments having at least one image are kept in the assessments). To be evaluated on these relevance assessments, official 2006 submissions are also filtered. Results are listed in Table 9. We are aware that making this filtering of officials submissions of INEX participants implies a decrease of the number of returned results, and thus biases the comparison between approaches. Table 9 however gives an idea of such a comparison.

After removing the textual fragments, our approach is now ranked fifth among all submissions. Results presented here are based on parameters fixed on the 2007 set, where relevance assessments are slightly different from the 2006 ones (see discussion in Section 4.4.4). By returning Images, Ancestors and Descendants instead of only returning Images and Descendants we would have had a MAep of 0.3106 for the filtered assessments (and thus we would have been ranked first).

5.2. Focused strategy: INEX MMF 2007

The same analysis was done for the focused strategy. Table 10 ranks the results of different INEX participants and our results using the *iP*[0.01] metric and the official assessments. Columns *%images* and *%Text. Frag.* are respectively the

Table 9
Ranking of our approach compared to INEX 2006 official participants, multimedia task, according to the filtered relevance assessments and using the *MAeP* metric.

Rank	MAeP	Organisation	Run Id	%Removed Elmts.
1	0.2641	Qutau	MMfragmentstitlePSname	54
2	0.2536	Qutau	MMfragmentstitlename	54
3	0.2469	Qutau	MMfragmentstitlePS	48
4	0.2419	Qutau	MMfragmentstitle	49
	0.2412		**Our approach**	0
5	0.2244	Qutau	MMfragmentsCAStitle	27
6	0.2098	Qutau	MMfragmentscastitlePS	37
7	0.1248	RMIT	zet-Gift-MMF-Mix-10	38
8	0.0185	RMIT	zet-Gift-MMF-Title-10	47
9	0.0139	Utwente	frag-art-title	0

Best results are in bold.

Table 10
Ranking of our approach compared to INEX 2007 official participants, multimedia task, according to the official relevance assessments and using the *iP*[0.01] metric.

Rang	*iP*[0.01]	Organisation	Run Id	%images	%Text. Frag.
	0.4821		**Our approach**	**100%**	**0%**
1	0.3389	Utwente	article-MM	0%	0%
2	0.3039	Qutau	CosFocused	31.29%	62.09%
3	0.2947	Qutau	CoFocused	3.21%	87.51%
4	0.2467	Utwente	starloglength-MM	2.55%	4.83%
5	0.0595	Utwente	starlognormal-MM	91.79%	0%

Best results are in bold.

Table 11
Ranking of our approach compared to INEX 2007 official participants, multimedia task, according to the filtered relevance assessments and using the *iP*[0.01] measure.

Rank	*iP*[0.01]	Organisation	Run Id	%images	%Removed Elmts.
	0.5543		**Our approach**	**100%**	
1	0.3171	Utwente	article-MM (Filt)	0%	0%
2	0.2165	Qutau	CosFocused (Filt)	82.74%	62.09%
3	0.2155	Utwente	starloglength-MM (Filt)	2.55%	4.83%
4	0.2003	Qutau	CoFocused (Filt)	25.75%	87.51%
5	0.0465	Utwente	starlognormal-MM (Filt)	91.79%	0%

Best results are in bold.

Table 12
Ranking of our approach compared to the best other systems (Multimedia and Adhoc) participated in the INEX 2007, according to non-filtered relevance assessments and using the *iP*[0.01] metric.

Rang	*iP*[0.01]	Organisation	Run Id	%images	%Text. Frag.
	0.4821		**Our approach**	**100**%	0%
1	0.4435	Mines	EMSE.boolean.-Prox200NF.0010	69.16%	96.39%
2	0.3389	Utwente	article-MM	0%	0%

Best results are in bold.

Table 13
Ranking of our approach compared to the best other systems (Multimedia and Adhoc) participated in the INEX 2007, according to filtered relevance assessments and using the *iP*[0.01] metric.

Rang	*iP*[0.01]	Organisation	Run Id	%images	%Removed Elmts
	0.5543		**Our approach**	**100%**	0%
1	0.4460	Mines	EMSE.boolean.-Prox200NF.0010(Filt)	69.16%	96.39%
2	0.3171	Utwente	article-MM (Filt)	0%	0%

Best results are in bold.

percentage of image elements or descendant elements of an image in the considered run and the percentage of elements that does not contain any images.

As we can see, using the *iP*[0.01] metric, our approach is ranked first, with an improvement of 42% compared to the best official run (*article − MM* of the *Utwente* university). Comparing the official submissions, we note that the best results (*article-MM* run of *Utwente*) are composed of entire documents and not document parts. Conclusions that can be drawn with our approach is the opposite: returning multimedia fragments leads to better results than returning whole documents.

Let us now compare results using the filtered relevance assessments. As for the *Thorough* strategy, official submissions of INEX 2007 multimedia task were filtered to be evaluated on these relevance assessments. Results are showed in Table 11.[12]

Our approach is ranked first with an improvement of 75% compared to the best official filtered run. The same conclusions than using the official relevance assessments can be drawn here.

5.3. Adhoc task versus Multimedia task in focused retrieval

In INEX 2007, Multimedia task queries were part of the Adhoc task query set. For this reason, official runs of adhoc task were evaluated in the Multimedia task framework and the surprising conclusion was that the adhoc runs were better than the Multimedia runs. This suggests that, in the case of a multimedia need, it is better to use an adhoc retrieval system than a multimedia retrieval system.

Using the official relevance assessments, the best adhoc results, according to the *iP*[0.01] metric, were obtained by the *Ecole des Mines de Saint-Etienne* (*EMSE.boolean.Prox200NF.0010*) which uses the proximity between the document terms and the query terms by taking into account the structure of documents (Beigbeder, 2006).

Tables 12 and 13 compare the best official adhoc results on Multimedia task 2007 (*Mines:EMSE.boolean.-Prox200NF.0010*), the best multimedia run (*Utwente:article-MM*) and our approach, using the official and the filtered relevance assessments.

Whatever the considered relevance assessments, our approach is ranked first with an improvement of respectively 9% and 24% over the best official adhoc run.

[12] The fact that the number of returned results is not the same for all runs is not a problem here (contrary to the comparison for INEX 2006), since the used metric (*iP*[0.01]) only consider results at first ranks.

6. Discussion and future work

The aim of this paper is to better study the importance of the text surrounding images for multimedia elements and fragments retrieval. To do so, we used the document structure, that allowed us to evaluate the participation degree of each textual part to the relevance of elements and fragments. Our approach first retrieves multimedia elements, and then uses these multimedia elements to evaluate the relevance of multimedia fragments.

Our results on INEX multimedia track test sets showed that the document structure helps to improve effectiveness in both tasks.

Concerning multimedia elements retrieval, structural information is used to weight the importance of different textual parts in the elements relevance. Thanks to the structure, we assigned more "weight" to textual descendants of the multimedia element than to textual descendants of its brothers and of its ancestors. Although all textual parts are useful, they should not all be taken into account with the same importance degree.

If we now focus on multimedia fragment retrieval, for both strategies *Thorough* and *Focused*, text position is taken into account when evaluating the initial score of a fragment: textual nodes which are far from the fragment less participate to its relevance. We also demonstrated that the use of the distance factor between the multimedia element and each associated multimedia fragments had a high impact on retrieval effectiveness. This means that all the context of a fragment should not be taken into account in the same way: some parts are more important than others, and they can be identified with their underlying structural information.

We can summarize the comparison of our method with the approaches of the official INEX participants as follows:

- Firstly, in almost all approaches, participants proposed to directly retrieve multimedia fragments (using some contextual factors). Our approach is different: we start to retrieve relevant multimedia elements and we then use them to retrieve relevant multimedia fragments. This technique outperforms results of official participants.
- Secondly, INEX 2007 organizers showed that using an adhoc XML retrieval system is better than using a specific multimedia retrieval system even when the user need had a multimedia character. Our experiments showed the contrary: the use of an XML multimedia retrieval system is better than the use of an XML adhoc retrieval system in the case of a multimedia need.

In the future, we aim at studying the effect on elements and fragments retrieval of other factors such as element size and links between elements. Moreover, we plan to study the combination of XML multimedia retrieval with content-based multimedia retrieval approaches.

References

Akbas, E., & Yarman-Vural, F. (2007). Automatic image annotation by ensemble of visual descriptors. In *The international conference on computer vision and pattern recognition, CVPR'07* (pp. 1–8). IEEE Computer Society.

Beigbeder, M. (2006). Structured content-only information retrieval using term proximity and propagation of title terms. In *INEX* (pp. 200–212).

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., & Rajagopalan, S. (1998). Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th international World Wide Web conference* (pp. 65–74).

Chakrabarti, S., Punera, K., & Subramanyam, M. (2002). Accelerated focused crawling through online relevance feedback. In *The eleventh international conference on World Wide Web, WWW'02* (pp. 148–159).

Chen, Z., Liu, W., Zhang, F., Li, M., & Zhang, H. (2001). Web mining for web image retrieval. *Journal of the American Society for Information Science and Technology*, 831–839.

Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML corpus. In *The 29th annual international ACM SIGIR conference on Research and development in information retrieval (Forum), SIGIR'06* (pp. 64–69).

Dunlop, M. (1991). *Multimedia information retrieval*. Thesis diploma, Computing Science Department, University of Glasgow.

Dunlop, M., & Rijsbergen, C. (1993). Hypermedia and free text retrieval. *Information Processing and Management*, 287–298.

Fan, L., & Li, B. (2006). A hybrid model of image retrieval based on ontology technology and probabilistic ranking. In *The 2006 international conference on web intelligence, WI'06* (pp. 477–480).

Advances in XML information retrieval and evaluation, (2006). In N. Fuhr, M. Lalmas, S. Malik, & G. Kazai (Eds.)*4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28–30, 2005, Revised Selected papers*, *Lecture Notes in Computer Science* (Vol. 3977). Springer.

Comparative evaluation of XML information retrieval systems, (2007). In N. Fuhr, M. Lalmas, & A. Trotman (Eds.)*5th International workshop of the initiative for the evaluation of XML retrieval, INEX 2006, Dagstuhl Castle, Germany, December 17–20, 2006, Revised Selected papers*, *Lecture Notes in Computer Science* (Vol. 4518). Springer.

Geva, S. (2005). GPX – gardens point XML IR at INEX 2005. In *INEX'05* (pp. 240–253).

Geva, S. (2006). GPX – gardens point XML IR at INEX 2006. In *INEX'06* (pp. 137–150).

Gong, Z., Hou, H. L., & Cheang, C. W. (2006). Web image indexing by using associated texts. *Knowledge and Information Systems*, 243–264.

Guglielmo, E., & Rowe, N. (1996). Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems, TOIS'96*, 237–267.

Haveliwala, T., Gionis, A., Klein, D., & Indyk, P. (2002). Evaluating strategies for similarity search on the Web. In *The eleventh international World Wide Web conference, WWW'02* (pp. 432–442).

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E., & Milios, E. (2006). Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems*, 55–73.

Iskandar, A., Pehcevski, J., Thom, J., & Tahaghoghi, S. (2006). Social media retrieval using image features and structured text. In *Proceedings of INEX 2006 workshop, Dagstuhl, Germany* (pp. 358–372).

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., & Robertson, S. (2007). INEX 2007 evaluation measures. In *INEX'07* (pp. 24–33).

Kong, Z., & Lalmas, M. (2005). XML multimedia retrieval. In *The 9th international symposium on string processing and information retrieval, SPIRE'05* (pp. 218–223).

Kong, Z., & Lalmas, M. (2007). Using XML logical structure to retrieve (multimedia) objects. In *European conference on digital libraries, ECDL'07* (pp. 100, 111).
Kong, Z., & Lalmas, M. (2007). Combining multiple sources of evidence in XML multimedia documents: An inference network incorporating element language models. In *29th European conference on information retrieval (Poster), ECIR'07* (pp. 716–719).
LaCascia, M., Sethi, S., & Sclaroff, S. (1998). Combining textual and visual cues for content-based image retrieval on the World Wide Web. In *IEEE workshop on content-based access of image and video libraries* (pp. 24–28).
Lalmas, M., Kazai, G., Kamps, J., Pehcevski, J., Piwowarski, B., & Robertson, S. (2006). Inex 2006 evaluation measures. In *Proceedings of INEX 2006 workshop, Dagstuhl, Germany* (pp. 20–34).
Lau, C., Tjondronegoro, D., Zhang, J., Geva, S., & Liu, Y. (2006). Fusing visual and textual retrieval techniques to effectively search large collections of wikipedia images. In *INEX'06* (pp. 345–357).
Mass, Y., & Mandelbrod, M. (2005). Experimenting various user models for XML retrieval. In *INEX'05 proceedings*.
Mihajlovic, V., Ramírez, G., Westerveld, T., Hiemstra, D., Blok, H., & Vries, A. (2005). TIJAH scratches INEX 2005: Vague element selection, image search, overlap, and relevance feedback. In: *Proceedings of INEX 2005, Dagstuhl, Allemagne* (pp. 72–87).
Müller, M., Kurth, F., Damm, D., Fremerey, C., & Clausen, M. (2007). Lyrics-based audio retrieval and multimodal navigation in music collections. In *Proceedings of the 11th European conference on digital libraries, ECDL'07* (pp. 112–123).
Moulin, C., Barat, C., Lemaître, C., Géry, M., Ducottet, C., & Largeron, C. (2010). Combining text/image in wikipediamm task 2009. In *Proceedings of the 10th international conference on cross-language evaluation forum: Multimedia experiments, CLEF'09* (pp. 164–171). Berlin, Heidelberg: Springer Verlag.
Noah, S., Azilawati, A., Sembok, T., & Meriam, S. (2008). Exploiting surrounding text for retrieving web images. *Journal of Computer Science*, 842–846.
Piotrowski, J. (2009). Top-down approach to image similarity measures. In L. Bolc, J. Kulikowski, & K. Wojciechowski (Eds.), *Computer vision and graphics. Lecture notes in computer science* (Vol. 5337, pp. 66–69). Berlin/Heidelberg: Springer.
Popescu, A., Grefenstette, G., & Moëllic, P.-A. (2008). Improving image retrieval using semantic resources. In M. Wallace, M. C. Angelides, & P. Mylonas (Eds.), *Advances in semantic media adaptation and personalization* (pp. 75–96). Berlin/Heidelberg: Springer.
Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(1), 17–30.
Sauvagnat, K. (2005). *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. Phd thesis, Paul Sabatier University, Toulouse, France.
Shen, H., Ooi, B., & Tan, K. (2000). Giving meanings to WWW images. In *The eighth ACM international conference on Multimedia, MULTIMEDIA'00* (pp. 39–47).
Srihari, R., Zhang, Z., Rao, A., Baird, H., & Chen, F. (2000). Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval Journal*, 245–275.
Tjondronegoro, D., Zhang, J., Gu, J., Nguyen, A., & Geva, S. (2006). Integrating text retrieval and image retrieval in XML document searching. In *Fuhr et al.* 2005 (pp. 511–524).
Torjmen, M., Pinel-Sauvagnat, K., & Boughanem, M. (2008). Evaluating the impact of image names in context-based image retrieval. In *Advances in multilingual and multimodal information retrieval, 9th workshop of the cross-language evaluation forum, CLEF'08* (pp. 756–762).
Torjmen, M., Pinel-Sauvagnat, K., & Boughanem, M. (2009). XML Multimedia Retrieval: From relevant textual information to relevant multimedia fragments. In *31th European Conference on Information Retrieval, ECIR'09* (pp. 150–161).
Torjmen, M., Pinel-Sauvagnat, K., & Boughanem, M. (2010). Using textual and structural context for searching multimedia elements. *International Journal of Business Intelligence and Data Mining, Special Issue on Beyond Multimedia and XML Streams Querying and Mining, 5*(4), 323–352.
Tsikrika, T., & Vries, A. (2009). CWI at the photo retrieval task of ImageCLEF 2009. In *Working notes of the 10th workshop of the cross-language evaluation forum, CLEF-campaign*.
Tsikrika, T., & Westerveld, T. (2007). Report on the INEX 2007 multimedia track. In *INEX* (pp. 410–422).
Tsikrika, T., & Westerveld, T. (2008). The inex 2007 multimedia track. In *Focused access to XML documents: 6th international workshop of the initiative for the evaluation of XML retrieval, INEX'07* (pp. 440–453).
Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly, R., Hiemstra, D., & Vries, A. (2007). Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In *INEX* (pp. 273–286).
van Zwol, R., Kazai, G., & Lalmas, M. (2005). INEX 2005 Multimedia Track. In *INEX* (pp. 497–510).
Volkmer, T., & Natsev, A. (2006). Exploring automatic query refinement for text-based video retrieval. In *IEEE international conference on multimedia and expo* (pp. 765–768).
Westerveld, T., & Zwol, R. (2006). The INEX 2006 multimedia track. In: *INEX'06* (pp. 331–344).
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83.
Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on association for computational linguistics, ACL 94* (pp. 133–138).
Zargayouna, H. (2004). Contexte et sTmantique pour une indexation de documents semi-structurTs. In *COnference en Recherche d'Information et Applications* (pp. 571–581).