

Towards a Structure-Based Multimedia Retrieval Model

Mouna Torjmen
IRIT
Paul Sabatier University
Toulouse, France
Mouna.Torjmen@irit.fr

Karen Pinel-Sauvagnat
IRIT
Paul Sabatier University
Toulouse, France
Karen.Sauvagnat@irit.fr

Mohand Boughanem
IRIT
Paul Sabatier University
Toulouse, France
Mohand.Boughanem@irit.fr

ABSTRACT

In this paper, we are interested in multimedia XML document retrieval, whose aim is to find relevant document components (i.e XML elements) that focus on the user needs. We propose to represent multimedia elements using not only textual information, but also hierarchical structure. Indeed, an XML document can be represented as a tree, whose nodes correspond to XML elements. Thanks to this representation, an analogy between XML documents and ontologies can be established. Therefore, to quantify the participation degree of each node in the multimedia element representation, we propose two measures using the ontology hierarchy. Another part of our model consists of defining the best window of multimedia fragments to be returned to the user. Through the evaluation of our model on the INEX 2006 Multimedia Fragments Task, we show the importance of using the document structure in multimedia information retrieval.

Categories and Subject Descriptors

H.3;3 [Information Search and Retrieval]:

General Terms

Theory, Performance, Experimentation

Keywords

semantic similarity measure, contextual multimedia retrieval, XML, hierarchical structure.

1. INTRODUCTION

The main difference between a non-textual media (image, sound, video) and text is that the meaning behind the non-textual media cannot be interpreted in the same way by different users, as it is not defined in a natural language. For example, the figure 1 can be interpreted differently by two users, the first describes it as a butterfly and the second, who has medical knowledge, as a cross section of spinal cord stained with cresyl violet.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR '08, October 30–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-312-9/08/10 ...\$5.00.



Figure 1: Example of an image that can be interpreted differently

Thus, the non-textual media meaning is not easy to represent because it can vary depending on person knowledge and search context. So, each non-textual media has its own way to describe a concept (for example, an image uses low level features as color and texture to describe something), but to communicate its meaning to someone else (without using the image itself), it is necessary to use a language known by everyone as written text or speech. Many studies have been carried out on textual-based multimedia retrieval [6, 40]. Our work follows this research field, using not only textual information but also structural information. Several works showed that logical structure of documents has an important role in providing effective retrieval in semi-structured (XML) textual documents [9, 8, 10]. Our aim here is to study the use of logical structure in multimedia retrieval.

In XML documents, a multimedia element¹ can be described by some textual content. These descriptions (annotations) are found generally in caption elements. Using solely the content of these captions to estimate the image relevance score is not sufficient for the following reasons: (1) captions are very specific in describing the image, whereas queries may tend to be more general. So vocabulary used in captions and queries cannot always be matched; (2) the textual content of the captions is generally small (a few terms), so the probability of matching caption terms to query terms is bound to be very low; (3) captions can be even absent in the XML multimedia document.

The best way to solve these problems is to use the textual content of elements that surround the multimedia element in the logical structure of the document in order to represent this multimedia element.

In this paper, we will use the analogy between XML documents and ontologies in order to improve multimedia re-

¹In this paper, the work described was carried out with images. Nonetheless, the approach can be applied to any other media.

trieval in XML documents. Indeed, an XML document can be represented as a tree, whose nodes correspond to XML elements. Thanks to this representation, we can consider the XML document tree as a simple ontology whose concepts are elements organized with the relationship "IsPartOf". Using a semantic similarity measure between concepts, we propose a measure to quantify the participation of each document node in the multimedia element representation.

The rest of this paper is organized as follows. Section 2 reviews the literature on multimedia retrieval in semi-structured documents and on semantic resources. Section 3 is a detailed presentation of our approach with some examples. An analogy between XML documents and ontologies is given in the first part of the section. Then, our measure to evaluate the degree of each node in the multimedia element representation is presented in the second part. After that, we present how our approach can be applied in two ways. Finally, we propose to use a window to specify which multimedia fragments must be returned to the user. Section 4 offers a detailed description and discussion of our experiments and results. Finally, section 5 concludes with possible directions for future work.

2. RELATED WORK AND BACKGROUND

2.1 Multimedia (image) retrieval in semi structured documents

The overload of multimedia documents and its heterogeneity including collections of photos, music and videos has been phenomenal in recent years and brought a renewed spurt of the research activity in multimedia information retrieval. Our approach can be applied on any media, but as we will use a collection containing images to evaluate it, we talk thus in this section about image retrieval.

Currently there are three approaches in image retrieval: content-based (CBIR), context-based and combination of the both. Content-based approach uses visual features for searching similar images to a query image [1]. The main limitation of these techniques is the well-known semantic gap between low-level visual features and the corresponding high level concepts [32]. Thus, visual features are generally not sufficient for searching similar images. A state of the art and new challenges of CBIR systems are presented in [21].

The main element used in context-based image retrieval techniques is the text, but other elements can be used such as hyperlinks. Most of the image search engines belongs to this kind of approach (For example Google Image).

The third type of approaches consists of combining the above methods in order to improve the performance of image retrieval systems [14].

We are mainly interested in the second approach where we use two context elements: text and structure. The integration of CBIR techniques in our approach will be the subject of future work. We present below some related works which relies on using structure in image retrieval.

Dividing documents into parts is already used in web image retrieval. For example, in [11] textual content of documents is divided into image caption, neighbouring image captions, the rest of text in the page, and the text in the pages pointing to that page.

Until 2005, when the INEX² evaluation campaign intro-

²INEX: Initiative for the Evaluation of XML Retrieval.

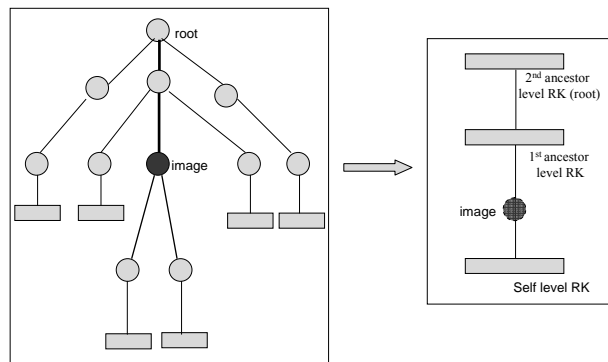


Figure 2: Dividing XML document into knowledge regions using only relationship between a multimedia element and root element

duced a new task called Multimedia Task [36], only a few studies were interested in multimedia retrieval in semi structured documents.

Some XML-related approaches [15, 34, 35] used a linear combination of evidences to merge the retrieval results of content-based image retrieval and text-based XML retrieval, but no definitive conclusions were done regarding the efficiency of the structure in multimedia retrieval. In addition, both textual and image queries were needed.

The approach proposed in [17, 18] was the first attempt to exploit document structure into multimedia retrieval. Textual content is divided into *Region Knowledge*³ (*RKs*). Self level *RK*: *RK* of the multimedia element; sibling level *RK*: *RK* of the sibling elements of the multimedia node; 1st ancestor level *RK*: *RK* of the first ancestor of the multimedia element excluding nodes having been already used; 2nd ancestor level *RK*; ...; N^{th} ancestor level *RK*.

The limitation of this approach is that document structure is used partially as we can see in figure 2: only the relationship between the multimedia element and the root element is used. Indeed, textual content of elements belonging to the same region *RK* are merged together and used as a single element.

Another approach in [33] makes use of the semantic structure and logical structure in XML documents on one hand, and their combination for representing and retrieving XML multimedia document content⁴ on the other hand. The logical structure is exploited by a bayesian network incorporating element based language models for the retrieval of a mixture of text and image. This approach is evaluated with a small collection (Lonely Planet of INEX Multimedia 2005) and showed its efficiency, but it must be evaluated within a larger collection (as the Wikipedia collection of INEX Multimedia Fragment task 2006-2007) to be validated.

The following section presents the benefits of using semantic resources in Information Retrieval (IR), since our model is based on the ontology hierarchy.

2.2 Effectiveness of semantic resources in IR

It is widely assumed that using semantic resources such

³The textual content of the multimedia object and elements hierarchically surrounding it.

⁴The multimedia content refers to any type of multimedia data or a mixture of this latter and text.

as ontologies and concept taxonomies improves IR performances [30].

With the overload of available information and heterogeneity of data, it is insufficient to find relevant information using only the document vocabulary. Keyword-based search methods suffer from several general limitations: a keyword in a document does not necessarily mean that the document is relevant, and relevant documents may not contain the explicit word. Synonyms lower recall rate, homonyms lower precision rate, and semantic relations such as hyponymy, meronymy, antonymy [7] are not exploited [13]. For these reasons, many works exploiting external semantic resources have been proposed in the literature.

Ontology is a formal and explicit representation of knowledge which consists of concept lexicon, concept properties, and relations among concepts [3].

In [4], authors proved the efficiency of ontologies in IR, particularly in query expansion strategies. Another work [39] proposed to integrate the semantic neighbourhood in indexing XML documents by connecting document terms with ontology concepts.

Ontologies are also used in generating automatic summaries, word sense disambiguation (*WSD*) [2], query relevance feedback, multilingual information retrieval, semantic web, image retrieval [12, 26] and video retrieval.

Many semantic measures on ontology 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 [31] classify semantic measures into three categories:

- **Edge Counting 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 [22, 27, 38].
- **Information Content Methods:** 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 [24, 28, 16].
- **Hybrid methods:** combine the above ideas [29, 20].

In our work, we are interested in the first type of similarity measure as we will consider a multimedia element as a concept. Therefore, no content can be used.

In this field, the measure of Wu-Palmer [38] has the advantage of being simple to implement and have good performances compared to the other similarity measures [23].

The similarity measure of Wu-Palmer, based on the edge counting method, is defined by the following expression:

$$Sim_{WP}(C1, C2) = \frac{2 * N3}{(N1 + N2 + 2 * N3)} \quad (1)$$

where $N1$ and $N2$ are the distance which separate concepts $C1$ and $C2$ from the root R , and $N3$ the distance that separates the closest common ancestor (CS) of $C1$ and $C2$ from the root R .

Our assumption for improving the accuracy of multimedia retrieval in semi-structured documents is based on using the notion of ontology for representing the XML document, so

Table 1: Analogy between an XML document and a simple ontology

Simple ontology	XML document tree representation
Ontology concepts	XML document nodes
"IsPartOf" relationship between concepts	Structural relationship between nodes
Similarity measure between two concepts :c1 et c2 $sim(c1,c2)$	Measure of participation of a non-multimedia node N in the multimedia node I representation $rep(I,N)$

as to let nodes of the document participate in the multimedia element relevance estimation. The participation degree of these nodes in the multimedia element representation is inspired from the Wu-Palmer measure.

3. A SEMANTIC REPRESENTATION OF MULTIMEDIA NODES USING TEXTUAL AND STRUCTURAL CONTEXT

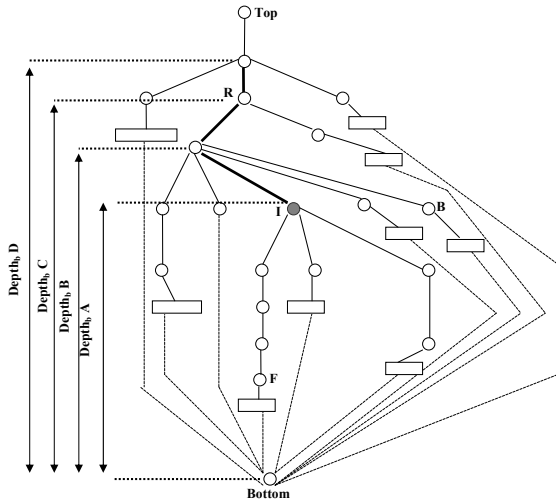
3.1 Analogy between XML documents and ontologies

The main intuition of our approach is that each textual node holds semantic information allowing to represent each multimedia node. Hence, each node must participate to semantically represent the multimedia element. The challenge is how to quantify the participation of each textual node in the semantic representation of the multimedia element?

Our main intuition is the following: we believe that nodes occurring close to the multimedia element (on level and on edges counting), contain more relevant information about it than ones on higher levels. It seems therefore that the higher the level is, the less nodes participate in the multimedia element score. Thus, descendants of the multimedia element must participate more than descendants of the first ancestor, the latter must participate more than descendants of the second ancestor, and so on...Root descendants (excluding nodes already used) have the lowest participation degree in the multimedia element participation.

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). 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 calculate how much each non-multimedia node in the document should participate in the representation of the multimedia element. We consider the multimedia node as a concept $C1$, and each other node of the document as another concept $C2$. Table 1 shows the analogy between an XML document and a simple ontology.



Depth_A : depth_b of the closest common ancestor between I and its descendants Des => depth_b(I,Des)
Depth_B : depth_b of the closest common ancestor between I and its first ancestor descendants (including the first ancestor itself) =>depth_b(I,1st ancestor).
Depth_C : depth_b of the closest common ancestor between I and its second ancestor descendants (including the second ancestor itself) =>depth_b(I,2nd ancestor).
Depth_D : depth_b of the closest common ancestor between I and its nth ancestor descendants (including the nth ancestor itself) =>depth_b(I,nth ancestor).

Figure 3: Relationships between nodes (concepts) and bottom

3.2 A multimedia element representation measure

In [39], the Wu-Palmer measure is used in a semantic indexing scheme for XML documents. Nevertheless, this measure is not very suitable for the proposed approach because brothers can be ranked before descendants of a concept whereas authors of [39] want to select all descendants of a concept before its brothers.

For example, let us consider the document tree in Figure3. We calculate the similarity between *B* (then *F*) and *I*.

$$Sim_{WP}(I, B) = \frac{2 * 3}{(1 + 1 + 2 * 3)} = 0.75 \quad (2)$$

$$Sim_{WP}(I, F) = \frac{2 * 4}{(4 + 0 + 2 * 4)} = 0.66 \quad (3)$$

The brother *B* of node *I* is returned before the descendant *F*.

To avoid this, authors proposed to penalize brothers by adding a function $spec(C1, C2)$ to the Wu-Palmer measure, which calculates the specificity of the two concepts (*C1* and *C2*) comparing to the lowest concept (*bottom*) (Figure 3).

$$spec(C1, C2) = depth_b(CS) * distance(CS, C_1) \quad (4)$$

$$* distance(CS, C_2)$$

when *CS* is the closest common ancestor of *C1* and *C2*. $depth_b(CS)$ is the maximum number of edges between *CS* and *bottom*. $distance(CS, C_i)$ is the distance (number of edges) between *CS* and *C_i*. The similarity measure becomes:

$$Sim_{WPSpec}(C1, C2) = \frac{2 * N3}{(N1 + N2 + 2 * N3 + spec(C1, C2))} \quad (5)$$

Using formula 5, the node *F* is selected before the node *B*:

$$Sim_{WPSpec}(I, B) = \frac{2 * 3}{(1 + 1 + 2 * 3 + 7 * 1 * 1)} = 0.4 \quad (6)$$

$$Sim_{WPSpec}(I, F) = \frac{2 * 4}{(4 + 0 + 2 * 4 + 6 * 4 * 0)} = 0.66 \quad (7)$$

Remember that in our approach we want to select concept descendants before brothers (including their descendants) and finally ancestors (including their descendants). The new formula partially satisfy our needs as an ancestor can be selected before a brother. To illustrate this problem, let us calculate the similarity between *I* and *R*:

$$Sim_{WPSpec}(I, R) = \frac{2 * 2}{(2 + 0 + 2 * 2 + 8 * 2 * 0)} = 0.66 \quad (8)$$

The ancestor *R* has a similarity score higher than the brother *B*.

To solve this problem, we propose to penalize ancestors by multiplying the denominator by $(depth_b)^{\beta}(CS)$ with $\beta > 1$.

$$Sim_{WPI_m}(C1, C2) = \quad (9)$$

$$\frac{2 * N3}{N1 + N2 + 2 * N3 + spec(C1, C2) + (depth_b)^{\beta}(CS)}$$

The use of this new factor penalizes ancestors compared to descendants and brothers. Indeed, the higher the level is, the lowest the score of concepts is: nodes of the first ancestor are selected before nodes of the second ancestor...nodes of the nth-1 ancestor are selected before nodes of the nth ancestor.

In our example, using the new formula with $\beta=2$, we obtain:

$$Sim_{WPI_m}(I, B) = \frac{2 * 3}{1 + 1 + 2 * 3 + 7 * 1 * 1 + 7^2} = 0.093 \quad (10)$$

$$Sim_{WPI_m}(I, R) = \frac{2 * 2}{2 + 0 + 2 * 2 + 8 * 2 * 0 + 8^2} = 0.052 \quad (11)$$

$$Sim_{WPI_m}(I, F) = \frac{2 * 4}{4 + 0 + 2 * 4 + 6 * 4 * 0 + 6^2} = 0.166 \quad (12)$$

So the descendant *F* is selected before the brother *B*, and this latter is selected before the ancestor *R*.

In our work, this measure is used to calculate the participation of each relevant node in the multimedia element representation. We define the multimedia element representation measure as follows:

$$Rep(I, N) = \quad (13)$$

$$\frac{2 * N3}{N1 + N2 + 2 * N3 + spec(I, N) + depth_b^{\beta}(CS)} * S_N$$

where *I* is an image, *N* is a relevant node which participate in the image representation, *CS* is the closest common ancestor and S_N the score of the node *N*, already calculated using a classical structured information retrieval system.

The final score of each image is calculated as follows:

The evaluation measure used in Multimedia track 2006 is "effort-precision/gain-recall". Effort-precision (ep) is calculated, at a given cumulated gain value (r), as follows:

$$ep[r] = \frac{i_{ideal}}{i_{run}} \quad (15)$$

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:

$$gr[i] = \frac{\sum_{j=1}^i specS(e_j)}{\sum_{j=1}^n specI(e_j)} \quad (16)$$

where i is the i -th 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 j -th element in the system ranking and $specI(e_j)$ is the specificity of the j -th element in the ideal ranking.

The non-interpolated mean average effort-precision, denoted $MAep$, is calculated by averaging the ep values obtained for each rank where a relevant document is returned. More details for this measure can be found in [19].

4.2 Results and discussion

In this paper, we evaluated our approach using textual nodes to represent image elements. Textual nodes scores (S_N in equation 13) are evaluated using a classical XML search engine [25]. The weights of terms are computed using the $tf*idf*idf$ formula that takes into account both the importance of terms in the whole collection of documents and in the collection of elements. More details about the approach can be found in [25].

The first experiments we conducted on our method concern the value of β used in equation 13. Figure 6 shows the effect of this parameter on the $MAep$ measure where the multimedia fragment window is Win_{Im} . We recall that β is used to penalize high ancestor nodes from lowest ancestor nodes: nodes of the first ancestor participate more than nodes of the second ancestor,...

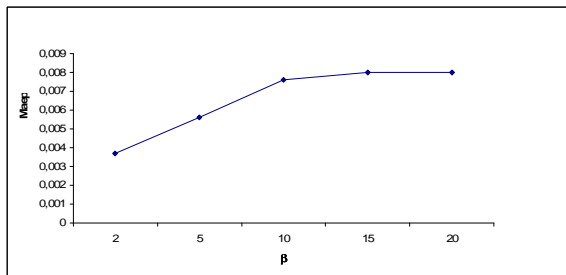


Figure 6: Effect of the β parameter in Equation13

We notice that $MAep$ increases when β increases in the range of values [2..15] interval. Beyond the point 15, a stability of performance is observed. We can conclude that giving a high importance to the penalization parameter of ancestors descendants improves results. Best results are obtained with $\beta \geq 15$. It means that $depth_b^\beta(CS)$ in Equation 13 is a very important factor: indeed, the other parameters are in this case almost negligible.

For the previous reasons, we proposed to evaluate the effect of $depth_b^\beta(CS)$ individually excluding the other factors. We thus defined the following formula:

$$Rep^{depth}(I, N) = \frac{1}{depth_b^\beta(CS)} * S_N \quad (17)$$

Figure 7 shows our results when varying the β factor in Equation 17.

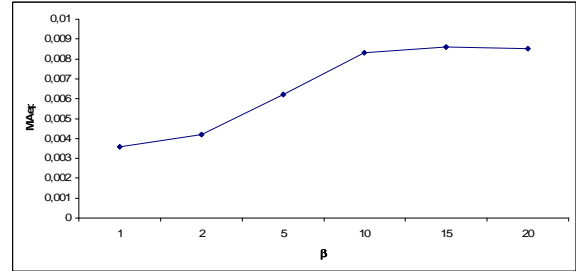


Figure 7: Effect of the β parameter in Equation17

As we can see, increasing β leads to increasing $MAep$, and as for Equation 13 a stability step is reached for $\beta = 15$. These results confirm our intuition that descendant nodes must participate more than 1^{st} ancestor descendants, these latter must participate more than 2^{nd} ancestor descendants...

We then compared our metrics to the original Wu-Palmer metric and its adaptation presented in [39]. Figure 8 shows the results we obtained with Sim_{WP} (Equation 1), Sim_{WPspec} (Equation 5), Sim_{WPIIm} (Equation 13) with $\beta = 15$ and $Rep^{depth}(I, N)$ (Equation 17) with $\beta = 15$. The window of multimedia fragments is fixed to include only multimedia elements (Win_{Im}).

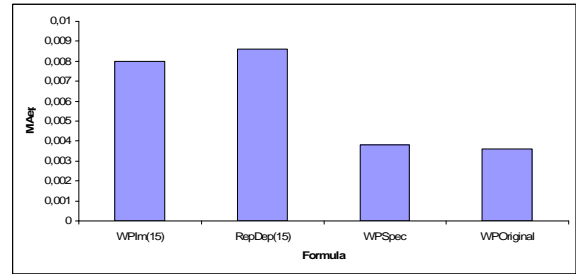


Figure 8: Comparison of $WPIIm(\beta = 15)$, $Rep^{depth}(I, N)(\beta = 15)$, $WPSpec$ and $WPOriginal$

First, we can confirm that our two proposed measures are better than $WPSpec$ and $WPOriginal$. The high performance of $WPIIm$ and $Rep^{depth}(I, N)$ can be explained by the $depth_b^\beta$ parameter. Secondly however, we cannot conclude on the best formula between Equation 13 and Equation 17 as they allow to obtain very similar results. Consequently, both equations are used in the rest of our experiments concerning the optimal multimedia fragments window.

As described above, the definition of the window is based on the hierarchical relationship between the multimedia element and its ancestors. In our experimental study, we evaluated 10 multimedia fragments windows as the document

average depth of the used collection is 7 : Win_{Im} = multimedia elements,

$Win_{Im-Desc} = \{\text{multimedia elements} + \text{descendants}\}$,

$Win_{Im-Desc-Asc} = \{\text{multimedia elements} + \text{descendants} + \text{ascendants}\}$,

$Win_{Im-Desc-Asc-Anc1} = \{\text{multimedia elements} + \text{descendants} + \text{ascendants} + 1^{st} \text{ ancestor descendants}\}$,

$Win_{Im-Desc-Asc-Anc2} = \{\text{multimedia elements} + \text{descendants} + \text{ascendants} + 2^{nd} \text{ ancestor descendants}\}$, ...,

$Win_{Im-Desc-Asc-Anc7} = \{\text{multimedia elements} + \text{descendants} + \text{ascendants} + 7^{th} \text{ ancestor descendants}\}$. To evaluate these 10 windows, we fixed β to its best value ($\beta = 15$) in both equations. The score of each element in the window is set to the score of the multimedia element evaluated with Equation 13 or 17. Figure 9 shows the performance results obtained for each of the windows for the two equations.

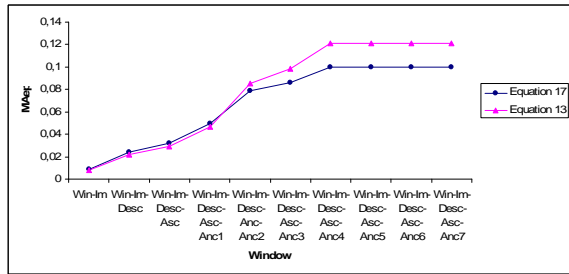


Figure 9: Comparison of the multimedia fragment window effect using Equation 13 and Equation 17

Results show that biggest windows allow to obtain better results. The best window is $Win_{Im-Desc-Asc-Anc5}$ for which returned elements are multimedia elements, their descendants, their ascendants, descendants of the 1st ancestor, ..., descendants of the 5th ancestor. We can conclude that a relevant multimedia fragment for a user is an XML tree composed of a multimedia element and other non-multimedia elements enclosed by the same node.

Another conclusion that can be drawn is that Equation 17 gives best results than Equation 13 using $Win_{Im-Desc-Asc-Anci}$ with $i \geq 2$ although it was the inverse using Win_{Im} . This can be explained by the fact that using only the $depth_b^\beta$ parameter gives the same importance to all the descendants of the same ancestor. For example, all descendants of the multimedia element will participate in the same way to its score. These results confirm that structure must be exploited not only between the multimedia element and its ascendants, but also between the different elements having the same ancestor.

We also compared our structure-based multimedia retrieval model to a classical XML information retrieval model [25].

Our model, using $\beta = 15$, Equation 13 and $Win_{Im-Desc-Asc-Anc-5}$, gives the best results (MAep=0.1214) comparatively to the search engine with optimal parameters (MAep=0.046). Thus in multimedia information retrieval, we can assume that identifying the relevant multimedia element in a first step, and defining the set of related non-multimedia elements in a second step, achieves better performance comparatively to the evaluation of XML elements without any multimedia specification. The last conclusion we can have is that the multimedia retrieval performance

varies depending on two aspects: the degree of penalization between ancestors descendants participating in the multimedia element representation, and the content of the multimedia fragment window. These two aspects are related to the structure hierarchy of XML documents.

Thanks to these experiments, our organization should have been ranked second in the official INEX 2006 Multimedia Fragment Task (best MAep = 0.1591). Experimentations showed the efficiency of our method but further evaluation is however needed. A scoring method need to be found for elements in the window ; since here a vey naive solution is used: all elements have the same score than the multimedia element used for the window. Moreover, multimedia fragment task 2006 provided only 9 topics. We plan to evaluate our method with the same task in 2007 where more topics are provided (20 topics) and to do significance tests on the latter set of topics (9 topics in 2006 are not enough to do significance tests).

5. CONCLUSIONS AND FUTURE WORK

In this paper, we exploited document structure in two ways: to retrieve relevant multimedia element in a first step, and to define a relevant multimedia fragment window of returned fragments in a second step.

We proposed a novel approach which consists of representing the multimedia element by other nodes containing relevant information. This approach investigates the use of textual information and semantic hierarchical relationship between XML elements. An extended measure of Wu-Palmer is defined to calculate the participation degree of each non-multimedia node in the representation. Results showed that textual information of the elements must contribute differently, and returning multimedia fragments based on hierarchical relationships between elements and the multimedia element allows to obtain good results. Therefore, XML document structure has an important role in multimedia retrieval.

In the future, we aim at evaluating our model in INEX 2007. Further studies are also needed in the definition of the multimedia fragment window.

Moreover, we plan to study the effect of using other similarity measures in the the multimedia element representation and to study other weighting scheme for some important terms (like for example terms in *caption* elements).

6. REFERENCES

- [1] Y. A. Aslandogan and C. T. Yu. Techniques and systems for image and video retrieval. *IEEE Trans. on Knowl. and Data Eng.*, pages 56–63, 1999.
- [2] M. Baziz, M. Boughanem, and N. Aussenac-Gilles. A Conceptual Indexing Approach based on Document Content Representation . In *CoLIS5: Glasgow, UK*, pages 171–186, June 2005.
- [3] M. Bertini, A. D. Bimbo, and C. Torniai. Automatic video annotation using ontologies extended with visual information. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 395–398, 2005.
- [4] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing & Management*, pages 866–886, July 2007.

- [5] L. Denoyer and P. Gallinari. The wikipedia xml corpus. *SIGIR Forum*, pages 64–69, June 2006.
- [6] H. Elghazel, K. Idrissi, A. Baskurt, and C. Ben Amar. Approche textuelle pour la recherche d’image. In *SETIT 2005*, Mar. 2005.
- [7] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.
- [8] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai. INEX workshop proceedings, Dagstuhl, Germany, 2005.
- [9] N. Fuhr, M. Lalmas, S. Malik, and Z. Szlavik. INEX workshop proceedings, Dagstuhl, Germany, 2004.
- [10] N. Fuhr, M. Lalmas, and A. Trotman. INEX workshop proceedings, Dagstuhl, Germany, 2006.
- [11] V. Harmandas, M. Sanderson, and M. D. Dunlop. Image retrieval by hypertext links. In *Proceedings of SIGIR-97*, pages 296–303, Philadelphia, US, 1997.
- [12] M. Hwang, H. Kong, S. Baek, and P. Kim. A method for processing the natural language query in ontology-based image retrieval system. In *Adaptive Multimedia Retrieval. AMR*, pages 1–11, 2006.
- [13] E. Hyvnen, S. Saarela, A. Styrman, and K. Viljanen. Ontology-based image retrieval. In *WWW (Posters)*, 2003.
- [14] D. A. Iskandar, J. Pehcevski, J. A. Thom, and S. M. Tahaghoghi. Combining image and structured text retrieval. In *Proceedings of INEX, Dagstuhl, Germany*, pages 365–372, 2005.
- [15] D. N. F. A. Iskandar, J. Pehcevski, J. A. Thom, and S. M. M. Tahaghoghi. Combining image and structured text retrieval. In *Proceedings of INEX Workshop, Dagstuhl, Germany*, pages 525–539, 2005.
- [16] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy, 1997.
- [17] Z. Kong and M. Lalmas. Xml multimedia retrieval. In *SPIRE*, pages 218–223, 2005.
- [18] Z. Kong and M. Lalmas. Using xml logical structure to retrieve (multimedia) objects. In *ECDL*, pages 100–111, 2007.
- [19] M. Lalmas, G. Kazai, J. Kamps, J. Pehcevski, B. Piwowarski, and S. Robertson. Inex 2006 evaluation measures. In *Proceedings of INEX Workshop, Dagstuhl, Germany*, pages 20–34, 2006.
- [20] C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. 1998.
- [21] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2:1–19, February 2006.
- [22] Y. Li, Z. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.*, pages 871–882, 2003.
- [23] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.
- [24] G. A. Miller and W. G. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, pages 1–28, 1991.
- [25] K. Pinel-Sauvagnat. *Flexible Information Retrieval model for semi-structured document collections*. PhD thesis, Paul Sabatier University, Toulouse, France, june 2005.
- [26] A. Popescu, G. Grefenstette, and P.-A. Moellic. Using semantic commonsense resources in image retrieval. In *SMAP ’06*, pages 31–36, Washington, DC, USA, 2006.
- [27] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, pages 17–30, 1989.
- [28] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, pages 95–130, 1999.
- [29] M. A. Rodríguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Trans. on Knowl. and Data Eng.*, pages 442–456, 2003.
- [30] P. Rosso, E. Ferretti, D. Jiménez, and V. Vidal. Proceedings of the Second International WordNet Conference—GWC 2004. In *Text Categorization and Information Retrieval Using WordNet Senses*, pages 299–304. Czech Republic, 2003.
- [31] T. Slimani, B. B. Yaghlane, and K. Mellouli. A new similarity measure based on edge counting. In *Proceedings of world academy of science, engineering and technology*, December 2006.
- [32] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 1349–1380, 2000.
- [33] Z. Szlávik, A. Tombros, and M. Lalmas. Feature- and query-based table of contents generation for xml documents. In *ECIR*, pages 456–467, 2007.
- [34] D. Tjondronegoro, J. Zhang, J. Gu, A. Nguyen, and S. Geva. Integrating text retrieval and image retrieval in xml document searching. In *Proceedings of INEX Workshop, Dagstuhl, Germany*, pages 511–524, 2005.
- [35] R. van Zwol. Multimedia strategies for 3-sdr, based on principal component analysis. In *Proceedings of INEX Workshop, Dagstuhl, Germany*, pages 540–553, 2005.
- [36] R. van Zwol, G. Kazai, and M. Lalmas. Inex 2005 multimedia track. In *Proceedings of INEX Workshop, Dagstuhl, Germany*, pages 497–510, 2005.
- [37] T. Westerveld and R. van Zwol. The inex 2006 multimedia track. In *Proceedings of INEX Workshop, Dagstuhl, Germany*, pages 331–344, 2006.
- [38] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.
- [39] H. Zargayouna. *Indexation sémantique de documents XML*. Phdthesis, Ecole Doctorale d’Informatique de Paris Sud, 2005.
- [40] C. Zhang, J. Y. Chai, and R. Jin. User term feedback in interactive text-based image retrieval. In *SIGIR ’05*, pages 51–58, 2005.