Une étude de l'impact de la structure sur la recherche multimedia

Mouna Torjmen, Karen Pinel-Sauvagnat

IRIT, Université Paul Sabatier
118, Route de Narbonne, 31400, Toulouse

RÉSUMÉ. *Cet article s'inscrit dans le cadre de la recherche XML multimedia, dont l'objectif est de trouver des fragments multimedia pertinents (c'est à dire des fragments XML contenant au moins un autre media que le texte). Dans des travaux précédents, nous avons proposé un modèle pour la recherche de fragments multimedia appliqué au media "image". Ce modèle consiste tout d'abord à trouver les images pertinentes et ensuite, à définir les fragments multimedia pertinents à partir de ces images. Dans cet article, nous nous intéressons plus particulièrement à la première partie du modèle où nous étudions l'impact de différents facteurs structurels pour la recherche d'images. Cette étude comparative est effectuée à travers une approche basée sur une analogie entre un document XML et une ontologie. Les facteurs sont évalués dans le cadre de la tâche Multimedia de campagne d'évaluation INEX 2007, et montrent l'intérêt de l'utilisation de la structure dans le processus de recherche multimedia.*

ABSTRACT. *In this paper, we are interested in XML multimedia retrieval, whose aim is to find relevant multimedia components(i.e XML fragments containing at least another media than text). The work described here is carried out with images, but can be extended to any other media. We proposed in previous work a multimedia fragment retrieval model which consists to retrieve in a first step relevant images and in a second step the best multimedia fragments through the retrieved images. In this paper, we are interested in the first step of our model. In fact, we study the impact of different structural factors on image retrieval. This comparative study is carried out through an approach based on an analogy between XML documents and ontologies. These factors are evaluated in the Multimedia Task of INEX 2007 and show the efficiency of using document structure in multimedia retrieval process.*

MOTS-CLÉS : *recherche d'information, fragment multimedia, texte, structure, image*

KEYWORDS: *information retrieval, multimedia fragment, text, structure, image*

1. Introduction

La recherche d'information dans des documents XML consiste à retrouver des fragments pertinents, c'est à dire des passages ou des éléments XML contenant des informations pertinentes. Bien que le média "*texte*" reste une composante dominante dans la majorité des documents XML, d'autres types de médias peuvent également être présents dans ces documents. Les études existantes concernant la recherche d'information multimédia ont montré qu'elle est loin d'être triviale dans le cas où l'utilisateur cherche une combinaison de médias (par exemple texte et image).

Dans cet article, le travail présenté est appliqué sur le média *"image"*. Toutefois, l'approche proposée peut être utilisée sur n'importe quel autre type de média.

La plupart des travaux existants dans le domaine de recherche d'images sont basés généralement soit sur le contenu textuel des documents contenant les images (Elghazel *et al.,* 2005) (Zhang *et al.,* 2005), soit sur les caractéristiques de bas niveau des images telles que la couleur et la texture (Lew *et al.,* 2006) (on parle alors de recherche d'images basée contenu -CBIR-). D'autres travaux proposent de combiner les deux pour utiliser les avantages de chaque approche (Iskandar *et al.,* 2005) (Tollari *et al.,* 2008). D'autres sources dévidence ont été récemment envisagées. Parmi elles on peut citer l'utilisation de ressources sémantiques comme les ontologies ou encore d'autres facteurs extraits des documents tels que les hyperliens. Dans cet article, nous proposons d'étudier l'impact de la structure sur la recherche d'images.

La principale différence entre la recherche XML adhoc et la recherche XML multimedia concerne les éléments retournés qui sont respectivement des fragments textuels et des fragments multimedia. Un fragment multimedia doit possèder un caractère multimedia, c'est à dire que les éléments retournés doivent être des éléments multimedia ou bien contenant au moins un élément multimedia (Tsikrika *et al.,* 2007b).

La plupart des techniques de recherche XML multimedia ne prennent pas en compte la spécificité multimedia d'une façon explicite : soit elles combinent les résultats XML adhoc avec les résultats de recherche d'images basée contenu (Iskandar *et al.,* 2005) (Tjondronegoro *et al.,* 2005) (van Zwol, 2005), soit elles filtrent les résultats XML adhoc en gardant ceux qui répondent au besoin multimédia (Tsikrika *et al.,* 2007a).

Dans (Torjmen *et al.,* 2008a), nous avons proposé un modèle qui prend en compte le caractère multimedia dans la recherche des fragments XML. Il consiste tout d'abord à rechercher les images pertinentes, et ensuite, à les utiliser pour trouver les bons fragments multimedia.

Dans cet article, nous évaluons et discutons plusieurs paramètres permettant de déterminer la pertinence des images. Nous abordons aussi quelques problématiques liées à l'évaluation des fragments multimedia et comment nous les avons résolues.

La suite de l'article est organisée comme suit : la section 2 présente un état de l'art sur la recherche de fragments XML multimedia. Dans la section 3, nous décri-

vons notre modèle, en se focalisant sur la première partie pour laquelle nous étudions l'impact de plusieurs facteurs sur l'évaluation de la pertinence des images. Des expérimentations et résultats sur la collection INEX 2007 sont présentés dans la section 4. Une discussion générale est menée dans la section 5, et enfin, quelques conclusions et perspectives sont décrites dans la section 6.

2. La recherche multimedia dans des documents semi-structurés

A l'origine, les systèmes de recherche d'information ont été conçus pour rechercher des documents entiers de type textuel, l'utilisateur devant lire toutes les informations des documents afin de trouver les parties qui l'intéressent. La recherche d'information structurée a apporté une réponse à ce problème, en utilisant la structure des documents et en renvoyant des éléments (noeuds) XML se focalisant sur le besoin de l'utilisateur.

Ces dernières années, avec le nombre croissant de média de type image, vidéo et son dans les documents, de nouvelles problématiques liées à l'inclusion de médias autre que le texte dans les documents semi-structurés sont apparues. Des fragments multimédia contenant à la fois du texte et un média autre que le texte doivent pouvoir être renvoyés aux utilisateurs.

Nos travaux se focalisent sur ce besoin, pour lequel nous décrivons quelques approches issues de l'état de l'art ci-dessous, avec des applications sur le média "*image*".

Jusqu'en 2005, où la campagne d'évaluation INEX [1] a donné naissance à une nouvelle tâche appelée tâche multimédia, offrant ainsi une plateforme d'évaluation de traitement de requêtes multimedia, peu de travaux se sont intéressés à la recherche multimedia (et plus précisément à la recherche d'images) dans des documents XML.

Parmi les premiers travaux proposés utilisant la structure XML pour la recherche d'éléments multimédia, citons ceux de (Kong *et al.*, 2005) (Kong *et al.*, 2007) qui consistent à diviser tout le contenu textuel du document XML en plusieurs *Region Knowledge* [2] *RKs* : *Self level RK* : *RK* du nœud multimedia ; *Sibling level RK* : *RK* des nœuds frères du nœud multimedia ; *1st ancestor level RK* : *RK* du premier ancêtre (parent) du nœud multimedia à l'exclusion du texte déjà utilisé ; *2nd ancestor level RK,..., Nth ancestor level RK*. Le modèle vectoriel est ensuite utilisé pour évaluer chaque *Region Knowledge*. Même si cette méthode exploite la structure verticale des documents, elle ne prend pas en compte la distribution des éléments contenus dans une même *Region Knowledge*.

D'autres travaux utilisent une combinaison linéaire des résultats obtenus par une recherche d'images basée sur le contenu (c'est à dire les caractéristiques de bas niveau des images) et une recherche textuelle. Dans (Iskandar *et al.*, 2006) par exemple, les auteurs ont proposé d'utiliser le système de recherche d'images par contenu *GIFT*

1. Initiative for the Evaluation of XML Retrieval. http ://inex.is.informatik.uni-duisburg.de/
2. Le contenu textuel de l'objet multimédia et des éléments l'entourant hiérarchiquement.

d'une part et le système de recherche textuel *Zettair* d'une autre part. La combinaison de ces deux systèmes n'a pas montré son intérêt dans la campagne d'évaluation INEX.

Une autre méthode proposée par l'équipe *CWI/UTwente* (Tsikrika *et al.*, 2007a) consiste à utiliser une méthode de recherche traditionnelle basée sur le modèle de langage en évaluant plusieurs priorités de longueur. Afin de respecter la spécificité multimedia, les résultats obtenus sont filtrés en ne gardant que les fragments contenant au moins une image. Ainsi, aucun traitement multimedia supplémentaire n'est effectué. Les meilleurs résultats retournés par cette méthode sont obtenus en ne renvoyant que des documents entiers.

Une autre approche proposée dans (Szlávik *et al.*, 2007) consiste à utiliser un réseau bayésien intégrant un modèle de langage, appliqué aux éléments et non aux documents, pour la recherche de texte et d'images. Cette méthode a été évaluée avec une petite collection (Lonely Planet d'INEX Multimedia 2005) et a montré son intérêt, même si des expérimentations avec une plus grosse collection (telle que la collection Wikipedia d'INEX, Tâche Multimedia Fragment 2006-2007) seraient nécessaires.

En conclusion, la recherche de fragments multimedia se réalise soit en combinant une recherche adhoc XML et une recherche d'images basée-contenu, soit par filtrage des résultats d'une recherche adhoc XML en ne gardant que les fragments contenant au moins une image. Peu de travaux tiennent compte à la fois de la structure des documents et de la spécificité multimédia. Nous présentons dans ce qui suit notre modèle qui vise à utiliser ces deux sources d'évidence.

3. Un modèle de recherche de fragments multimedia basée sur l'information textuelle et structurelle des documents

Dans (Torjmen *et al.*, 2009), nous avons proposé un modèle dédié à la recherche de fragments multimedia. La recherche s'effectue en deux étapes : (1) recherche des éléments images en utilisant le contenu textuel et structurel des documents, (2) détermination des fragments multimedia pertinents à partir de ces images. Le défi ici est de sélectionner le meilleur fragment multimedia qui doit être retourné.

3.1. *Représentation textuelle et structurelle d'éléments multimedia dans des documents semi-structurés*

Un document XML peut être représenté par un arbre où la racine est le document, les nœuds internes sont les nœuds représentant les éléments ou les attributs, et les nœuds feuilles sont les nœuds contenant les valeurs (texte, nom-image).

Dans des travaux précédents, nous avons proposé deux approches pour définir les éléments multimedia pertinents. La première (Torjmen *et al.*, 2009) consiste à évaluer un score pour les images en utilisant trois sources d'évidence : ses descendants, ses frères et ses ascendants ayant déjà des scores de pertinence précalculés par un sys-

tème de recherche XML adhoc. L'inconvénient de cette méthode est qu'elle est très dépendante du système de recherche adhoc utilisé.

La deuxième (Torjmen *et al.,* 2008a) consiste à représenter l'image via les nœuds textuels en se basant sur une analogie entre un document XML et une ontologie.

Cette dernière approche est basée sur deux intuitions : (1) chaque nœud textuel porte des informations permettant de représenter sémantiquement une image. Par conséquent, chaque élément textuel contenant des informations pertinentes doit participer à représenter l'image sémantiquement ; (2) certains nœuds textuels du document doivent participer plus que d'autres dans la représentation de l'image. En effet, l'apport de chaque nœud dans la représentation de l'image doit se calculer en fonction de la position hiérarchique de ce nœud par rapport à l'image.

La question qui s'impose derrière cette idée est : comment calculer la participation de chaque nœud textuel dans la représentation sémantique de l'image ?

Afin de répondre à la première intuition, pour utiliser l'information textuelle du document XML, nous avons calculé un score de pertinence pour chaque nœud feuille à partir d'un système de recherche XML classique, basé sur la formule *tf*idf*ief*.

Pour prendre en compte la deuxième intuition, nous avons utilisé l'information structurelle du document XML. La représentation arborescente d'un document XML nous permet de le considérer comme une ontologie très simplifiée où les nœuds sont les concepts qui sont organisés hiérarchiquement avec la relation ***est partie de***. Par exemple, *section* ***est partie de*** *article* et *paragraphe* ***est partie*** de *section*.

L'idée consiste à transposer une mesure de similarité sémantique utilisée entre les termes d'une ontologie pour calculer l'apport de chaque nœud à la représentation de l'image. Nous considérons le nœud image comme un concept C1, et le nœud à utiliser comme un autre concept C2 (Torjmen *et al.,* 2008b) (Torjmen *et al.,* 2008a).

Etant donné que l'image peut ne pas avoir ou en avoir très peu de contenu textuel, nous nous intéressons aux mesures de similarité basées sur les arcs et pas sur le contenu. Plusieurs mesures de similarités basées sur le nombre d'arcs entre les concepts sont proposées dans la littérature telles que celle de (Rada *et al.,* 1989), celle de (Hirst *et al.,* 1997) et celle de (Wu *et al.*, 1994).

La mesure de Wu-Palmer (Wu *et al.*, 1994), prenant en compte la position des concepts par rapport à la racine de l'ontologie est à la fois la plus simple à implémenter et la plus performante (Lin, 1998). Elle est définie comme suit :

$$Sim_{WP}(C_1, C_2) = \frac{2 * N_3}{(N_1 + N_2 + 2 * N_3)} \quad [1]$$

où N_1 et N_2 sont le nombre d'arcs qui séparent C_1 et C_2 de leur ascendant commun le plus spécifique *C*. N_3 est le nombre d'arcs qui séparent *C* de l'élément racine.

Cette mesure n'a cependant pas montré son intérêt dans des travaux précédents (Torjmen *et al.*, 2008a). Dans ce qui suit, nous proposons d'autres facteurs permet-

tant de prendre en compte la différence d'importance des nœuds dans l'arbre du document. Nous souhaiterions que les descendants de l'image participent plus que les descendants de son premier ancêtre puisqu'ils sont les plus spécifiques [3], que les descendants du premier ancêtre participent plus que les descendants du deuxième ancêtre puisqu'ils ont une forte probabilité de partager le même sujet avec l'image, etc. Les nœuds qui participeraient le moins à la représentation de l'image sont les descendants de l'élément racine puisqu'ils sont les plus loins de l'image.

La mesure de Wu-Palmer a été utilisée dans l'indexation sémantique des documents XML dans (Zargayouna, 2004). Cependant, les auteurs ont constaté qu'elle représente une limite car il est possible d'avoir la similarité entre un concept et son fils inférieure à la similarité entre ce concept et son frère alors qu'il était envisagé de ramener tous les fils d'un concept avant ses frères.

Pour éviter cela, les auteurs ont proposé de pénaliser les scores des frères en ajoutant une fonction $spec(C_1, C_2)$ qui calcule la spécificité de deux concepts par rapport au concept le plus bas (*bottom*) (voir Figure 1).

$$Sim_{WP}(C_1, C_2) = \frac{2 * N_3}{(N_1 + N_2 + 2 * N_3 * spec(C_1, C_2))} \quad [2]$$

$$où\ spec(C_1, C_2) = depth_b(C) * N_1 * N_2 \quad [3]$$

avec *C* est l'ancêtre commun le plus spécifique, $depth_b$ est le nombre maximum d'arcs qui séparent *C* de *bottom* (figure 1) et N_1 (N_2) est la distance en nombre d'arcs entre *C* et C_1 (C_2).

Comme le montre la figure 1, le facteur $depth_b$ utilise la structure hiérarchique verticale afin de différencier la participation des descendants de chaque ancêtre de l'image. Dans la figure 1, les descendants *F, M et S* de l'image *I* ont un $depth_b$ ($Depth_b$I) plus petit que celui de *B* ($Depth_b$B). Ce facteur $depth_b$ semble donc adéquat pour prendre en compte notre intuition dans la représentation sémantique de l'image.

Cependant, l'utilisation seule de ce facteur comme information structurelle conduit à égaliser la participation des descendants du même ancêtre de l'image. Afin de palier cet inconvénient, nous avons décidé de conserver aussi les facteurs N_1 et N_2 privilégiant ainsi les nœuds descendants les plus proches de l'image : plus les nœuds textuels sont loins de l'image, moins ils participent à sa représentation.

Prenons l'exemple des nœuds textuels *H* et *K* dans la figure 1, ils participent avec le même $depth_b$ ($Depth_b$B) et la même distance entre l'image *I* et l'ancêtre commun *B* (N_2^I) dans la représentation du nœud image *I*, mais le nœud *K* participe plus que le nœud *H* puisque que la distance N_1^K est plus petite que la distance N_1^H.

3. Les éléments images, outre le fait de donner l'URL du fichier image concerné, peuvent contenir d'autres éléments trés spécifiques, comme le nom de l'image et sa légende.

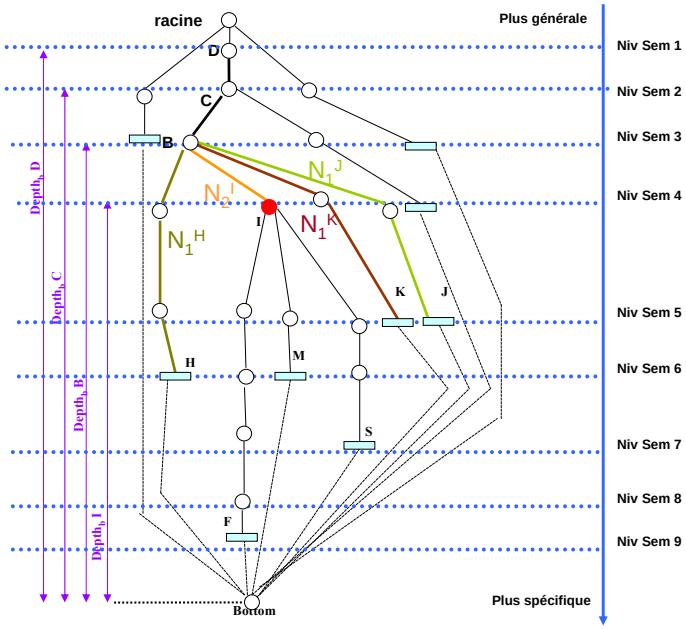

Figure 1. *Représentation d'un élément image en se basant sur l'information structurelle*

Dans nos travaux, nous proposons d'utiliser cette mesure pour calculer la participation de chaque nœud pertinent dans la représentation sémantique de l'image. Nous définissons la mesure de représentation sémantique comme suit :

$$Rep_{SPEC}(I, NT) = \frac{S_{NT}}{depth_b(C) * N_1 * N_2} \quad [4]$$

où *I* est le nœud image, *NT* est le nœud textuel qui participe à la représentation de l'image et S_{NT} est le score du nœud *NT*, calculé à l'aide d'un système de recherche d'information structurée classique et C est l'ancêtre commun entre *I* et *NT*.

Chacun des facteurs de cette formule sera évalué séparément dans la partie *Evaluation* (section 4.3).

Le score final de chaque image est calculé comme suit :

$$S(I) = \sum_{i=1}^{|NT|} Rep(I, NT_i) \quad [5]$$

avec NT_i un nœud textuel du document et $|NT|$ le nombre des noeuds textuels du document contenant l'image.

3.2. *Recherche de fragments multimédia à travers les images*

En recherche multimedia, le besoin utilisateur peut être un media comme une image, ou bien un fragment de document contenant au moins une image (Tsikrika *et al.*, 2007b). Par conséquence, dans notre cas, les résultats à retourner à l'utilisateur ne sont pas obligatoirement des images, mais ils peuvent être aussi des fragments multimedia (image + texte pertinent).

La problématique ici est de décider quels éléments doivent être retournés en se focalisant sur le besoin de l'utilisateur (*Focused Retrieval* dans la terminologie INEX). Les éléments retournés doivent être les plus exhaustifs et spécifiques possibles et ne doivent pas être imbriqués les uns dans les autres. Ce type de recherche suppose que l'utilisateur préfère l'élément (un seul) le plus pertinent d'un sous arbre pertinent (Kamps *et al.*, 2007).

Pour réaliser cet objectif, une méthode a déjà été proposée dans (Torjmen *et al.*, 2009). Elle consiste à définir pour chaque image retrouvée dans la première étape, un ensemble de fragments composé de l'image elle-même et ses ancêtres.

Le score de chaque ancêtre S_a^{im} (des images) est calculé en fonction de son score normalisé obtenu par un système adhoc XML (S_{Adhoc}) et des scores normalisés des images elles-même (S_{im}) contenues par cet ancêtre. La combinaison de ces deux scores se fait à travers une combinaison linéaire :

$$S_a^{im} = \gamma * S_{Adhoc} + (1 - \gamma) * \sum_{i=1}^{|NI|} S_{im_i} \quad [6]$$

avec γ un pivot $\in$[0..1] et $|NI|$ le nombre d'images contenues dans l'ancêtre *a*.

4. Evaluation

Pour calculer un score de pertinence des nœuds textuels des documents, nous avons utilisé le système XFIRM (Pinel-Sauvagnat *et al.*, 2004) (Sauvagnat *et al.*, 2006). Ce système est aussi utilisé pour calculer un score de pertinence pour les nœuds ancêtres des images, utilisé dans l'équation 6.

4.1. *INEX : Collection et mesures d'évaluation*

INEX (Initiative for the Evaluation of XML Retrieval) est actuellement la seule campagne d'évaluation des différents systèmes de recherche d'information pour des documents XML. Le but principal d'INEX est de promouvoir l'évaluation de la recherche sur des documents XML en fournissant une collection de test, des procédures d'évaluation et un forum pour permettre aux différentes organisations participantes de comparer leurs résultats. La collection de test consiste en un ensemble de documents XML, requêtes et jugements de pertinence. Le langage de requêtes utilisé dans INEX est NEXI (Trotman *et al.*, 2005). Nous nous intéressons ici à la tâche multimedia qui

a eu lieu en 2007 pour la troisième fois, et plus particulièrement à la sous tâche Multimedia Fragments qui consiste à retrouver des fragments XML multimedia (contenant au moins une image). Des détails concernant cette tâche sont présentés dans (Westerveld *et al.*, 2006) (Tsikrika *et al.*, 2007b). La collection de cette tâche est la collection XML Wikipedia (Denoyer *et al.*, 2006), comprenant plus de 650 000 documents.

En 2007, 19 requêtes sont fournies pour la tâche Fragments Multimedia. Ces requêtes comportent plusieurs parties : une représentation textuelle par simples mots clés, une représentation textuelle et structurelle en ajoutant des contraintes structurelles, et finalement une représentation multimedia en ajoutant par exemple des concepts ou des images exemples.

Dans les travaux présentés dans cet article, seule la représentation textuelle simple des requêtes est utilisée.

La première partie de notre modèle consiste à retrouver les images pertinentes à partir d'une base de jugement de pertinence composée seulement d'images. Pour évaluer cette partie, nous avons créé une nouvelle base de jugements de pertinence à partir de la base originale de fragments multimedia, et ceci en gardant seulement les éléments images. Cette partie est évaluée grâce à la moyenne de la précision moyenne (MAP), en utilisant l'outil *trec-eval*.

La deuxième partie de notre approche consiste à retrouver des fragments multimedia pertinents. Elle est évaluée avec les mesures officielles de la tâche Fragments multimedia d'INEX 2007 (Kamps *et al.*, 2007). Deux mesures sont utilisées :

– ***La précision interpolée selon quatre niveaux de rappel sélectionnés :*** $iP[jR], j \in [0.00, 0.01, 0.05, 0.1]$

La précision à un rang r est défini comme suit :

$$P[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)} \quad [7]$$

où p_r (p_i) est la partie du document assignée au rang r ($i <= r$) dans la liste de résultats L_q des parties de documents retournées par un système de recherche pour une requête q.

$rsize(p_r)$ est la taille du texte pertinent contenu dans p_r en nombre de caractères et $size(p_r)$ est la taille du texte totale contenu dans p_r en nombre de caractères.

Le rappel à un rang r est défini comme suit :

$$R[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{Trel(q)} \quad [8]$$

où $Trel(q)$ est la quantité totale du texte pertinent pour une requête q.

La mesure de précision interpolée $iP[x]$ est la suivante :

$$iP[x] = \begin{cases} max_{1 \leq r \leq |L_q|}(P[r] \wedge R[r] \geq x) \;\; if\ x \leq R[|L_q|] \\ \\ 0 \qquad\qquad\qquad\qquad\qquad if\ x \succ R[|L_q|] \end{cases} \quad [9]$$

où $R[|L_q|]$ est le rappel pour tous les documents restitués.

– ***La moyenne des précisions moyennes interpolées selon 101 niveaux de rappel***.

Supposons que nous avons n requêtes, $MAiP$ est calculée comme suit :

$$MAiP = \frac{1}{n}. \sum_{t} AiP(t) \qquad [10]$$

où Aip est la précision moyenne interpolée.

4.2. *Problèmes liés aux jugements de pertinence*

Alors que la différence principale entre la recherche XML adhoc et la recherche XML multimedia est que cette dernière a pour objectif de retourner des fragments documentaires pertinents contenant au moins une image (Tsikrika *et al.*, 2007b), les jugements de pertinence fournis par la campagne d'évaluation INEX 2007, tâche Fragments Multimedia ne respectent pas cette spécificité multimedia. En effet, nous avons constaté que 84,71% de ces jugements concernent des fragments purement textuels (c'est à dire ne contenant aucune image).

Par conséquent, nous ne pouvons pas évaluer la deuxième partie de notre méthode avec cette base de jugements de pertinence. Afin de palier cet inconvénient, nous avons décidé de filtrer ces jugements de pertinence en ne gardant que les fragments ayant au moins une image pertinente.

Conjointement à ce filtrage des jugements de pertinence, nous avons filtré les runs officiels des participants d'INEX 2007 puisque quelques uns renvoient également des fragments textuels purs, et ceci afin de les comparer à notre modèle. Suite à ce filtrage des runs officiels des participants d'INEX, le nombre de résultats retournés par leurs systèmes est diminué, et par conséquent, la comparaison selon la mesure MAiP n'est plus significative. Afin d'effectuer tout de même une comparaison, nous avons décidé de tracer les courbes Rappel/Précision interpolées selon les niveaux de rappel [0.00..0.05] et [0.1..1]. Nous nous intéressons plus particulièrement aux précisions dans les premiers niveaux de rappel puisque le nombre de résultats retournés n'est plus le même pour tous les runs.

4.3. *Résultats de la représentation sémantique des images par les nœuds textuels*

Afin d'étudier l'efficacité de la mesure de similarité Wu-Palmer (Equation 1) dans nos travaux ainsi que l'importance de chaque facteur de la formule 4, nous avons évalué tout d'abord le contenu textuel seul sans utiliser les facteurs de structure (Figure 2, formule *Cont-Text*), et ceci en sommant simplement les scores des nœuds textuels évalués par le système *XFIRM*. Ainsi, les images du même document auront tous le même score qui est la somme des scores du contenu textuel.

Nous avons ensuite évalué les scores des images en multipliant la mesure de Wu-Palmer représentée dans l'équation 1 par la formule *Cont-Text*. La valeur *MAP* de

cette mesure est représentée dans la figure 2 sous le nom *Cont-Text-Wu-Palmer*. Nous constatons que cette mesure permet une amélioration de 14.48% en MAP.

Pour évaluer l'impact de chaque facteur structurel sur la pertinence des images, nous les avons évalués séparemment en les multipliant par le score des nœuds textuels pré-calculé avec le système *XFIRM*.

La figure 2 montre les résultats des différents facteurs selon la mesure *MAP*.

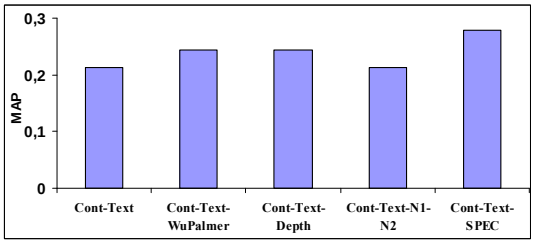

Figure 2. *Comparaison des différents facteurs pour la représentation des images*

Cont-Text-Depth consiste à multiplier le score de contenu textuel (*Cont-Text*) par le facteur $1/depth_b$. *Cont-Text-N1-N2* consiste à multiplier le score de contenu textuel (*Cont-Text*) par le facteur $1/(N_1 * N_2)$, et *Cont-Text-SPEC* consiste à multiplier le score de contenu textuel (*Cont-Text*) par les facteurs $1/depth_b$ et $1/(N_1 * N_2)$ (c'est à dire à utiliser l'équation 4).

Nous constatons tout d'abord que le facteur $1/depth_b$ joue un rôle important dans l'amélioration des résultats (*Cont-Text-Depth*). Ceci illustre l'importance de différencier la participation des nœuds textuels selon leur hiérarchie verticale avec l'image.

Nous constatons d'autre part que le facteur $1/(N_1 * N_2)$ seul (*Cont-Text-N1-N2*) n'apporte pas d'amélioration au contenu textuel, alors que c'est le cas en le multipliant par $1/depth_b$. Ces résultats confirment aussi notre intuition que plus la distance entre le nœud textuel et l'image est petite, plus ce nœud textuel doit participer à la représentation de l'image.

Finalement, en comparant les deux formules Wu-Palmer (*Cont-Text-WuPalmer*) et *spec* (*Cont-Text-SPEC*), nous constatons que cette dernière apporte une amélioration significative (+14.42% en *MAP*).

Le reste de nos expérimentations est basé sur les résultats obtenus par la formule *Cont-Text-SPEC* (Equation 4).

4.4. *Résultats de la recherche de fragments multimédia à travers les images*

Avant d'évaluer notre modèle, nous avons évalué les runs officiels des participants de la tâche Fragment Multimedia d'INEX 2007 avec la nouvelle base de jugements de

pertinence filtrée (en ne gardant que les fragments Multimedia, c'est à dire les fragments contenant au moins une image). Nous avons évalué également le meilleur run adhoc selon la tâche Multimedia en utilisant la mesure MAiP (run *MeilleurRunAdhoc-Indstaint-MAiP*), et le meilleur run adhoc selon la tâche Multimedia en utilisant la mesure iP[0.01] (run *MeilleurRunAdhoc-Mines-iP[0.01]*).

En effet, la différence majeure entre la tâche adhoc et la tâche multimedia dans INEX 2007 étant que les fragments retournés doivent avoir un caractère multimedia dans le second cas, les requêtes de la tâche Multimedia font partie de l'ensemble de requêtes de la tâche Adhoc. Les runs adhoc ont été ainsi évalués dans la cadre de la tâche Multimedia. Les courbes de rappel/précisions interpolées de ces 7 runs sont présentées sur la figure 3. Le meilleur run au niveau de rappel 0.01 est *MeilleurRunAdhoc-Mines-iP[0.01]*.

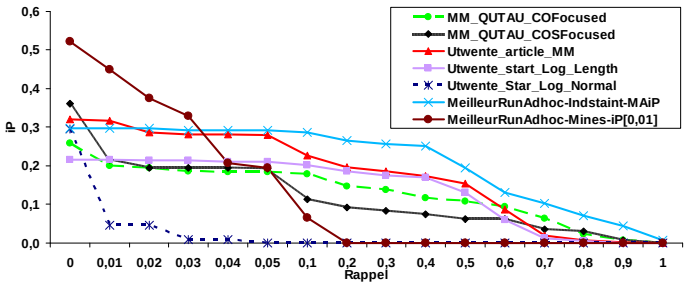

Figure 3. *Comparaison des runs officiels de la tâche Multimedia et des meilleurs runs Adhoc à INEX 2007 selon la nouvelle base de jugements filtrée*

La figure 4 donne les résultats de quelques uns de nos runs, à savoir le meilleur run en renvoyant à la fois des images ou des ancêtres d'images (run *Gamma1-Ancêtres-Images*), le meilleur run en ne renvoyant que des ancêtres d'images et jamais les images elles mêmes (*Gamma1-Ancêtres*), le run obtenu avec $\gamma = 0$, dans lequel seul le score des images est utilisé pour le calcul du score des ancêtres (run *Gamma0*), et enfin le run où nous utilisons seulement la première partie de notre modèle (c'est à dire que seuls les éléments images sont retournés) (*RunImages*).

Nous avons mené plusieurs expérimentations, non présentées ici, pour déterminer la meilleure valeur de γ dans l'équation 6. Selon la mesure officielle d'INEX 2007 (iP[0.01]), le meilleur valeur de γ est 1, lorsque nous renvoyons à la fois des images ou des ancêtres, ou lorsque nous renvoyons seulement des éléments ancêtres (runs *gamma1-Ancêtres-Images* et *gamma1-Ancêtres*). Ceci signifie que l'utilisation seule du score calculé par le système XFIRM est meilleure que la combinaison des deux scores. L'utilisation seule des scores des images n'a pas donné des bons résultats (run *gamma0*). En effet, la mesure iP [0.01] se dégrade de 37.26% par rapport à l'utilisation du score de XFIRM seul en renvoyant des images ou des ancêtres (run *gamma1-Ancêtres-Images*) et de 27.43% dans le cas où seulement des ancêtres d'images sont renvoyés (run *gamma1-Ancêtres*).

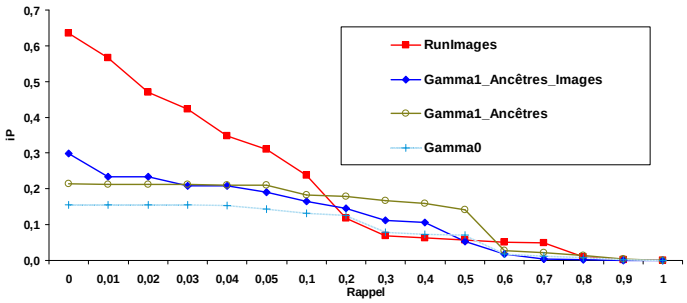

Figure 4. *Comparaison des différents runs de notre modèle*

Ceci peut être expliqué comme suit : chaque image va contribuer aux scores des ancêtres avec le même score (si un document contient une seule image, tous les ancêtres de cette dernière auront le même score). Dans nos expérimentations, si deux ancêtres ont le même score, on renvoie celui qui possède le plus haut niveau, et par conséquent, dans le cas de $\gamma = 0$, l'ancêtre qui possède plus d'images sera classé le premier. Ce comportement va ainsi pousser à renvoyer toujours l'élément ayant le plus haut niveau ("article").

Nous concluons de ces résultats que le score des images n'est pas bien utilisé dans l'évaluation de pertinence des ancêtres. En effet, la participation du score d'une image doit être différent d'un ancêtre à un autre. D'autres expérimentations sont nécessaires pour déterminer la bonne façon de combiner les scores des ancêtres précalculés par XFIRM et les scores des images obtenues dans l'étape 1 de notre modèle.

En comparant les deux runs *gamma1-Ancêtres-Images* et *gamma1-Ancêtres*, nous concluons que renvoyer à la fois des images ou des ancêtres d'images donne des résultats meilleurs que renvoyer seulement des ancêtres. Ceci signifie que l'image toute seule peut satisfaire correctement le besoin de l'utilisateur.

Enfin, en comparant les courbes selon les niveaux de rappel [0.0..0.1], nous constatons que les meilleures précisions interpolées sont obtenues avec le run qui ne renvoie que des images (*RunImages*). Ce résultat, qui peut paraître surprenant, ne l'est en fait pas lorsque on examine les jugements de pertinence. En effet, pour la collection INEX 2007, le pourcentage de fragments multimedia ayant une pertinence supérieure à celle de l'image qu'ils contiennent est de 3.48% seulement. Ceci montre encore un autre problème des jugements de pertinence d'INEX 2007 : la plupart des jugements de pertinence de fragments ont été basés sur la pertinence des images et non sur la pertinence des fragments multimedia (image + texte). Ceci implique que cette base de jugements est plus appropriée pour comparer des systèmes de recherche d'images dans des documents semi-struturées que pour comparer des systèmes de recherche de fragments multimédia. Afin d'évaluer d'une manière satisfaisante la deuxième partie de notre modèle, d'autres expérimentations sont donc nécessaires avec une autre collection prévilégiant suffisamment des fragments multimedia des éléments images.

4.5. *Discussion*

Une conclusion intéressante de ces expérimentations est que l'information structurelle permet bien d'améliorer la recherche d'images dans les documents semi-structurés (31% d'amélioration selon la mesure MAP dans l'étape 1 de notre modèle).

En se basant sur les figures 3 et 4, nous pouvons conclure également que l'utilisation d'une méthode spécifiée multimedia donne des résultats meilleurs qu'une méthode sans spécification multimedia dans le cas de recherche multimedia. Selon la mesure officielle d'INEX 2007, tâche Multimedia, nous obtenons le premier rang avec une amélioration de 78.74% par rapport au meilleur run Multimedia et 26.26% par rapport au meilleur run adhoc (iP[0.01]=0.5668 Versus iP=[0.01]=0.3171 pour le meilleur run multimedia et iP[0.01]=0.4489 pour le meilleur run adhoc). Ceci montre que les sytèmes XML adhoc, même adaptés à la recherche multimédia, ne sont pas suffisants pour répondre aux besoins multimédia des utilisateurs. Cela montre également que partir du contenu textuel et structurel est plus intéressant que chercher des fragments adhocs pertinents et de les filtrer selon un caractère multimédia.

Comme nous l'avons mentionné ci-dessus, la comparaison de notre modèle avec les autres selon la mesure MAiP n'est plus significative étant donné que le nombre de résultats n'est plus le même pour tous les runs. Cependant, à titre indicatif, notre meilleur run (runImages) sur la base de jugements de pertinence filtrée obtient une MAiP de 0.0932, alors que le meilleur run adhoc selon la mesure MAiP (*MeilleurRunAdhoc-Indstaint-MAiP*) obtient 0.1724.

Ces moins bons résultats peuvent être expliqués de la façon suivante : de bons éléments sont retournés par notre modèle aux premiers niveaux de rappel alors qu'ils sont retournés dans des niveaux de rappel plus élevés par le run adhoc, or, l'interpolation des précisions ne pénalise pas les systèmes renvoyant les bons résultats dans des niveaux de rappel éloignés par rapport aux systèmes renvoyant les bons résultats dans les premiers niveaux de rappel.

Afin d'illustrer ce problème, nous avons tracé respectivement la courbe Rappel/Précision non interpolée et la courbe Rappel/Précision interpolée de la requête 529 selon les deux runs (voir figure 5).

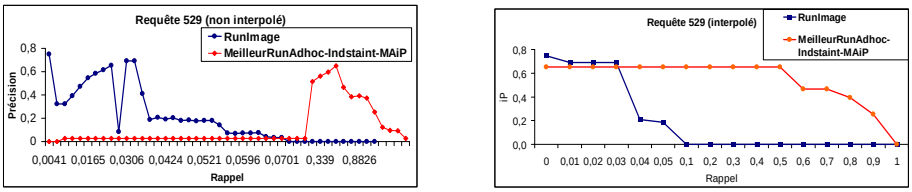

Figure 5. *Courbe Rappel/Précision non interpolée et interpolée de la requête 529*

La mesure MAiP montre ainsi ses limites : la mesure MAP (pas d'interpolation) serait mieux appropriée pour évaluer la performance globale des différents systèmes.

5. Conclusion et perspectives

Dans cet article, nous avons évalué l'impact de différents facteurs pour juger la pertinence des éléments multimedia dans des documents XML. Les expérimentations ont prouvé que l'information structurelle permet d'améliorer significativement la performance de recherche. De plus, les évaluations ont montré quelques problèmes liés à la base de jugement de pertinence. En effet, 84.71% des fragments jugés pertinents sont des fragments purement textuels ce qui ne respecte pas la spécificité de la tâche multimedia. D'autre part, la plupart des jugements de pertinence prévilégient les éléments images par rapport aux fragments multimedia (image et texte). Ceci ne permet pas de comparer les méthodes de recherche de fragments multimédia, mais plutôt des méthodes de recherche d'images dans des documents structurés.

Les perspectives envisageables à nos travaux consistent d'une part à améliorer la deuxième partie de notre modèle en testant d'autres méthodes de combinaison du score de l'image et du score des noeuds ancêtres, et d'une autre part à étudier autres facteurs pour la recherche des images tels que les liens et le nom de l'image.

6. Bibliographie

Denoyer L., Gallinari P., « The Wikipedia XML corpus », *SIGIR Forum 2006, p 64-69*, 2006.

Elghazel H., Idrissi K., Baskurt A., Amar C. B., « Approche textuelle pour la recherche d'image. », *3rd International Conference SETIT'05*, 2005.

Fuhr N., Lalmas M., Malik S., Kazai G., « INEX 2005 », 2005.

Fuhr N., Lalmas M., Trotman A., « INEX 2006 », 2006.

Fuhr N., Lalmas M., Trotman A., Kamps J., « INEX 2007 », 2007.

Hirst G., St-Onge D., « Lexical Chains as representation of context for the detection and correction malapropisms », 1997.

Iskandar D. N. F. A., Pehcevski J., Thom J. A., Tahaghoghi S. M. M., « Combining Image and Structured Text Retrieval », *INEX'05*, p. 525-539, 2005.

Iskandar D. N. F. A., Pehcevski J., Thom J. A., Tahaghoghi S. M. M., « Social Media Retrieval Using Image Features and Structured Text », in *INEX, p358-372* (Fuhr *et al.*, 2006), 2006.

Kamps J., Pehcevski J., Kazai G., Lalmas M., Robertson S., « INEX 2007 Evaluation Measures », in *INEX, p 24-33* (Fuhr *et al.*, 2007), 2007.

Kong Z., Lalmas M., « XML Multimedia Retrieval », *SPIRE 2005, p 218-223*, 2005.

Kong Z., Lalmas M., « Using XML Logical Structure to Retrieve (Multimedia) Objects », *ECDL 2007, p 100-111*, 2007.

Lew M. S., Sebe N., Djeraba C., Jain R., « Content-based multimedia information retrieval : State of the art and challenges », *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, p. 1-19, 2006.

Lin D., « An Information-Theoretic Definition of Similarity », *Proceedings of 15th International Conference On Machine Learning*, 1998.

Pinel-Sauvagnat K., Boughanem M., Chrisment C., « Searching XML documents using relevance propagation », *Symposium on String Processing and Information Retrieval (SPIRE'04)*, p. 242-254, 2004.

Rada R., Mili H., Bicknell E., Blettner M., « Development and application of a metric on semantic nets », *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, n° 1, p. 17-30, 1989.

Sauvagnat K., Boughanem M., Chrisment C., « Answering content and structure-based queries on XML documents using relevance propagation », *Journal Information Systems*, vol. 31, n° 7, p. 621-635, 2006.

Szlávik Z., Tombros A., Lalmas M., « Feature- and Query-Based Table of Contents Generation for XML Documents », *ECIR 2007, p 456-467*, 2007.

Tjondronegoro D., Zhang J., Gu J., Nguyen A., Geva S., « Integrating Text Retrieval and Image Retrieval in XML Document Searching », in *INEX'05, p 511-524* (Fuhr *et al.*, 2005), 2005.

Tollari S., Mulhem P., Ferecatu M., Glotin H., Detyniecki M., Gallinari P., Sahbi H., Zhao Z.-Q., « A Comparative Study of Diversity Methods for Hybrid Text and Image Retrieval Approaches », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of CLEF*, 2008.

Torjmen M., Pinel-Sauvagnat K., Boughanem M., « Towards a structure-based multimedia retrieval model », *ACM International Conference MIR'08*, 2008a.

Torjmen M., Pinel-Sauvagnat K., Boughanem M., « Une métrique pondérée pour la recherche textuelle d'images dans des documents semi-structurés », *CORIA'08*,, p. 55-70, 2008b.

Torjmen M., Pinel-Sauvagnat K., Boughanem M., « XML Multimedia Retrieval : From relevant textual information to relevant multimedia fragments », *ECIR'09*, 2009.

Trotman A., Sigurbjörnsson B., Fuhr N., Lalmas M., Malik S., Szlavik Z., « Narrowed Extended XPath I (NEXI) », *Lecture Notes in Computer Science*, vol. 3493, Springer Verlag, Heidelberg, p. p 16-40, 2005.

Tsikrika T., Serdyukov P., Rode H., Westerveld T., Aly R., Hiemstra D., de Vries A. P., « Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah », in *INEX, p273-286* (Fuhr *et al.*, 2007), 2007a.

Tsikrika T., Westerveld T., « Report on the INEX 2007 Multimedia Track », in *INEX, p 410-422* (Fuhr *et al.*, 2007), 2007b.

van Zwol R., « Multimedia Strategies for ^{3}-SDR, Based on Principal Component Analysis », in *INEX'05, p540-553* (Fuhr *et al.*, 2005), 2005.

Westerveld T., Zwol R. V., « The INEX 2006 Multimedia Track », in *INEX, p 331-344* (Fuhr *et al.*, 2006), 2006.

Wu Z., Palmer M., « Verb semantics and lexical selection », *Proceedings of the 23nd Annual Meetings of the Associations for Computational Linguistics*, p. 133-138, 1994.

Zargayouna H., « Contexte et sémantique pour une indexation de documents semi-structrés », *COnference en Recherche d'Information et Applications*, p. 571-581, Mars, 2004.

Zhang C., Chai J. Y., Jin R., « User term feedback in interactive text-based image retrieval », *SIGIR'05*, p. 51-58, 2005.