Exploitation des liens pour la recherche d'images dans des documents XML

Hatem Awadi* — **Mouna Torjmen****

** Laboratoire REDCAD, Ecole Nationale d'Ingénieurs de Sfax (ENIS)*
Université de Sfax, 3000 Tunisie
awadi.hatem@gmail.com

*** IRIT SIG-RI*
118 route de Narbonne, 31 062 Toulouse
torjmen@irit.fr

RÉSUMÉ. *En recherche d'images basée sur le contexte, la technique la plus utilisée est celle basée sur le texte entourant l'image. Cependant, d'autres approches sont proposées telles que l'exploitation des ressources sémantiques pour enrichir la description textuelle de l'image, l'exploitation de la structure des documents, etc. Dans cet article, nous étudions l'exploitation des liens entre éléments des documents semi-structurés de type XML pour la recherche d'images. Plus précisément, nous proposons d'étendre l'ensemble des éléments images retrouvés et ceci à travers les liens. Par la suite, nous proposons de recalculer les scores des images en tenant compte de la pondération des liens. De cette manière, nous attendons une amélioration des résultats par réordonnancement et par ajout d'autres éléments images pertinents.*

ABSTRACT. *In context-based image retrieval, the most popular technique is based on the text surrounding the image. However, other approaches have been proposed such as exploiting semantic resources to enhance the textual description of the image, the exploitation of document structure, etc. In this paper, we study the exploitation of the links between elements in semi-structured XML documents to retrieve images. More precisely, we propose to extend through links the list of image elements already retrieved. Thereafter, we propose to re-compute scores of images by taking account of the link weights. In this manner, we think to obtain an improvement of results by re-ranking and by adding other relevant image elements.*

MOTS-CLÉS : *Documents XML, analyse des liens, recherche d'images basée contexte.*

KEYWORDS: *XML documents, link analysis, context based image retrieval.*

1. Introduction

Dans la littérature, un lien est essentiellement un pointeur (ou une citation) d'une page vers une autre. L'intuition derrière l'utilisation des liens dans la recherche d'information est que ces liens ne sont pas mis au hasard : l'auteur d'une page ne cite une autre page que s'il la considère pertinente. Un des avantages de l'utilisation des liens est que les pages pertinentes, même si elles ne contiennent pas de mots de la requête, peuvent être retournées à l'utilisateur.

Bien que plusieurs travaux exploitant les liens ont été développés pour la recherche d'information classique telles que *PageRank* (Brin *et al.*, 1998) et *HITS* (Kleinberg, 1998), peu sont proposés pour le cas des images. Dans cet article, nous nous sommes intéressés à étudier l'impact des liens pour la recherche d'images dans un cadre plus sémantique que le Web puisque les documents traités sont des documents XML : la sémantique de ces documents provient du fait que les balises ne sont pas des balises de forme comme dans le cas du HTML, mais des balises de contenu sémantique. Notre idée de base consiste à utiliser une liste primaire de résultats (éléments images[1]) obtenue par un système de recherche d'images dans des documents XML, telle que le système de (Torjmen, 2009), et compléter par la suite cette liste en ajoutant d'autres éléments grâce aux liens entrants et sortants. Pour chacun de ces éléments, nous proposons de ne pas utiliser le document entier, mais plutôt de définir la meilleure partie du document (*région*) qui concerne l'image. Ensuite, nous pondérons les liens entre les différentes régions en fonction de la structure hiérarchique de la région. Ces poids des liens sont par la suite utilisés dans l'algorithme HITS appliqué aux régions (et non aux documents). Enfin, les scores autorités résultats de cet algorithme sont combinés avec les scores initiaux des images pour obtenir des scores finaux.

Le reste de cet article est organisé comme suit : dans la section 2, nous présentons quelques travaux d'exploitation des liens pour la recherche d'images. Nous détaillons dans la section 3 notre proposition, et enfin nous concluons dans la section 4.

2. Etat de l'art

Dans la littérature, peu de travaux qui exploitent les liens pour la recherche d'images ont été développés. Pour le cas de la recherche d'images sur le Web, nous pouvons citer le système *PicASHOW* (Lempel *et al.*, 2002) qui applique des approches basées sur la co-citation et des méthodes inspirées de *PageRank* pour calculer les scores des images. Les pages Web sont considérées comme unités atomiques et les liens entre les pages sont traités de la même manière. Ce système a été amélioré par la suite en intégrant une technique de pondération des liens (Voutsakis *et al.*, 2005). Cependant, les pages Web ne contiennent pas seulement des liens informationnels, mais aussi des liens de navigation et des liens publicitaires, ce qui dégrade les performances. De plus, une page peut contenir plusieurs sujets ce qui implique qu'il faut

1. Un élément image est défini dans nos travaux par le fragment XML contenant la référence du fichier de l'objet multimédia ainsi que la description textuelle associée (caption,...)

définir l'ensemble des liens informationnels à utiliser pour déterminer la pertinence de chaque image. Pour résoudre ces problèmes, les auteurs du système *iFind* (Cai *et al.*, 2007) ont proposé de segmenter les pages Web en blocs. Les algorithmes d'analyse des liens tels que *HITS* et *PageRank* sont appliqués par la suite au niveau des blocs.

Pour le cas de la recherche d'images dans des documents structurés, à notre connaissance, le seul travail proposé est celui de Kong et Lalmas (Kong *et al.*, 2004) et qui consiste à combiner la structure hiérarchique et les liens avec l'information textuelle des documents pour la recherche d'éléments multimédias. Les liens, selon les auteurs, sont divisés en deux types : *XLink* qui représente un rapport explicite entre ressources ou parties de ressources dans les documents XML, et les liens visuels qui sont calculés par une similarité de bas niveau entre objets multimédias.

3. Proposition d'exploitation des liens pour la recherche d'images

Dans les documents XML, nous n'avons pas le problème des liens bruits puisque tous les liens sont informationnels, c'est à dire ils ramènent vers des informations pertinentes. Pour exploiter ces liens dans la recherche d'images, nous proposons d'adapter l'algorithme HITS (Kleinberg, 1998) afin de réordonner les résultats de recherche et de retourner d'autres images pertinentes non retrouvées initialement. Les différentes étapes de notre approche sont présentées dans ce qui suit.

3.1. *Définition d'une région d'image*

Nous proposons de ne pas utiliser tous les liens du document, mais seulement les liens qui appartiennent aux éléments de l'entourage de l'image (*région*). Une région est définie par l'ensemble des éléments les plus proches de l'élément image, qui sont susceptibles d'être reliés au même sujet. La façon la plus simple de définir une région est de considérer l'élément image lui même et ses éléments descendants puisqu'ils sont les plus spécifiques. Cependant, cette région peut être très petite (dans le cas où il n'y a pas une description textuelle pour l'image par exemple). Pour cela, il vaut mieux monter et étudier les différents niveaux. La figure 1 montre un exemple de différentes régions possibles d'une image.

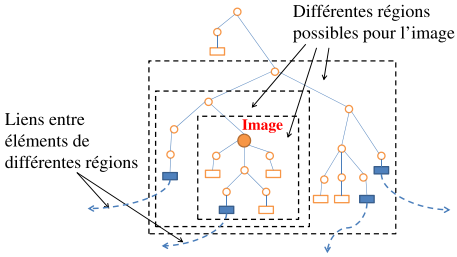

Figure 1. *Exemple de régions d'une image*

Une fois, la meilleure région est déterminée, chaque image aura une seule région alors que chaque région peut avoir une ou plusieurs images.

3.2. *Extension de l'ensemble des éléments images*

Nous proposons par la suite d'étendre l'ensemble initial des régions par les régions de la collection les plus similaires. Afin de calculer la similarité entre deux régions, plusieurs mesures basées sur les liens peuvent être utilisées telles que la co-citation (Small, 1973), Bibliographic Coupling (Kessler, 1963), etc. Dans nos travaux, nous proposons d'utiliser la mesure basée sur la co-citation puisqu'elle a montré son intérêt dans le cas de pages web (Calado *et al.*, 2006).

On désigne par P_r l'ensemble des régions qui pointent vers une région r. La similarité entre deux régions r_1 et r_2 en utilisant la mesure de co-citation est calculée comme suit :

$$cocitation(r_1, r_2) \quad = \quad \frac{P_{r_1} \bigcap P_{r_2}}{\mid P_{r_1} \bigcup P_{r_2} \mid} \qquad [1]$$

L'ensemble obtenu des régions liées entre elle est considéré comme un graphe, où les nœuds sont les régions et les arcs sont les liens entre eux.

3.3. *Pondération des liens entre les différentes régions*

Dans une même région, les liens ne doivent pas avoir la même importance : les liens appartenant aux descendants de l'image sont les plus spécifiques à cette image et par conséquent, ils doivent avoir les poids les plus élevés. La figure 2 représente une région d'image où les nœuds L et K sont des nœuds contenants des liens vers d'autres régions. Dans cet exemple, le lien du nœud L doit avoir plus d'importance que le lien du nœud K.

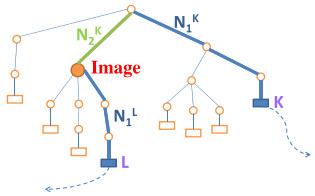

Figure 2. *Relation entre les nœuds contenant des liens et l'image*

Pour calculer l'importance d'un lien l pour une image im, nous proposons la formule suivante :

$$Imp(l, im) \quad = \quad \frac{1}{N_1 + (e^{N_2} - 1)} \qquad [2]$$

où N_1 est le nombre d'arcs qui séparent le nœud contenant le lien l de son ancêtre commun avec le nœud image im. N_2 est le nombre d'arcs qui séparent le nœud image im de son ancêtre commun avec le nœud contenant le lien l.

Dans l'exemple de la figure 2, la distance N_2 pour le lien du nœud L est égale à 0. On a recours à une fonction exponentielle pour la distance N_2 afin de favoriser l'importance des nœuds descendants de l'image par rapport aux autres nœuds les plus éloignés. Un nœud contenant un lien peut avoir plus qu'une valeur d'importance dans le cas où il y a plus d'une image dans la même région. Nous prenons dans ce cas la plus grande valeur d'importance pour chaque lien. La valeur maximale retournée par l'équation [2] est égale à 1. Cette valeur tend vers 0 lorsque N_2 augmente.

3.4. *Calcul des scores des images issus des liens*

Après avoir calculé les poids d'importance des liens, nous pouvons adapter l'algorithme HITS à notre graphe des régions interconnectées. Un lien d'une région r_i vers une région r_j est noté par $r_i \mapsto r_j$.
Pour chaque région r_i, nous calculons deux scores *autorité* ($a(r_i)$) et *hub* ($h(r_i)$) comme suit :

$$a(r_i) = \sum_{\substack{r_j \mapsto r_i \\ l_k, im_k \in r_j}} Imp(l_k, im_k) \times h(r_j) \quad [3]$$

$$h(r_i) = \sum_{\substack{r_i \mapsto r_j \\ l_k, im_k \in r_i}} Imp(l_k, im_k) \times a(r_j) \quad [4]$$

avec $a(r_i)$ et $h(r_i)$ sont initialement égales à 1. l_k et im_k sont respectivement un lien et une image de la région r_i. Bien que l'utilisation seule des scores issus des liens nous permette de trouver des images pertinentes même si elles existent dans des régions ne contenant pas des termes de la requête, elle présente quelques limites : (1) il est possible de trouver des régions contenant des images pertinentes mais qui ne sont pas pointées par d'autres régions ; (2) une région non pertinente ayant un score *hub* élevé peut influencer les scores *autorité* des régions pointées. Pour ces raisons, nous proposons de faire une combinaison linéaire entre le score initial ($S_{initial}(im)$) d'une image et le score autorité ($a(r_i)$) de la région contenant cette image. Le score final SF d'une image im devient donc :

$$SF(im) = \alpha \times \frac{S_{initial}(im)}{Max(S_{initial})} + (1-\alpha) \times \frac{a(r_i)}{Max(a(r))}, \quad [5]$$

où α désigne un paramètre de combinaison. $Max(S_{initial})$ et $Max(a(r))$ sont utilisés pour normaliser les valeurs des scores entre 0 et 1.

4. Conclusion

Dans cet article, nous avons proposé d'exploiter les liens pour améliorer la recherche d'images dans des documents semi-structurés. Plus précisément, nous avons adapté l'algorithme de *HITS*, appliquée au niveau régions, en ajoutant un poids d'importance des liens dans la région, calculé en fonction de la position hiérarchique de ces liens par rapport l'élément image. L'utilisation de notre approche peut nous permettre de retrouver des nouveaux éléments multimédia et de réordonner les résultats d'une façon plus appropriée. Comme perspectives, nous envisageons évaluer notre proposition dans le cadre de la campagne d'évaluation INEX 2006 et 2007, tâche multimédia.

5. Remerciement

Nous tiendrons à remercier Mr. Maher Ben Jemaa, Maitre assistant à l'ENIS, pour son aide et ses conseils durant ce travail de recherche.

6. Bibliographie

Brin J., Page L., « The anatomy of a large-scale hypertextual web search engine », *in Proc. 9th Annual ACM-SIAM Symposium Discrete Algorithms*, vol. 30, n° 1-7, p. 107-117, 1998.

Cai D., He X., Wen J.-R., Ma W.-Y., Zhang H.-J., « Clustering and Searching WWW Images Using Link and Page Layout Analysis », *ACM Transactions on Multimedia Computing, Communications and Applications*, 2007.

Calado P., Cristo M., Gonçalves M. A., de Moura E. S., Ribeiro-Neto B., Ziviani N., « Link-Based Similarity Measures for the Classification of Web Documents », *American Society for Information Science and Technology*, vol. 57, n° 2, p. 208-221, 2006.

Kessler M. M., « Bibliographic coupling between scientific papers », *American Documentation*, vol. 14, n° 1, p. 10-25, 1963.

Kleinberg J. M., « Authoritative sources in a hyperlinked environment », *WWW7 / Computer Networks*p. 668-677, janvier, 1998.

Kong Z., Lalmas M., « Integrating XLink and XPath to Retrieve Structured Multimedia Documents in Digital Libraries », *RIAO*, p. 571-581, 2004.

Lempel R., Soffer A., « PicASHOW : Pictorial Authority Search by Hyperlinks On the Web », *ACM Transactions on Information Systems*, vol. 20, n° 1, p. 1-24, 2002.

Small H. G., « Co-citation in the scientific literature : A new measure of relationship between two documents », *Journal of the American Society for Information Science*, vol. 24, n° 4, p. 265-269, 1973.

Torjmen M., Approches de Recherche Multimedia dans des Documents Semi-Structurés : Utilisation du contexte textuel et structurel pour la sélection d'objets multimedia, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 2009.

Voutsakis E., Petrakis E. G., Milios E., « Weighted Link Analysis for Logo and Trademark Image Retrieval on the Web », *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.