Combinaison d'information visuelle et textuelle pour la recherche d'information multimédia

Cédric Lemaître, Christophe Moulin, Cécile Barat, Christophe Ducottet
Université de Lyon, F-69003, Lyon, France
Université de Saint-Étienne, F-42000, Saint-Étienne, France
CNRS UMR5516, Laboratoire Hubert Curien
Christophe.Moulin@univ-st-etienne.fr, Cecile.Barat@univ-st-etienne.fr
ducottet@univ-st-etienne.fr, Cedric.Lemaitre@univ-st-etienne.fr

Résumé – Nous présentons dans cet article un modèle de représentation de documents multimédia combinant des informations textuelles et des descripteurs visuels. Le texte et l'image composant un document sont chacun décrits par un vecteur de poids $tf.idf$ en suivant une approche "sac-de-mots". Le modèle utilisé permet d'effectuer des requêtes multimédia pour la recherche d'information. Notre méthode est évaluée sur la base imageCLEF'08 pour laquelle nous possédons la vérité de terrain. Plusieurs expérimentations ont été menées avec différents descripteurs et plusieurs combinaisons de modalités. L'analyse des résultats montre qu'un modèle de document multimédia permet d'augmenter les performances d'un système de recherche basé uniquement sur une seule modalité, qu'elle soit textuelle ou visuelle.

Abstract – This paper presents an approach to model the content of multimedia documents using textual information combined with image features. Text and image are processed separately using a same bag-of-words approach and a $tf.idf$ weighting scheme. Two vectors of textual and visual terms are obtained and linearly combined. The proposed model allows to perform multimedia queries in a retrieval process. We evaluate it on the ImageCLEF'08 collection of about 150'000 Wikipedia documents. Experiments with different image descriptors and different modalities combinations are studied. Results prove that a multimedia model outperforms a text only one, which encourages to use multiple modalities rather than a single one.

1 Introduction

Au cours des dernières années, le développement des systèmes de communication a entraîné une explosion du nombre de collections multimédia. La richesse et la diversité de ces collections rend l'accès à l'information utile de plus en plus difficile. Il devient essentiel de développer des méthodes d'indexation et de recherche adaptées à la nature des documents prenant en compte les différentes modalités contenues dans les documents (texte, image, vidéo, etc.). Des collections standards comme TREC, ImagEval ou encore ImageCLEF, pour lesquelles une "vérité-de-terrain" existe, permettent d'évaluer de tels systèmes.

Le plus grand nombre des systèmes traitant des documents multimédia n'exploite que la partie textuelle de ces documents. L'approche standard consiste à représenter un texte sous forme de "sac-de-mots" [1] et à associer à chaque mot un poids caractérisant sa fréquence par la méthode $tf.idf$. L'enjeu aujourd'hui est d'étendre les systèmes aux autres modalités, notamment aux images puisqu'elles sont très présentes au sein des collections multimédia. Une démarche naturelle consiste à utiliser la même représentation à base de "sac-de-mots" afin de modéliser l'image. Cette approche a déjà montré son efficacité notamment pour des applications d'annotations d'images ou de recherche d'objets au sein de grandes collections [2].

Nous proposons ici un modèle de représentation de documents multimédia qui combine à la fois des informations visuelles et textuelles contenues dans le document. La pertinence de notre modèle est évaluée sur une tâche de recherche d'information. Il s'agit de comparer les résultats obtenus avec notre modèle à ceux obtenus avec une seule modalité, textuelle ou visuelle. Cette démarche fait suite à nos travaux effectués dans le cadre de la compétition ImageCLEF'08 [3].

Dans cet article, nous présentons tout d'abord notre modèle de représentation de document multimédia et son utilisation dans le contexte de la recherche d'information. Ensuite, nous montrons des résultats d'application obtenus sur la collection ImageCLEF'08.

2 Modèle de représentation des documents multimédia

Le modèle de document multimédia proposé consiste à décrire le texte et les images à l'aide de termes textuels et visuels. Les deux modalités sont d'abord traitées séparément en utilisant pour chacune l'approche "sac-de-mots". Elles sont alors représentées sous forme d'un vecteur de poids $tf.idf$ caractérisant la fréquence de chacun des mots visuels ou textuels. Utiliser un même mode de représentation pour les deux modalités permet de les combiner par une méthode de fusion tardive et d'effectuer ensuite des requêtes multimédia pour retrouver de

l'information. Cette méthodologie générale est présentée à la figure 1.

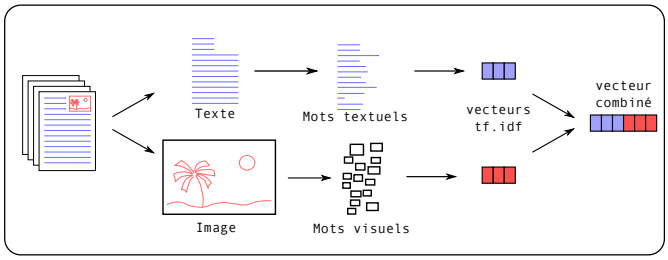

FIG. 1 – Représentation d'un document multimédia

2.1 Représentation de la modalité textuelle

Pour représenter un document textuel sous la forme d'un vecteur de poids, il est tout d'abord nécessaire de définir un index de termes textuels ou vocabulaire. Pour cela, nous appliquons d'abord une lemmatisation de Porter et une suppression des mots vides à l'ensemble des documents. L'indexation est ensuite réalisée au moyen du logiciel Lemur[1]. Chaque document est alors représenté, suivant le modèle de Salton [1], comme un vecteur de poids $\mathbf{d_i^T} = (w_{i,1}, ..., w_{i,j}, ..., w_{i,|T|})$ où $w_{i,j}$ représente le poids du terme t_j au sein d'un document d_i. Ce poids est calculé comme le produit des deux facteurs $tf_{i,j}$ et idf_j. Le facteur $tf_{i,j}$ correspond à la fréquence d'apparition du terme t_j dans le document d_i et le facteur idf_j mesure l'inverse de la fréquence du terme dans l'ensemble du corpus. Ainsi, le poids $w_{i,j}$ est d'autant plus élevé que le terme t_j est fréquent dans le document d_i et rare dans l'ensemble du corpus.

Pour le calcul des facteurs tf et idf, nous utilisons les formulations définies par Robertson [4] :

$$tf_{i,j} = \frac{k_1 n_{i,j}}{n_{i,j} + k_2(1 - b + b\frac{|d_i|}{d_{avg}})}$$

où $n_{i,j}$ est le nombre d'occurrences du terme t_j dans le document d_i, $|d_i|$ est la taille du document d_i et d_{avg} est la taille moyenne de tous les documents du corpus. k_1, k_2 et b sont trois constantes qui prennent les valeurs respectives 1, 1 et 0.5.

$$idf_j = \log \frac{|D| - |\{d_i | t_j \in d_i\}| + 0.5}{|\{d_i | t_j \in d_i\}| + 0.5}$$

où $|D|$ est la taille du corpus et $|\{d_i | t_j \in d_i\}|$ est le nombre de documents du corpus où le terme t_j apparaît au moins une fois.

Une requête textuelle q_k pouvant être considérée comme un document texte très court, elle peut, elle aussi, être représentée par un vecteur de poids. Ce vecteur, noté $\mathbf{q_k^T}$, sera calculé avec les formules de Roberson mais avec $b = 0$.

Pour calculer le score de pertinence d'un document d_i vis à vis d'une requête q_k, nous appliquons la formule donnée par Zhai dans [5] et définie par :

$$score_T(q_k, d_i) = \sum_{j=1}^{|T|} d_{i,j}^T \, q_{k,j}^T$$

2.2 Représentation de la modalité visuelle

La représentation de la modalité visuelle s'effectue en deux phases : la création d'un vocabulaire visuel et la représentation de l'image à l'aide de ce vocabulaire.

Le vocabulaire V de la modalité visuelle est obtenu en utilisant l'approche "sac-de-mots" [6]. Le processus comporte trois étapes : le choix de régions ou points d'intérêts, la description par le calcul d'un descripteur de points ou de régions et le regroupement des descripteurs en classes constituant les mots visuels. Nous utilisons deux approches différentes pour les deux premières étapes.

La première approche utilise un découpage régulier de l'image en n^2 imagettes. Ensuite un descripteur couleur de dimension 6, noté Meanstd, est obtenu pour chaque imagette, en calculant la moyenne et l'écart-type des composantes normalisées $\frac{R}{R+G+B}$, $\frac{G}{R+G+B}$ et $\frac{R+G+B}{3\times255}$ où R, G et B sont les composantes couleurs.

La seconde approche utilise la caractérisation des images par des régions d'intérêt détectées par les MSER [7] et représentées par leurs ellipses englobantes (selon la méthode proposée par [8]). Ces régions sont ensuite décrites par le descripteur de Sift [9].

Pour la troisième étape, le regroupement des classes est effectué en appliquant l'algorithme k-means sur l'ensemble des descripteurs de manière à obtenir k clusters de descripteurs. Chaque centre de cluster représente alors un mot visuel.

La représentation d'une image utilise le vocabulaire précédemment défini pour calculer un vecteur de poids $\mathbf{d_i^V}$ exactement comme pour la modalité textuelle. Pour obtenir les mots visuels de l'image, on commence par calculer les descripteurs sur les points ou les régions de cette image, puis on associe, à chaque descripteur, le mot du vocabulaire le plus proche au sens de la distance euclidienne.

2.3 Modèle de Fusion

A partir de nos deux vocabulaires T et V, nous calculons un score qui est une combinaison linéaire des scores obtenus pour chaque modalité :

$$score(q_k, d_i) = \alpha score_V(q_k, d_i) + (1 - \alpha) score_T(q_k, d_i)$$

Ce score correspond en fait à un produit scalaire entre deux vecteurs représentant respectivement la requête et le document et comportant chacun une partie textuelle et une partie visuelle. Le paramètre α permet de pondérer la quantité d'information véhiculée par chaque modalité.

[1] http://www.lemurproject.com

3 Evaluation expérimentale

3.1 Données de test et critères d'évaluation

La pertinence de notre modèle est évaluée sur la collection fournie pour la compétition ImageCLEF'08[10]. Cette collection est composée de 151519 documents extraits de Wikipedia. Chaque document est composé d'une image et d'une partie texte. Les images sont très hétérogènes en taille et en contenu. Elles peuvent correspondre à des photographies, des dessins ou des graphiques. La partie texte est relativement courte avec une moyenne de 33 mots par document.

Le but de la tâche de recherche d'information est de retourner pour les 75 requêtes fournies par ImageCLEF'08 une liste de documents pertinents. Toutes les requêtes possèdent une partie textuelle, mais plusieurs ne possèdent pas d'image requête. Afin d'avoir une partie visuelle pour chaque requête, nous utilisons les deux premières images pertinentes retournées par notre système lorsque nous utilisons la partie textuelle seule. Ceci correspondrait à un retour de pertinence fait par l'utilisateur du système. Le critère de précision moyenne (Mean Average Precision - MAP), qui est un critère classique en recherche d'information, est ensuite utilisé pour évaluer la pertinence des résultats.

3.2 Résultats et discussion

Pour montrer l'apport de l'utilisation de notre modèle multimédia par rapport à un modèle uniquement textuel ou visuel, nous avons réalisé des expérimentations utilisant une seule modalité, textuelle ou visuelle, puis des expérimentations combinant deux modalités, la modalité texte plus un descripteur visuel, ceci pour les deux descripteurs visuels Meanstd et Sift présentés précédemment. Le vocabulaire textuel est composé d'environ 200000 mots alors que les deux vocabulaires visuels en comportent 10000.

La table 1 récapitule les valeurs de MAP obtenues pour chaque expérimentation. On constate d'une part que l'utilisation de la modalité visuelle seule quel que soit le descripteur utilisé conduit à de moins bons résultats que l'utilisation de la modalité textuelle seule. D'autre part, combiner un descripteur visuel avec le texte permet d'améliorer les performances de recherche obtenues avec le descripteur textuel seul. Ces observations globales sont confirmées par les courbes de précision / rappel présentées sur la figure 2.

Une analyse détaillée par requête montre que, pour certaines, les premiers résultats retournés par la modalité visuelle sont meilleurs que pour la modalité textuelle. A titre d'illustration, les figures 3, 4 et 5 présentent les résultats obtenus pour la requête "Blue Flower".

On peut ajouter, concernant les performances obtenues avec une modalité visuelle, que le découpage régulier de l'image associé au descripteur couleur Meanstd se révèle plus robuste que l'association Mser + Sift. Nous expliquons ce comportement par des problèmes de clustering. Avec le descripteur couleur, nous travaillons avec 6 paramètres caractéristiques et 4 millions d'imagettes à regrouper en mots de vocabulaire. Avec le descripteur Sift, nous disposons de 128 paramètres caractéristiques et 54 millions d'imagettes. Dans le deuxième cas, les imagettes se répartissent de manière très irrégulière dans l'espace des descripteurs, à cause de l'utilisation des Mser, de la grande dimension et de la grande quantité de données. Cette situation est très défavorable aux algorithmes de clustering tels que kmeans[11]. Par ailleurs, il a été montré dans [12, 13] que les descripteurs des régions les plus denses de l'espace des paramètres ne sont pas nécessairement les plus informatifs.

TAB. 1 – Résultats de précision moyenne obtenus pour différentes modalités

Type de modalité	MAP
Sift	0.1057
Meanstd	0.1326
Texte	0.2554
Fusion : Texte+Sift	0.2826
Fusion : Texte+Meanstd	0.3071

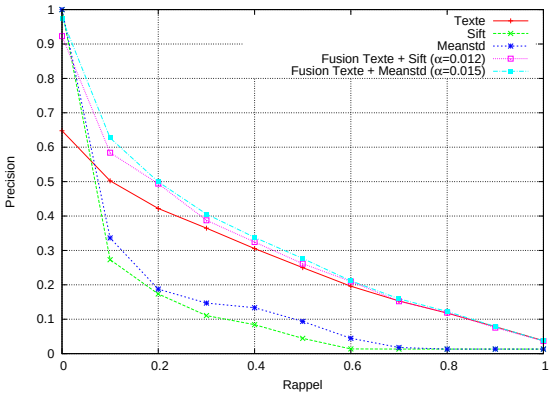

FIG. 2 – Courbes précision/rappel pour différentes modalités (texte seul, visuel seul et fusion texte/visuel) et deux descripteurs visuels

4 Conclusion

Nous avons présenté dans cet article un modèle de représentation des documents multimédia contenant à la fois des informations visuelles et des informations textuelles. Ce modèle fusionne tardivement ces deux informations représentées en utilisant une approche "sac-de-mots".

Les performances du système d'indexation et de recherche ont été validées sur la base ImageCLEF'08 pour laquelle nous possédons la vérité de terrain. Les résultats expérimentaux encouragent à l'utilisation d'un modèle multimédia, comme celui proposé, pour une tâche de recherche d'information dans une

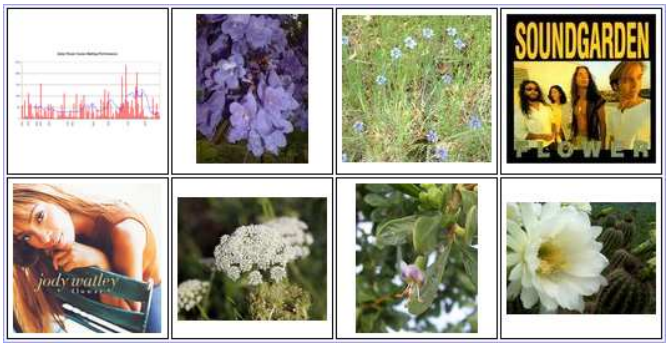

FIG. 3 – Résultats obtenus avec la modalité textuelle pour la requête "Blue Flower", les images 2 et 3 sont celles retenues pour la requête visuelle.

FIG. 4 – Résultats obtenus avec le descripteur visuel Meanstd à partir des images de la requête "Blue Flower"

collection multimédia. En effet, la fusion des deux modalités textuelle et visuelle permet à chaque fois d'accroître les performances du système. Larlus [11] propose une méthode de clustering qui permet de quantifier uniformément l'espace contrairement au kmeans qui se focalise sur les espaces denses. Cette méthode pourrait être utilisée pour améliorer notre système lors de la création du vocabulaire visuel.

Références

[1] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communication ACM*, 18(11) :613–620, 1975.

[2] J. Sivic and A. Zisserman. Video Google : A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.

[3] C. Moulin, C. Barat, M. Géry, C. Ducottet, and C. Largeron. Ujm at imageclefwiki 2008. In *Working note paper of ImageCLEFwiki 2008*, 2008.

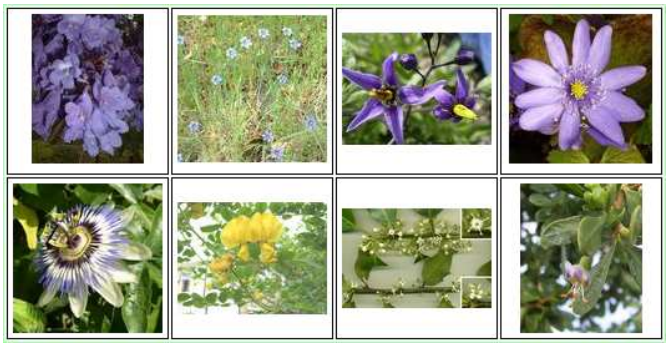

FIG. 5 – Résultats obtenus pour la fusion de modalité pour la requête "Blue Flower"

[4] S. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at trec-3. In *Text REtrieval Conference*, pages 21–30, 1994.

[5] C. Zhai. Notes on the lemur tfidf model. Technical report, Carnegie Mellon University, 2001.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV'04 workshop on Statistical Learning in Computer Vision*, pages 59–74, 2004.

[7] J. Matas, O. Chum, U Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, pages 384–393. BMVA, September 2002.

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65 :43–72, 2005.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[10] T. Tsikrika and J. Kludas. Overview of the wikipediaMM task at ImageCLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, sep 2008 (printed in 2009).

[11] Diane Larlus, Gyuri Dorkó, and Frédéric Jurie. Création de vocabulaires visuels efficaces pour la catégorisation d'images. In *Reconnaissance des Formes et Intelligence Artificielle*, 2006.

[12] F. Jurie and W. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[13] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *ICCV*, pages 281–288, 2003.