Impact de l'information visuelle pour la Recherche d'Images par le contenu et le contexte

Christophe Moulin, Christine Largeron, Mathias Géry

Université de Lyon, F-42023, Saint-Étienne, France
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
{ Christophe.Moulin, Christine.Largeron, Mathias.Gery }@univ-st-etienne.fr

RÉSUMÉ. *Les documents multimédia composés de texte et d'images sont de plus en plus présents grâce à Internet et à l'augmentation des capacités de stockage. Cet article présente un modèle de représentation de documents multimédia qui combine l'information textuelle et l'information visuelle. En utilisant une approche par sac de mot, un document composé de texte et d'image peut être décrit par des vecteurs correspondant à chaque type d'information. Pour une requête multimédia donnée, une liste de documents pertinents est retournée en combinant linéairement les résultats obtenus séparément sur chaque modalité. Le but de cet article est d'étudier l'impact, sur les résultats, du poids attribué à l'information visuelle par rapport à l'information textuelle. Des expérimentations, réalisées sur la collection multimédia ImageCLEF extraite de l'encyclopédie Wikipedia, montrent que les résultats peuvent être améliorés après une première étape d'apprentissage de ce poids.*

ABSTRACT. *Multimedia documents are increasingly used which involve to develop model to that kind of data. In this paper we present a multimedia model which combines textual and visual information. Using a bag of words approach, we can represent a textual and visual document with a vector for each modality. Given a multimedia query, our model lets us linearly combine scores obtained for each modality and return a list of relevant retrieved documents. This article aims at studying the influence of the weight given to the visual information according to the textual one. Experiments on the multimedia ImageCLEF collection extracted from Wikipedia show that results can be improved by learning this weight parameter.*

MOTS-CLÉS : *Recherche d'Images, Modèle de représentation, Mots visuels, Apprentissage*

KEYWORDS: *Images Retrieval, Document modelling, Visual words, Marchine Learning*

1. Introduction

Avec l'augmentation croissante des capacités de production, de stockage et de diffusion des documents multimédia, l'accès à l'information utile devient de plus en plus difficile. Notamment, dans le contexte du Web, l'exploitation des documents pose un vrai challenge de recherche d'information (RI) multimodale, et il devient essentiel de développer des outils d'organisation et de recherche de documents multimédia.

La plupart des systèmes de recherche d'information visant à traiter des documents multimédia développés actuellement peuvent être répartis en différentes catégories, selon qu'ils exploitent l'information textuelle, l'information visuelle ou qu'ils combinent les deux.

Dans la première catégorie, les images sont indexées uniquement à partir du texte associé à l'image (le nom du fichier, la légende, le texte entourant l'image dans le document, etc.), sans tenir compte des caractéristiques intrinsèques de l'image. On parle alors de Recherche Image basée sur le texte (*Text based Image Retrieval*). C'est le cas par exemple des outils de recherche d'images proposés par les moteurs de recherche généralistes du Web, dont les résultats ne sont pas toujours concluants, ou encore de certains moteurs du Web spécialisés dans la recherche d'images comme Picsearch[1].

Dans la seconde catégorie, seul le contenu visuel de l'image, défini par des paramètres locaux de couleur, de texture ou de forme, est utilisé. Il est alors question de recherche d'images basée sur le contenu (*CBIR : Content Based Image Retrieval*) (Smeulders *et al.*,, 2000 ; Lew *et al.*,, 2006). Par exemple, QBIC, le système précurseur d'IBM (Flickner *et al.*,, 1995), propose à l'utilisateur de retrouver des images à partir d'une requête exprimée à l'aide de ces mêmes paramètres de couleur, de texture, ou de forme. Les systèmes donnant les meilleurs résultats sont ceux dans lesquels la recherche s'effectue à l'aide d'une image exemple fournie par l'utilisateur ("Search by image", cf. par exemple QBIC ou plus récemment le moteur TinEye[2]) ou encore à l'aide d'une image construite par l'utilisateur. Ainsi, certains systèmes proposent à l'utilisateur de dessiner un croquis de l'image recherchée ("Search by sketch", cf. par exemple les moteurs Gazopa[3] et Retrievr[4]), d'autres enfin proposent à l'utilisateur de disposer sur un canevas vierge des icônes correspondant à des concepts préalablement identifiés dans la base d'images. Mais ces systèmes ont un inconvénient, car les utilisateurs ne disposent pas toujours d'une image de référence, et les langages de requêtes basés sur les caractéristiques visuelles ne sont pas toujours très intuitifs, ce qui rend leur utilisation difficile.

Enfin, la dernière catégorie regroupe les systèmes visant à traiter simultanément les descripteurs visuels et textuels. On peut citer le système PicHunter (Cox *et al.*,, 2000) qui vise à prédire l'objectif des utilisateurs en fonction de leurs actions, le système

1. Picsearch : http ://www.picsearch.com
2. TinEye : http ://www.tineye.com/
3. Gazopa : http ://www.gazopa.com/
4. Retrievr : http ://labs.systemone.at/retrievr/

Picitup[5] qui propose de définir une requête textuelle et de filtrer ensuite les résultats à l'aide d'éléments visuels (une image, une catégorie, une couleur, une forme, etc.), ou encore le système ImageRover (Sclaroff *et al.*,, 1997 ; La Cascia *et al.*,, 1998).

Les premiers travaux visant à combiner information visuelle et information textuelle se sont avérés prometteurs (Barnard *et al.*,, 2003 ; Datta *et al.*,, 2008). Cependant, comme les caractéristiques visuelles restent de bas niveau, elles ne sont pas adaptées à la recherche d'image grand public, dans la mesure où l'utilisateur rencontre souvent des difficultés pour exprimer sous cette forme son besoin d'information. Il préfère ainsi formuler ses requêtes par quelques mots clés. Pour développer des systèmes de recherche exploitant texte et image, il importe donc de réduire le fossé sémantique qui existe entre les objets et leurs représentation visuelle (Smeulders *et al.*,, 2000). Une direction de recherche, dans ce sens, consiste en l'emploi d'ontologies visuelles (Snoek *et al.*,, 2006) ; une autre proposée récemment par Tollari vise à associer des mots clés et des informations visuelles (Tollari *et al.*,, 2007 ; Tollari *et al.*,, 2009).

Les conclusions tirées de ces travaux antérieurs nous ont conduit, dans un premier temps, à proposer un système qui, à partir d'un premier ensemble de documents retournés en réponse à une requête textuelle, permet d'enrichir cette dernière en combinant information textuelle et visuelle, de façon à aboutir à un second ensemble de résultats (Moulin *et al.*,, 2008). Les termes visuels peuvent être ajoutés aux termes textuels de la requête initiale de façon entièrement automatique ou semi automatique en demandant un retour à l'utilisateur sur les premiers résultats produits. Ces premières expériences ont confirmé l'intérêt de combiner information visuelle et textuelle. Le but du travail présenté dans cet article est d'approfondir la question du poids à accorder à l'information visuelle par rapport à celui accordé à l'information textuelle. Nous proposons dans ce travail d'apprendre automatiquement ce poids à partir d'une collection d'apprentissage. Le second objectif de cet article est de vérifier dans quelle mesure le poids optimal accordé à chaque type d'information varie selon le type de requête, et dans quelle mesure le choix d'un poids spécifique à chaque requête permet d'améliorer de façon significative les résultats. En effet, comme l'a noté Tollari, pour des concepts tels que "animal" ou "véhicule", l'information visuelle est moins importante car sous ces concepts peuvent se cacher des éléments visuellement très différents.

Après avoir décrit dans la section suivante le modèle de document combinant texte et image que nous proposons, nous présentons les expérimentations réalisées dans le cadre d'une tâche de RI sur la collection ImageCLEF dans la section 3, puis les résultats de ces expérimentations dans la section 4. Les conclusions et perspectives de ce travail[6] font l'objet de la dernière section.

5. http ://www.picitup.com/picitup

6. Ce travail a été réalisé dans le cadre du projet Web Intelligence de la région Rhône-Alpes (http ://www.web-intelligence-rhone-alpes.org).

2. Modèle de Recherche d'Information multimodale

2.1. *Architecture globale*

Le modèle de recherche d'information multimodale que nous proposons comporte plusieurs modules, comme l'indique la figure 1 qui décrit son architecture globale. Le premier est consacré à l'indexation des documents de la collection D et des requêtes, composés à la fois d'une partie textuelle et d'images. Le contenu textuel, comme le contenu visuel de chaque document est décrit sous forme de sacs de mots. Le second module, vise à déterminer, pour une requête donnée, un score par document et par type de contenu. Ce score sera d'autant plus élevé que le contenu du document, relativement au type d'information considéré (visuel ou textuel), correspondra à celui de la requête. Enfin, le dernier module vise à combiner linéairement les scores obtenus pour chaque type d'information de façon à déterminer les documents répondant le mieux à la requête.

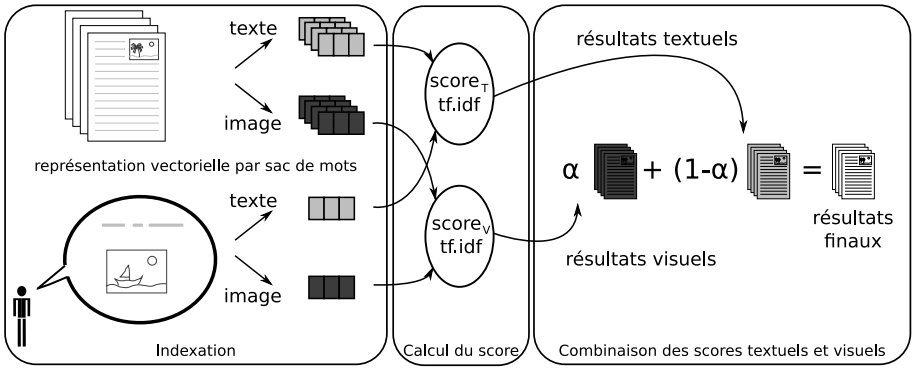

Figure 1. *Architecture globale du modèle de RI multimodale*

2.2. *Modèle de représentation textuelle*

Étant donnée une collection de documents D et $T = \{t_1, ..., t_j, ..., t_{|T|}\}$ l'ensemble des termes/mots présents dans les documents de D, chaque document d_i est représenté comme un vecteur de poids $\vec{d_i} = (w_{i,1}, ..., w_{i,j}, ..., w_{i,|T|})$ où $w_{i,j}$ est le poids du terme t_j dans le document d_i. Ce poids est calculé à l'aide d'une formule de type $tf.idf$ (cf. figure 2). La fréquence du terme (term frequency) $tf_{i,j}$ met en valeur la fréquence relative du terme t_j dans le document d_i. Il s'agira, par exemple, de la variation de la formule BM25 (Robertson *et al.*,, 1994) implantée dans le système Lemur (Zhai,, 2001) :

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

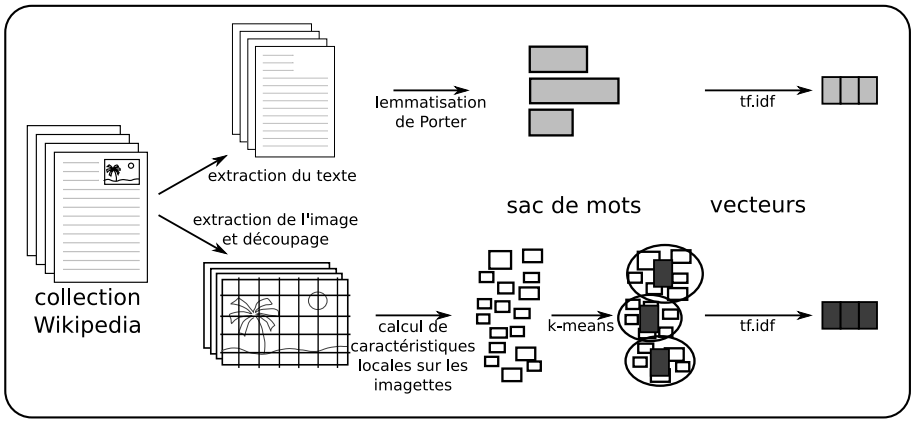

Figure 2. *Indexation basée sur des vecteurs de termes textuels et visuels.*

où b est une constante, $t_{i,j}$ est l'occurrence du terme t_j dans le document d_i, $|d_i|$ est la taille du document et d_{avg} est la taille moyenne des documents de D. La taille d'un document correspond au nombre de termes dans ce document.

La fréquence inverse de document (inverse document frequency) idf_j mesure l'importance du terme t_j sur l'ensemble de la collection D.

$$idf_j = \log \frac{|D| - df_j + 0,5}{df_j + 0,5}$$

où $|D|$ est le nombre de documents dans la collection et df_j est le nombre de documents de D dans lequel le terme t_j apparaît au moins une fois.

Le poids $w_{i,j}$ est obtenu en multipliant $tf_{i,j}$ et idf_j. Ce poids aura une valeur d'autant plus élevée que le t_j apparait fréquemment dans le d_i, mais peu dans les autres documents de D.

Soit q_k une requête composée de quelques termes. Si l'on considère cette dernière comme un court document, q_k peut alors être représentée sous la forme d'un vecteur de poids. Afin de trier les documents par ordre de pertinence pour une requête donnée, un score est calculé par :

$$score_T(q_k, d_i) = \sum_{t_j \in q_k} tf_{i,j} idf_j tf_{k,j} idf_j$$

2.3. *Modèle de représentation visuelle*

Le contenu visuel d'un document est représenté, comme son contenu textuel, sous forme vectorielle. Ceci nécessite la définition d'un vocabulaire visuel $V =$

$\{v_1, ..., v_j, ..., v_{|V|}\}$ utilisé, comme pour la partie textuelle, pour associer à chaque image un vecteur de poids comportant autant de composantes qu'il y a de mots visuels dans V.

La construction du vocabulaire V, réalisée en utilisant l'approche par sac de mots (Csurka *et al.*,, 2004), s'effectue en deux étapes (cf. figure 2). Dans la première, chaque image de la collection D est découpée en imagettes, décrites à l'aide d'un descripteur visuel. Le découpage utilisé, est un découpage régulier en 16×16 imagettes. Un minimum de 8×8 pixels est imposé pour définir une imagette. Le descripteur retenu pour décrire chaque imagette est le descripteur SIFT (Lowe,, 2004). D'autres descripteurs ont été utilisés pour décrire les images (Moulin *et al.*,, 2008), mais seul celui qui donne les résultats les plus intéressants est présenté dans cet article. L'étape suivante consiste à classer automatiquement, à l'aide de l'algorithme des k-means, les vecteurs de descripteurs visuels associés aux imagettes figurant dans les documents de la collection. Les centres des classes sont retenus comme mots du vocabulaire visuel. On peut remarquer que cette étape est assimilable à la constitution de l'index sur la partie textuelle. Sur la partie visuelle, elle permet de passer de paramètres visuels de bas niveau (couleur, texture) à un niveau de représentation plus général sous forme de mots visuels, caractéristiques par exemple d'une texture ou d'une couleur, mais qui n'ont pas encore un pouvoir d'expression comparable à celui du vocabulaire textuel.

Pour représenter une image, associée à un document ou à une requête, il suffit de la découper en 16×16 imagettes, de décrire ces imagettes à l'aide du descripteur SIFT et d'associer chacune au mot visuel de V le plus proche en terme de distance euclidienne. Ainsi, de même qu'il est possible de dénombrer le nombre d'occurrences $t_{i,j}$ d'un mot textuel t_j dans un document d_i, il est possible de dénombrer le nombre d'occurrence $v_{i,j}$ d'un mot visuel v_j : il s'agira du nombre d'imagettes de l'image affectées à la classe ayant le mot visuel v_j comme centre. Ainsi, comme dans le modèle textuel, une image est représentée à l'aide d'un vecteur de poids calculés à l'aide des formules $tf.idf$ en remplaçant $t_{i,j}$ par $v_{i,j}$ et t_j par v_j.

De même que pour la partie textuelle, il est possible de calculer un score de pertinence $score_V(q_k, d_i)$ d'un document d_i pour une requête image donnée q_k en considérant les termes visuels v_j figurant dans cette requête et dans les images du document par :

$$score_V(q_k, d_i) = \sum_{v_j \in q_k} tf_{i,j} idf_j * tf_{k,j} idf_j$$

2.4. *Combinaison des informations textuelles et visuelles*

Le score global d'un document d_i pour une requête donnée q_k est calculé simplement en combinant linéairement les scores obtenus respectivement sur les modalités visuelles et textuelles :

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

où α est un paramètre permettant d'accorder plus ou moins d'importance à l'information visuelle par rapport à l'information textuelle.

3. Expérimentations

Afin d'évaluer notre modèle de représentation de documents multimédia, nous avons effectué plusieurs expérimentations sur la collection ImageCLEF (Tsikrika *et al.*,, 2008 ; Tsikrika *et al.*,, 2009). Le but est d'évaluer l'impact de la prise en compte de l'information visuelle sur les résultats d'un SRI multimodale. Cela nécessite d'étudier l'influence du paramètre de fusion α. Après une description de la collection et du protocole expérimental, nous analyserons les résultats obtenus.

3.1. *Description de la collection*

La collection ImageCLEF est une collection multimédia composée de 151 519 documents XML provenant de l'encyclopédie Wikipedia. Les documents sont composés d'une seule image accompagnée d'un texte court. Les images sont de tailles hétérogènes aux formats JPEG et PNG. Elles peuvent correspondre aussi bien à des photos, qu'à des dessins ou des peintures. Le texte court décrit généralement l'image, mais peut également contenir des informations relatives à l'utilisateur qui a fourni l'image ou sur les droits d'utilisation de cette dernière. Les principales caractéristiques de la collection qui a été utilisée dans le cadre de la compétition ImageCLEF 2008 et 2009 (Tsikrika *et al.*,, 2008 ; Tsikrika *et al.*,, 2009) sont présentées dans la table 1.

Tableau 1. *Collection ImageCLEF 2008 et 2009*

	2008	2009
Nombre de documents	151 519	
Nombre moyen de mots textuels par document	33	
Nombre de requêtes	75	45
Nombre moyen d'images par requête	1,97	1,84
Nombre moyen de mots textuels par requête	2,64	2,93

Chaque année un ensemble différent de requêtes a été fourni :

– Le premier correspond aux 75 requêtes fournies par la compétition ImageCLEF 2008. Toutes les requêtes 2008 possèdent une partie textuelle composée de quelques mots, mais ne possèdent pas forcément de partie visuelle. Pour chaque requête les 2 premières images pertinentes retournées par une interrogation sur la base de la modalité texte ont été utilisées pour définir la partie visuelle. Ceci correspondrait à un retour de pertinence utilisateur. Ce jeu de requête sera utilisé comme un échantillon d'apprentissage pour calculer les paramètres de notre modèle.

– Le deuxième ensemble est composé des 45 requêtes de la compétition ImageCLEF 2009. Chaque requête 2009 est composée d'une partie textuelle et d'une partie visuelle sous la forme d'environ 2 images par requête.

3.2. *Mesures d'évaluation*

Nous avons utilisé plusieurs mesures classiques (MAP, $P10$, $iP[0,1]$) pour évaluer les performances de notre SRI multimodale. Soient $Q = \{q_1, ..., q_k, ..., q_{|Q|}\}$ l'ensemble des requêtes d'une collection et $D_k = \{d_{k,1}, ..., d_{k,i}, ...d_{k,|D_k|}\}$ l'ensemble des documents pertinents pour une requête q_k. La liste des N_k documents retrouvés par le système pour la requête q_k est une liste triée par score de pertinence. Pour la compétition ImageCLEF, N_k est limité à 1000 documents. Le rang r correspondra au r^e document retrouvé par le système parmi les N_k documents.

La précision $P_k(N)$ correspond à la proportion de documents pertinents pour la requête q_k parmi les N premiers documents retrouvés triés par score de pertinence. Le rappel $R_k(N)$ correspond à la proportion de documents pertinents retrouvés pour la requête q_k sur le nombre de documents pertinents à retrouver.

$$P_k(N) = \frac{\sum_{r=1}^{N} \mathrm{rel}_k(r)}{N} \qquad R_k(N) = \frac{\sum_{r=1}^{N} \mathrm{rel}_k(r)}{|D_k|}$$

où $\mathrm{rel}_k(r)$ est une fonction binaire de pertinence qui vaut 1 si le document au rang r est pertinent pour la requête q_k, 0 sinon.

La précision moyenne AP_k est calculée par :

$$AP_k = \frac{\sum_{r=1}^{N_k} (P_k(r) \times \mathrm{rel}_k(r))}{|D_k|}$$

Pour évaluer les performances de notre modèle, nous utilisons 3 mesures d'évaluation. La première (MAP : Mean Average Precision) correspond à la moyenne sur l'ensemble des requêtes de la précision moyenne AP_k. La seconde ($P10$) correspond à la précision pour 10 documents retournés. La dernière ($iP[0,1]$) correspond à la précision au rappel de point 0,1.

$$MAP = \frac{\sum_{k=1}^{|Q|} AP_k}{|Q|} \quad P10 = \frac{\sum_{k=1}^{|Q|} P_k(10)}{|Q|} \quad iP[0,1] = \frac{\sum_{k=1}^{|Q|} iP_k[0,1]}{|Q|}$$

Avec :

$$iP_k[0,1] = \begin{cases} \max_{1 \leq r \leq N_k}(P_k(r) | R_k(r) \geq 0,1) & \text{si } 0,1 \leq R_k(N_k) \\ 0 & \text{sinon} \end{cases}$$

3.3. *Protocole expérimental*

Afin d'évaluer l'impact de la prise en compte de l'information visuelle sur une tâche de recherche d'information, et par conséquent, étudier l'influence du paramètre α, plusieurs expérimentations ont été effectuées.

3.3.1. *Apprentissage du paramètre α*

Dans un premier temps, les requêtes de la collection 2008 on été utilisées comme une base d'apprentissage afin de calculer α_g^{2008} (resp. α_g^{2009}), la valeur de α qui optimise globalement la qualité les résultats sur ImageCLEF 2008 (resp. ImageCLEF 2009). La mesure d'évaluation considérée est le MAP, qui est la mesure principale de la compétition ImageCLEF. Le paramètre α_g^{2008} a ensuite été utilisé pour paramétrer le système et traiter toutes les requêtes de la collection 2009 dans le cadre de la compétition ImageCLEF 2009. Une des premières questions à laquelle les expérimentations doivent permettre de répondre concerne la possibilité d'apprendre efficacement le paramètre du modèle sur un ensemble de requêtes pour traiter d'autres requêtes. En d'autres termes : est-ce qu'il est possible d'approcher la valeur optimale α_g^{2009} par apprentissage sur 2008. Une comparaison de α_g^{2008} par rapport à α_g^{2009} permettra de conclure sur la pertinence d'un tel apprentissage.

3.3.2. *Stabilité du paramètre α_g par rapport à la mesure d'évaluation*

Le deuxième objectif de nos expérimentations est de vérifier si l'importance de l'information visuelle par rapport à l'information textuelle reste la même selon que l'on privillégie une recherche exhaustive retournant un nombre important de résultats, pas nécessairement tous pertinents, ou au contraire une recherche précise renvoyant moins de résultats, mais avec moins d'erreurs. Pour ce faire, nous nous proposons d'étudier la variation des valeurs de α_g en fonction de différents critères d'évaluation à savoir les mesures $P10$ et $iP[0,1]$, qui mettent l'accent sur la précision d'une part, et la mesure MAP qui donne aussi de l'importance au rappel, d'autre part.

3.3.3. *Optimisation du paramètre α par rapport à la requête*

Troisièmement, nous avons expérimenté notre modèle en fonction du type de requêtes traitées. En effet, certaines requêtes ont un profil plutôt textuel (par exemple : *"people with dogs"*, *"street musician"*), et d'autres ont un profil plus visuel (par exemple : *"red fruit"*, *"real rainbow"*). Il est donc intéressant d'observer le comportement du système requête par requête.

Cette approche, dite locale, vise à calculer α_k, la valeur de α optimisée pour une requête q_k. La valeur moyenne et l'écart-type des différentes valeurs de α_k nous permettront de conclure sur la variation du paramètre α en fonction des requêtes, et sur l'intérêt d'explorer des méthodes pour estimer la valeur α_k optimale pour une nouvelle requête soumise par l'utilisateur.

Ici aussi, il est intéressant d'étudier l'optimisation de α en fonction de la mesure d'évaluation considérée, et donc de calculer les valeurs de α_k optimisées selon différentes mesures (MAP, $P10$ et $iP[0,1]$).

3.3.4. *Approche globale vs locale*

Dans l'approche globale, on étudie, les variations, en fonction de la valeur du paramètre α, de la mesure MAP_α (resp. $P10_\alpha$, $iP[0,1]_\alpha$) afin de l'optimiser. On note α_g et on appelle valeur optimale globale, la valeur du paramètre qui maximise MAP_α (resp. $P10_\alpha$, $iP[0,1]_\alpha$) :

$$\alpha_g = \alpha | MAP_{\alpha_g} = max\{MAP_\alpha, \alpha \in [0,1]\}$$

α_g est ensuite utilisé pour toutes les requêtes. Dans le cadre de la compétition ImageCLEF, la valeur α_g a été calculé sur l'ensemble des requêtes de la collection 2008 et a ensuite été utilisé pour les requêtes de la collection 2009.

Dans l'approche locale, on souhaite optimiser les mesures AP_k, $P_k(10)$ et $iP_k[0,1]$ pour chaque requête q_k. À chaque requête q_k correspond une valeur α_k qui optimise la mesure AP_k. La mesure MAP_{α_l} correspond à la moyenne des AP_k optimisée et est calculée par :

$$MAP_{\alpha_l} = \frac{\sum_{k=1}^{|Q|} AP_k | \alpha = \alpha_k}{|Q|}$$

Il est alors possible de calculer la moyenne (μ_{α_l}) et l'écart-type (σ_{α_l}) des différentes valeurs optimisées de α_k par requête.

3.4. *Paramétrage du système*

Le paramétrage du système a consisté en l'utilisation des paramètres par défaut du logiciel Lemur (Zhai,, 2001). La valeur par défaut du paramètre k_1 de la fonction de pondération BM25 d'un terme dans un document ou une requête est 1. Étant donné que $|d_k|$ et d_{avg} ne sont pas définis pour une requête q_k, pour le calcul de $tf_{k,j}$, le paramètre b est défini à 0. Pour le $tf_{i,j}$ d'un document d_i et d'un terme t_j, ce paramètre b est fixé à $0,5$. Par ailleurs, aucun anti-dictionnaire n'a été utilisé, et une lemmatisation des mots est réalisée à l'aide de l'algorithme de Porter (Porter,, 1980).

4. Résultats

4.1. *Apprentissage du paramètre* α

Le MAP est une mesure globale qui correspond à la moyenne des précisions moyennes obtenues pour chaque requête. Cette mesure est utilisée pour classer les

résultats des participants d'ImageCLEF. La table 2 résume les différents résultats obtenus pour le MAP en fonction des modalités (texte/image) utilisées et de la méthode d'optimisation.

Tableau 2. *Résultats sur la collection ImageCLEF 2009 (mesure MAP)*

Run	MAP	Gain / texte seul
Texte seul	0,1667	
Visuel seul	0,0085	-94,90%
Texte+visuel (α_g^{2008})	0,1903	+14,16%
Texte+visuel (α_g^{2009})	0,1905	+14,28%

Selon la mesure MAP, l'utilisation de la modalité visuelle seule conduit sans surprise, à de mauvais résultats (MAP de 0,0085) par rapport à ceux obtenus en utilisant seulement le texte (MAP de 0,1667). Comme le montre la figure 3, augmenter l'importance de la partie visuelle permet d'améliorer significativement les résultats, notamment en prenant une valeur de α proche de 0,1. Il ne faut pas cependant lui donner trop d'importance, en prenant une valeur de α supérieure à 0,1, au risque de les dégrader. Notons cependant que les valeurs de α n'étant pas normalisées, il est difficile de les interpréter directement. Seule l'amélioration mise en évidence par les mesures d'évaluation permet de juger l'apport de l'information visuelle.

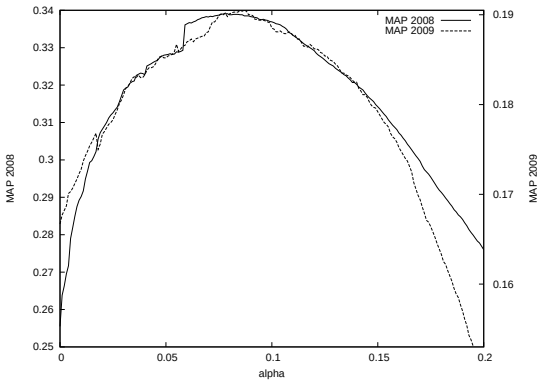

Figure 3. *Variation de la mesure MAP en fonction du paramètre α pour 2008 et 2009.*

Le paramètre α_g^{2008} calculé sur la collection d'apprentissage 2008 a amélioré les résultats obtenus avec le texte seul sur 2009 de 14,16% (MAP de 0,1903), ce qui est très satisfaisant en comparaison avec les résultats optimaux (MAP de 0,1905), obtenus avec une valeur α_g^{2009} optimisée sur la collection de test 2009 elle même. Les courbes similaires du MAP en fonction de α et les valeurs de $\alpha_g^{2008} = 0,084$ et $\alpha_g^{2009} = 0,085$ montrent pour la mesure MAP une bonne robustesse du paramètre α_g au changement de collection. L'apprentissage de α_g est donc envisageable.

4.2. *Stabilité du paramètre α_g par rapport à la mesure d'évaluation*

Pour des mesures d'évaluation plus spécifiques comme les mesures orientées précision $P10$ et $iP[0,1]$, le paramètre α semble moins stable que pour la mesure MAP, surtout sur la collection 2009, comme le montrent les figure 4 et 5 (rappelons que $P10$ et $iP[0,1]$ sont des moyennes, alors que MAP est une moyenne de moyenne).

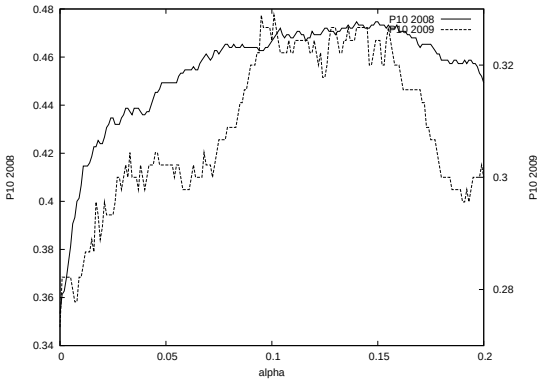

Figure 4. *Variation de la mesure $P10$ en fonction du paramètre α pour 2008 et 2009.*

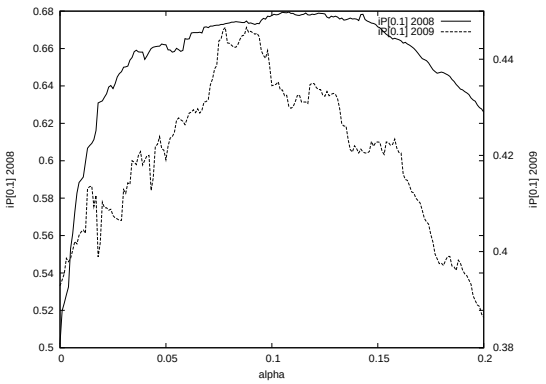

Figure 5. *Variation de la mesure $iP[0,1]$ en fonction du paramètre α pour 2008 et 2009.*

Pour ces deux mesures, la valeur du paramètre α apprise sur 2008 ($P10$: α_g^{2008} = 0,140 ; $iP[0,1]$: α_g^{2008} = 0,108) diffère du paramètre α optimal pour 2009 ($P10$: α_g^{2009} = 0,095 ; $iP[0,1]$: α_g^{2009} = 0,078).

Néanmoins, la pondération de l'information visuelle par le biais du paramètre α_g^{2008}, même relativement différent de la valeur optimale α_g^{2009}, permet tout de même d'améliorer significativement les résultats pour la mesure $P10$ comme pour la mesure

$iP[0,1]$, comme le montrent les tables 3 et 4. En effet, pour la mesure $P10$, on observe une amélioration de 19,54% et pour la mesure $iP[0,1]$ une augmentation de 9,49%.

Tableau 3. *Résultats sur la collection ImageCLEF 2009 (mesure $P10$)*

Run	$P10$	Gain / texte seul
Texte seul	0,2733	
Visuel seul	0,0178	-93,49%
Texte+visuel (α_g^{2008})	0,3267	+19,54%
Texte+visuel (α_g^{2009})	0,3289	+20,34%

Tableau 4. *Résultats sur la collection ImageCLEF 2009 (mesure $iP[0,1]$)*

Run	$iP[0,1]$	Gain / texte seul
Texte seul	0,3929	
Visuel seul	0,0160	-95,93%
Texte+visuel (α_g^{2008})	0,4302	+9,49%
Texte+visuel (α_g^{2009})	0,4466	+13,67%

4.3. *Approche globale vs locale : optimisation de α par rapport à la requête*

L'utilisation d'un paramètre α_k spécifique pour chaque requête q_k est une approche plus difficile à mettre en place que l'approche globale, car elle nécessite de déterminer a priori la valeur de α_k pour chaque nouvelle requête ; ce qui reste un problème ouvert.

Toutefois, dans l'hypothèse où ce verrou serait levé, elle offre une marge de progression très importante avec une amélioration potentielle de 29,99% (resp. 52,87%, 39,14%) pour la mesure MAP (resp. $P10$, $iP[0,1]$), comme le montre la table 5. Mais la mise en place de cette approche locale parait d'autant plus difficile qu'il existe une forte disparité des valeurs de α_k en fonction des requêtes comme l'indique l'écart-type σ_{α_l} observé pour les 3 mesures d'évaluation.

5. Conclusion

Dans cet article nous avons présenté un modèle de recherche d'information multimodale qui combine linéairement l'information textuelle et visuelle des documents multimédia. Ces informations sont représentées sous forme de vecteurs grâce à l'utilisation d'une approche sac de mots. Leur combinaison est basée sur un paramètre α qui permet de pondérer l'importance accordée au contenu visuel des documents par rapport au contenu textuel.

Tableau 5. *σ_{α_l} : optimisation de α_k pour chaque requête*

	Run		Gain / texte seul	μ_{α_l}	σ_{α_l}
MAP	Texte seul	0,1667			
	Texte+visuel (α_l)	0,2167	+29,99%	0,080	0,063
$P10$	Texte seul	0,2733			
	Texte+visuel (α_l)	0,4178	+52,87%	0,055	0,058
$iP[0,1]$	Texte seul	0,3929			
	Texte+visuel (α_l)	0,5467	+39,14%	0,083	0,072

Nos expérimentations montrent qu'il est possible d'apprendre avec succès une valeur α_g^{2008} de ce paramètre, à l'aide d'une collection d'apprentissage ImageCLEF 2008. La valeur apprise diffère dans certains cas de la valeur optimale α_g^{2009} (quand les mesures $P10$ et $iP[0,1]$ sont considérées), mais elle permet dans tous les cas d'améliorer significativement les résultats selon MAP, $P10$ et $iP[0,1]$.

L'utilisation d'un paramètre α_k spécifique pour chaque requête semble être une perspective intéressante qui permettrait d'améliorer encore les résultats. Pour apprendre ce paramètre, une première approche consisterait à différencier les requêtes en 3 catégories : visuelles, textuelles et neutres. L'utilisation de la longueur des requêtes textuelles peut-être envisagée car elle semble être corrélée avec la particularité des requêtes (Tsikrika *et al.,* 2009). Une autre idée serait d'analyser les mots visuels des images extraites des premiers résultats textuels en faisant l'hypothèse qu'ils sont représentatifs de la requête et que de ce fait, ils sont porteurs d'information. La distribution de ces mots visuels permettrait d'évaluer un α_k spécifique à la requête.

Par ailleurs, la fusion que nous utilisons est une combinaison linéaire plutôt simple qui reste à améliorer, même si elle donne déjà de bons résultats. D'autres méthodes de fusion sont donc à considérer comme les fusions précoces qui combinent l'information avant le calcul des résultats sur les modalités séparées ou d'autres fusions non-linéaires, ou basées sur le rang plutôt que sur le score des documents.

6. Bibliographie

Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D. M., Jordan M. I., « Matching words and pictures », *The Journal of Machine Learning Research*, vol. 3, p. 1107-1135, 2003.

Cox I. J., Miller M. L., Minka T. P., Papathomas T. V., Yianilos P. N., « The Bayesian image retrieval system, PicHunter : theory, implementation, and psychophysical experiments », *Image Processing, IEEE Transactions on Image Processing*, vol. 9, n° 1, p. 20-37, 2000.

Csurka G., Dance C., Fan L., Willamowski J., Bray C., « Visual categorization with bags of keypoints », *ECCV'04 workshop on Statistical Learning in Computer Vision*, p. 59-74, 2004.

Datta R., Joshi D., Li J., Wang J. Z., « Image retrieval : Ideas, influences, and trends of the new age », *ACM Computing Surveys*, 2008.

Flickner M., Sawhney H. S., Ashley J., Huang Q., Dom B., Gorkani M., Hafner J., Lee D., Petkovic D., Steele D., Yanker P., « Query by Image and Video Content : The QBIC System », *IEEE Computer*, vol. 28, n° 9, p. 23-32, 1995.

La Cascia M., Sethi S., Sclaroff S., « Combining textual and visual cues for content-based image retrieval on the world wide web », *CBAIVL '98 : Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, IEEE Computer Society, p. 24, 1998.

Lew M. S., Sebe N., Djeraba C., Jain R., « Content-based multimedia information retrieval : State of the art and challenges », *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 2, n° 1, p. 1-19, 2006.

Lowe D., « Distinctive image features from scale-invariant keypoints », *International Journal of Computer Vision*, vol. 60, n° 2, p. 91-110, 2004.

Moulin C., Barat C., Géry M., Ducottet C., Largeron C., « UJM at ImageCLEFwiki 2008 », *CLEF, Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, vol. 5706 of *Lecture Notes in Computer Science*, Springer, p. 779-786, 2008.

Porter M. F., « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.

Robertson S. E., Walker S., Hancock-Beaulieu M., Gull A., Lau M., « Okapi at TREC-3 », *Text REtrieval Conference*, p. 21-30, 1994.

Sclaroff S., Taycher L., La Cascia M., « ImageRover : A Content-Based Image Browser for the World Wide Web », *Proceedings. IEEE Workshop on Content-Based Access of Image and Video Libraries*, p. 2, 1997.

Smeulders A. W. M., Worring M., Santini S., Gupta A., Jain R., « Content-Based Image Retrieval at the End of the Early Years », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 12, p. 1349-1380, 2000.

Snoek C. G. M., Worring M., Gemert J. C. V., mark Geusebroek J., Smeulders A. W. M., « The challenge problem for automated detection of 101 semantic concepts in multimedia », *in Proceedings of the ACM International Conference on Multimedia*, p. 421-430, 2006.

Tollari S., Detyniecki M., Marsala C., Fakeri-Tabrizi A., Amini M.-R., Gallinari P., « Exploiting Visual Concepts to Improve Text-Based Image Retrieval », *in Proceedings of European Conference on Information Retrieval (ECIR)*, 2009.

Tollari S., Glotin H., « Web Image Retrieval on ImagEVAL : Evidences on visualness and textualness concept dependency in fusion model », *in Proceedings of the ACM International Conference on Image and Video Retrieval (ACM CIVR)*, 2007.

Tsikrika T., Kludas J., « Overview of the wikipediaMM task at ImageCLEF 2008 », *in* , C. Peters, , D. Giampiccol, , N. Ferro, , V. Petras, , J. Gonzalo, , A. Peñas, , T. Deselaers, , T. Mandl, , G. Jones, , M. Kurimo (eds), *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.

Tsikrika T., Kludas J., Overview of the wikipediaMM task at ImageCLEF 2009, Technical report, 10th Workshop of the Cross-Language Evaluation Forum, 2009.

Zhai C., Notes on the Lemur TFIDF model, Technical report, Carnegie Mellon University, 2001.