Exploitation du contenu visuel pour améliorer la recherche textuelle d'images en ligne

Sabrina Tollari — Marcin Detyniecki — Ali Fakeri-Tabrizi — Christophe Marsala — Massih-Reza Amini — Patrick Gallinari

Université Pierre et Marie Curie - Paris 6, UMR CNRS 7606 - LIP6
4 place Jussieu, 75252 PARIS cedex 05, France, prénom.nom@lip6.fr

RÉSUMÉ. *Les moteurs de recherche d'images sur le web utilisent principalement l'information textuelle associée aux images afin de retrouver les images pertinentes, tandis que le contenu visuel, moins sémantique et plus coûteux en temps de calcul, est très peu utilisé dans la phase "en ligne". Nous proposons une chaîne de traitements complète proposant deux façons efficaces et peu coûteuses d'utiliser le contenu visuel des images dans la phase en ligne. La première façon propose d'améliorer la précision des résultats retrouvés en filtrant les résultats textuels en fonction des concepts visuels détectés dans la requête textuelle. Pour cela, nous apprenons les concepts visuels à l'aide de forêts d'arbres de décision flous. Ce travail montre une nette amélioration des résultats lorsque l'on utilise les concepts apparaissant explicitement dans la requête. La deuxième façon propose d'améliorer la diversité des résultats pertinents obtenus afin de mieux satisfaire le besoin d'information de l'utilisateur. Pour cela, nous utilisons un partitionnement de l'espace visuel. Nous montrons que cette approche est effectivement efficace.*

ABSTRACT. *Web image search engines tend use the text information associated to the images to find the relevant ones. The visual content is rarely used in the "on-line" phase. We propose a complete processing chain exhibiting two efficient ways to use the visual content on the fly. The first method focuses on improving the retrieval precision by filtering text based results, using visual concepts identified in the (text) query. The visual concepts were previously learned (off-line) using Forest of Fuzzy Decision Trees. There is a clear score improvement when concepts explicitly mentioned in the query are used. The second method focuses on the diversity of the results. We propose to partition the visual space and show that it is actually effective.*

MOTS-CLÉS : *recherche d'images, concepts visuels, arbre de décision flou, diversité, ImageClef*

KEYWORDS: *images retrieval, visual concepts, fuzzy decision trees, diversity, ImageClef*

RSTI - DN – 13/2010. Recherche d'information, pages 189 à 211

1. Introduction

Les moteurs de recherche d'images sur le web utilisent principalement des informations textuelles, telles que le titre de la page web, le nom de l'image, le texte adjacent, pour tenter de « comprendre » le sens de l'image. Cependant, le texte d'une page web n'est pas toujours en rapport avec le contenu visuel de l'image. De plus, l'utilisateur préfèrant souvent exprimer son besoin d'information à l'aide de quelques mots-clés, il est difficile de trouver des liens avec l'information visuelle contenue dans les images. Il est donc intéressant de trouver des méthodes qui permettent de vérifier l'adéquation visuelle de l'image avec le texte de la requête posée par l'utilisateur. Dans (Yavlinsky *et al.*, 2006), des concepts visuels sont utilisés pour raffiner visuellement les résultats obtenus, cependant, l'utilisateur doit choisir manuellement le concept visuel à appliquer. Nous proposons dans cet article d'étudier une méthode qui permet de choisir *automatiquement* le concept visuel à appliquer. Par exemple, la figure 1 montre un utilisateur qui pose la requête *animal swimming*, notre système détecte automatiquement que la requête contient le mot *animal*, et donc qu'il doit vérifier que chaque image résultat de la requête textuel contient bien visuellement un animal.

D'un autre côté, quand un utilisateur pose une requête, ce qui l'intéresse c'est d'avoir des documents qui soient, certes tous pertinents, mais aussi qui soient les plus dissimilaires les uns des autres (Song *et al.*, 2006, Chen *et al.*, 2006, Zhai *et al.*, 2003, Deselaers *et al.*, 2009b). Par exemple, si l'utilisateur cherche des photographies d'animaux en train de nager pour illustrer un article sur la faculté de nager des animaux, toutes les images d'animaux en train de nager seront certes pertinentes, mais afin qu'il ait un aperçu, dès le début de sa recherche, de toute la diversité des images pertinentes, il serait intéressant de lui fournir dans les premiers résultats, des images qui contiennent toutes des animaux différents. Par exemple, dans la figure 1, l'utilisateur est plus intéressé par retrouver dans les deux premiers résultats un poisson rouge et un canard, plutôt que deux poissons rouges. Cette diversification des résultats selon le critère *animal* peut permettre à l'utilisateur de trouver plus rapidement ce qu'il recherche. Dans cet article, nous proposons une méthode de diversification basée sur les informations visuelles des images, et nous la comparons avec les résultats obtenus par une diversification aléatoire.

En résumé, dans cet article, nous proposons une chaîne de traitements complète utilisant deux méthodes efficaces et peu coûteuses utilisant le contenu visuel des images dans la phase en ligne. Afin d'évaluer cette chaîne de traitements, nous utilisons les corpus de deux tâches de la campagne internationale CLEF 2008. La première tâche appelée *Visual Concept Detection Task (VCDT)* (Deselaers *et al.*, 2009a) est une tâche de détection de concepts visuels. Lors de cette campagne, notre système de détection de concepts visuels est arrivé troisième sur 11 équipes. La deuxième appelée ImageCLEFphoto (Arni *et al.*, 2009) est une tâche de recherche d'images basée sur les informations textuelles et visuelles, et propose d'étudier les problèmes soulevés par la diversité.

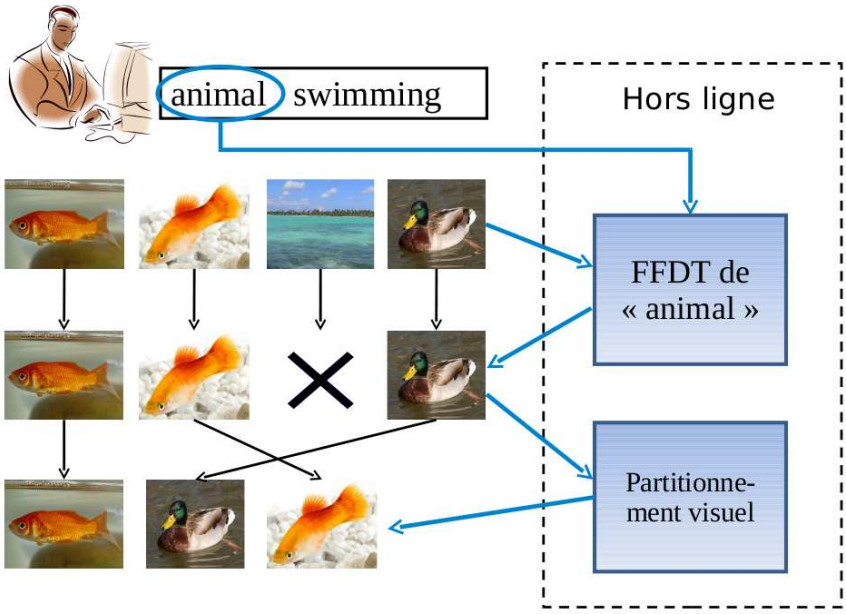


Figure 1. *Schéma de la chaîne de traitement. L'utilisateur pose une requête sous la forme de quelques mots-clés. Dans la première étape, le système détecte dans la requête un mot correspondant à un concept visuel (ici, le concept visuel est* animal*). Il applique alors un filtrage basé sur les scores de confiance fournis par la forêt d'arbres de décision (FFDT) correspondant au concept. Dans la deuxième étape, les résultats restants sont réordonnés pour être dissimilaires. Pour cela, notre système utilise un partitionnement visuel. Idéalement, les premières images obtenues doivent contenir un animal différent*

Dans la section suivante, nous détaillons la méthode de détection des concepts visuels proposée, puis nous utilisons l'analyse des cooccurrences pour détecter des relations d'exclusion et d'implication, et nous discutons les résultats obtenus par notre méthode dans la tâche VCDT. Dans la troisième section, nous expliquons comment utiliser les concepts visuels de VCDT pour améliorer la recherche d'images basée sur le texte, puis nous présentons notre méthode de diversité visuelle basée sur le partitionnement de l'espace et enfin nous discutons les résultats obtenus. Dans la dernière section, nous concluons.

2. Détection « hors ligne » de concepts visuels dans les images

Cette première partie de notre chaîne de traitements s'effectue dans la phase « hors ligne ». Elle consiste à extraire les descripteurs visuels, puis à apprendre à détecter les concepts visuels dans les images, et enfin, à améliorer les détections obtenues par analyse des cooccurrences des concepts.

2.1. *Détection de concepts visuels à l'aide de forêts d'arbres de décision*

L'annotation automatique d'images est un problème typique d'apprentissage inductif. L'apprentissage inductif consiste à passer du spécifique vers le général en construisant un modèle de connaissances complexes (comme une base de règles) à partir d'un ensemble d'exemples qui constitue une base d'apprentissage. Chaque exemple de cette base est décrit à l'aide d'un ensemble d'attributs et est associé à une classe (le concept, l'annotation proprement dite) à reconnaître.

Une des méthodes classiques dans ce domaine utilise les arbres de décision (*Decision Trees (DT)*) comme modèle de reconnaissances complexes. A partir d'une base d'apprentissage donnée, un arbre de décision est construit, de sa racine vers ses feuilles, par partitionnements successifs de l'ensemble d'apprentissage en sous-ensembles. Chaque partition est réalisée au moyen d'un test sur la valeur d'un des attributs. L'attribut le plus discriminant vis-à-vis des valeurs de la classe est sélectionné à l'aide d'une mesure d'information et est utilisé pour définir un nœud dans l'arbre. Les valeurs de cet attribut servent alors pour libeller les arcs issus de ce nœud. Le processus de construction s'arrête lorsque tous les exemples du sous-ensemble courant possèdent la même classe. Une feuille de l'arbre est alors construite et celle-ci est libellée par la valeur de cette classe (Quinlan, 1986, Marsala *et al.*, 1997). Une fois construit, un arbre de décision peut être utilisé de deux façons : pour caractériser les classes (en examinant les attributs et leurs valeurs qui ont été retenus dans l'arbre) ou pour classer de nouveaux exemples qui sont alors injectés dans l'arbre en suivant les chemins relatifs à leur propre valeur d'attribut. La feuille atteinte fournit alors la classe qui peut leur être associée.

Les arbres de décision classiques rencontrent des difficultés pour traiter des données numériques ou imprécises. L'introduction de la logique floue a permis de réduire ces difficultés. Les arbres de décision flous ont ainsi été proposés pour une meilleure prise en compte des valeurs numériques et des imprécisions des données d'apprentissage (Marsala *et al.*, 1997). Outre le fait qu'un arbre de décision flou utilise des valeurs d'attributs flous pour libeller ses arcs, il fournit aussi une classification graduelle des exemples. Lors de sa classification, un nouvel exemple sera associé non plus à une seule classe (comme dans le cas classique), mais à un ensemble de classes, chacune associée à un degré d'appartenance.

En présence de grands ensembles (en termes de dimension et de taille) de données non équilibrées, il est intéressant de combiner plusieurs arbres de décision flous, et

ainsi d'obtenir une forêt d'arbres de décision flous (*Forest of Fuzzy Decision Trees (FFDT)*) (Marsala *et al.*, 2006). De plus, pour un nouvel exemple à classer, la combinaison des résultats de classification de plusieurs arbres de décision flous offre le moyen d'obtenir un degré de confiance à chacune des classes. On peut citer aussi, par exemple, les forêts aléatoires (*random forests*) (Shotton *et al.*, 2008) qui ont aussi permis d'obtenir de bons résultats en annotation de documents.

Les forêts d'arbres de décision flous sont donc intéressantes en présence de bases d'apprentissage fortement déséquilibrées (et où le déséquilibre peut être induit par un nombre de concepts à reconnaître dans le problème étudié). Ainsi, durant la phase d'apprentissage, une forêt de n arbres de décision flous est apprise pour chaque concept. Chaque arbre F_j de la forêt est construit en utilisant un sous-ensemble d'apprentissage T_j. Chaque sous-ensemble est un ensemble équilibré constitué d'images de l'ensemble d'apprentissage sélectionnées aléatoirement (Marsala *et al.*, 2006).

Durant la phase de classification, chaque image I est classée par chaque arbre F_j de la forêt associée à un concept C. On obtient alors un degré $d_j \in [0, 1]$ pour l'image I. Ce degré représente la présence du concept C pour l'arbre F_j. Pour chaque image I, on peut donc ainsi déterminer n degrés $d_j,\ j = 1 \ldots n$. Tous les degrés sont ensuite agrégés par un vote majoritaire : $d = \sum_{j=1}^{n} d_j$ afin de déterminer si une image peut contenir le concept C selon la forêt d'arbres de décision flous construite pour la reconnaissance de ce concept. Dans notre approche, nous utilisons un seuil t tel que $t \leq n$: l'image I contient le concept C si $d \geq t$.

2.2. *Découverte de relations entre concepts par analyse des cooccurrences*

Les arbres de décision apprennent chaque concept de manière indépendante. Cependant, les concepts sont reliés entre eux. Par exemple, une scène ne peut pas être simultanément à l'intérieur (*indoor*) et à l'extérieur (*outdoor*) ; si l'on observe qu'il y a des nuages (*cloudy*), on peut en déduire qu'il y le concept ciel (*sky*). Dans cette section, nous proposons d'utiliser l'analyse des cooccurences pour déterminer automatiquement les relations entre les concepts. Une fois que nous avons découvert une relation, nous avons besoin d'une règle pour résoudre les conflits d'annotation. Cette règle doit prendre en compte les degrés de confiance donnés par les FFDTs. Par exemple, chaque image sera annotée par *outdoor* avec un certain degré et par *indoor* avec un autre degré. Cependant, les concepts *outdoor* et *indoor* ne peuvent apparaître simultanément. Pour trouver les règles à appliquer, nous étudions deux types de relations entre les concepts : les exclusions et les implications.

2.2.1. *Exclusions*

Pour découvrir automatiquement les *exclusions* entre concepts, nous étudions les concepts qui n'apparaissent jamais ensemble. Pour cela, nous calculons la matrice de cooccurrences COOC entre les concepts. Comme il peut y avoir du bruit (erreurs d'annotation), nous utilisons un seuil α pour décider quels couples de concepts n'ap-

paraissent jamais ensemble. Quand nous savons quels concepts sont reliés, nous appliquons une règle de résolution en fonction des degrés de confiance fournis par les FFDTs. Nous avons choisi une règle qui, pour les concepts mutuellement exclusifs, éliminent (c'est-à-dire donnent un degré de confiance de zéros) les étiquettes ayant le plus faible degré de confiance. Par exemple, si *outdoor* a un degré de confiance de 42/50 et *indoor* a un degré de 20/50, alors le degré de confiance de *indoor* sera mis à zéro. Pour chaque image de test, soit d(I,C) le degré de l'image I pour le concept C, nous appliquons l'algorithme suivant :

Pour chaque couple (A,B) tel que $COOC(A, B) \leq \alpha$ (*découverte*)

Si d(I,A) $<$ d(I,B) alors d(I,A)=0 sinon d(I,B)=0 (*règle de résolution*)

où COOC est la matrice de cooccurrences des concepts.

2.2.2. *Implications*

Pour découvrir les *implications*, nous étudions, par définition de l'implication, les cooccurrences entre l'absence d'un concept et la présence d'un autre concept. La matrice de cooccurrences résultante COOCNEG est asymétrique, ce qui reflète le fait qu'un concept implique un autre concept, mais cela n'est pas réciproque. La règle de résolution utilisée suppose que si un concept implique un autre concept, alors le degré de confiance de ce dernier doit être au moins égal au premier. Comme il peut y avoir du bruit, nous utilisons un seuil β pour décider quel concept implique un autre concept.

Pour chaque image de test I, soit d(I,C) le degré de l'image I pour le concept C, nous appliquons l'algorithme suivant :

Pour chaque couple (A,B) tel que $COOCNEG(A, B) \leq \beta$ (*découverte*)

d(I,B)=max(d(I,A),d(I,B)) (*règle de résolution*)

où COOCNEG est la matrice asymétrique de cooccurrences entre un concept et la négation d'un autre concept.

2.3. *Expérimentations et résultats*

2.3.1. *Corpus*

Nous appliquons notre méthode de détections de concepts visuels au corpus de la tâche *Visual Concept Detection Task (VCDT)* (Deselaers *et al.*, 2009a) de la campagne internationale d'évaluation CLEF 2008. Cette tâche correspond à un problème de classification multiclasse multi-étiquette. Le corpus de VCDT contient 1 827 images d'apprentissage et 1 000 images de test. Il y a 17 concepts visuels. Une image d'apprentissage est annotée en moyenne par 5,4 concepts (entre 0 (2 images) et 11 concepts par image). Un concept annote en moyenne 584 images d'apprentissage (entre 68 et 1 607 images d'apprentissage par concept).

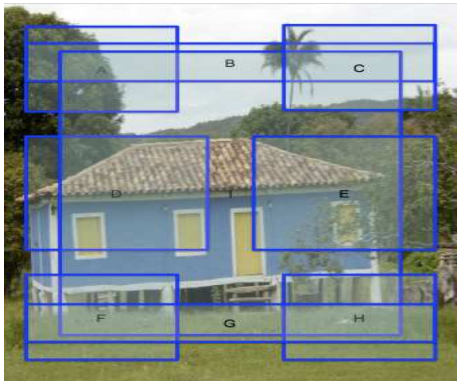


Figure 2. *Les images sont segmentées en 9 régions*

2.3.2. *Descripteurs visuels*

La couleur est généralement bien adaptée pour discriminer des concepts généraux, tels que : *indoor, outdoor, water, day, night*, c'est pourquoi nous avons choisi d'utiliser des descripteurs basés uniquement sur la couleur. Afin d'obtenir une information liée à la disposition spatiale des objets dans les images, nous segmentons les images en 9 régions qui se chevauchent (voir figure 2). Pour chaque région, nous calculons un histogramme HSV. Le nombre de dimensions de l'histogramme reflète l'importance de la région. La région centrale représente le thème de l'image. La région en haut et celle du bas sont intéressantes pour les concepts visuels généraux, tels que le ciel, le soleil, la végétation, la mer... Les autres régions sont décrites en termes de différences de couleurs entre la gauche et la droite. L'idée est de rendre explicite les symétries. En effet, les objets peuvent apparaître d'un côté ou de l'autre de l'image. Or étant donné que les arbres de décision ne sont pas capables de découvrir automatiquement ce genre de relations, l'utilisation de ces différences permet de leur donner la possibilité de tenir compte de ces symétries. Au final, chaque image est représentée par un vecteur de valeurs numériques.

2.3.3. *Mesures de performances*

Lorsque l'on crée un système de classification, on souhaite avant tout que ce système génère peu d'erreurs. Ainsi, il est préférable que parmi les documents qu'il retrouve, peu soient non pertinents, et que parmi les documents qui n'ont pas été retrouvés, il y en ait peu qui soient pertinents. Dans le premier cas, on évalue le système par son *taux de fausses acceptations* (ou *False Acceptation Rate (FAR)*) qui mesure le taux de documents non pertinents qui ont été retrouvés. Dans le second cas, on évalue le système par son *taux de faux rejets* (ou *False Reject Rate (FRR)*) qui mesure le taux

Méthode	Rang	EER	Gain
Multi-scale, regular grid, patch-based images features and a Fisher-Kernel Vector classifier	1	16.65	+51 %
Patch-based bag-of-visual words approach using a log-linear classifier	3	20.45	+40 %
FFDT	**4**	**24.55**	**+28** %
FFDT+Exclusion	11	27.37	+19 %
FFDT+Exclusion+Implication	12	27.32	+19 %
Global and local features with SVMs and random forest classifiers	24	32.09	+5 %
Moyenne des 53 runs des participants	-	33.92	-
Classifieurs aléatoires	-	50.17	-48 %

Tableau 1. *Comparaisons des méthodes utilisées, des rangs du meilleur run et des scores de quelques participants à la tâche VCDT 2008 (EER : Equal Error Rate). La moyenne des participants correspond à la moyenne des scores EER des 53 runs soumis par les 11 équipes participantes à la tâche VCDT en 2008. Le gain est calculé par rapport à la moyenne des participants. Notre run FFDT est arrivé quatrième sur 53 runs soumis (troisième équipe sur 11 équipes)*

de documents pertinents qui ont été oubliés[1]. Pour calculer ces taux de performance, il faut que le système prenne une décision sur le degré de confiance minimale qui est requis pour un document afin de pouvoir être retrouvé. Selon l'application pour laquelle le système a été créé, le seuil de décision ne sera pas le même.

Une première mesure classique d'évaluation des systèmes de classification est donnée par le *Equal Error Rate (EER)*. Cette mesure choisit pour seuil de décision le point pour lequel le FAR est égal au FRR. Les deux taux étant liés, quand le nombre de documents retrouvés augmentent alors le FAR diminue, et le FRR augmente. Cela permet de trouver un compromis entre les deux taux d'erreurs. Pour calculer ce point, on ordonne les degrés de confiance obtenus par les documents de test du plus fort degré au plus faible, puis on augmente le nombre de documents retrouvés jusqu'à obtenir un FAR égal au FRR. L'avantage de cette mesure est quelle détermine automatiquement le seuil de décision. Son inconvénient est que ce seuil n'est pas forcément adapté à l'application visée.

Une deuxième mesure d'évaluation des systèmes de classification est donnée par le *Normalized Score (NS)*. C'est une mesure qui a déjà été utilisée par de nombreux

1. Soit N le nombre d'images de l'ensemble de test, n le nombre d'images pertinentes pour le concept C, r le nombre d'images pertinentes retrouvées, et w le nombre d'images non pertinentes retrouvées, alors $FAR = \frac{w}{r+w} = 1 - précision$ est le nombre de documents non pertinents retrouvés sur le nombre total de documents retrouvés, et FRR=$\frac{n-r}{N-r-w}$ est le nombre de documents pertinents non retrouvés sur le nombre total de documents non retrouvés.

modèles d'annotation automatique (Barnard *et al.*, 2003). Ce score peut être calculé quelle que soit l'application visée. Pour un concept C donné, le *Normalized Score (NS)* peut être évalué ainsi :

$$NS = \frac{r}{n} - \frac{w}{N-n}$$

où N est le nombre d'images de l'ensemble de test, n est le nombre d'images pertinentes pour le concept C, r est le nombre d'images pertinentes retrouvées, et w est le nombre d'images non-pertinentes retrouvées. Ce score correspond donc à la somme de la *sensibilité* (appelée aussi *rappel*) et de la *spécificité* moins 1. Il varie entre -1 et 1. Le score vaut 1 quand le système ne commet aucune erreur, il vaut -1 quand toutes les images sont mal annotées, et il vaut 0 quand toutes les images sont annotées par le concept C.

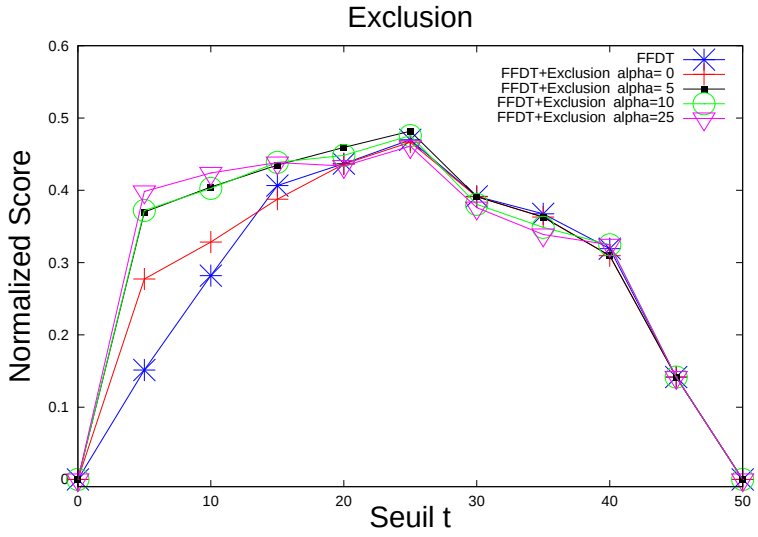
2.3.4. *Résultats obtenus lors de la campagne d'évaluation ImageCLEF*

Le tableau 1 compare les résultats obtenus en 2008 dans la tâche VCDT de la campagne d'évaluation ImageCLEF. Ces résultats sont extraits de (Arni *et al.*, 2009). Notre méthode basée sur les FFDTs est arrivée quatrième sur 53 résultats soumis (troisième équipe sur 11 équipes internationales). Les résultats obtenus montrent l'efficacité des FFDTs. Cependant, l'utilisation des règles d'exclusion et d'implication diminue les scores EER. L'EER n'étant pas une mesure adaptée pour étudier l'influence des règles d'exclusion et d'implication en fonction du seuil de décision choisie pour une application donnée, nous avons par la suite choisie d'utiliser le *Normalized Score*.

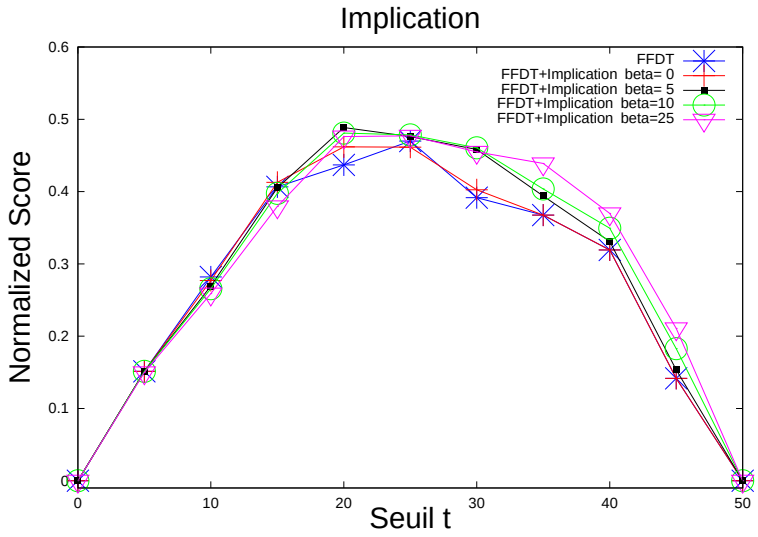
2.3.5. *Utilisation des relations d'exclusion et d'implication*

Une étape préliminaire avant d'extraire des concepts visuels est d'étudier les valeurs de cooccurrence entre les concepts pour découvrir les relations d'exclusion et d'implication. Pour les 17 concepts, il y a 136 valeurs de cooccurence. Ces valeurs varient de 0 (non-cooccurrences) à 1443 (sur les 1827 images d'apprentissage). Comme il peut y avoir du bruit, nous avions fixé, lors de notre participation à la tâche VCDT, le seuil α à la valeur 5. Ce seuil avait été déterminé grâce à la distribution des valeurs de cooccurrence de l'ensemble d'apprentissage. Si $\alpha = 5$, cela signifie que deux concepts sont considérés exclusifs si moins de 5 images sont annotées par les deux concepts. La figure 3(a) montre que cette valeur du seuil maximise en effet les résultats pour $t = 25$, mais que pour un seuil de décision $t \leq 15$, il peut être intéressant de prendre une valeur de α plus grande. De même, nous avions fixé $\beta = 5$ (un concept implique un autre concept si au maximum 5 images d'apprentissage ne sont pas annotées par le premier concept, et en même temps annotées par le second concept). La figure 3(b) confirme que les meilleurs scores à $t = 25$ sont obtenus pour $\beta = 5$, mais que pour $t \geq 30$, il peut être intéressant de prendre une valeur de β plus grande. Enfin, ces deux figures montrent que prendre $\alpha = 0$ ou $\beta = 0$ ne donne jamais de meilleurs résultats que pour $\alpha = 5$ ou $\beta = 5$, et que donc prendre en compte les erreurs d'annotations permet d'améliorer sensiblement les résultats.

Pour $\alpha = 5$ et $\beta = 5$, notre système a automatiquement découvert 25 relations d'exclusion et 12 relations d'implication (voir figure 4). Nous avons trouvé toutes



(a) Influence de α en fonction du seuil t



(b) Influence de β en fonction du seuil t

Figure 3. *Influence des paramètres α et β sur le* Normalized Score (NS) *en fonction du seuil t de décision (les degrés de confiance des arbres varient entre 0 et 50)*

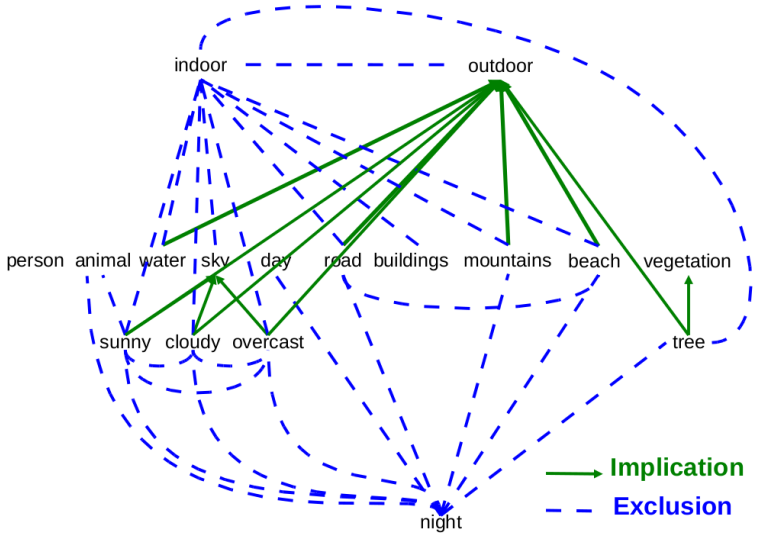


Figure 4. *Schéma montrant les relations d'exclusion et d'implication entre les concepts automatiquement découverts*

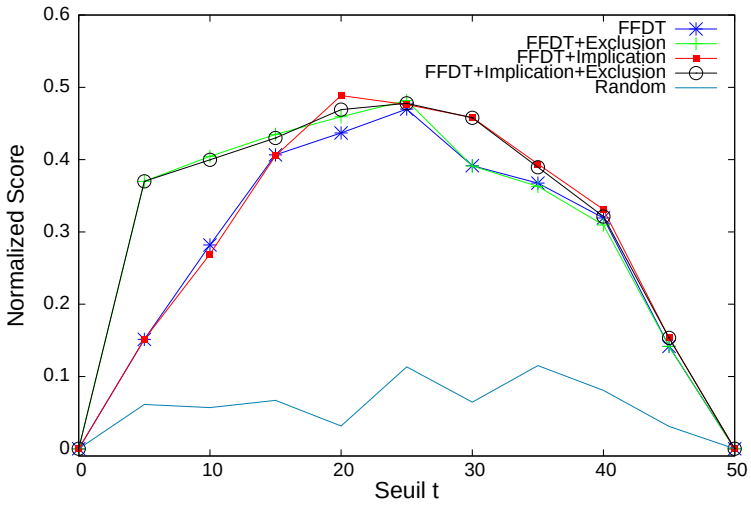


Figure 5. *Courbes de scores NS en fonction du seuil de décision (les degrés de confiance fournis par les arbres varient entre 0 et 50) avec* $\alpha = 5$ *et* $\beta = 5$

les relations évidentes (par exemple, les concepts *indoor* et *outdoor* sont exclusifs ; un arbre implique de la végétation), ainsi que quelques relations moins triviales. Par exemple, les concepts *sunny* et *animal* sont trouvés exclusifs. Cette relation peut être expliquée par le fait que, lorsqu'une personne annote une image, son attention peut se focaliser sur un objet (ou un animal) et ne pas préter attention au fait que le ciel soit ensoleillé, cette information étant jugée secondaire ou inintéressante.

Finalement, la figure 5 compare les scores NS obtenus par les FFDTs seuls ou avec les régles d'implication et d'exclusion. Pour $t = 25$, toutes les méthodes donnent quasiment les mêmes résultats. Pour ce seuil que nous utiliserions classiquement pour prendre une décision, l'intérêt d'utiliser les règles d'implication et d'exclusion n'est donc pas concluant. Cependant, nous notons que les meilleurs résultats sont obtenus pour $t = 20$ par l'application des règles d'implication aux degrés de confiance des FFDT. De plus, pour un seuil de décision $t \leq 15$, il est préférable d'utiliser les exclusions ; pour $t \geq 30$, il vaut mieux utiliser les implications. Finalement, la méthode FFDT+Implication+Exclusion donne globalement les meilleurs résultats.

3. Exploitation « en ligne » du contenu visuel lors de la recherche d'images

Dans la première partie de notre chaîne de traitement, nous avons extrait les descripteurs visuels et calculé le degré de confiance de chaque concept pour chaque image. Ces prétraitements sont les parties les plus coûteuses en temps de calcul de notre chaîne de traitement, mais ils sont effectués hors ligne. Dans la deuxième partie, nous utilisons en ligne les concepts visuels en fonction de la requête textuelle de l'utilisateur, puis nous utilisons le partitionnement de l'espace visuel pour réordonner les résultats afin d'obtenir des résultats visuellement dissimilaires.

3.1. *Utilisation de concepts visuels pour améliorer la recherche d'images basée sur le texte*

De nombreux travaux (Barnard *et al.*, 2003, Monay *et al.*, 2004, Datta *et al.*, 2008, Tollari *et al.*, 2007, Rege *et al.*, 2008, Lienhart *et al.*, 2009) montrent que combiner des informations visuelles et textuelles améliorent la recherche d'images, mais la plupart de ces travaux se concentrent sur la fusion précoce ou tardive de ces informations ou sur l'annotation des images. Quelques travaux intéressants (Yavlinsky *et al.*, 2006, Popescu *et al.*, 2007) proposent d'utiliser des concepts visuels et/ou une ontologie pour étendre les requêtes afin d'améliorer la recherche d'images, mais les modèles proposés ne sont pas toujours pleinement automatique. Nous proposons d'utiliser une méthode pour utiliser automatiquement les concepts visuels appris par les FFDTs pour améliorer la recherche d'images basée uniquement sur le texte. La difficulté principale est de déterminer comment utiliser les concepts visuels dans le cas où les seules informations que l'on peut utiliser sont le nom du concept, les descripteurs visuels, et la requête composée de quelques mots-clés.

A l'aide des FFDTs apprises pour chaque concept (voir section 2) et des descripteurs visuels de chaque image, nous pouvons donner un degré de confiance qu'un certain concept apparaisse dans une nouvelle image. Il reste donc à trouver un moyen de faire la correspondance entre la requête et le (ou les) concept(s) que l'on veut détecter dans les images.

Premièrement, si le nom du concept apparaît directement dans les mots de la requête (méthode DIRECTE), nous proposons de filtrer les images ordonnées par la recherche textuelle en fonction du degré qu'elles obtiennent pour ce concept.

Deuxièmement, si le nom du concept apparaît dans les mots de la requête ou dans une liste de synonymes des mots de la requête donnés par WordNet (Fellbaum, 1998) (méthode WN), nous proposons également de filtrer les images ordonnées par la recherche textuelle en fonction du degré qu'elles obtiennent pour ce concept. Par exemple, la requête 5 d'ImageCLEFphoto 2008 est « *animal swimming* ». En utilisant la méthode DIRECTE, le système détermine automatiquement qu'il doit utiliser la FFDT du concept *animal*. Si, de plus nous utilisons WordNet (méthode WN), le système détermine automatiquement qu'il doit utiliser les FFDTs des concepts *animal* et *water* (car d'après WordNet, un synonyme pour *swimming* est « *water sport, aquatics* ».

Pour chaque requête, nous déterminons la liste ordonnée des images pertinentes selon le modèle de langue (LM) ou selon le modèle TD-IDF utilisé sur le texte. Puis, à l'aide des FFDTs, nous réordonnons les 50 premières images de chaque requête ainsi : le système parcourt les images retrouvées du rang 1 au rang 50. Si le degré d'une image est inférieur à un seuil t, alors cette image est réordonnée à la fin de la liste des 50 premières images. De cette façon, les images pertinentes se trouvent toujours dans les 50 premiers résultats.

3.2. *Promouvoir la diversité en utilisant le partitionnement de l'espace visuel*

Pour une requête donnée, les documents similaires sont naturellement ordonnés à des rangs proches. Quand un utilisateur pose une requête, ce qui l'intéresse c'est d'avoir des documents qui soient certes tous pertinents, mais aussi qui soient les plus dissimilaires les uns des autres. La diversité est définie dans (Lemire *et al.,* 2008) comme "a high level of heterogeneity in a collection of entities, that is, if the entities in the collection are different from each other". Les modèles traditionnels de recherche d'information supposent que la pertinence d'un document est indépendant de la pertinence des autres documents. Cependant, pour une requête donnée, l'utilité d'un document dépend des documents que l'utilisateur a déjà vu. En effet, comme l'indique (Zhai *et al.,* 2003) : "a relevant document may be useless to a user if the user has already seen another document with the same content".

Ce problème peut être posé comme un problème de recherche de sous-thèmes (*subtopic retrieval problem*) (Zhai *et al.,* 2003). La recherche de sous-thèmes consiste à retrouver des documents couvrant le maximum de sous-thèmes pour un thème gé-

néral donné. Afin de promouvoir la diversité, la plupart des systèmes de recherche d'images proposés, comme par exemple (Maisonnasse *et al.,* 2009, Ah-Pine *et al.,* 2009), utilisent une stratégie en 2 étapes : d'abord, une recherche d'images traditionnelle est effectuée, puis les N meilleurs résultats pour chaque requête sont réordonnés en fonction d'un partitionnement. D'autres systèmes essayent de construire le résultat final de manière incrémentale. Comme dans la stratégie précédente, une recherche traditionnelle est effectuée et le premier document pertinent est ajouté à la réponse finale, puis, à chaque itération, le document le plus dissimilaire aux documents retrouvés non encore sélectionnés est ajouté à la réponse finale (Zhang *et al.,* 2008, Ferecatu *et al.,* 2008). Cette approche est appelée approche par dissimilarité. Comme indiqué dans (Ziegler *et al.,* 2005, Tollari *et al.,* 2009), augmenter la diversité diminue souvent la pertinence des résultats. C'est la raison pour laquelle (Deselaers *et al.,* 2009b, Zhang *et al.,* 2008, Chen *et al.,* 2006) posent le problème de trouver des résultats pertinents et divers comme un problème d'optimisation jointe.

Nous avons choisi d'utiliser une méthode de diversité basée sur le partitionnement. Les techniques de partitionnement sont étudiées depuis de nombreuses années. Deux approches sont généralement proposées : le partitionnement des données et le partitionnement de l'espace. La première approche nécessite beaucoup de temps de calcul et doit être adaptée à la distribution des premières images résultats d'une requête donnée comme dans (Inoue *et al.,* 2008). La seconde approche, comme elle est faite indépendamment des données, est souvent moins efficace, mais peut-être appliquée de manière très rapide. Nous avons choisi de réaliser un partitionnement de l'espace visuel fondé sur l'histogramme Hue de l'espace HSV. Pour chaque image, nous binarisons l'histogramme Hue. Chaque vecteur binaire correspond à un cluster. En fonction du nombre de dimensions nh de l'histogramme, nous obtiendrons 2^{nh} clusters possibles (les clusters ne seront pas forcement tous instantiés, car certains pourraient ne correspondre à aucune donnée).

Pour chaque requête, les images sont classées en deux listes. Le système parcours les images dans l'ordre : des plus pertinentes vers les moins pertinentes. Si une image appartient à aucun des clusters des images plus pertinentes qu'elle, alors cette image est ajoutée à la fin de la première liste. Si une image appartient au même cluster qu'une image plus pertinente qu'elle, alors cette image est ajoutée à la fin de la deuxième liste. Au final, nous obtenons dans la première liste uniquement des images avec des clusters différents. Nous concaténons ensuite la première liste et la deuxième liste. L'image de rang 1 est toujours au rang 1 ; l'image de rang 2 se retrouve soit au rang 2 si son cluster est différent du cluster de l'image de rang 1, soit à la position $nbcv + 1$ si son cluster est identique (avec $nbcv$ le nombre de clusters visuels), et ainsi de suite. Nous remarquons que plus nous aurons de clusters, et moins il y aura de changement dans l'ordre des images. Nous appelons cette méthode : DIVVISU.

Pour avoir un point de comparaison, nous proposons également une méthode de diversification « naïve » qui consiste à permuter aléatoirement les N premiers résultats. Nous appelons cette méthode : DIVALEA.

3.3. *Expérimentations et résultats*

3.3.1. *Corpus*

Nous utilisons le corpus la tâche ImageCLEFphoto (Arni *et al.*, 2009) de la campagne d'évaluation CLEF 2008. Ce corpus contient 20k images généralistes et 39 *topics*. Chaque *topic* est composé d'un titre, d'une partie narrative, de 3 images correspondant au *topic*, ainsi que d'un élément indiquant sur quel critère doit être appliqué la diversité (<CLUSTER>). Par exemple, le premier *topic* est :

```
<TITLE>church with more than two towers</TITLE>
<CLUSTER>city</CLUSTER>
<NARR>Relevant images will show a church, cathedral or a mosque with
three or more towers. Churches with only one or two towers are not
relevant. Buildings that are not churches, cathedrals or mosques are
not relevant even if they have more than two towers.</NARR>
```

Les 39 requêtes doivent être dérivées de chacun des *topics*. Il y a 17 critères (parfois appelés sous-thèmes (*subtopics*) (Zhai *et al.*, 2003)) de diversités différents : *animal, bird, city, city/nationalpark, composition, country, group composition, landmark location, sport, state, statue, tourist attraction, vehicle type, venue, volcano, weather condition*. La plupart de ces critères correspondent à des lieux. Par exemple, pour la première requête, le critère de diversité est *city*. Pour cette requête, ce critère contient 5 clusters ($n_c = 5$) : *Moscow, Saint Petersburg, Melbourne, Sydney, Bolshaya Reka*. Chaque image du corpus est associée à une légende contenant le titre de l'image, sa date de création, sa localisation, le nom du photographe, une description sémantique du contenu de l'image (déterminée par le photographe) ainsi que de notes additionnelles.

3.3.2. *Mesures de performances*

Les mesures classiquement utilisées en recherche d'information sont généralement la précision et le rappel. De plus, afin de combiner ces deux mesures, la F1-mesure est généralement utilisée. Cependant, dans notre cas, nous souhaitons retrouver non seulement les documents pertinents, mais aussi retrouver, dans les premiers résultats, les documents pertinents qui sont les plus différents les uns des autres en fonction du critère de diversité choisi. C'est pourquoi les mesures de performances utilisées sont : la précision à 20 (P20), le *cluster recall* à 20 (CR20) (Zhai *et al.*, 2003) et la F1-mesure appliquée au P20 et au CR20. Soit $nbpr(n)$ le nombre de documents pertinents retrouvés parmi les n premiers documents retrouvés, la précision à 20 peut être définie ainsi :

$$P20 = \frac{nbpr(20)}{20}.$$

Le *cluster recall* à 20 (CR20) (appelé aussi *S-recall*) (Zhai *et al.*, 2003) a pour but de mesurer le nombre de clusters différents présents dans les 20 premiers résultats. Soit n_c le nombre de clusters différents pour une requête donnée, soit $nbcpr(n)$ le

Texte	Méthode	Moyenne sur 39 requêtes P20 (gain %)	Moyenne sur 39 requêtes CR20 (gain %)	Moyenne des requêtes modifiées Nb topics	Moyenne des requêtes modifiées P20 (gain %)	Moyenne des requêtes modifiées CR20 (gain %)
LM	-	0.185(-)	0.247(-)	11	0.041(-)	0.090(-)
				25	0.148 (-)	0.254 (-)
	DIRECTE	0.195(+6)	0.257(+4)	11	0.077(+88)	0.126(+40)
	WN	0.176(-5)	0.248(+1)	25	0.134(-9)	0.257(+1)
TF-IDF	-	0.250(-)	0.300(-)	11	0.155(-)	0.161(-)
				25	0.210(-)	0.305(-)
	DIRECTE	0.269(+8)	0.313(+5)	11	0.223(+44)	0.209(+30)
	WN	0.260(+4)	0.293(-2)	25	0.226(+8)	0.294(-4)

Tableau 2. *Comparaison des méthodes DIRECTE et WN. Par la méthode DIRECTE, seulement 11 requêtes sont modifiées. Par la méthode WN, 25 requêtes sont modifiées*

nombre de clusters différents couverts par les documents pertinents retrouvés parmi les n premiers documents retrouvés pour cette requête, alors le CR20 est défini ainsi :

$$CR20 = \frac{nbcpr(20)}{n_c}.$$

La dernière mesure utilisée est la F1-mesure définie ainsi dans notre cas :

$$\mathrm{F1-mesure} = 2 \times \frac{P20 \times CR20}{P20 + CR20}.$$

3.3.3. *Utilisation des concepts visuels par correspondance directe et par WordNet*

Dans cette section, nous donnons, puis discutons les scores obtenus par l'utilisation des concepts visuels par les méthodes présentées dans la section 3.1. Nous réalisons d'abord une recherche d'images basée uniquement sur le texte. Pour cela, nous construisons les requêtes en utilisant les éléments du titre ainsi que les phrases de la balise `<NARR>` qui ne contiennent pas le mot *not*. Pour décrire chaque image, nous prenons en compte le texte contenu dans tous les éléments de la légende. Nous appliquons ensuite un modèle classique de langue (LM) ainsi que le modèle TF-IDF.

Pour déterminer si une image contient un concept visuel, nous choisissons de fixer le seuil t à la médianne de tous les degrés obtenus par un concept donné (cette valeur varie de 7.3 (*overcast*) à 28.8 (*outdoor*). Nous n'avons pas utilisé dans cette partie les règles d'exclusion et d'implication.

Le tableau 2 montre que, en moyenne sur tous les topics, la méthode DIRECTE améliore la précision à 20 documents (P20) de +8 % par rapport au TF-IDF et de +6 % par rapport au LM, tandis que la méthode WN améliore le P20 du TF-IDF de +4 %, mais diminue de -5 % le P20 du LM. Comme les méthodes DITECTE et WN dépendent de la présence du nom du concept dans la requête textuelle et que certaines

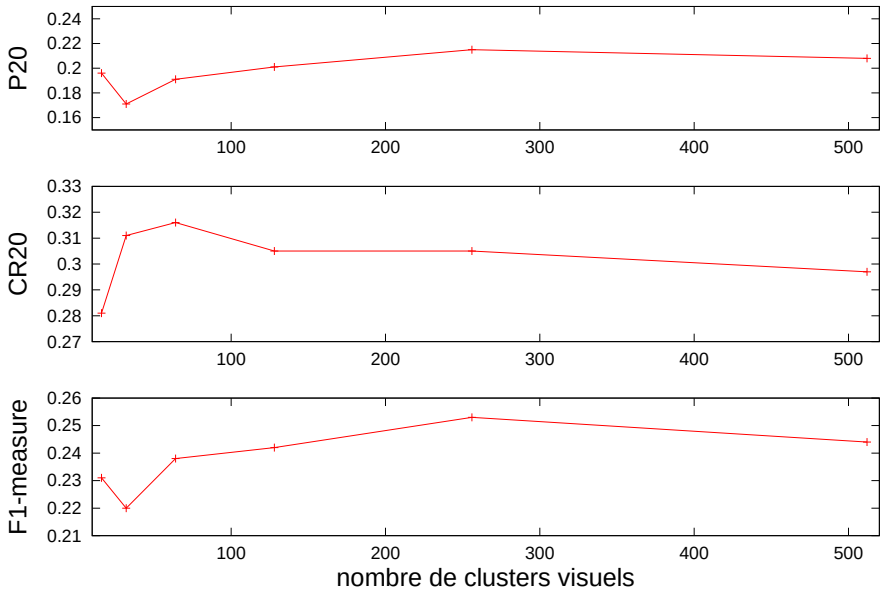


Figure 6. *Influence du nombre de clusters sur les scores P20, CR20 et F1-measure obtenus par diversification DIVVISU des résultats du modèle TF-IDF*

requêtes ne contiennent aucun des noms des 17 concepts, les résultats de certaines requêtes ne sont pas modifiés. La méthode DIRECTE modifie seulement 11 requêtes, tandis que la méthode WN modifie 25 requêtes. C'est pourquoi nous séparons, dans le tableau 2, les résultats en 3 groupes. Nous remarquons une amélioration des scores de P20 de +44 % par rapport au TF-IDF (+30 % par rapport au LM) pour les 11 requêtes modifiées par la méthode DIRECTE, mais une amélioration de seulement +8 % et une diminution de -9 % en utilisant WordNet. Nous en déduisons que la méthode DIRECTE permet d'améliorer sensiblement les résultats obtenus par le texte seul, mais que par contre l'utilisation de WordNet n'est pas adaptée pour ce genre de tâche. Par exemple, d'après WordNet, le concept visuel *person* n'est pas dans la liste des synonymes du mot *people*. Nous ne pouvons donc trouver de lien entre les requêtes recherchant des personnes et le concept *person*. Nous avons également étudié d'autres types de relations (hyponymie, hypernymie...), mais les résultats obtenus étaient globalement inférieurs à ceux obtenus par synonymie.

3.3.4. *Diversification des résultats par partitionnement visuel*

Notre méthode de diversification, présentée dans la section 3.2, est basée sur un partitionnement de l'espace visuel, et non pas sur des informations sémantiques qui pourraient être utiles comme, par exemple, une liste de villes pour le critère *city*. Nous

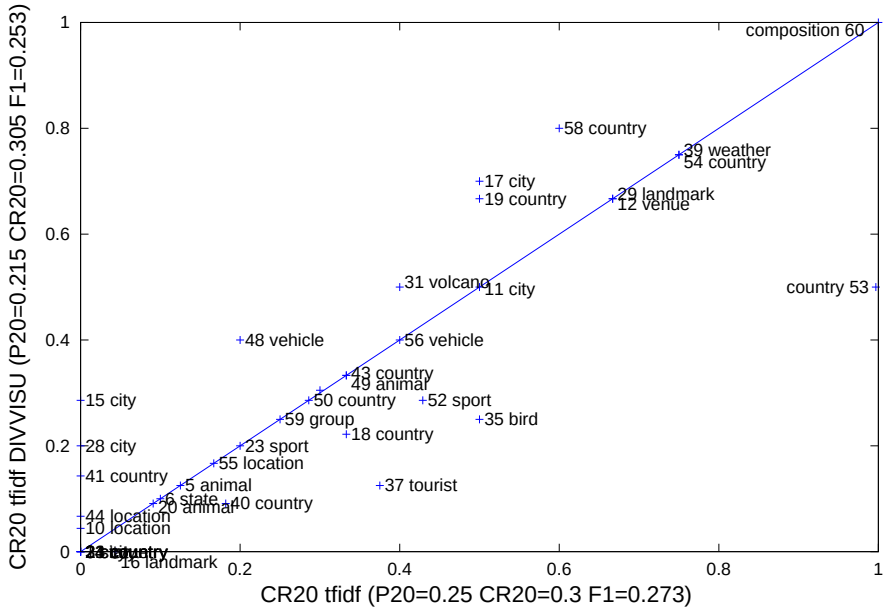


Figure 7. *Comparaison des scores CR20 du TF-IDF et du TF-IDF diversifié par DIV-VISU avec $nh = 8$ (256 clusters visuels). Chaque point correspond à une requête (le nombre correspond au numéro de la requête et le mot associé correspond au critère de diversification demandé dans le topic)*

savons que nos résultats seront sous-optimaux, car la diversité visuelle n'entraîne pas forcement la diversité sémantique. Nous supposons cependant que des photographies de cathédrales prises dans une même ville seront assez visuellement similaires.

Pour pouvoir étudier l'influence du nombre de clusters visuels, nous avons choisi de faire varier le nombre de dimensions de l'histogramme Hue de 4 à 9 dimensions. La quantité d'information initiale étant ainsi toujours la même. Nous obtenons un nombre théorique de clusters variant de 16 à 512, et un nombre de clusters instanciés légèrement inférieur. La figure 6 montre l'évolution des scores P20, CR20 et F1-measure en fonction du nombre de clusters de l'espace visuel. Pour un nombre de clusters égal à 16, la diversification n'est pas complètement effectuée, car certaines images ayant des clusters similaires aux images de rangs plus élevés se retrouvent toujours dans les 20 premiers résultats. Pour un nombre de clusters supérieur à 32, le P20 chute brutalement et puis remonte doucement vers le P20 du TF-IDF sans toutefois l'atteindre. A l'inverse, le CR20 augmente atteignant son maximum pour un nombre de clusters de 64, puis diminue pour atteindre le CR20 du TF-IDF. En moyenne, la meilleur valeur de F1-measure est obtenu pour un nombre de clusters de 256 (soit une dimension de l'espace de 8).

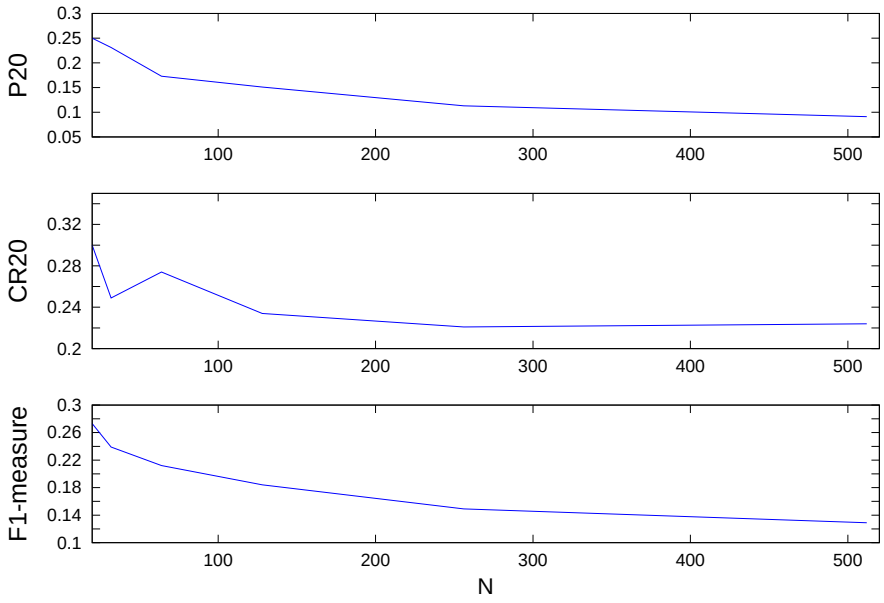


Figure 8. *Permutation aléatoire des N premiers documents obtenus par TF-IDF*

La figure 7 montre les scores CR20 de TF-IDF comparés aux scores CR20 de TF-IDF+DIVVISU. Notre méthode DIVVISU, qui est basée uniquement sur de l'information visuelle, ne semble pas être adaptée spécialement pour un type de critère de diversification. En effet, il n'y a aucun critère de diversification qui donne de bons résultats seulement pour TF-IDF+DIVVISU. Par exemple, pour le critère *city*, il y a autant de requêtes qui ont un meilleur résultat par TF-IDF seul, que de requêtes qui ont un meilleur résultat par TF-IDF+DIVVISU.

Pour étudier la difficulté de garder des scores forts lorsque l'on effectue une diversification des résultats, la figure 8 montre la baisse des scores P20, CR20 et F1-mesure en fonction du nombre de documents permutés aléatoirement. La figure 9 compare les scores de diversification pour les méthodes DIVVISU (à 256 clusters) et DIVALEA (permutation aléatoire des 40 premiers documents) par rapport aux scores TF-IDF sans diversification (NO DIV). Nous remarquons que les deux méthodes proposées donnent des scores P20 largement inférieurs au P20 du TF-IDF, mais DIVALEA diminue le CR20, tandis que DIVVISU l'améliore légèrement (+2 %). Nous en concluons que notre méthode DIVVISU améliore légèrement les résultats de diversification, mais que, comme de nombreuses autres méthodes (voir (Tollari *et al.*, 2009)), elle diminue le P20.

Le tableau 3 compare notre méthode DIVVISU basée sur un partitionnement de l'espace visuel à une méthode, appelée VisKmeans, basée sur le partitionnement des

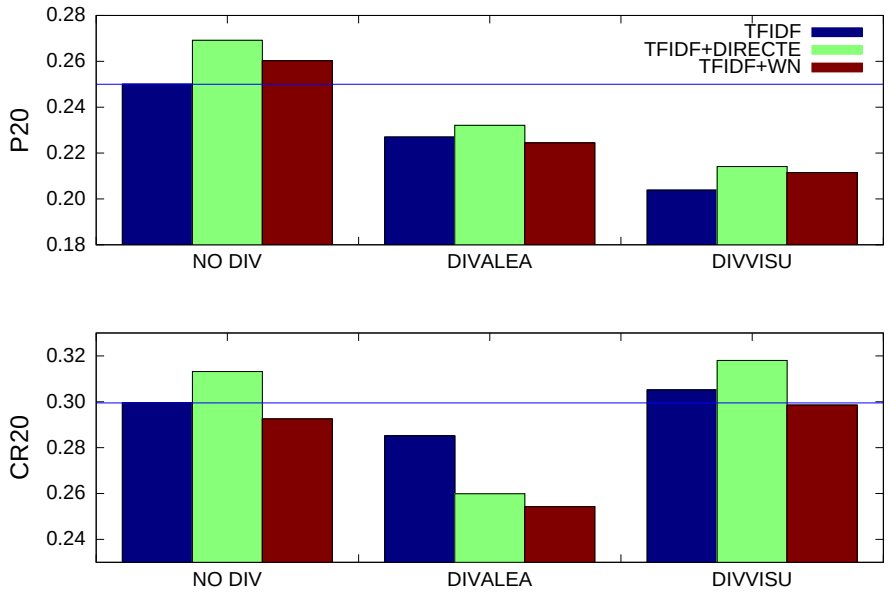


Figure 9. *Comparaison des méthodes de diversification (NODIV : sans diversification, DIVALEA : diversification aléatoire des 40 premiers documents, DIVVISU : diversification par partitionnement de l'espace visuel (avec $nh = 8$)). Pour chaque méthode de diversification, la première barre correspond au résultat du TF-IDF seul, la deuxième au résultat du TF-IDF filtré par les FFDTs à l'aide de la méthode DIRECTE et la troisième au TF-IDF filtré par la méthode WN*

données, proposée dans (Maisonnasse *et al.*, 2009). Dans cette méthode, les clusters visuels sont construits à l'aide d'un Kmeans dans un espace visuel à 4 608 dimensions, le nombre de clusters étant fixé à 500, il y a en moyenne 40 images par cluster. Ces deux méthodes, ainsi que la méthode DIVALEA, sont appliquées dans (Tollari *et al.*, 2009) à un ensemble de 26 runs proposés par différents participants à la campagne ImageCLEFphoto 2008. Parmi ces 26 runs, 3 runs utilisent seulement le texte associé aux images, 3 seulement le contenu visuel, 9 le texte et le visuel, 1 est un run où les requêtes sont étendues manuellement, et 10 sont des runs dit idéaux. Ces runs idéaux sont construits en prenant, pour chaque requête, les documents pertinents déterminés à partir des vérités terrains, et en permutant aléatoirement ces documents 10 fois. Les runs idéaux ont donc une précision proche de 1, car tous les documents sont pertinents, mais ils n'ont pas forcemment un très bon cluster recall. Les résultats de la comparaison montrent que notre méthode DIVVISU obtient globalement un meilleur CR20 que VisKmeans, mais un plus faible P20. La différence de scores entre les deux méthodes DIVVISU et VisKmeans ne semble pas être significative. On remarque cependant que DIVVISU et VisKmeans ne sont pas adaptées pour les runs visuel seul.

	NO DIV		DIVALEA		DIVVISU		VisKmeans	
nb clusters	-	-	-	-	256		500	
nb dims	-	-	-	-	8		4608	
runs	P20	CR20	P20	CR20	P20	CR20	P20	CR20
texte seul	0.265	0.325	0.243	0.303	0.246	**0.333**	0.253	**0.333**
visuel seul	0.198	**0.262**	0.138	0.227	0.119	0.252	0.141	0.247
texte-image	0.339	0.407	0.277	0.366	0.284	0.406	0.306	**0.412**
idéaux	0.994	0.738	0.993	0.764	0.993	**0.787**	0.993	0.767
26 runs	0.580	0.514	0.548	0.504	0.548	**0.533**	0.560	0.528

Tableau 3. *Comparaison des scores P20 et CR20 obtenus par différentes méthodes de diversité appliquées sur différents types de runs (3 runs texte seul, 3 runs visuel seul, 9 runs texte-image, 10 runs idéaux) ; DIVALEA : diversification aléatoire des 40 premiers documents*

Notre méthode DIVVISU qui ne nécessite aucun calcul et qui utilise un petit espace visuel est donc intéressante pour une approche « en ligne », et pourra être intéressante pour introduire rapidement de la diversité.

4. Conclusion

Dans cet article, nous proposons une chaîne de traitement complète qui propose deux solutions efficaces et peu coûteuses pour utiliser le contenu visuel des images dans un système de recherche d'images « en ligne ». Nous nous intéressons particulièrement à deux difficultés.

La première est l'exploitation de concepts visuels pour améliorer la recherche d'images basée uniquement sur le texte. Pour tenter de résoudre cette difficulté, nous apprenons « hors ligne » des forêts d'arbres de décision flous qui nous donnent des degrés de confiance que les concepts soient présents dans une image. Puis, en fonction des termes de la requête, nous filtrons « en ligne » les images dont les degrés de confiance correspondant aux concepts visuels de la requête sont trop faibles. Les résultats montrent une nette amélioration des scores pour les requêtes qui contiennent explicitement le nom d'un concept. Nous en déduisons que la difficulté principale est de déterminer quel concept appliquer pour une requête qui ne contient pas explicitement de concept.

La seconde est la diversification des résultats pertinents. Nous proposons d'utiliser le partitionnement de l'espace visuel afin d'obtenir très rapidement le cluster visuel d'une image, puis de garder dans les 20 premiers documents uniquement des images qui ont des clusters différents. Notre méthode augmente légèrement les scores de diversité, et a l'avantage de pouvoir être utilisée très simplement sans lourd calcul.

Dans nos futurs travaux, nous souhaitons d'abord améliorer nos règles de résolutions (exclusion et implication) pour obtenir de meilleurs résultats de classification, puis les utiliser dans la tâche de recherche d'images. En effet, nous avons fixé un seuil de décision t à la médiane de tous les degrés obtenus, or cette valeur varie de 7.3 à 28.8, l'utilisation de règles d'exclusion dans la tâche de recherche d'images devrait, d'après la figure 5, améliorer nos résultats. Nous souhaitons également les comparer à des méthodes utilisant des règles d'association.

Nous souhaitons également étendre notre travail sur l'extension de requêtes avec utilisation des concepts visuels. La construction d'un thésaurus adapté pour permettre le passage d'une requête exprimée en langage naturel à une requête exprimée avec des concepts visuels est une étape préliminaire. Puis nous envisageons d'utiliser l'analyse de la sémantique latente (LSA) pour rapprocher les requêtes en langage naturel et les concepts visuels dans un même espace latent afin d'étendre efficacement les requêtes.

Remerciement : ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche (ANR) portant la référence ANR-06-MDCA-002 (projet AVEIR).

5. Bibliographie

Ah-Pine J., Clinchant S., Csurka G., Liu Y., « XRCE's participation in ImageCLEF 2009 », *Working Notes for the CLEF 2009 workshop*, 2009.

Arni T., Clough P., Sanderson M., Grubinger M., « Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum, Revised Selected Papers*, LNCS 5706, 2009.

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

Chen H., Karger D. R., « Less is more : probabilistic models for retrieving fewer relevant documents », *ACM SIGIR*, p. 429-436, 2006.

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

Deselaers T., Deserno T. M., « The Visual Concept Detection Task in ImageCLEF 2008 », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum, Revised Selected Papers*, LNCS 5706, 2009a.

Deselaers T., Gass T., Dreuw P., Ney H., « Jointly Optimising Relevance and Diversity in Image Retrieval », *ACM Conference on Image and Video Retrieval (CIVR)*, 2009b.

Fellbaum C., *WordNet - An Electronic Lexical Database*, Bradford books, 1998.

Ferecatu M., Sahbi H., « TELECOM ParisTech at ImageClefphoto 2008 : Bi-Modal Text and Image Retrieval with Diversity Enhancement », *Working Notes for the CLEF 2008 workshop*, 2008.

Inoue M., Grover P., « Effects of Visual Concept-based Post-retrieval Clustering in ImageCLEFphoto 2008 », *Working Notes for the CLEF 2008 workshop*, 2008.

Lemire D., Downes S., Paquet S., « Diversity in open social networks », 2008. Published online.

Lienhart R., Romberg S., Hörster E., « Multilayer pLSA for multimodal image retrieval », *CIVR '09 : Proceeding of the ACM International Conference on Image and Video Retrieval*, ACM, New York, NY, USA, p. 1-8, 2009.

Maisonnasse L., Mulhem P., Gaussier E., Chevallet J.-P., « LIG at ImageCLEF 2008 », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum, Revised Selected Papers*, LNCS, p. 704-711, 2009.

Marsala C., Bouchon-Meunier B., « Forest of fuzzy decision trees », *International Fuzzy Systems Association World Congress*, vol. 1, p. 369-374, 1997.

Marsala C., Detyniecki M., « TRECVID 2006 : Forests of fuzzy decision trees for high-level feature extraction », *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

Monay F., Gatica-Perez D., « PLSA-based image auto-annotation : constraining the latent space », *ACM Multimedia*, ACM Press, New York, NY, USA, p. 348-351, 2004.

Popescu A., Millet C., Moëllic P.-A., « Ontology driven content based image retrieval », *CIVR '07 : Proceedings of the 6th ACM international conference on Image and video retrieval*, p. 387-394, 2007.

Quinlan J., « Induction of Decision Trees », *Machine Learning*, vol. 1, p. 81-106, 1986.

Rege M., Dong M., Hua J., « Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering », *WWW '08 : Proceeding of the 17th international conference on World Wide Web*, ACM, New York, NY, USA, p. 317-326, 2008.

Shotton J., Johnson M., Cipolla R., « Semantic texton forests for image categorization and segmentation », *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1-8, 2008.

Song K., Tian Y., Gao W., Huang T., « Diversifying the image retrieval results », *ACM Multimedia*, ACM, New York, NY, USA, p. 707-710, 2006.

Tollari S., Glotin H., « Web Image Retrieval on ImagEVAL : Evidences on visualness and textualness concept dependency in fusion model », *ACM Conference on Image and Video Retrieval (CIVR)*, p. 65-72, 2007.

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 the Cross-Language Evaluation Forum, Revised Selected Papers*, LNCS 5706, 2009.

Yavlinsky A., Heesch D., Rüger S. M., « A Large Scale System for Searching and Browsing Images from the World Wide Web », *CIVR 2006*, p. 537-540, 2006.

Zhai C. X., Cohen W. W., Lafferty J., « Beyond independent relevance : methods and evaluation metrics for subtopic retrieval », *ACM SIGIR*, p. 10-17, 2003.

Zhang M., Hurley N., « Avoiding monotony : improving the diversity of recommendation lists », *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, p. 123-130, 2008.

Ziegler C., McNee S. M., Konstan J., Lausen G., « Improving recommendation lists through topic diversification », *WWW*, p. 22-32, 2005.