

Mémoire de stage d'initiation à la recherche scientifique

Octobre 2005 - Septembre 2006

Master 2 Recherche
Intelligence, Interaction et Information

**Apports réciproques des
informations textuelles et
visuelles par analyse de la
sémantique latente pour la
recherche d'information**

Clément Fleury

CLIPS - MRIM

I²R - IPAL

385, rue de la Bibliothèque 21, Heng Mui Keng Terrace
38 041, Grenoble 119 613, Singapore

13 juin 2006

Remerciements

Je voudrais exprimer ma gratitude envers toutes les personnes qui ont fait en sorte que mon stage se passe pour le mieux.

Je remercie particulièrement mon superviseur en France, Philippe Mulhem, ainsi que mes encadrants à Singapour, Jean-Pierre Chevallet et Joo Hwee Lim, pour avoir su me donner les pistes dont j'avais besoin afin de mener à bien mon projet tout en me laissant la liberté de pouvoir choisir mes objectifs.

Je me dois de témoigner ma reconnaissance envers toute l'équipe MRIM du laboratoire CLIPS et principalement les doctorants qui ont su me conseiller sur la méthodologie à adopter pour la rédaction de ce mémoire.

Enfin, je sais gré à toute l'équipe du laboratoire IPAL et principalement Nicolas Maillot ainsi que Vlad Valea pour l'aide précieuse qu'ils m'ont apportés dans l'implémentation et l'expérimentation de mon projet.

Résumé

L'indexation des documents contenant textes et images est un sujet de recherche devenu essentiel depuis une dizaine d'années. Cependant, combiner deux médias si différents de manière efficace est une tâche *a priori* difficile. Les travaux précédents semblent indiquer que la combinaison texte-image à travers une analyse de la sémantique latente est un nouveau modèle de recherche d'information prometteur.

L'objectif spécifique de cette contribution est de mesurer l'amélioration apportée par l'utilisation combinée des informations textuelles et visuelles contenues dans les documents, face à l'utilisation seule d'une modalité. Nous tâcherons de montrer l'efficacité de l'utilisation de l'indexation par la sémantique latente pour la fusion de ces médias.

À notre connaissance, cette étude est la première à s'intéresser au rôle du couple image-texte avec indexation par sémantique latente pour l'indexation d'une base de documents de taille significative. Cette contribution pourrait être le point de départ d'un nouveau système d'extension d'annotation d'images ou d'autres projets plus ambitieux combinant texte et image au niveau sémantique.

Table des matières

1	Introduction	1
1.1	Entrée en matière	1
1.2	Problématique	2
1.3	Contribution	3
1.4	Plan du mémoire	4
2	État de l’art	7
2.1	Modèles de recherche d’information	9
2.2	Recherche d’information sur le texte	10
2.2.1	Modèles de langue	10
2.2.2	Modèles basés sur des connaissances	10
2.2.3	Modèle vectoriel appliqué au texte	11
2.3	Recherche d’information sur les images	12
2.3.1	Recherche d’information basée sur les annotations	12
2.3.2	Recherche d’information basée sur le contenu	13
2.4	Modèle vectoriel	22
2.4.1	Pondération des termes	22
2.4.2	Indexation par sémantique latente	23
2.5	Recherche d’information sur documents contenant texte et images	25
3	Modèle de fusion	27
3.1	Prétraitements	27
3.1.1	Texte	28
3.1.2	Image	28
3.2	Indexation	30
3.2.1	Pondération des termes	30
3.2.2	Fusion	31
3.2.3	Indexation par sémantique latente	31
3.3	Traitement de la requête	33
3.3.1	Prétraitements	33

TABLE DES MATIÈRES

3.3.2	Projection dans l'espace de sémantique latente	33
3.3.3	Fonction de correspondance	34
4	Expérimentation	35
4.1	Évaluation des SRI	35
4.1.1	Base de documents	36
4.1.2	Pertinence	36
4.1.3	Évaluation	36
4.2	Protocole	38
4.3	Résultats expérimentaux	39
4.3.1	Exemples	39
4.3.2	Courbes rappel-précision	43
4.3.3	Mean Average Precision	46
4.4	Conclusions	48
4.5	Discussion	49
4.5.1	Exemples	49
4.5.2	Résultats	51
5	Conclusion générale	55
5.1	Contribution	55
5.2	Travaux futurs	56
A	Généralités	57
A.1	Structures d'accueil	57
A.2	Déroulement du stage	58

Chapitre 1

Introduction

1.1 Entrée en matière

Depuis une décennie, les bases de documents multimédia se multiplient et leur taille augmente de manière spectaculaire. Parmi les médias les plus stockés à l'heure actuelle, on trouve en première position le texte et les images loin devant les autres médias.

En effet, Internet, plus grand ensemble de documents accessibles au public, compte plus de 40 millions de sites actifs d'après Netcraft [Netcraft06], dont la plupart sont composés de pages contenant texte et images, d'après *HOW MUCH INFORMATION 2003 ?* [Lyman03]. On voit de même augmenter le nombre et la taille des bases de dossiers médicaux, comprenant souvent une description textuelle du cas médical et des images de différentes natures. Avec l'apparition et la démocratisation des appareils photographiques numériques, les bases d'images personnelles annotées telles que *flickr* [flickr] se sont multipliées.

Ce contexte est extrêmement récent (moins d'une dizaine d'années), cependant déjà très problématique de par la taille croissante des bases de documents. C'est pourquoi les systèmes efficaces de recherche de documents contenant texte et images deviennent de plus en plus nécessaires.

Il est souvent admis qu'*une image vaut cent mots*. En effet, si l'on voulait décrire l'image d'une pomme, il faudrait en donner la couleur, la taille, la position, le contexte dans lequel elle apparaît... Mais de la même manière, on pourrait dire qu'*un mot vaut cent images* ! Car si l'on voulait représenter le mot « pomme » de manière visuelle avec le même degré de conceptualisation, il faudrait au moins une image par variété, par couleur, par taille...

Le paradoxe ainsi soulevé s'explique par le fait que le texte se place à un

niveau d'abstraction élevé, tandis qu'une image représente l'instance d'un ou plusieurs concepts.

Le constat que l'on peut alors faire est que les modalités textuelles et visuelles étant de plus en plus souvent réunies dans les documents, on ne peut plus se contenter d'utiliser des systèmes de recherche basés uniquement sur l'une ou l'autre.

Mais créer un système combinant les deux pour l'indexation et la recherche de documents multimodaux est un problème loin d'être évident à résoudre de par la distance sémantique séparant ces deux modalités.

Dans le domaine textuel, l'analyse de la sémantique latente est une méthode statistique permettant de découvrir les liens sémantiques existants entre les mots.

Nous pensons et nous allons tâcher de montrer dans cette étude que cette technique appliquée aux informations extraites du texte et des images permet d'obtenir de meilleurs résultats que l'utilisation seule de l'une ou l'autre des modalités, ou que la combinaison des deux sans analyse de la sémantique latente.

1.2 Problématique

Le domaine de la RI¹ a pour objectif l'élaboration de modèles et de systèmes permettant de répondre de manière pertinente au besoin d'information exprimé par un utilisateur, à partir d'une base de documents. Nous nous placerons dorénavant dans le contexte de documents composés d'une partie textuelle et d'une partie visuelle.

L'hypothèse faite dans la plupart des livres, des actes de conférence et des articles issus du domaine de la RI est que si l'on trouve un moyen permettant à un ordinateur d'avoir une interprétation et une connaissance sémantique des données qu'on lui fournit, il sera capable de répondre de manière pertinente au besoin d'information exprimé par un utilisateur.

Ainsi, une des difficultés majeures que se pose le domaine de la RI est de traverser le *fossé sémantique* existant entre une *image* et son *sens*. C'est-à-dire, à partir d'une image, retrouver ce qu'elle cherche à *exprimer*. Naturellement, ce problème est insoluble dans l'absolu, puisque le sens porté par une image est fortement dépendant de l'individu qui l'interprète. De plus, si l'on voulait qu'une machine soit capable de résoudre un tel problème, elle devrait

¹RI : Recherche d'Information

1.3 Contribution

être capable d'acquérir une quantité de connaissance phénoménale.

Nous ne prétendons pas résoudre ici ce qui est l'équivalent de la quadrature du cercle pour la Recherche d'Information, mais nous allons simplement prouver par l'expérience que, dans le cadre de notre étude, l'utilisation du texte associé à une image permet d'augmenter le sens porté par le document global.

Un autre problème essentiel en Recherche d'Information concerne la *synonymie* et la *polysémie* entre mots. Ainsi, le concept de « *maison* » pourra être instancié dans les mots *maison*, *demeure* ou *domicile* (cas de synonymie), tandis que le mot *sphinx* pourra faire référence aux concepts « *papillon* », « *statue* » où « *chat* » (cas de polysémie).

Ici encore, la résolution de ce problème est un domaine de recherche à part entière. Cependant, les instances *maison*, *demeure* et *domicile* du concept « *maison* » seront souvent visuellement assez semblables, et les concepts « *papillon* », « *statue* » et « *chat* » associés au mot *sphinx* seront visuellement très différents.

Il semble donc possible que les images permettent de rapprocher les synonymes et distinguer le sens des mots polysémiques. Nous allons tenter de montrer que l'utilisation des informations issues de l'analyse des images associées au texte permettra de résoudre une partie du double problème inhérent à la langue naturelle.

Par ailleurs, la méthode LSA^2 (Deerwester [Deerwester90]), ou LSI^3 , nous semble être appropriée à la combinaison des informations venant de différentes modalités. En effet, cette technique qui a déjà prouvé son efficacité en indexation textuelle, avec Landauer *et al.* [Landauer98], comme en indexation d'images, avec Parks *et al.* [Praks03].

1.3 Contribution

La contribution apportée au domaine par cette étude est de deux ordres. Tout d'abord, nous mesurerons l'influence réciproque et combinée des informations textuelles et visuelles extraites de documents contenant les deux modalités textuelle et visuelle. Puis nous évaluerons l'amélioration apportée par *Latent Semantic Analysis* dans la combinaison des deux sources d'information.

La mesure de ces influences sera faite de manière *relative*. Autrement dit,

²LSA : Latent Semantic Analysis

³LSI : Latent Semantic Indexing

les résultats absolus de chacun des modèles ne seront pas significatifs, mais leur rapport deux à deux le seront.

Pour cela, nous allons mesurer, comparer et analyser les performances, pour les tâches d'indexation et de recherche, des six modèles de Recherche d'Information représentés dans le tableau 1.1.

Les modèles IMAGE, TEXTE et IMAGE&TEXTE s'intéresseront aux

modèle	texte	image	LSA
IMAGE	non	oui	non
TEXTE	oui	non	non
IMAGE&TEXTE	oui	oui	non
IMAGE&LSA	non	oui	oui
TEXTE&LSA	oui	non	oui
IMAGE&TEXTE&LSA	oui	oui	oui

TAB. 1.1 – Description des modèles étudiés.

informations issues respectivement uniquement du texte, uniquement des images, et à la fois du texte et des images, contenus dans les documents, sans appliquer la méthode LSA.

Les modèles IMAGE&LSA, TEXTE&LSA et IMAGE&TEXTE&LSA utiliseront les données respectivement uniquement textuelles, uniquement visuelles, et textuelles et visuelles, issues des documents, en utilisant la méthode LSA.

1.4 Plan du mémoire

La problématique que l'on s'attachera à résoudre étant posée, nous ferons une revue critique d'un ensemble de travaux sur le domaine de la recherche d'information textuelle, la recherche d'images basée sur le contenu et un ensemble d'études ayant traité de problème de la fusion de ces deux médias.

La suite sera consacrée à la modélisation et l'implémentation du système de recherche d'information créé dans le but de répondre à la problématique posée. Les phases de pré-traitement, d'indexation et de traitement de la requête seront explicitées en détail.

1.4 Plan du mémoire

Enfin, nous décrirons l'expérimentation mise en place afin de tester le système développé. Les objectifs et le protocole expérimental seront exposés avant de présenter les résultats obtenus et leur interprétation.

Chapitre 2

État de l'art

Le domaine de la Recherche d'Informations s'attache à répondre à trois types de tâches :

recherche Étant donné un document requête et une base de documents, retrouver les documents les plus *sémantiquement proches* de celle-ci.

classification Étant donné un document et un ensemble de classes de documents *sémantiquement homogènes*, affecter à ce document la classe contenant les documents qui lui sont le plus *sémantiquement proches*. Bien que la classification automatique de documents ne soit pas explicitement une tâche de recherche, les concepts manipulés et les techniques utilisées sont les mêmes, c'est pourquoi cette tâche est associée au domaine de la Recherche d'Information.

indexation Cette tâche est particulière, car elle n'existe que dans le but de réaliser une des deux tâches précédentes. Cependant, c'est un domaine de recherche à part entière qui mérite d'être cité ici. Étant donné un ensemble de documents, affecter à chacun d'eux un descripteur qui peut être soit une classe, si l'objectif final est la classification, soit un vecteur dans un espace de caractéristiques, si l'objectif final est la recherche.

Le contexte dans lequel nous nous plaçons est celui de la tâche d'indexation dans le but d'effectuer des recherches de documents. Afin d'effectuer cette tâche, le domaine a comme moyen la construction de systèmes permettant d'indexer et de retrouver les documents qu'il juge pertinents à une requête émise par un utilisateur. La structure typique d'un SRI¹ est décrite dans la figure 2.1, page 8.

Une décomposition fonctionnelle d'un système de RI comporte typiquement deux parties principales :

¹SRI : Système de Recherche d'Information

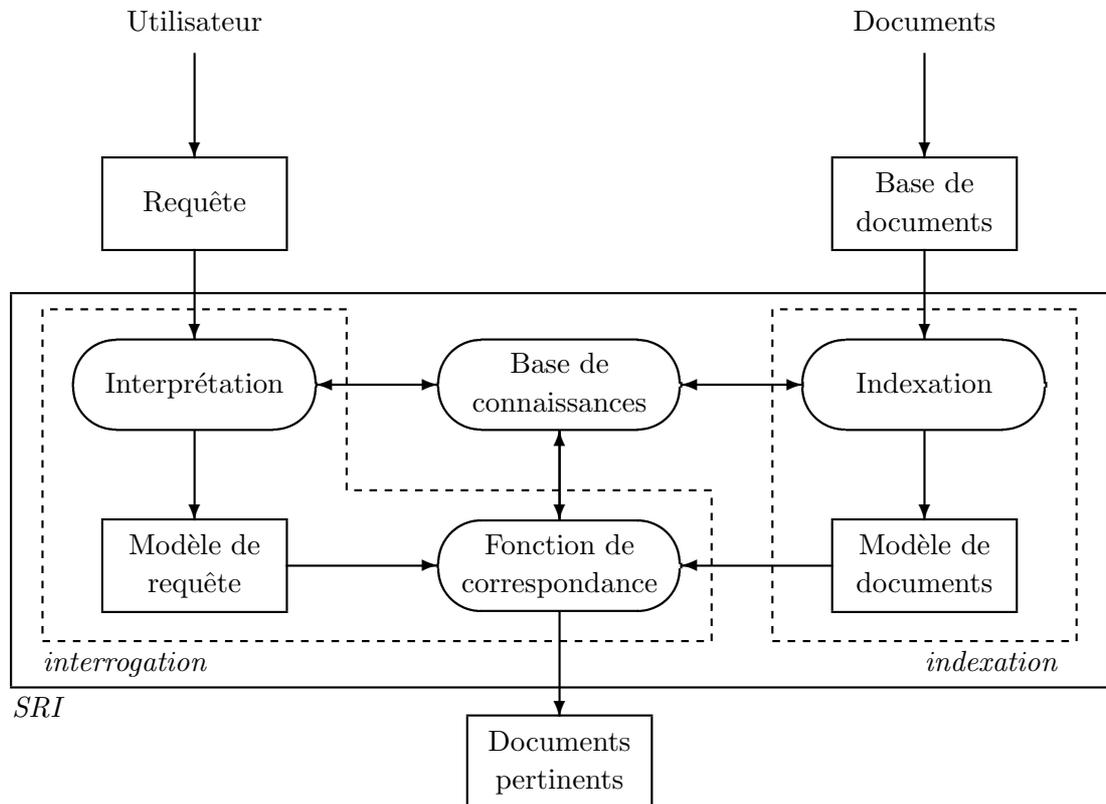


FIG. 2.1 – Architecture d'un système de recherche d'information.

- une partie *indexation*, qui correspond fondamentalement à un traitement des documents du corpus afin de leur donner une représentation manipulable par le système
- une partie *interrogation* correspondant à la partie en charge de gérer l'interaction avec l'utilisateur, incluant essentiellement le traitement de la requête et la présentation des résultats

Notre intérêt dans cette étude est focalisé sur la partie *indexation* des modèles de RI. Pour une introduction générale au domaine de la RI, se référer à Frakes *et al.* [Frakes92] ou au cours de Jian-Yun Nie [Nie06, Introduction], pour une vue d'ensemble détaillée du domaine, consulter vanRijsbergen [VanRijsbergen79] ou Salton et McGill [SaltonMcGill83].

2.1 Modèles de recherche d'information

Un MRI² est composé d'une modélisation du contenu des documents, d'un modèle de représentation du contenu des requêtes, ainsi que d'une fonction de correspondance sur les modèles de documents et de requêtes. Ces modèles correspondent aux représentations internes qu'a le système des documents, de la requête et du jugement de pertinence de l'utilisateur.

L'étude des modèles de recherche d'information a un long passé. Au fil des décennies, un grand nombre de modèles ont été proposés et testés (Jones et Willett [JonesWillett97]).

Les approches et méthodologies développées ont été très diverses et aucun modèle de recherche particulier ou unifié n'a prouvé qu'il ne pouvait être surpassé par un autre modèle. Cependant, on peut noter deux grandes catégories de modèles.

D'une part, des études théoriques sur la modélisation des documents ont mené à l'élaboration de différents types de modèles logiques et probabilistes (Salton *et al.* [Salton82], van Rijsbergen [VanRijsbergen86], Fuhr [Fuhr92], Robertson *et al.* [Robertson81], Wong et Yao [WongYao95]). Ces approches ont été principalement appliquées à la recherche d'information sur documents textuels, bien que rien n'empêche d'imaginer une application sur des images.

D'autre part, un grand nombre d'études a porté sur différentes variantes du modèle vectoriel aussi bien du côté texte que du côté image (Salton et Buckley [SaltonBuckley88], [SaltonBuckley90], Singhal, Buckley et Mitra [Singhal96]). De plus, ce modèle permet de formaliser de la même manière les informations issues du texte et celles issues des images, ce qui rend leur combinaison plus aisée. Notre étude se placera donc dans ce type de modèle.

Par ailleurs, certains modèles dits adaptatifs ou dynamiques intègrent totalement le comportement de l'utilisateur. Une technique souvent utilisée est l'utilisation du *retour de pertinence*, implicite ou non, afin de raffiner la perception qu'a le système du besoin d'information de l'utilisateur (Salton et Buckley [SaltonBuckley90], Harman [Harman92]).

Cependant, dans le cadre de notre étude, l'utilisateur n'est pas un paramètre de comparaison des modèles étudiés. Nous modéliserons son jugement de pertinence par la *vérité terrain* associée à la base de documents.

²MRI : modèle de recherche d'information

2.2 Recherche d'information sur le texte

Ce type de SRI considère des documents contenant du texte, qui peut être structuré ou libre, multilingue ou non. Une requête émise par un utilisateur est aussi de nature textuelle.

2.2.1 Modèles de langue

Récemment, les modèles de langue ont présenté une approche originale et ont donné des résultats prometteurs (Ponte et Croft [PonteCroft98], Berger et Lafferty [BergerLafferty99], Miller, Leek et Schwartz[Miller99]). La méthode est intéressante car elle rapproche le problème de la recherche d'information avec celui de l'estimation de modèle (au sens probabiliste) de langue, qui à déjà été étudié intensivement dans d'autres domaines tels que la reconnaissance de la parole.

L'idée de base est d'associer à chaque document un modèle de langue, *i.e.* une densité de probabilité de l'apparition des mots du vocabulaires dans le document. Une requête est alors modélisée comme une conjonction d'événements que sont les mots qui la composent. La fonction de correspondance entre une requête et un document est alors la vraisemblance de la requête étant donné le modèle du document.

Ce modèle est très bien adapté au domaine textuel, mais n'a jamais été appliquée aux images à notre connaissance.

2.2.2 Modèles basés sur des connaissances

D'autres approches basées sur l'analyse syntaxique ou sémantique de la langue naturelle ont été proposées afin de résoudre les problèmes de synonymie et polysémie (Voorhees [Voorhees93], Sanderson [Sanderson94], Krovetz *et al.* [KrovetzCroft92]). Ces modèles nécessitent une grande quantité de connaissances sous formes de thésaurus, dictionnaires, lexiques (tels que WordNet [Al-Halimi98]), réseaux sémantiques. . .

Ils ne peuvent donc être utilisés que dans le cadre d'un domaine d'application spécifique, ce qui ne sera pas notre cas. De plus, Sanderson [Sanderson94] et Voorhees [Voorhees99] concluent qu'une telle approche n'apporte d'amélioration significative par rapport à l'approche vectorielle que dans un nombre très limité de cas.

Les modèles présentés ci-dessus sont par nature destinés au traitement des informations textuelles. L'utilisation d'un tel modèle dans notre cas com-

2.2 Recherche d'information sur le texte

pliquerait sérieusement le travail de combinaison de ces informations avec des informations issues d'images.

2.2.3 Modèle vectoriel appliqué au texte

En revanche, le modèle vectoriel est un modèle commun aux deux domaines et les résultats qu'il permet d'obtenir sont généralement assez bons, comme le montre Salton *et al.* [Salton75] et le système SMART de Salton. C'est donc celui que nous allons choisir pour le traitement du texte.

Le principe général du modèle vectoriel appliqué au texte est d'associer à chaque document un vecteur de termes dans l'espace du vocabulaire. Le vocabulaire peut être spécifique à un domaine, multilingue, très vaste ou au contraire très restreint. . . Toujours est-il que moins le vocabulaire est ambigu (peu de polysémie et de synonymie), plus le modèle sera efficace. L'idéal étant d'avoir un vocabulaire composé uniquement de concepts, ce qui nécessite une grande quantité de connaissance et de traitement.

Il existe cependant un moyen relativement simple de réduire la taille du vocabulaire tel qu'il est lorsque l'on extrait simplement tous les mots contenus dans tous les documents. Le traitement du texte se décompose classiquement en deux étapes :

antidictionnaire Cette phase de prétraitement consiste en l'élimination du vocabulaire des mots ne portant pas de sens, tels que les pronoms et les déterminants en français. Cette phase permet d'éliminer le *bruit* de la langue naturelle.

lemmatisation Dans cette étape, les mots restants sont formatés pour en retrouver *forme canonique*, *i.e.* la racine. Par exemple, les majuscules sont minuscules, la terminaison des verbes éliminées, les noms communs sont ramenés à la forme masculin singulier et les préfixes et suffixes sont tronqués. Plusieurs algorithmes de lemmatisation ont été proposés, les plus utilisés étant ceux de Porter [Porter80] et de Savoy [Savoy93]. Cette méthode est particulièrement efficace pour résoudre les problèmes de synonymie. Cependant, elle peut renforcer les problèmes de polysémie en associant à une seule racine un mot ayant plusieurs sens. L'hypothèse faite alors est que les problèmes de polysémie sont plus rares que les problèmes de synonymie.

Les documents et la requête sont exprimés sous forme de vecteur dans un espace de mots. La fonction de correspondance entre un document et la requête est la distance entre ces deux vecteurs.

Sans connaissance supplémentaire, il est difficile d'élever le niveau conceptuel des termes extraits. À ce niveau, les termes du vocabulaire se trouvent entre les niveau sémantique des concepts et celui des mots. Les termes sont alors appelés *mots clés*.

Une fois ces mots clés extraits des documents, la base de documents est représentée sous la forme d'une matrice mots clés - documents, appelée *matrice mot clé-document*. Une ligne de cette matrice représente les coordonnées d'un document dans l'espace des mots clés, la valeur d'un élément étant le nombre d'occurrences d'un mot clé dans le document. Et une colonne représente un mot clé dans l'espace des documents, la valeur d'un élément étant alors le nombre d'occurrences du mot clé dans un document.

Une requête subit typiquement exactement le même prétraitement et la recherche consiste simplement en un classement des documents les plus proches, au sens d'une norme choisie préalablement, de la requête.

Les aspects du modèle vectoriel indépendants de la nature des documents seront présentés plus en détail dans la partie 2.4, page 22.

2.3 Recherche d'information sur les images

Le domaine de la recherche d'images se divise en deux classes qui étaient encore bien distinctes quelques années auparavant. Ces deux catégories se distinguent fondamentalement par la manière dont est vue la recherche d'image. Inoue [Inoue04] propose une comparaison intéressante entre ces deux familles de systèmes que sont les ABIR³ et les CBIR⁴.

Dans le premier cas, les images sont associées à des annotations (texte environnant, description, date de création, auteur...) et une requête à un tel système se fait par le texte. Tandis que dans le second cas, le système ne se soucie pas du contexte, mais uniquement au contenu des images et une requête se fait par l'image.

2.3.1 Recherche d'information basée sur les annotations

Un système de ce type considère donc des documents de type images annotées ou prises dans un contexte textuel et une requête est de forme textuelle.

³ABIR : Annotation Based Image Retrieval

⁴CBIR : Content Based Image Retrieval

2.3 Recherche d'information sur les images

La plupart des systèmes de recherche d'images accessibles au grand public ([GoogleImages], [flickr]...) se basent sur des informations provenant d'annotations de l'image et sont totalement indépendants du contenu de celles-ci. On sait que Google Images indexe les images du web en fonction du texte qui les entoure et Flickr indexe les images de sa base de données en fonction des mots-clés que les utilisateurs affectent à leurs images.

Dans les deux cas, une requête est purement textuelle. Inoue [Inoue04] justifie la popularité de ce type de systèmes par leur facilité d'utilisation et leur efficacité. En effet, si les annotations sont bien construites, la recherche devient une simple recherche textuelle. Le niveau sémantique du texte étant assez élevé, il est alors possible de soumettre des requêtes complexes, avec noms propres ou scènes particulières.

Une recherche de ce type portera donc sur le *sens* des images recherchées.

Cependant, rien ne garantit que les annotations de l'image sont sémantiquement proches du contenu de l'image. La qualité du système est donc entièrement dépendante de la qualité des annotations. C'est un risque que l'on peut prendre dans le cas d'images personnelles, mais qui est risqué dans le cas d'Internet où des images illégales mal indexées pourraient ne pas être filtrées et diffusées de cette manière.

Nous choisirons donc d'utiliser le contexte textuel des images pour les informations sémantiques que l'on peut en tirer. Mais conscients des lacunes que l'utilisation seule de ces informations présente, nous ne nous en contenterons pas et utiliserons en plus les informations contenues dans les images elles-mêmes.

2.3.2 Recherche d'information basée sur le contenu

Les documents sont ici des images seules, sans autre information disponible que ce qu'elles contiennent, et les requêtes sont de même nature.

Contrairement aux systèmes grand public décrits ci-dessus, les systèmes issus de la recherche s'appuient essentiellement sur les informations contenues dans les images. On peut citer parmi ces systèmes QBIC⁵ d'IBM ([Flickner95]), VisualSEEk de l'université de Columbia ([SmithChang96]), Photobook de MIT Media Lab ([Pentland96]), MARS⁶ de l'Université de Californie ([Mehrotra97]).

Mais un tel système sera jugé efficace par les utilisateurs si les réponses qu'il lui apporte satisfont son besoin d'information. Or le seul indice sur ce

⁵QBIC : Query By Image Content

⁶MARS : Multimedia Analysis and Retrieval System

besoin d'information dont le système dispose est l'image requête et comme le *sens* d'une image est totalement dépendant de la personne qui l'interprète, l'hypothèse faite par grande majorité des systèmes est que plus une image est *visuellement ressemblante* à la requête, mieux elle répond au besoin d'information de l'utilisateur.

Naturellement, cette hypothèse est largement discutable. Cependant, le *fossé sémantique* étant toujours infranchi, il est à présent quasiment toujours impossible d'extraire la sémantique, ou plutôt *les* sémantiques, d'une image. Par conséquent, la ressemblance *visuelle*, et non *sémantique*, est le seul moyen de mesurer la pertinence d'une image à l'heure actuelle.

Reste à définir ce que l'on appelle *ressemblance visuelle*. Ici encore, les avis divergent. Certains considéreront les informations de l'image entière, d'autres extrairont les informations depuis un ensemble de sous-fenêtres, d'autres tenteront de trouver et de décrire les points d'intérêt de l'image et enfin, certains s'attacheront à tenter de découvrir et reconnaître les objets contenus dans l'image.

Dans notre objectif de fusionner texte et image, il nous semble important que les aspects visuels et textuels soient de nature comparable. Nous allons donc utiliser le modèle vectoriel appliqué aux images, qui est le modèle le plus largement utilisé dans ce domaine. Dans ce but, nous allons tenter de construire un *vocabulaire visuel*, comparable au concept de vocabulaire vu dans la partie 2.2.3, page 11. Le concept de *thésaurus visuel* a été introduit par Picard [Picard95] dans le domaine de la vidéo, Lim [Lim01] propose l'idée moins ambitieuse et cependant très efficace de *vocabulaire visuel* pour l'indexation d'images.

Dans quasiment tous les systèmes, le schéma d'indexation des images reste le même :

- échantillonner l'image en imagerie
- extraire un ensemble de caractéristiques de ces zones
- parfois, regrouper ces zones en classes

[Smeulders00] propose une revue intéressante, critique et assez complète du domaine CBIR.

Échantillonnage

Cette étape consiste en la sélection et le découpage de l'image d'origine en zones d'où certaines caractéristiques visuelles seront extraites. L'idée principale est d'arriver à ce que les zones choisies apportent le plus d'informa-

2.3 Recherche d'information sur les images

tions et soient le plus discriminantes possibles, l'idéal étant d'avoir un objet présent par zone.

Sans échantillonnage Cette méthode conserve l'image entière. Elle est très mauvaise, car les objets présents dans l'image ne sont pas distingués lors de l'extraction des caractéristiques.

Ceci peut toutefois convenir pour une base d'images très petite et où les images sont *visuellement* assez éloignées les unes des autres et où un seul objet se trouve dans l'image, concordance de propriétés somme toute improbable.

Échantillonnage régulier Cette technique définit un quadrillage fixe de l'image. L'image d'origine peut être découpée en plusieurs imasettes de manière régulière ([PicardMinka95]), pouvant se superposer ([Lim01]) et être multi-échelles. Les caractéristiques sont extraites sur chaque imasette.

Dans la version la plus simple, un objet caractéristique de l'image peut se trouver découpé dans plusieurs imasettes. Lors de l'extraction des caractéristiques, son influence sera faible pour chaque imasette et donc faible pour l'image entière. Une grande quantité d'information sera alors perdue.

La version avec superposition des imasettes permet de résoudre en partie le problème précédent en augmentant l'indépendance du système à une translation des objets entre deux images.

L'échantillonnage multi-échelle augmente souvent la qualité des résultats en augmentant l'indépendance du système à la taille des objets dans l'image. Dans cette méthode, les caractéristiques sont calculées sur plusieurs échantillonnages de tailles différentes.

Un système combinant échantillonnage avec superposition et multi-échelle sera donc très peu sensible à la position et à la taille des objets dans l'image.

Échantillonnage aléatoire Cette méthode, utilisée par [Maree05] pour la classification de photographies, consiste en la sélection d'imasettes de position, de taille et d'orientation aléatoire. Cette méthode permet d'avoir une très grande invariance du système à la translation, la taille et à la rotation des objets dans les images.

Toutes les méthodes de sélection de zones où les caractéristiques seront extraites étudiées jusqu'ici sont indépendantes du contenu des images.

Ces méthodes sont robustes et assez efficaces pour une complexité de mise en oeuvre faible, cependant le choix de la taille des imasettes, du décalage lors de la superposition, du nombre d'échelles lors de l'utilisation de la méthode multi-échelle et du nombre d'imasettes lors de l'échantillonnage aléatoire est

totalemment empirique et dépendant de la base de documents.

Détection de points d'intérêt

Une méthode dépendante des informations contenues dans les images est la détection de points d'intérêts. Les points d'intérêt d'une image sont souvent des points où une des caractéristiques de l'image change brutalement.

Les différences de gaussiennes, le détecteur SIFT⁷ ([Lindberg98]) et le détecteur de Harris (HarrisStephens88) sont des algorithmes permettant de retrouver les points d'intérêts d'une image.

Les zones autour des points d'intérêt portent beaucoup d'information. De plus, les zones détectées dépendent directement du contenu des images.

La détection de ces zones et l'extraction de caractéristiques de ces zones assure donc de tirer un maximum d'information locale sur les images. C'est pourquoi cette technique est très utilisée pour la reconnaissance d'objet de forme fixe en traitement d'images.

Mais les informations locales sur un certain nombre de points ne sont pas suffisantes pour décrire l'image dans sa globalité.

Segmentation

Cette méthode dépend aussi directement des données. Il existe deux types de segmentation :

Segmentation forte L'objectif d'un algorithme de segmentation forte est de découper l'image d'origine en zones telles qu'une zone contienne la totalité d'un et un seul objet.

Segmentation faible Un algorithme de segmentation faible est plus permissif, son objectif est de découper l'image en zones telles que chaque zone soit incluse dans un objet.

Ces algorithmes sont basés sur un critère d'homogénéité de la zone. [ForsythFleck99] utilise la couleur, [Carson97] et [MirmehdiPetrou00] utilisent couleur et texture comme critères d'homogénéité des régions.

L'hypothèse faite dans le cas de la segmentation forte est qu'un objet possède au moins une caractéristique visuelle homogène sur sa surface, ce qui est très rare en pratique. La segmentation faible permet d'éliminer cette hypothèse et de résoudre en partie le problème posé par l'occlusion partielle d'un objet.

Ces algorithmes sont donc utilisés pour la reconnaissance et la détection

⁷SIFT : Scale-Invariant Feature Transform

2.3 Recherche d'information sur les images

d'objets dans les images. Cependant ils restent, à l'heure actuelle, généralement assez peu fiables et la segmentation forte n'est efficace qu'appliquée à un domaine précis avec des techniques spécifiques à ce domaine.

L'objectif ici n'étant pas de modéliser le meilleur SRI possible, mais de comparer différents SRI, nous choisirons la méthode simple d'échantillonnage régulier des images lors de l'indexation.

Extraction des caractéristiques

Une fois l'ensemble de zones défini, il s'agit de les décrire. L'objectif des descripteurs est de discriminer au mieux les objets différents tout en faisant en sorte qu'un type d'objet ait une description le moins sensible aux variations du contexte de l'image (transformations géométriques, variations de luminosité. . .) que possible.

Le problème de discriminer deux objets différents, mais visuellement semblables est l'équivalent du problème de la polysémie transposé au domaine visuel. De même, le problème d'identifier deux images visuellement très différentes, mais représentant le même objet est l'équivalent du problème de la synonymie transposé au domaine visuel.

Nous cherchons à résoudre au mieux ce double problème afin rapprocher le niveau sémantique exprimé par les images de celui exprimé par texte, dans le but de comparer ces modalités de manière aussi pertinentes que possible. Cependant, nous ne chercherons pas à utiliser les algorithmes les plus complexes afin de nous focaliser sur le sujet de notre étude, et non sur un objectif de performance du SRI modélisé, qui n'est pas notre but ici.

Il existe trois principaux types de descripteurs. La couleur, la texture et les contours s'appliquent aux régions issues de l'échantillonnage ou de la segmentation de l'image, tandis que les descripteurs locaux s'utilisent sur les points d'intérêts détectés dans l'image.

Descripteurs globaux Les descripteurs globaux parcourent la région en extrayant les caractéristiques voulues. Le résultat est un vecteur décrivant la zone du point de vue de la caractéristique considérée.

Couleur Cette caractéristique est très importante, car c'est la composante principale que l'œil humain perçoit et elle possède un lien étroit avec la sémantique des objets. Cependant, il existe un grand nombre de manières de représenter cette information.

Le problème posé par cette caractéristique est que la couleur d'un objet telle qu'un observateur (humain ou électronique) la perçoit dépend non seulement de la couleur intrinsèque de l'objet, mais aussi de l'éclairage auquel il

est soumis, de la nature dont il est constitué, du point de vue de l'observateur et de l'observateur lui-même. L'objectif d'un descripteur de couleur en RI est donc de faciliter l'interprétation de cette caractéristique en faisant en sorte de rendre la représentation de la couleur la moins dépendante du contexte que possible.

Pour cela, il existe une grande variété d'espaces de représentation dont les plus utilisés sont résumés ici.

RGB ⁸Cette manière de représenter la couleur est extrêmement basique, puisqu'aucun traitement n'est nécessaire, mais particulièrement sensible aux changements de conditions et peu discriminante. C'est-à-dire qu'un même objet pourra avoir plusieurs descriptions différentes selon le contexte et deux objets différents pourront avoir la même description.

Cependant, cela peut suffire lorsqu'il s'agit d'images représentant des images, donc où le contexte n'intervient pas. Le système QBIC ([Flickner95]) utilise cette représentation de la couleur.

HSV ⁹Cette représentation est plus sophistiquée. En effet, elle nécessite un changement de base de l'espace *RGB* vers l'espace *HSV*. Les trois composantes de cet espace sont la teinte (*hue*), la saturation (*saturation*) et la valeur (*value*) de la couleur. La teinte est la couleur intrinsèque de l'objet, la saturation est la pureté de la couleur perçue et la valeur mesure combien la couleur perçue est sombre ou claire.

Ce domaine de représentation est plus efficace que RGB, car il est indépendant du contexte lumineux de l'objet donc un objet aura une représentation plus homogène. Il a été notamment utilisé dans les systèmes MARS ([Mehrotra97]) et VisualSEEk ([SmithChang96]).

L*a*b* La représentation *CIE 1976 L*a*b**, définie par la *Commission Internationale d'Éclairage* est souvent considérée comme la plus proche de la perception humaine. En effet, elle a été construite de façon à ce que la distance euclidienne entre deux vecteurs dans cet espace modélise le plus fidèlement possible la différence entre deux couleurs telle que l'oeil humain la perçoit. Les composantes de cet espace ont le sens suivant : *L* indique combien la couleur est sombre, *a* indique la quantité de bleu et *b* indique la quantité de jaune.

Cette description de la couleur est très intéressante dans le cadre de la RI, de par ses propriétés sémantiques, mais les formules de transformation étant non linéaires, le temps de calcul peut être trop long.

⁸RGB : Red Green Blue

⁹HSV : Hue Saturation Value

2.3 Recherche d'information sur les images

Nous choisirons d'utiliser la représentation qui nous semble être le meilleur compromis entre simplicité et efficacité : la représentation RGBL. Cette représentation inclut l'espace RGB classique, plus la moyenne de ces valeurs (*i.e.* le niveau de gris). Le descripteur utilisé sera un *histogramme de couleur* de la zone dans le domaine RGBL. Les résultats donnés par cette simple représentation vont nous permettre de faire une première discrimination des imagerie extraites, en séparant celles qui présentent des différences au niveau des couleurs.

Cependant, l'utilisation seule des informations sur la couleur n'est pas assez discriminant. En effet, une forte présence de bleu dans une imagerie s'ignifie-t-elle la présence d'eau, de ciel, ou d'autre chose ? Nous avons donc besoin d'autres descripteurs.

Texture Cette caractéristique est aussi largement utilisée dans la plupart des CBIR (Lim [Lim01], QBIC [Flickner95], PhotoBook [Pentland96], MARS [Mehrotra97]). On appelle *texture* la variation répétitive de l'intensité lumineuse résultant de propriétés physiques ou photométriques des objets observés. L'Homme est sensible aux textures, mais il est difficile de les définir formellement. Les principales méthodes d'extraction de textures proposées se classent en deux catégories :

Méthodes statistiques Ces méthodes s'appuient sur la construction de matrices de co-occurrence ou d'autocorrélation représentant les distributions spatiales des intensités de l'image. Ces méthodes ont été étudiées par Gottlieb *et al.* [GottliebKreyszig90] et utilisées par Lin *et al.* [Lin97].

Méthodes génératives Ces méthodes consistent en l'application de filtres à base d'ondelettes sur l'image (Li *et al.* [LiWang03]). Le filtre le plus souvent utilisé étant le filtre de Gabor. Cet algorithme est efficace pour décrire un grand nombre de texture de manière très discriminante, comme le montre Jain *et al.* [JainFarrokhnia90].

D'autres techniques géométriques s'intéressent à la variation du gradient de luminosité. C'est le cas par exemple du descripteur *histogramme de contours* inclus dans la norme MPEG-7. Won *et al.* [Won02] donnent une description détaillée de ce descripteur.

La quantité d'information portée par la texture est grande et permet de distinguer des objets que les informations issues de la couleur seule n'ont pas suffi à différencier. En effet, si le ciel et la mer peuvent avoir approximativement la même couleur, les cirrus présents dans le ciel et les vaguelettes sur la mer présenteront des textures différentes qui, idéalement, feront que ces images seront bien différenciées.

Nous utiliseront le filtre de Gabor qui a prouvé son efficacité dans le domaine des CBIR.

Descripteurs locaux Ce type de descripteurs s'appliquent aux points d'intérêts extraits de l'image d'origine. Il existe une grande variété de descripteurs présentant chacun des propriétés d'invariance aux transformations géométriques affines de l'image, aux variations lumineuses et aux changements de points de vue, comme Mindru *et al.* [Mindru99], Mikolajczyk *et al.* [MikolajczykSchmid01] et Lowe [Lowe04] le proposent.

Le descripteur SIFT issu de l'idée originale de Lindeberg [Lindberg98] et décrit par Lowe [Lowe04] est notamment très efficace, car il présente une invariance intrinsèque aux changements d'échelles et une normalisation permet de le rendre invariant aux changements de luminance. Sivic *et al.* montrent l'efficacité de ce descripteur dans la découverte et la recherche d'objets dans les vidéos [Sivic04] et les images [Sivic05]. De plus, son implémentation est très simple, puisqu'elle consiste simplement en un calcul d'histogramme de gradients sur de petites fenêtres autour du point d'intérêt.

Cependant, ces descripteurs sont plutôt adaptés à une tâche de recherche d'objet particulier, et nous pensons que l'utilisation de descripteurs globaux est suffisante pour une première confrontation des modèles définis dans notre problématique.

Les descripteurs choisis pour le traitement des images contenues dans les documents sont les descripteurs classiquement utilisés en CBIR, choisis à la fois pour leur pouvoir de discrimination et leur simplicité.

L'utilisation de descripteurs plus complexes améliorerait certes les résultats des modèles basés uniquement sur les images et peut-être ceux des modèles combinant texte et images dans l'absolu, mais nous rappelons que notre objectif est relatif. C'est-à-dire que le chiffre significatif de notre étude sera le rapport entre les résultats donnés par les modèles combinant texte et image, *par rapport* à ceux donnés par les modèles n'utilisant que les images.

Regroupement des vecteurs de caractéristiques

À la suite de la phase d'extraction des caractéristiques, chaque zone composant chaque image de la base de documents est décrite par un vecteur de caractéristiques. La quantité de cas de *synonymie* entre ces vecteurs est alors très grande. En effet, étant donné un objet présent dans plusieurs images, chaque zone contenant une partie de cet objet aura un vecteur de caractéristique différent (sauf cas exceptionnel). Il faut alors trouver un moyen

2.3 Recherche d'information sur les images

de regrouper ces vecteurs en un ensemble homogène faisant référence à l'objet lui-même plutôt à une instance de cet objet. Cette phase est appelée *clustering*.

L'étape analogue à celle-ci dans le domaine textuel serait la lemmatisation (*cf.* 2.2.3, page 11). En effet, cette étape va chercher à rassembler un ensemble de vecteurs proches (mots ayant la même racine) en un représentant unique (racine du mot) alors appelé *mot clé visuel* (mot clé).

L'hypothèse faite ici est que deux vecteurs caractéristiques proches représentent le même objet, ce qui est idéalement vrai si les extracteurs de caractéristiques sont assez discriminants et invariants aux changements du contexte de l'objet.

Regroupement hiérarchique Cette méthode de regroupement est récursive, dans le sens où la création d'un nouveau groupe de documents se fait en fonction des groupes déjà existants.

La version *agglomérante* de cette méthode considère initialement chaque élément comme un groupe, puis associe les groupes les plus proches (au sens d'une norme donnée) jusqu'à n'obtenir qu'un nombre de groupes défini à l'avance. Cette technique a été largement utilisée dans de nombreux domaines, comme par Eisen *et al.* [Eisen98] pour la bioinformatique.

La version *divisante* considère initialement l'ensemble de tous les éléments comme un groupe, puis divise chaque groupe en deux jusqu'à obtenir le nombre de groupes voulu. Savaresi *et al.* [SavaresiBoley04] comparent l'efficacité cette méthode à des méthodes de partitionnement.

Regroupement par partitionnement Contrairement au regroupement hiérarchique, cette technique de regroupement peut être vue comme itérative. Il existe plusieurs implémentations de cette technique, dont la plus répandue est *k-means*.

Cette méthode commence par affecter aléatoirement chaque élément à un groupe, puis calcule les coordonnées de l'isobarycentre de chaque groupe. Les éléments sont ensuite réaffectés au groupe dont l'isobarycentre est le plus proche (au sens d'une norme donnée). À la suite de quoi, les coordonnées de l'isobarycentre de chaque groupe sont mises à jour, etc. Le critère d'arrêt de cet algorithme est le plus souvent un seuil sur le nombre d'éléments changeant de groupe entre deux itérations.

L'algorithme *fuzzy-c-means* est la version floue de l'algorithme précédent. L'algorithme sous-jacent commun à toutes ces approches est l'algorithme

EM¹⁰ (McLachlan *et al.* [McLachlanKrishnan97])

L'algorithme *k-means* a prouvé son efficacité dans de nombreux domaines et permet une interprétation intuitive des groupes générés. Nous allons donc intégrer cette méthode dans notre modèle. Ceci permettra d'élever le niveau sémantique des informations issues de l'image au rang de *mots clés visuels*, plus proches des *mots clés* issus du texte.

À la fin de la phase de prétraitement des images, on peut construire une matrice mot clé-document, tout comme dans le cas du modèle vectoriel textuel 2.2.3, 11. Une ligne représentera les coordonnées d'un document dans l'espace des mots clés visuels, un élément représentant le nombre d'occurrences d'un mot clé visuel dans le document. Tandis qu'une colonne représente un mot clé visuel dans l'espace des documents, un élément étant alors le nombre d'occurrences de ce mot clé visuel dans un document.

De la même manière que pour le texte, une image requête subit le même prétraitement qu'une image indexée. Et la recherche consiste en un classement des documents par distance à la requête croissante, au sens d'une norme donnée.

2.4 Modèle vectoriel

Nous avons vu précédemment comment obtenir une description vectorielle des documents, textuels ou visuels, et des requêtes, textuelles ou visuelles. La forme obtenue pour ces deux modalités est la même, une matrice mot clé-document, nous pouvons alors nous abstraire de leur origine en appelant indifféremment *document* un texte ou une image, *mot clé* un mot clé textuel ou visuel et *requête* une requête textuelle ou visuelle.

2.4.1 Pondération des termes

Après prétraitement, une matrice mot clé-document représente pour chaque document, un histogramme des mots clés. L'importance des mots clés est donc bien prise en compte, si l'on fait l'hypothèse intuitive que plus un mot clé est représenté dans un document, plus il est représentatif de ce document.

Une première observation simple permet de mesurer la nécessité d'une normalisation de ces histogrammes. En effet, si d'une part le document d_1

¹⁰EM : Expectation Maximisation

2.4 Modèle vectoriel

contient 3 occurrences du mot clé m_1 et 1 occurrence du mot clé m_2 et d'autre part le document d_2 contient 3 occurrences du mot clé m_1 et 10 occurrence du mot clé m_2 , on ne peut pas dire que le mot clé m_1 est aussi représentatif du document d_1 que du document d_2 .

Par conséquent, une première normalisation de la matrice mot clé-document est de considérer la *fréquence* des mots clés au sein des documents. On appelle cette normalisation tf ¹¹. Elle permet de représenter les documents indépendamment de leur taille.

Si l'on observe un document indépendamment des autres documents de la base et sans connaissance supplémentaire, tf est la représentation rendant le mieux compte de l'importance d'un mot clé pour un document. Cependant, il est possible d'utiliser les informations fournies par les autres documents de la base afin de pondérer l'importance des mots clés. En effet, si un mot clé apparaît uniformément dans tous les documents, il ne sera pas plus représentatif d'un document que d'un autre. Le pouvoir de discrimination d'un tel mot clé est donc nul et son importance doit donc être réduite.

Chaque mot clé sera donc pondéré par un poids inversement proportionnel au nombre de documents qui en contiennent une occurrence. On appelle cette pondération idf ¹². Elle permet de donner plus d'importance aux mots clés les plus discriminants.

Cette méthode est quasiment systématiquement utilisée dans tous les systèmes basés sur le modèle vectoriel (Salton *et al.* [SaltonBuckley88]). Nous allons nous aussi utiliser cette pondération afin de rendre les résultats plus significatifs.

2.4.2 Indexation par sémantique latente

Une technique intéressante dans le but de résoudre les problèmes de synonymie et polysémie est d'observer le contexte dans lequel apparaissent les mots clés au sein des documents.

Par exemple, si un grand nombre de documents contiennent le mot clé textuel *nuage* et le mot clé visuel *zone blanche uniforme* et un grand nombre de documents contiennent le mot clé textuel *neige* et le mot clé visuel *zone blanche uniforme*, et si les mots clés *nuage* et *neige* apparaissent rarement dans les mêmes documents, il est possible d'inférer que les concepts qu'ils représentent sont différents, donc que le mot clé *zone blanche uniforme* peut être associé à deux concepts différents. On constate donc qu'il est possible de résoudre certains problèmes de polysémie en acceptant que le sens d'un mot

¹¹ tf : term frequency

¹² idf : inverted document frequency

clé puisse dépendre non seulement des autres mots clés apparaissant dans le même document, mais aussi des co-occurrences sur tous les documents de la base.

D'autre part, si un grand nombre de documents contiennent le mot clé textuel *pomme de terre* et le mot clé visuel *zone jaune clair tachetée de marron*, et si ce mot clé visuel n'apparaît qu'en la présence de l'un de ces mots clés textuels, il est possible d'inférer que le mot clé visuel *zone jaune clair tachetée de marron* représente un concept unique dont les mots clés textuels *pomme de terre* et *patate* ne sont que des instances. On constate donc qu'il est aussi possible de résoudre certains problèmes de synonymie.

Les exemples donnés ci-dessus expliquent très exactement le fonctionnement théorique de la méthode de LSI¹³. Cette technique considère que deux mots apparaissant fréquemment dans le même contexte, même s'ils n'apparaissent jamais dans les mêmes documents, auront le même sens et seront donc considérés comme des synonymes. Réciproquement, un mot apparaissant fréquemment dans deux contextes différents aura deux sens différents et sera donc considéré comme polysémique.

Ainsi, lors du traitement d'une requête, si celle-ci contient un mot dont un synonyme a été détecté dans le vocabulaire, le synonyme sera aussi considéré comme faisant partie, de manière implicite, de la requête. Réciproquement, un mot considéré comme polysémique apparaissant dans la requête prendra le sens le plus probable en fonction du contexte formé par les autres mots de la requête.

Cette technique non-supervisée est statistique et basée sur une approximation de la matrice mot clé-document par une matrice de rang inférieur. Ceci permet de regrouper les mots clés en *sacs de mots clés* par propagation sémantique implicite (Deerwester [Deerwester90]).

Cette méthode est très proche de ce que certains psychologues pensent être l'apprentissage humain (Landauer *et al.* [LandauerDumais97]) et a été appliquée avec succès dans le domaine de la recherche d'informations sur des documents textuels multilingues (Rehder *et al.* [Rehder97], Dumais *et al.* [Dumais97]). Si l'on considère ces deux aspects, tout laisse à penser que LSA permettra de combiner efficacement le texte et les images.

En effet, du point de vue de l'analogie avec la psychologie humaine, le premier moyen de reconnaissance d'objets par les enfants, alors qu'ils n'ont pas de vocabulaire, est le principe d'association. C'est-à-dire qu'on apprend à un bébé à reconnaître les objets en les lui montrant (image) tout en prononçant le nom de l'objet (mot). On peut aussi faire une analogie entre la RI

¹³LSI : Latent Semantic Indexing

2.5 Recherche d'information sur documents contenant texte et images

sur documents multilingues et la RI sur documents multimodaux. En effet, dans le cas qui nous intéresse, l'image et le texte décrivent la même scène dans d'un document, on peut donc considérer que l'un est la traduction de l'autre dans un niveau sémantique différent.

2.5 Recherche d'information sur documents contenant texte et images

Après avoir décrit les différents outils à disposition pour traiter notre problématique et sélectionné ceux qui nous semblent les plus intéressants, nous allons passer en revue un ensemble de travaux ayant abordé le problème spécifique de la fusion des informations textuelles et visuelles pour l'indexation et la recherche.

K. Barnard et D. Forsyth *et al.* 2001 2003 [BarnardForsyth01] [Barnard03]

Des travaux récents associant explicitement image et texte ont été faits à l'université de Californie, à Berkeley, par Barnard et Forsyth. Ils utilisent un algorithme de segmentation des images pour l'annotation automatique.

Les caractéristiques utilisées sont la taille des régions extraites, leur position, la moyenne et variance de RGB, L^*a^*b et $(\frac{R}{R+G+B}, \frac{G}{R+G+B})$ sur chaque région, la moyenne et variance de seize descripteurs de texture. Leur système modélise les statistiques d'occurrence et mot clé-document des mots et des caractéristiques visuelles grâce à un modèle probabiliste.

Si cette approche a fourni de bons résultats pour certains types d'images, les auteurs avouent que le système est largement dépendant de la qualité de la segmentation des images, qui est rarement bonne dans le cas d'une base d'images hétéroclites.

S. Sclaroff, M. la Cascia et S. Sethi 1998 [Sclaroff98]

Ces travaux de Sclaroff, la Cascia et Sethi du département d'informatique de l'université de Boston, Massachusets, proposent de retrouver la sémantique du texte grâce à LSA. D'autre part, les informations visuelles sont extraites grâce aux histogrammes de couleur et de contour. Ces caractéristiques visuelles sont regroupées grâce à la méthode d'analyse en composantes principales.

Cependant, les deux modalités ne sont pas combinées à travers LSA. Nous pensons que les résultats obtenus en appliquant LSA sur les informations venant des deux sources seront meilleurs qu'en appliquant

LSA séparément sur chaque source.

T. Westerveld 2000 [Westerveld00] Westerveld, du département informatique de l'université de Twente, Hollande, applique la méthode LSA aux informations issues des deux modalités. Les caractéristiques utilisées sont l'histogramme HSV et le filtre Gabor. Il n'applique pas de méthode de regroupement des vecteurs caractéristiques.

La conclusion de Westerveld est que la combinaison à travers LSA des deux modalités n'est pas toujours plus performante que l'utilisation séparée de l'une ou l'autre, mais il ne donne pas de résultats chiffrés permettant de faire une comparaison quantitative des modèles. Le fait qu'il n'observe pas d'amélioration significative peut être dû à l'absence de regroupement des caractéristiques visuelles qui permet d'élever le niveau sémantique des informations visuelles.

Nous allons montrer, en nous basant sur des résultats qualitatifs, que dans notre cas cette stratégie améliore considérablement les résultats.

R. Zhao et W. Grosky 2002 [ZhaoGrosky02] Zhao, de l'université de New York, et Gorsky, de l'université de Michigan, utilisent LSA sur les informations textuelles et visuelles combinées. Leur descripteur visuel, *anglogram* est particulièrement efficace pour décrire la forme, la couleur et la structure des images.

Les résultats qu'ils obtiennent semblent très bons à première vue, mais leur base de test ne comprenait que 20 documents. Leurs résultats ne sont donc pas significatifs, mêmes s'ils tendent à montrer que la démarche est efficace. Notre expérimentation sera faite sur un corpus de 4500 documents, ce qui permettra d'exposer des résultats interprétables.

Notre contribution au domaine est donc de montrer l'avantage que procure l'utilisation combinée du texte et des images en utilisant la méthode LSA, sur une base de documents non spécialisée et de taille permettant de tirer des conclusions significatives.

Chapitre 3

Modèle de fusion

Dans cette section, nous allons présenter et expliciter chaque partie des modèles que nous allons comparer. La plupart de ces parties seront communes à chaque modèle. Les seuls paramètres que nous allons faire varier seront l'utilisation ou non de LSA, la prise en compte ou non du texte et la prise en compte ou non des images.

3.1 Prétraitements

Nous nous plaçons dans le cadre du modèle vectoriel. Dans ce modèle, la phase de prétraitement consiste en la transformation des documents tels qu'ils existent dans la base en un vecteur les représentant au sein du système.

Cette partie est entièrement dépendante de la base de documents utilisée, puisqu'elle est l'interface entre celle-ci et le système. Pour être le plus général possible, nous considérerons l'éventualité où chaque document est constitué d'une ou plusieurs parties textuelles et d'une ou plusieurs images.

Partant de ces données, nous allons leur appliquer plusieurs transformations qui vont permettre de construire deux matrices mot clé-document, l'une textuelle et l'autre visuelle, modélisant la base de documents au sein du modèle.

Nous noterons :

$$\left\{ \begin{array}{l} \mathcal{D} = \text{ensemble des documents de la base} \\ \mathcal{T} = \text{ensemble des textes de la base} \\ \mathcal{V}_T = \text{vocabulaire des textes de la base} \\ \forall d \in \mathcal{D} : \mathcal{T}_d = \text{ensemble des textes associés au document } d \\ \mathcal{I} = \text{ensemble des images de la base} \\ \forall d \in \mathcal{D} : \mathcal{I}_d = \text{ensemble des images associées au document } d \\ |.| = \text{taille de l'ensemble ou du vecteur .} \end{array} \right.$$

3.1.1 Texte

L'étape de prétraitement du texte réside en l'extraction des mots clés textuels. Un lemmatiseur tel que l'algorithme de Porter [Porter80] est utilisé pour cette phase du processus, *c.f.* partie 2.2.3, page 11.

Une fois chaque texte de chaque document traité, le lemmatiseur nous donne un ensemble de termes que nous allons considérer comme des mots clés. Ces mots clés constitueront notre vocabulaire textuel \mathcal{V}_T . Nous allons alors simplement compter les mots clés présents dans chaque document afin de construire la matrice mot clé-document textuelle M^T .

Nous définissons donc la matrice M^T , de taille $|D| \times |\mathcal{V}_T|$ de la manière suivante :

$$\forall d \in \mathcal{D}, m \in \mathcal{V}_T : M_{d,m}^T = \text{nombre d'occurrences du mot clé } m \text{ dans les textes } \mathcal{T}_d$$

Par exemple, le premier document de notre base est composé des mots *city*, *mountain*, *sky* et *sun*, indexés respectivement par les nombres 1, 2, 3 et 4. On aura donc :

$$M_{1,:}^T = \underbrace{(1, 1, 1, 1, 0, \dots, 0)}_{|\mathcal{V}_T|}$$

3.1.2 Image

De la même manière que pour le texte, le prétraitement des images consiste en l'extraction des mots clés visuels. Pour cela, nous allons découper chaque image de manière régulière mono-échelle. Puis nous extraieront les histogrammes de couleur RGBL et la texture grâce aux filtres de Gabor.

Ce choix d'échantillonnage et de caractéristiques est dicté par les résultats observés dans l'état de l'art. Enfin, les vecteurs de caractéristiques seront regroupés en classes avant de construire la matrice mot clé-document visuelle M^I .

Échantillonnage

Nous appliquerons un échantillonnage régulier mono-échelle. Nous découperons ainsi chaque image en 5×5 fenêtres. Ceci permettra de donner plus d'importance aux objets de taille moyenne dans l'image.

Nous noterons :

$$\begin{cases} \mathcal{F} = \text{ensemble des fenêtres extraites des images de la base} \\ \forall i \in \mathcal{I} : \mathcal{F}_i = \text{ensemble des fenêtres issues de l'image } i \end{cases}$$

3.1 Prétraitements

Dans notre cas, nous choisirons la valeur suivante :

$$|\mathcal{F}_{i,1}| = 25$$

Extraction des caractéristiques

Les caractéristiques que nous avons choisi d'extraire de chaque fenêtre sont l'histogramme de couleur RGBL, dont les canaux sont représentés par un histogramme de 32 classes (*bins*), et les coefficients issus de filtres de Gabor basés sur 5 échelles et 6 orientations. RGBL apporte une grande quantité d'informations pour une complexité faible et les filtres de Gabor sont insensibles à la rotation des objets et aux changements de luminosité.

En plus de ces caractéristiques purement visuelles, nous considérons les coordonnées relatives de la fenêtre d'où sont extraites les caractéristiques, conservant ainsi la structure interne de l'image. Cette information nous semble pertinente, car elle pourrait permettre de différencier une zone bleue se trouvant en haut de l'image, représentant le ciel, d'une zone bleue en bas de l'image, représentant la mer.

Nous noterons :

$$\left\{ \begin{array}{l} \mathcal{C} = \text{ensemble des histogrammes de couleurs extraits} \\ \forall f \in \mathcal{F} : c_f = \text{histogramme de couleurs sur la fenêtre } f \\ \mathcal{G} = \text{ensemble des coefficients de Gabor extraits} \\ \forall f \in \mathcal{F} : g_f = \text{coefficients de Gabor sur la fenêtre } f \\ \mathcal{P} = \text{ensemble des positions des fenêtres extraites} \\ \forall f \in \mathcal{F} : p_f = \text{coordonnées relatives de la fenêtre } f \end{array} \right.$$

Avec les valeurs suivantes :

$$\left\{ \begin{array}{l} \forall f \in \mathcal{F} : |c_f| = 128 \\ \forall f \in \mathcal{F} : |g_f| = 60 \\ \forall f \in \mathcal{F} : |p_f| = 2 \end{array} \right.$$

Regroupement des vecteurs de caractéristiques

À ce niveau, chaque fenêtre est décrite par un vecteur caractéristique de taille 190 et le nombre total d'images sur le corpus \mathcal{D} est de $25 \times |\mathcal{D}|$. Chacun de ces vecteurs est normalisés en moyenne à 0 et en variance à 1, de manière à les rendre plus facilement comparables.

Cette phase de clustering est centrale car elle nous permet d'élever le niveau sémantique des informations visuelles. C'est de plus à la suite de cette étape que nous pourrons regrouper les images en ensembles visuellement homogènes.

On applique ensuite l'algorithme *k-means* décrit en 2.3.2, page 21, sur ces vecteurs afin de constituer notre vocabulaire visuel. La distance choisie pour l'exécution de l'algorithme est la distance euclidienne et nous avons choisi de créer le même nombre de mots clé visuels que de mots clés textuels afin que texte et image aient le même poids lors de leur fusion.

Nous noterons :

$\mathcal{V}_I =$ ensemble des groupes de vecteurs caractéristiques

Avec :

$$|\mathcal{V}_I| = |\mathcal{V}_T|$$

Comme ça a été le cas pour le texte, nous allons construire la matrice mot clé-document entre les images et les mots clés visuels. À ce niveau, chaque imagerie est associée à un mot clé visuel. Nous pouvons alors compter, pour chaque document, le nombre d'occurrences de chaque mot clé visuel.

La matrice mot clé-document M^I , de taille $|\mathcal{D}| \times |\mathcal{V}_I|$ de la manière suivante :

$$\forall d \in \mathcal{D}, m \in \mathcal{V}_I : M_{d,m}^I = \text{nombre d'occurrences du mot clé visuel } m \text{ dans le document } d, \text{ i.e. dans les images } i \in \mathcal{I}_d$$

3.2 Indexation

Le principe de cette phase est de traiter les matrices mot clé-document issues du texte et des images afin de construire les vecteurs représentant chaque document dans un nouvel espace de représentation. Hormis la dimension des matrices, cette étape est totalement indépendante de la base de documents utilisée.

3.2.1 Pondération des termes

Dans chacun des modèles étudiés, nous transformons les matrices mot clé-documents calculées lors de la phase de prétraitement en pondérant chaque élément par $tf * idf$. Le calcul de ces coefficients sur la matrice mot clé-document (textuelle ou visuelle) M , de dimensions $|\mathcal{D}| \times |\mathcal{V}|$ se fait de la manière suivante :

$$\forall d \in \mathcal{D}, \forall m \in \mathcal{V} : \begin{cases} tf_{d,m} = \frac{M_{d,m}}{\sum_{j=0}^{|\mathcal{V}|} M_{d,j}} \\ idf_{d,m} = \frac{|\mathcal{D}|}{|d' \in \mathcal{D} : m \in d'|} \\ tf * idf_{d,m} = tf_{d,m} \times \log(idf_{d,m}) \end{cases}$$

3.2 Indexation

Les valeurs de la matrice mot clé-document sont alors remplacées par les valeurs de tf^*idf .

3.2.2 Fusion

Nous allons appliquer la fusion des deux modalités à ce niveau, juste avant d'appliquer LSA. Les deux matrices mot clé-documents M^T et M^I seront simplement concaténées dans le sens vertical. La taille de la matrice M^{TI} ainsi obtenue sera donc de taille $|\mathcal{D}| \times (|\mathcal{V}^T| + |\mathcal{V}^I|)$

En effet, cette *fusion* a du sens, car les mêmes traitements ont été appliqués aux deux modalités et leur niveau d'abstraction relativement à celui des données initiales sont les mêmes. Un document sera alors représenté pour moitié par un ensemble de *mots clés textuels* issus de M^T et pour moitié par des *mots clés visuels* issus de M^I .

Nous conservons par ailleurs les matrices M^T et M^I afin de pouvoir comparer les résultats obtenus par chaque modalité indépendamment et par la fusion des deux.

3.2.3 Indexation par sémantique latente

Cette phase du processus d'indexation n'est appliquée qu'aux modèles IMAGE&LSA, TEXTE&LSA et IMAGE&TEXTE&LSA (*c.f.* tableau 1.1, page 4). Les trois matrices M^T , M^I et M^{TI} seront transformées. La matrice M désignera par la suite indifféremment l'une d'entre elles et l'ensemble \mathcal{V} désignera indifféremment \mathcal{V}^T , \mathcal{V}^I et \mathcal{V}^{TI} .

Le vocabulaire \mathcal{V} contient *a priori* un ensemble de synonymes et de mots polysémiques qui n'ont pas pu être détectés ou résolus lors de la phase de prétraitement. La méthode LSI suppose qu'il existe une relation entre mots et documents gouvernée par un paramètre caché qui est le *sens*. L'objectif de cette étape est de réunir les *mots* du vocabulaire \mathcal{V} en ensembles sémantiquement homogènes.

L'idée est d'approximer la matrice mot clé-document pondérée M par une matrice de rang inférieur M_k . Ceci est fait en décomposant la matrice d'origine en valeurs singulières (SVD¹) et en ne conservant que les $k < rang(M)$ premiers vecteurs propres. Documents et mots deviennent alors des points de cet espace réduit.

La requête, vecteur d'occurrences de mots de taille $|\mathcal{V}|$, est projetée dans cet espace réduit et peut être considérée comme un *pseudo document*. La fonction de correspondance est alors simplement une distance entre la re-

¹SVD : Singular Value Decomposition

quête et les documents de la base dans ce nouvel espace sémantique.

Pour les détails de calcul, se référer à l'article de Deerwester *et al.* [Deerwester90].

On a :

$M =$ matrice mot clé-document pondérée, de rang r , de dimensions $|\mathcal{D}| \times |\mathcal{V}|$

Décomposition :

$$M = US^tV$$

avec :

$$\begin{cases} U^tU = {}^tVV = Id_{|\mathcal{V}|} \\ {}^tUU = V^tV = Id_{|\mathcal{D}|} \\ S = \text{diag}(\sigma_1 \dots \sigma_{\min(|\mathcal{D}|, |\mathcal{V}|)}) \end{cases}$$

avec :

$$\begin{cases} \forall i \in [2; \min(|\mathcal{D}|, |\mathcal{V}|)] : \sigma_{i-1} \geq \sigma_i \\ \forall i \in [r + 1; \min(|\mathcal{D}|, |\mathcal{V}|)] : \sigma_i = 0 \end{cases}$$

Cette transformation permet de représenter M comme un produit de deux informations différentes. La première information est relative aux documents et la seconde aux mots.

En utilisant les k plus grandes valeurs propres de M et en tronquant les matrices U et V en conséquence on obtient la meilleure approximation de rang k de M . Cette réduction de dimension permet de capturer l'information importante en éliminant en partie le *bruit*, *i.e.* traiter la synonymie et la polysémie telle que définies en 2.4.2, page 23.

Nous choisissons de prendre k égal au nombre de thèmes traités par les documents. Les mots synonymiques seront ainsi regroupés dans la même classe, car ils font partie du même thème, et les mots polysémiques. Ce nombre peut être évalué *a priori*, ou défini par la base de documents utilisée.

Le nombre optimal de dimensions dépend directement de la base utilisée. Comme nous ne soucions pas des performances absolues des systèmes, mais de leur différence de performance deux à deux, nous faisons ce choix qui n'est certainement pas optimal, mais permettra de pouvoir comparer les différents systèmes.

On a alors :

$$\begin{cases} U_k = \text{matrice } U \text{ tronquée, de dimensions } |\mathcal{V}| \times k \\ V_k = \text{matrice } V \text{ tronquée, de dimensions } |\mathcal{D}| \times k \\ S_k = \text{matrice } S \text{ tronquée, de dimensions } k \times k \\ M_k = U_k S_k^t V_k \end{cases}$$

La matrice M_k est alors de même dimension que M . Les colonnes de cette

3.3 Traitement de la requête

matrice représentent les coordonnées des documents dans l'espace des mots en prenant en compte les relations sémantiques existant entre ceux-ci. Par exemple, un mot absent d'un document aura une valeur non nulle dans ce document si un de ses synonymes y est présent.

3.3 Traitement de la requête

Une fois les documents de la base indexés, il s'agit d'interpréter la requête soumise de manière à ce qu'elle soit comparable aux documents tels qu'ils sont représentés au sein du système. Les mêmes traitements que ceux appliqués aux documents seront appliqués à la requête.

3.3.1 Prétraitements

Les mots clés textuels sont extraits du texte et leur nombre d'occurrences sont comptés de la même manière qu'expliqué en 3.1.1, page 28. La normalisation $tf*idf$ est appliquée de la même manière.

Pour les images, les imasettes sont extraites et leurs caractéristiques sont calculées. Puis chaque imasette est associée au mot clé visuel calculé en 3.1.2, page 29, le plus proche au sens de la norme euclidienne. Pour chaque imasette, le nombre d'occurrences de chaque mot clé visuel est compté et la normalisation $tf*idf$ est appliquée.

Nous noterons :

$$\begin{cases} q^T = \text{requête textuelle, de taille } |\mathcal{V}^T| \\ q^I = \text{requête visuelle, de taille } |\mathcal{V}^I| \\ q^{TI} = \text{requête bi-modale, de taille } |\mathcal{V}^{TI}| \end{cases}$$

3.3.2 Projection dans l'espace de sémantique latente

Dans cette partie, le vecteur désigné par q peut faire référence indifféremment à q^T , q^I ou q^{TI} . La requête exprimée de manière vectorielle et prétraitée doit encore être projetée dans l'espace réduit afin de pouvoir être comparée aux documents au niveau sémantique défini par LSA.

Pour cela, nous calculons le *pseudo document* q_k représenté par la requête de la manière suivante :

$$q_k = {}^t q U_k$$

3.3.3 Fonction de correspondance

Les lignes de la matrice $V_k S_k$ peuvent être interprétées comme les coordonnées des documents dans l'espace réduit. Nous utilisons alors la similarité *cosinus*² entre q_k et les lignes de $V_k S_k$ afin de retrouver les documents les plus *sémantiquement* proches de la requête.

Nous obtenons finalement les six modèles de recherche d'images annotées tels que décrits dans le tableau 1.1, page 4. Les seuls paramètres variants entre chaque modèle sont l'utilisation du texte, l'utilisation de l'image et l'utilisation de LSA.

Chapitre 4

Expérimentation

Après avoir conçu puis implémenté les différents modèles décrits dans le tableau 1.1, page 4, nous allons tâcher de les évaluer et de comparer leur résultats deux à deux et par rapport à l'état de l'art afin de pouvoir mesurer l'apport de chacun des paramètres.

Nous allons présenter la mesure de référence permettant de juger de manière objective les SRI, puis nous décrirons la base de documents utilisée et le protocole expérimental adopté. Nous comparerons nos résultats aux objectifs que nous nous seront préalablement fixés et aux résultats de l'état de l'art, puis nous discuterons ces résultats.

4.1 Évaluation des SRI

L'hétérogénéité des modèles de RI et la diversité de leurs implémentations que sont les SRI font qu'il est difficile de définir un critère objectif permettant de les comparer. De plus, il faut que ce critère modélise au mieux le but initial de la RI qui est la satisfaction du besoin d'information de l'utilisateur.

Les comparaisons les plus courantes sont basées sur une approche *boîte noire* des systèmes afin de s'abstraire du modèle et de son implémentation. Ainsi, deux modèles différents restent comparables.

Idéalement, le jugement d'un SRI devrait être fait par des utilisateurs finaux, puisque telle est la cible des SRI. Une évaluation devrait réunir un échantillon représentatif de la population ciblée par le système et leur demander de juger le système selon la pertinence des documents qu'il retourne.

Cependant, une telle expérimentation demanderait énormément de temps et de moyens financiers. La pertinence de l'utilisateur est donc habituellement

modélisé relativement à la base de documents utilisée.

4.1.1 Base de documents

Nous utiliserons pour notre expérimentation la base de documents Corel Image, composée de 4500 documents. Chaque document est associé à un thème. La base contient 50 thèmes différents auxquels sont associés 90 documents. Les thèmes sont très variés.

Chaque document est composé d'une image et d'un nombre de mots clés compris entre 1 et 5. Les images et les annotations peuvent être très différentes au sein d'une classe, comme elles peuvent être similaires entre deux classes différentes. Les images restent cependant du type photographies personnelles, *i.e.* pas d'images médicales ou militaires et les images n'ont pas été modifiées par ordinateur, prises par des professionnels, *i.e.* pas de flou, de sur- ou sous-exposition et les cadrages sont bons. De même, le vocabulaire est très restreint, seulement 374 mots, afin d'être assez général.

Nous avons donc :

$$\begin{cases} |\mathcal{D}| = |\mathcal{T}| = |\mathcal{I}| = 4500 \\ |\mathcal{V}_T| = |\mathcal{V}_I| = 374 \\ \forall d \in \mathcal{D} : 0 \leq |\mathcal{I}_d| \leq 5 \\ \forall d \in \mathcal{D} : |\mathcal{I}_d| = 1 \end{cases}$$

4.1.2 Pertinence

Nous modéliserons la pertinence utilisateur en admettant que si une requête fait partie du thème t , tous les documents classés dans ce thème auront une pertinence de 1 (*i.e.* documents totalement pertinents), tandis que tous les autres documents auront une pertinence de 0 (*i.e.* documents absolument pas pertinents).

Naturellement, cette modélisation est loin d'être parfaite (*c.f.* partie 4.5, page 49). Cependant elle permet d'avoir une vérité de référence et un ensemble de travaux ont été publiés en se référant à cette mesure. Cette pertinence de référence est appelée *vérité terrain*.

4.1.3 Évaluation

Étant donné une base de documents et une *vérité terrain*, une mesure de la qualité des SRI communément admise et utilisée, dans le cadre de la tâche de recherche de documents, est le tracé de la courbe *rappel-précision*. Cette courbe est incluse dans le carré $[0; 1] \times [0; 1]$ et la fonction

4.1 Évaluation des SRI

$precision(rappel)$ est décroissante. L'objectif est de faire en sorte que cette fonction soit constante, *i.e.* $\forall rappel \in [0; 1] : precision(rappel) = 1$. Autrement dit, l'objectif est de retrouver tous les documents pertinents en premier.

Pour la requête q , le *rappel* $R(q)$ est défini comme étant la proportion de documents pertinents au sens de la vérité terrain $\mathcal{P}(q)$ parmi tous les documents retrouvés par le système $\mathcal{D}_r(q)$.

$$R(q) = \frac{|\mathcal{D}_r(q) \cap \mathcal{P}(q)|}{|\mathcal{P}(q)|} \in [0; 1]$$

Le rappel permet de mesurer la capacité qu'a le système à retrouver *tous* les documents pertinents de la base. Ce facteur est important pour un SRI, car il relate la quantité de documents pertinents que le système proposera à l'utilisateur.

Avoir un rappel faible signifie retrouver peu de documents pertinents. Ceci est dangereux, car si la pertinence utilisateur a mal été modélisée et que parmi $\mathcal{D}(q)$ seuls peu de documents sont effectivement pertinents pour un utilisateur, ceux-ci ont moins de chance d'être retrouvés par le système. Le besoin d'information de l'utilisateur a donc moins de chance d'être satisfait.

La *précision* du système pour la requête q , relativement à la vérité terrain $\mathcal{P}(q)$ et aux documents retrouvés $\mathcal{D}_r(q)$

$$P(q) = \frac{|\mathcal{D}_r(q) \cap \mathcal{P}(q)|}{|\mathcal{D}_r(q)|} \in [0; 1]$$

La précision permet de mesurer la capacité qu'a le système à ne retrouver *que* les documents pertinents de la base. Ce facteur est moins important si la métrique de l'espace des documents est continue non binaire.

Par exemple, dans le modèle booléen, il n'y a pas de notion de distance entre documents. Le système décide si un document *est* ou *n'est pas* pertinent et ne retourne que ceux qu'il juge pertinents.

Il n'est donc pas possible de classer ces documents par ordre de pertinence décroissante. Les documents réellement pertinents peuvent alors se trouver n'importe où parmi les documents retrouvés.

Alors, avoir une faible précision revient pour un tel système à noyer les documents réellement pertinents parmi une grande quantité de bruit, ce qui rend le système totalement inutile.

En revanche, pour les modèles incluant une mesure continue de la notion de pertinence, il est possible de classer les documents retournés par le

système du plus pertinent au moins pertinent.

En supposant, et c'est quasiment toujours le cas dans la pratique, que les utilisateurs cherchent les documents qui leur sont pertinents parmi les documents retrouvés par le système de manière non-aléatoire (par exemple de gauche à droite et de haut en bas), il suffit de leur présenter dans l'ordre qu'ils attendent.

C'est l'hypothèse faite par la plupart des moteurs de recherche sur Internet, qui n'hésitent pas à proposer plusieurs millions de documents par requête. La précision de ces systèmes est alors très faible, mais cela n'impacte pas la qualité globale.

La courbe rappel précision d'un système est obtenue en calculant la moyenne sur un certain nombre de requêtes des valeurs de précision et de rappel en ne considérant que le premier document retrouvé, puis en incrémentant le nombre de documents considérés. Une fois les courbes rappel-précision obtenues, il existe plusieurs moyens de résumer la qualité d'un SRI en une valeur. Nous utiliserons le MAP^1 , qui n'est rien d'autre que l'aire sous la courbe rappel-précision. Cette mesure donne un bon aperçu de la qualité globale du système.

4.2 Protocole

Le protocole adopté pour cette expérimentation est le suivant :

1. Nous soumettrons à chacun des six systèmes définis dans le tableau 1.1, page 4, un document de la base sur dix, soit 450 requêtes au total et 9 requêtes par classe de documents.
2. Nous associons à chaque couple requête-document la distance calculée par chaque système, *i.e.* la pertinence estimée du document par rapport à la requête, $p_{\text{système}}(q, d)$.
3. Nous utilisons ensuite l'outil d'évaluation de SRI *Trec Eval* [TrecEval]. Cet outil est largement utilisé dans la communauté RI et de nombreuses publications utilisent ce système pour mesurer la qualité de leurs résultats.

Pour chaque requête-document, Trec Eval compare la pertinence estimée par le système à la pertinence utilisateur. Il en déduit les courbes rappel-précision et la valeur du *map*, entre autres, pour chaque requête et calcule la moyenne de ces résultats sur la totalité des requêtes soumises.

¹MAP : Mean Average Precision

4.3 Résultats expérimentaux

Nous allons présenter dans cette section les résultats obtenus par chacun des modèles étudiés. Les résultats seront présentés sous trois formes :

- L’affichage des 5 premiers documents retrouvés pour quelques requêtes typiques, afin de se donner une idée concrète de la tâche entreprise.
- Les courbes rappel précision, afin d’avoir une représentation visuelle globale de la qualité des SRI.
- Le *MAP*, afin d’avoir une valeur chiffrée de la qualité globale des SRI.

Les résultats exposés ici sont obtenus en faisant la moyenne des résultats sur toutes les 450 requêtes soumises.

4.3.1 Exemples

Nous présentons ici les 5 premiers documents retrouvés par chaque système pour une requête tirée au hasard.

Modèle IMAGE

c.f. figure 4.1, page 40.

Pour cette requête, on constate que le résultat n’est pas bon du point de vue de la pertinence. En effet, seul un document parmi les quatre premiers retrouvés fait partie de la classe de la requête. En revanche, on peut observer que les images, si elles ne sont pas ressemblantes, restent cohérentes (présence de vert, de marron et un peu de rouge, textures fines).

Modèle TEXTE

c.f. figure 4.2, page 41.

Dans ce cas, les résultats sont plutôt bons. Les premiers documents font tous partie de la classe de la requête. Les mots clés associés aux images sont très similaires.

Modèle IMAGE&TEXTE

c.f. figure 4.2, page 41.

Les documents retrouvés par ce système sont les mêmes et dans le même ordre que ceux du système TEXTE.

rang	classe	image	texte
requête	12		flowers plants pond water
1	48		log reptile tree
2	15		building house street
3	9		crab food market
4	12		flowers garden tree

FIG. 4.1 – Premiers documents retrouvés par le modèle IMAGE

Modèle IMAGE&LSA

c.f. figure 4.3, page 42.

Les résultats semblent meilleurs ici que dans le cas du model IMAGE simple, car 2 documents sur quatre font partie de la classe de la requête.

Modèle TEXTE&LSA

c.f. figure 4.4, page 43.

4.3 Résultats expérimentaux

rang	classe	image	texte
requête	12		flowers plants pond water
1	12		flowers garden pond water
2	12		flowers garden pond tree
3	12		flowers garden pond tree
4	12		garden pond tree water

FIG. 4.2 – Premiers documents retrouvés par le modèle TEXTE

Ici, les mots *flowers* et *plants* sont présents dans les documents retrouvés, or ils indexent aussi des documents d'autre classes que la requête. Les résultats observés sont donc moins bons.

Modèle IMAGE&TEXTE&LSA

c.f. figure 4.5, page 44.

rang	classe	image	texte
requête	12		flowers plants pond water
1	12		garden grass house tree
2	10		buddha decoration statue temple
3	15		bridge house street
4	12		flowers garden house roofs

FIG. 4.3 – Premiers documents retrouvés par le modèle IMAGE&LSA

Ici encore, les résultats paraissent mauvais. En effet, les mots *pond* et *water* ont été comme éludés par le système, et la mare (*pond, water*) présente dans l'image requête n'est présente dans aucune des images retrouvées.

Mais ces résultats ne montrent que les tous premiers documents pour une seule requête. Ils n'ont donc pas valeur de vérité générale. Comme nous allons le voir, nous aurions pu exposer des exemples bien meilleurs ou d'autres plus mauvais.

Quoi qu'il en soit, l'objectif de ces exemples n'est pas de tirer des conclu-

4.3 Résultats expérimentaux

rang	classe	image	texte
requête	12		flowers plants pond water
1	35		flowers plants
2	27		flowers leaf water
3	35		butterfly flowers insect plants
4	49		log blooms maui plants

FIG. 4.4 – Premiers documents retrouvés par le modèle TEXT&LSA

sions sur la qualité des systèmes, mais de montrer la difficulté de la tâche abordée dans cette étude.

4.3.2 Courbes rappel-précision

Nous présentons ici les courbes moyennes rappel-précision de chaque système sur les 450 requêtes équiréparties dans les 50 classes de documents de la base.

Expérimentation

rang	classe	image	texte
requête	12		flowers plants pond water
1	35		flowers plants
2	49		log blooms maui plants
3	35		flowers glass plants
3	35		butterfly flowers insect plants

FIG. 4.5 – Premiers documents retrouvés par le modèle IMAGE&TEXT&LSA

Nous allons étudier ces courbes en essayant d'en extraire qualitativement l'influence de chaque paramètre.

4.3 Résultats expérimentaux

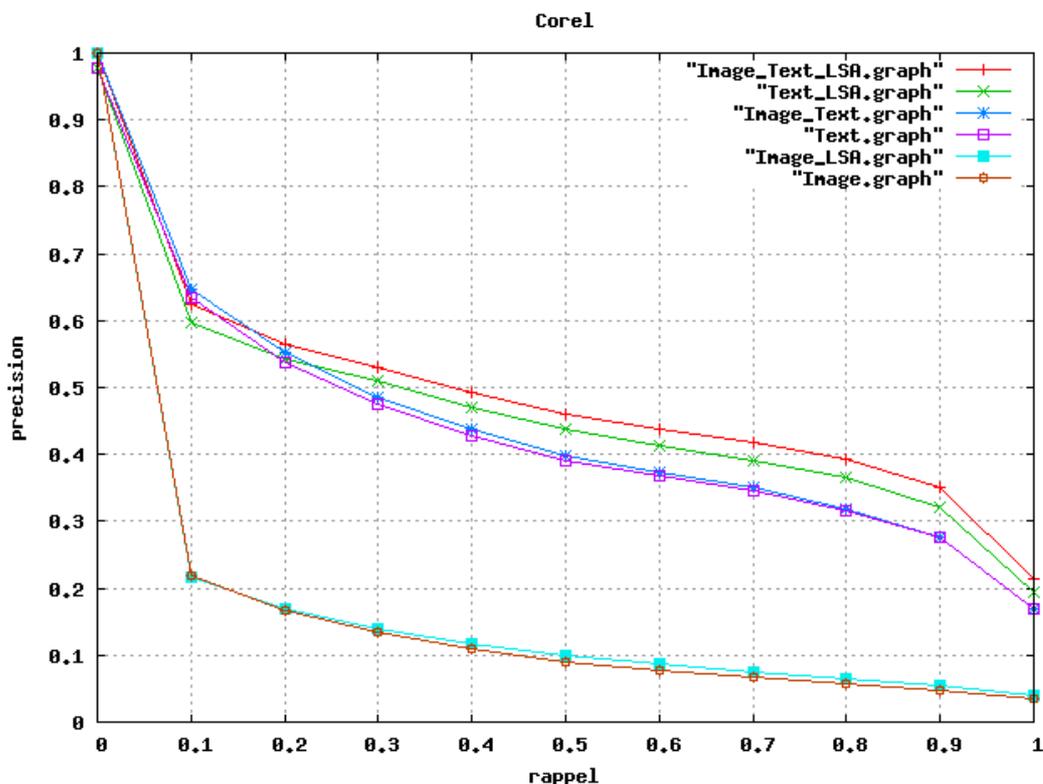


FIG. 4.6 – Courbes rappel-précision sur toute la base

Influence du texte

Nous remarquons tout d'abord la grande différence de résultats entre les systèmes utilisant le texte et les autres. En effet, la courbe des systèmes IMAGE et IMAGE&LSA sont très en dessous des autres courbes.

Du fait que les courbes IMAGE&TEXTE et IMAGE&TEXTE&LSA sont nettement au dessus respectivement des courbes IMAGE et IMAGE&LSA, nous pouvons déduire que l'apport du texte dans les systèmes fusionnant les deux modalités est très grande.

Par ailleurs, la différence est plus grande entre les courbes IMAGE&TEXTE&LSA et IMAGE&LSA qu'entre les courbes IMAGE&TEXTE et IMAGE. Nous en déduisons donc que la contribution du texte est plus efficace lorsque LSA est utilisé.

Influence de l'image

Nous avons déjà remarqué la pauvreté des résultats issus des systèmes purement basés sur les informations visuelles comparativement aux autres systèmes.

Cependant, nous observons tout de même que les courbes IMAGE&TEXTE et IMAGE&TEXTE&LSA sont au dessus respectivement des courbes TEXTE et TEXTE&LSA, ce qui tend à montrer que les images ont une influence positive sur les résultats des systèmes uniquement basés sur le texte.

De plus, la différence entre les courbes IMAGE&TEXTE&LSA et TEXTE&LSA est bien plus grande qu'entre les courbes IMAGE&TEXTE et TEXTE. L'amélioration apportée par les images est donc d'autant plus grande que la méthode LSA est appliquée.

Influence de LSA

Nous constatons que les courbes des systèmes utilisant LSA sont au dessus de leurs homologues sans LSA. Cependant cette différence varie grandement entre les systèmes.

En effet, pour ce qui est des systèmes IMAGE, l'amélioration observée en ajoutant LSA est infime. Dans le cas du système TEXTE, la différence est plus sensible. Mais force est de constater que la plus grande différence est atteinte entre les courbes IMAGE&TEXTE et IMAGE&TEXTE&LSA, ce qui tend à prouver que LSA est d'autant plus utile que les deux modalités sont utilisées.

Ces observations qualitatives étant faites, il est nécessaire de les chiffrer afin de pouvoir mesurer objectivement l'influence de chaque paramètre.

4.3.3 Mean Average Precision

Le tableau 4.1, page 47, montre les valeurs moyennes de MAP de chaque modèle étudié dans cette contribution.

Nous allons extraire de ces chiffres des informations objectives sur l'apport des données issues des images d'une part, du texte d'autre part. Nous verrons de plus quelle est l'amélioration apportée par l'utilisation de l'indexation par sémantique latente.

4.3 Résultats expérimentaux

modèle	MAP
IMAGE	0.1194
TEXTE	0.4107
IMAGE&TEXTE	0.4263
IMAGE&LSA	0.1256
TEXTE&LSA	0.4413
IMAGE&TEXTE&LSA	0.4694

TAB. 4.1 – MAP pour chacun des modèles

Influence du texte

L'influence apportée par le traitement de la partie textuelle des documents peut être mesurée en calculant les deux valeurs suivantes :

$$\frac{MAP_{IMAGE\&TEXTE} - MAP_{IMAGE}}{MAP_{IMAGE}} = \frac{0.4263 - 0.1194}{0.1194} = 257\%$$

$$\frac{MAP_{IMAGE\&TEXTE\&LSA} - MAP_{IMAGE\&LSA}}{MAP_{IMAGE\&LSA}} = \frac{0.4694 - 0.1256}{0.1256} = 274\%$$

L'amélioration apportée par le texte est donc très importante. En effet, ces informations permettent de multiplier par plus de 2,5 les performances des SRI ne les utilisant pas. Lorsque un texte ou des annotations relatives aux images sont disponibles au sein des documents il est donc essentiel de les utiliser.

Influence de l'image

Nous pouvons observer l'influence relative des informations visuelles en calculant les valeurs suivantes :

$$\frac{MAP_{IMAGE\&TEXTE} - MAP_{TEXTE}}{MAP_{TEXTE}} = \frac{0.4263 - 0.4107}{0.4107} = 3,8\%$$

$$\frac{MAP_{IMAGE\&TEXTE\&LSA} - MAP_{TEXTE\&LSA}}{MAP_{TEXTE\&LSA}} = \frac{0.4694 - 0.4413}{0.4413} = 6,4\%$$

Ainsi, utiliser les informations visuelles, lorsqu'elles sont disponibles et sont sémantiquement liées au texte contenu dans le même document, est bénéfique dans chacun des modèles proposés. L'amélioration apportée par ces informations est en effet de l'ordre de 5% en moyenne sur les 450 requêtes soumises aux systèmes développés.

Influence de LSA

Nous allons maintenant exhiber les chiffres montrant l'effet de l'utilisation de la méthode LSA lors de la fusion des modalités textuelles et visuelles. Pour cela, nous calculons les valeurs suivantes :

$$\frac{MAP_{IMAGE\&TEXTE\&LSA} - MAP_{IMAGE\&TEXTE}}{MAP_{IMAGE\&TEXTE}} = \frac{0.4694 - 0.4263}{0.4263} = 10,1\%$$

$$\frac{MAP_{TEXTE\&LSA} - MAP_{TEXTE}}{MAP_{TEXTE}} = \frac{0.4413 - 0.4107}{0.4107} = 7,5\%$$

$$\frac{MAP_{IMAGE\&LSA} - MAP_{IMAGE}}{MAP_{IMAGE}} = \frac{0.1256 - 0.1194}{0.1194} = 5,2\%$$

Ces valeurs nous confirment d'une part ce que l'état de l'art nous avait suggéré, qui est que LSA permet d'améliorer considérablement les performances des SRI pour l'indexation et la recherche. En effet, les résultats des systèmes TEXTE et IMAGE sont supérieurs d'au moins 5% à ceux respectivement des systèmes TEXTE&LSA et IMAGE&LSA.

Nous observons de plus que l'influence de l'utilisation de LSA est d'autant plus grande que l'on utilise deux modalités. En effet, l'amélioration apportée par LSA sur le texte et les images combinés est plus de 1,5 fois supérieure à l'amélioration sur le texte seul, et près de 2 fois supérieure à celle observée sur les images seules.

4.4 Conclusions

La première conclusion que l'on peut tirer de cette expérience est qu'en moyenne, sur la base que nous avons utilisé, l'utilisation conjointe des modalités textuelles et visuelles est considérablement plus efficace que la simple utilisation de l'une ou l'autre des modalités. Notamment, nous avons montré que si la contribution des informations textuelle comptait pour une très large partie de la qualité des résultats, l'apport des informations visuelles n'est pas à négliger.

De plus, nous avons vu que l'indexation par sémantique latente était une méthode efficace sur chacune des modalités, ce qui ne fait que confirmer les résultats observés dans l'état de l'art.

Par surcroît, nous avons montré en nous appuyant sur une expérience significative, que l'amélioration apportée par LSA est encore plus importante lors de la combinaison des deux modalités, ce qui confirme notre intuition initiale, *c.f.* introduction 1.1, page 1.

4.5 Discussion

La base Corel utilisée est assez hétérogène. C'est-à-dire qu'il est possible d'avoir une grande variance de documents au sein d'une même classe, et que plusieurs classes puissent contenir des documents très similaires. Cependant, certaines classes sont très bien définies.

Une classe bien définie est homogène, *i.e.* les documents sont assez similaires entre eux, et singulière, *i.e.* peu de documents d'autres classes ressemblent à ceux de celle-ci. Les classes ayant une grande variance de documents ont pour effet de réduire considérablement le rappel des résultats. Enfin, une classe de documents peut posséder une classe *jumelle*, *i.e.* contenant des documents très similaires. Ceci aura pour effet de faire baisser la précision des systèmes.

Il est alors à noter que la disparité de qualité des résultats est assez grande en fonction de la classe de la requête, et que les performances des systèmes s'en trouvent affectées. Dans le cas de classes bien définies, les performances sont bien meilleures que ceux présentés précédemment, tandis que dans le cas défavorable où la requête est tirée d'une classe mal définie, les performances sont largement diminuées.

Nous allons illustrer ceci en exhibant les résultats moyens des systèmes étudiés sur 9 requêtes issues de 4 classes typiques des cas de classe bien formée, de classe ayant une grande variance et de classes jumelles.

Nous noterons B la classe bien formée, V la classe ayant une grande variance interne, J_1 et J_2 les classes jumelles

4.5.1 Exemples

Nous présentons ici les 5 documents de chaque classe afin de se rendre compte de ce que nous appelons classe bien définie B, classe présentant une forte variance V et classe J_1 possédant une classe jumelle J_2 .

La classe B apparaît bien sémantiquement cohérente, la couleur verte et la texture des feuilles sont dominantes dans quasiment toutes les images et beaucoup de mots en commun, comme *flowers*, *garden* et *house*... De plus, la couleur verte et la texture des feuilles ainsi que des mots comme *house* sont très particuliers à cette classe. La classe V apparaît quant-à elle très hétérogène, avec des documents sémantiquement très différents, des images visuellement très éloignées et très peu de mots communs à dix documents. Les classes J_1 et J_2 sont, comme les exemples le montrent, sémantiquement, visuellement et textuellement très proches.

Expérimentation

classe B	classe V	classe J_1	classe J_2
 <p>flowers garden house tree</p>	 <p>clouds sky tower</p>	 <p>jet plane sky</p>	 <p>jet plane sky</p>
 <p>garden tree lawn tree</p>	 <p>food people pots</p>	 <p>jet plane runway</p>	 <p>jet plane runway</p>
 <p>flowers garden pond tree</p>	 <p>garden plants tree water</p>	 <p>jet plane smoke</p>	 <p>jet plane smoke</p>
 <p>flowers garden landscape tree</p>	 <p>castle night reflection water</p>	 <p>clouds jet plane sky</p>	 <p>clouds jet plane</p>
 <p>flowers garden landscape tree</p>	 <p>flowers pots wall</p>	 <p>plane prop sky</p>	 <p>clouds plane prop</p>

FIG. 4.7 – Exemples de documents de classes de type B, V et J

4.5 Discussion

4.5.2 Résultats

Nous allons maintenant observer les résultats de chaque systèmes sur chacune des classes décrites ci-dessus, *c.f.* figures 4.8, 4.9, 4.10 et 4.11, pages 51, 52, 54 et 54, ainsi que le tableau 4.2, page 51.

modèle	MAP sur B	MAP sur V	MAP sur J_1	MAP sur J_2
IMAGE	0.2643	0.0856	0.2132	0.0957
TEXTE	0.4997	0.0515	0.5069	0.7665
IMAGE&TEXTE	0.5623	0.0810	0.5539	0.5693
IMAGE&LSA	0.3121	0.0855	0.2539	0.0950
TEXTE&LSA	0.6207	0.0541	0.5205	0.6679
IMAGE&TEXTE&LSA	0.6599	0.0621	0.6813	0.6016

TAB. 4.2 – MAP pour chacun des modèles sur les classes B, V, J_1 et J_2

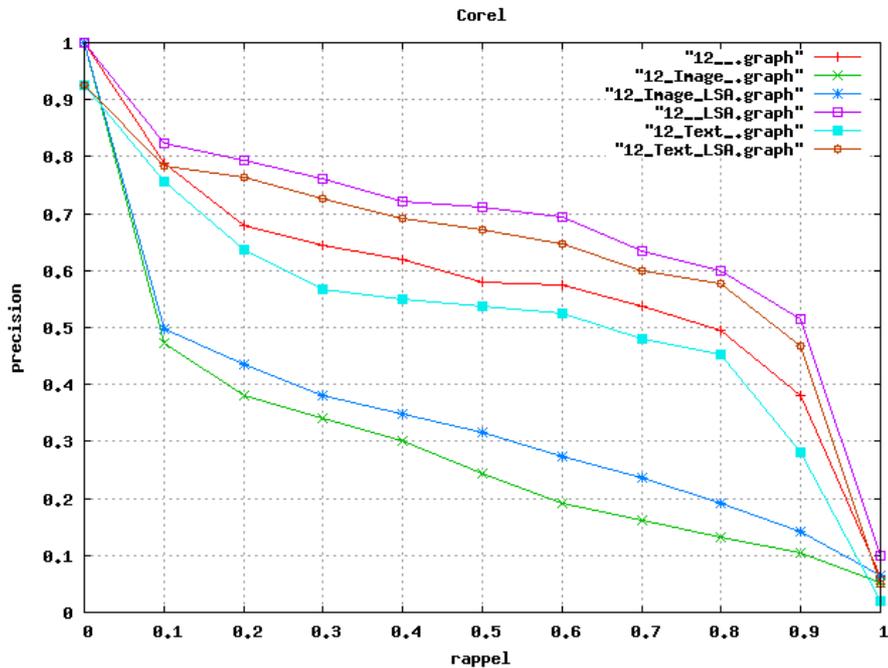


FIG. 4.8 – Courbes rappel-précision de la classe B

La figure 4.8, page 51, nous indique que dans le cas favorable d'une requête issue d'une classe bien définie, les résultats observés dans le cas général sont amplifiés. En effet, la fusion des deux modalités est 112% meilleure que

les images seules et 12,5% meilleure que le texte seul sans LSA et 111% meilleure que les images seules et 6,3% meilleure que le texte seul sans LSA, *c.f.* tableau 4.2, page 51.

L'amélioration apportée par le texte est donc moins grande que dans le cas général, qui était de l'ordre de 250%. De plus, l'amélioration apportée par les images est quasiment deux fois supérieure à la moyenne (5,5%). Ceci s'explique par le fait que la qualité des SRI n'utilisant que les images est en moyenne deux fois plus faible que dans le cas de la classe B.

Par ailleurs, l'influence positive de LSA pour la fusion est encore plus flagrante que dans le cas général. En effet, LSA améliore de 17,3% la fusion des deux modalités dans le cas de la classe B contre seulement 7,5% en moyenne sur toute la base.

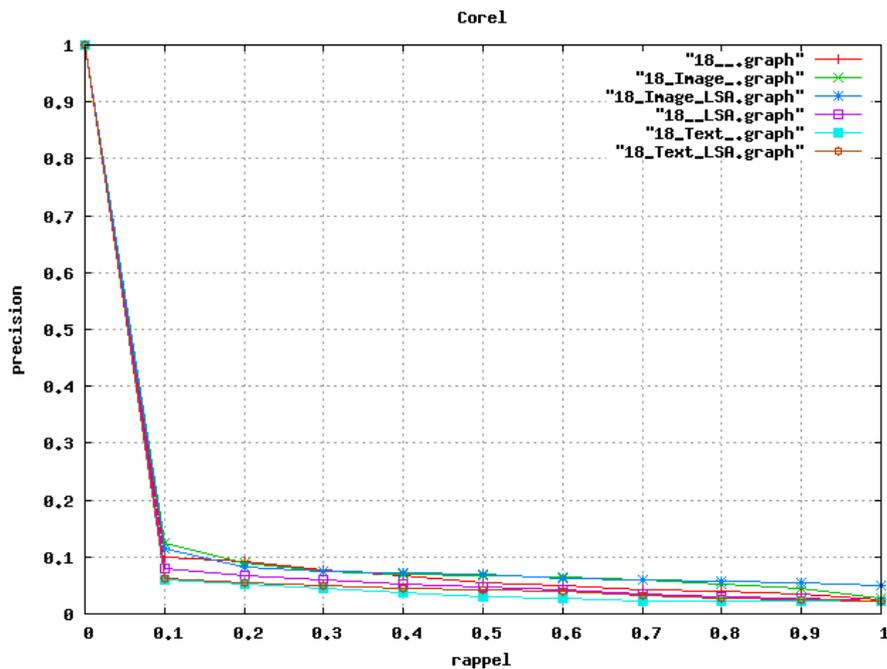


FIG. 4.9 – Courbes rappel-précision de la classe V

Dans le cas de la classe V (*c.f.* figure 4.9, page 52, et tableau 4.2, page 51), aucune statistique n'est significative. Cependant, cette expérience permet de montrer l'importance de la modélisation de la pertinence utilisateur et de la difficulté de la tâche. En effet, les documents au sein de cette classe sont tellement dissemblables que l'on peut se dire qu'il est impossible qu'un utilisateur soumettant une requête issue de cette classe soit satisfait par tous les autres documents de cette même classe. Le point faible du modèle est

4.5 Discussion

alors la représentation interne de l'utilisateur.

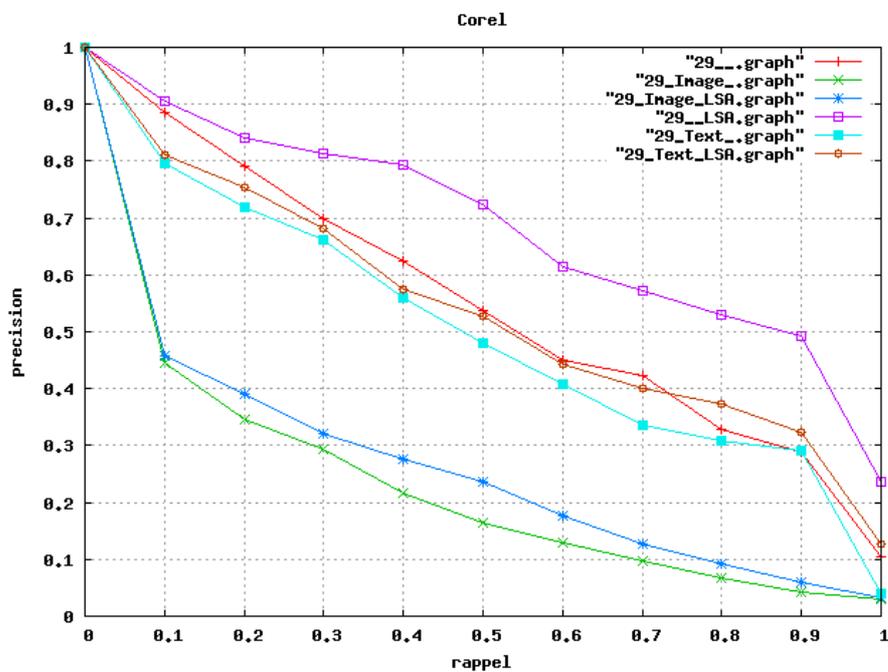
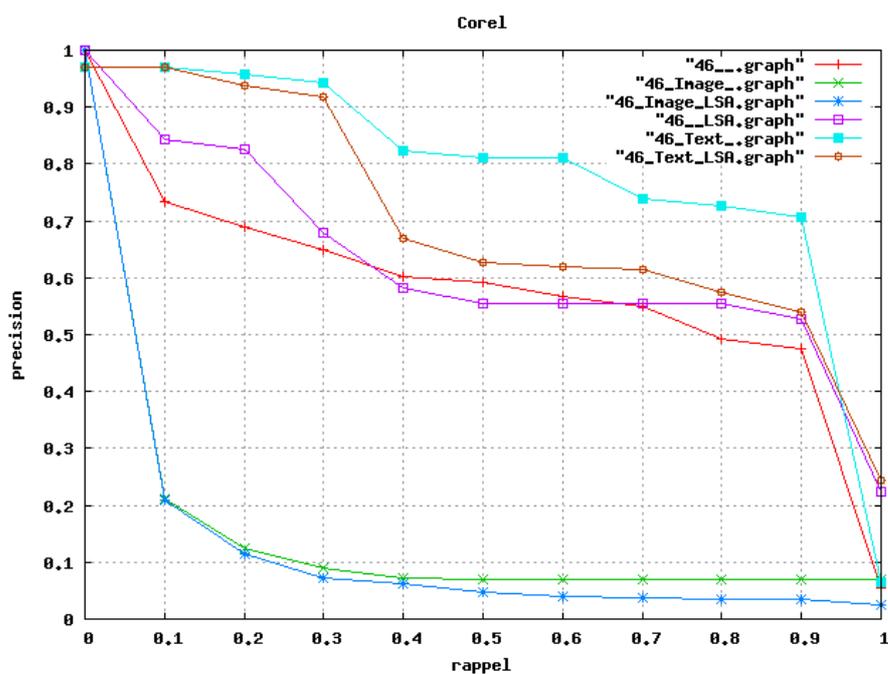
Cependant, dans le cas précis de la classe V utilisée en exemple, bien que les documents soient indéniablement dissemblables, ils ont néanmoins une sémantique commune. En effet, tous les documents de cette classe présentent une scène dont le lieu est la France. Cet exemple est un cas typique d'échec dans la tentative de franchir le *fossée sémantique*.

Cette expérience montre que partant de données ayant un niveau sémantique faible, même si l'on considère les mots clés issus du texte, il est impossible au jour d'aujourd'hui d'en extraire le sens profond sans connaissances extérieures. En effet, en observant les documents de cette classe, il est aisé pour un français de remarquer que ce sont tous des scènes typiquement françaises. Ceci car il sait, entre autre, que la tour Eiffel est à Paris et que Paris est la capitale de la France, il a une vague idée de la vie telle qu'elle peut être chez les personnes âgées vivant en zone rurale, il sait à quoi ressemble « Le bassin aux nymphéas » de Claude Monet et il sait que celui-ci a peint cette œuvre à Giverny en France, il sait que la France compte de nombreux châteaux et enfin que les géraniums sont les fleurs très populaires en France (*c.f.* documents pris en exemple pour la classe V dans la figure 4.7, page 50). Mais une personne, aussi intelligente soit-elle, n'ayant jamais entendu parler de la France pourra difficilement croire que les documents de cette classe ont un lien quelconque entre eux.

Dans le cas des classes jumelles (*c.f.* figures 4.10 et 4.11, pages 54 et 54, et tableau 4.2, page 51), les résultats sont cohérents avec la moyenne, *i.e.* la fusion et LSA améliorent les résultats, pour l'une d'elles (J_1 dans notre cas), et en totale contradiction avec les résultats attendus pour l'autre. Ceci s'explique par le fait que lorsqu'une requête issue d'une classe de ce type est émise, le système trouvera que les documents de chacune des deux classes sont pertinents, tandis que la vérité terrain est que la classe d'origine du document est la seule classe pertinente. Les résultats disjoints de ces deux classes n'ont alors pas de sens.

Le problème soulevé par cette expérience est celui de la dépendance de la pertinence des résultats en fonction de la qualité de la base de documents. En effet, considérer deux classes jumelles séparément n'a pas de sens et introduit un biais dans les résultats.

Cependant, il faut aussi accepter et prendre en compte qu'une base de test ne soit pas parfaitement construite. En effet, comme nous l'avons indiqué dans la partie 2.3, page 12, lors de la mise en place fonctionnelle d'un SRI, et d'autant plus que cette application est collaborative, comme c'est le cas de Flickr [flickr], rien ne garantit la cohérence des annotations avec les images.

FIG. 4.10 – Courbes rappel-précision de la classe J_1 FIG. 4.11 – Courbes rappel-précision de la classe J_2

Chapitre 5

Conclusion générale

5.1 Contribution

Dans cette étude, nous nous sommes proposé de mesurer dans quelle proportion l'utilisation conjointe du texte et des images améliore les résultats des systèmes pour la tâche de recherche de documents présentant ces deux modalités. Nous avons de plus voulu répondre à la question de savoir dans quelle mesure la méthode d'indexation par sémantique latente apporte un gain significatif lors de la fusion de ces deux formes d'information.

Aucune étude n'a, à notre connaissance, encore apporté de résultats objectifs, chiffrés et significatifs à ces questions.

Afin de répondre à la problématique posée, nous avons construit six différents modèles de recherche d'information sur documents bi-modaux texte et image. Ces modèles s'appuient sur des méthodes classiques issues de la recherche d'informations sur documents textuels d'une part et du domaine de la recherche d'images basée sur le contenu d'autre part. Nous avons de plus implémenté la méthode LSI que nous avons appliquée à chaque modalité ainsi qu'à leur fusion.

Les paramètres que nous avons fait varier entre les différents modèles de RI testés sont l'utilisation ou non des informations textuelles, l'utilisation ou non des informations visuelles et l'application ou non de la méthode LSA. Nous avons alors soumis 450 requêtes, issues de la base de documents Corel comptant 4500 documents bi-modaux, à chacun des systèmes implémentés, puis comparé les résultats obtenus.

Le bilan global confirme notre intuition initiale. En effet, non seulement la fusion des modalités textuelle et visuelle apporte un bénéfice significatif aux performances des SRI n'utilisant qu'une modalité, mais de plus LSI agit

comme un catalyseur sur cette fusion en augmentant considérablement le gain issu de cette fusion.

Cependant, nous avons exhibé des cas de classes de documents « pathologiques » pouvant rendre les résultats des modèles proposés erratiques. En effet, les résultats obtenus restent dépendants de la modélisation de la pertinence utilisateur et de sa cohérence avec la base de documents. Dans le cas d'une bonne modélisation, nous avons observé que les résultats obtenus pouvaient être largement meilleurs que la moyenne. Au contraire, dans un cas défavorable, les performances des systèmes fusionnant les modalités textuelle et visuelle ou de ceux utilisant LSA peuvent être bien plus faibles.

5.2 Travaux futurs

Les perspectives envisagées dans le futur seront tout d'abord d'améliorer les performances du modèle fusionnant les modalités textuelle et visuelle avec LSA en utilisant des techniques de prétraitement plus sophistiquées du côté visuel, tels que l'utilisation de SIFT et d'une segmentation des images. Nous pensons également utiliser l'algorithme de clustering *fuzzy c-mean* afin de considérer la possibilité qu'une image puisse appartenir en partie à plusieurs classes, *i.e.* représenter plusieurs objets.

Nous envisageons appliquer cette méthode au problème de l'annotation automatique d'images. L'idée est alors d'associer à une image requête un ensemble de mots sensés décrire la scène qu'elle représente. De même, nous comptons aborder l'illustration automatique de texte, qui est le problème dual du précédent, grâce au modèle de fusion par LSA. Ce problème se pose de la manière suivante : étant donné un texte, lui associer un ensemble d'images l'illustrant.

Plus prospectivement, nous envisageons employer notre modèle pour une tâche totalement originale que serait la génération automatique d'images à partir d'un texte, et non simplement une recherche d'images existant dans une base de documents. Cette approche générative de l'illustration de texte n'a à notre connaissance encore jamais été abordée.

Ceci nous semble accessible avec un modèle assez performant côté visuel, permettant d'atteindre un niveau sémantique supérieur (*e.g.* extraction et localisation d'objets), et un grand nombre de documents afin que LSA puisse associer plus efficacement *mots clés textuels* et *visuels*. L'idée serait qu'une requête du type « arbres montagne ciel » soumise en entrée du système génère une image contenant une zone verte en bas de l'image, une zone blanche au milieu et une zone bleue en haut.

Annexe A

Généralités

Quelle que soit la qualité de l'enseignement que l'on suit en tant qu'étudiant, cela ne nous prépare pas à la réalité du monde de la recherche. C'est pourquoi il est important que nous nous familiarisions avec la notion de recherche scientifique dans une première expérience sous forme de stage d'initiation.

A.1 Structures d'accueil

Ces quelques mois de stage en laboratoire m'ont permis d'apprendre beaucoup sur le fonctionnement d'un laboratoire et les méthodes de travail d'un chercheur. En effet, j'ai eu l'occasion d'observer de l'intérieur deux instituts de recherche.

La première partie de mon stage s'est déroulée d'Octobre à Mars dans l'équipe MRIM¹ du laboratoire CLIPS².

Le laboratoire CLIPS conduit des recherches dans les domaines des interactions Homme-machine, des interactions humaines médiées par la machine et du traitement des informations issues de l'Homme (langue, parole, images). Il est dirigé par Catherine Garbay, secondée par Catherine Berrut elle-même membre de l'équipe MRIM. Ce laboratoire comprend 10 équipes dont l'équipe MRIM.

Le principal domaine de recherche de l'équipe MRIM est la modélisation formelle des SRI multimédia. Ce thème est étudié selon les axes de la modélisation des données multimédia pour le filtrage ou la recherche d'informations, la définition de systèmes personnalisés de filtrage ou de recherche d'informa-

¹MRIM : Modélisation et Recherche d'Information Multimédia

²CLIPS : Communication Langagière et Interaction Personne-Système

tions et l'évaluation des systèmes.

Mon tuteur à Grenoble, Philippe Mulhem, se situe principalement dans le premier axe de recherche. Il a beaucoup travaillé pour le domaine de l'indexation et la recherche d'images personnelles, de documents multimédias et de vidéos.

Le laboratoire IPAL³ est une collaboration franco-singapourienne comprenant 5 permanents et une dizaine de stagiaires de différentes nationalités.

Les axes de recherche de ce laboratoire sont l'analyse et la recherche d'images et de documents multilingues et multimedia. L'IPAL existe dans les faits depuis 1998 et à été créé officiellement en 2000. Philippe Mulhem en a été le premier directeur de 1998 à 2003, suivi de mon tuteur français à Singapour, Jean-Pierre Chevallet. J'ai aussi eu l'occasion de travailler avec Joo-Hwee Lim, co-directeur de l'IPAL et représentant de l'institut singapourien I²R⁴ au sein du laboratoire.

Jean Pierre Chevallet a contribué au domaine de la recherche d'information notamment au niveau des documents multilingues. Joo Hwee Lim est quand-à lui issu du domaine de la recherche et de l'indexation d'images. Ils ont notamment remporté l'évaluation CLEF 2005 portant sur une base de dossiers médicaux composés d'images et de texte.

J'ai trouvé particulièrement intéressante la différence entre la recherche du point de vue français, plutôt théorique et tournée vers le public scientifique, et la recherche du point de vue singapourien, tournée vers les applications pratique des projets développés.

A.2 Déroulement du stage

La première période, d'Octobre à Mars, passée dans l'équipe MRIM, à été assez peu productive mais très enrichissante du point de vue de la découverte du milieu de la recherche. En effet, le sujet et les objectifs de mon stage n'étaient encore pas totalement fixés et ils n'allaient l'être qu'une fois à Singapour.

J'ai donc mis ces premiers mois à profit en me documentant sur les aspects généraux du domaine et plus précisément la problématique que j'allais avoir à résoudre. J'ai de plus participé à plusieurs réunions avec les autres membres de l'équipe afin de me tenir au courant de leurs travaux et d'en tirer des informations utiles. Les doctorants et les stagiaires de l'équipe MRIM

³IPAL : Image Processing and Applications Laboratory

⁴I²R : Institute for Infocomm Research

A.2 Déroulement du stage

organisaient des réunions hebdomadaires qui m'ont permis d'obtenir des réponses à des questions d'ordre technique, pratique ou encore organisationnel.

La seconde partie de mon stage, de Mars à juin à Singapour, s'est déroulée au sein du laboratoire IPAL. Ces deux mois de travail intensif m'ont permis de fixer définitivement les objectifs de mon projet et de faire une étude bibliographique plus ciblée et plus complète. Enfin, j'ai pu confronter mes idées issues de la lecture d'articles à la dure réalité expérimentale.

Cette expérience m'a permis non seulement d'utiliser et d'améliorer certaines compétences techniques lors de la phase d'expérimentation du projet, mais surtout d'exploiter et de développer des capacités plus intellectuelles nécessaires à la vocation de chercheur, telles que l'imagination et la créativité lors des phases de modélisation du système, d'interprétation et d'exploitation des résultats obtenus.

Index

- Échantillonnage, 14
- Échantillonnage aléatoire, 15
- Échantillonnage régulier, 15

- ABIR, 12
- Antidictionnaire, 11

- bin, 29

- CBIR, 13
- CIE, 18
- classification, 7
- CLIPS, 57
- clustering, 21

- Détecteur de Harris, 16
- Détection de points d'intérêt, 16
- Différence de gaussiennes, 16

- HSV, 18

- I²R, 58
- Idf, 23
- indexation, 7, 8
- IPAL, 58

- L*a*b*, 18
- Lemmatisation, 11
- LSA, 3, 31
- LSI, 3, 23, 24, 31

- MAP, 38
- MARS, 13
- Matrice mot clé-document, 12
- Modèle vectoriel, 11, 22

- Modèles basés sur des connaissances, 10
- Modèles de langue, 10
- mot clé, 12
- Mot clé visuel, 21
- MRI, 9
- MRIM, 57

- Photobook, 13
- précision, 37
- Pseudo document, 33

- QBIC, 13

- rappel, 37
- recherche, 7
- RGB, 18
- RI, 2

- Sac de mots clés, 24
- Segmentation, 16
- Segmentation forte, 16
- SIFT, 16
- SRI, 7
- SVD, 31

- Tf, 23
- tf*idf, 22
- Trec Eval, 38

- vérité terrain, 36
- Vocabulaire visuel, 14

Bibliographie

- [Al-Halimi98] R. Al-Halimi, R.C. Berwick, J.F.M. Burg, M. Chodorow, C. Fellbaum, J. Grabowski, S. Harabagiu, M.A. Hearst, G. Hirst, D.A. Jones, R. Kazman, K.T. Kohl, S. Landes, C. Leacock, G.A. Miller, K.J. Miller, D. Moldovan, N. Nomura, U. Priss, P. Resnik, D. St-Onge, R. Teng, R.P. van de Riet, E. Voorhees, 1998 : WordNet : An Electronic Lexical Database, Cognitive Science Laboratory Princeton University
- [Barnard03] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D.M. Blei et M.I. Jordan, 2003 : Matching words and pictures, in *Journal of machine learning research* Vol.3 p.1107-1135
- [BarnardForsyth01] K. Barnard et D. Forsyth, 2001 : Learning the Semantics of Words and Pictures, in *International Conference on Computer Vision* Vol.2, p.408-415
- [BergerLafferty99] A. Berger et J. Lafferty, 1999 : Information retrieval as statistical translation, in *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval* p.222-229
- [Carson97] C. Carson, S. Belongie, H. Greenspan et J. Malik, 1997 : Region-Based Image Querying, *Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97)*
- [Deerwester90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer et R. Harshman, 1990 : Indexing by latent semantic analysis in *Journal of the American Society for Information Science* Vol.41 No.6 p.391-407
- [Dumais97] S.T. Dumais, T.A. Letsche, M.L. Littman et T.K. Landauer, 1997 : Automatic Cross-Language Retrieval Using Latent Semantic Indexing, in *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*
- [Eisen98] M.B. Eisen, P.T. Spellman, P.O. Brown et D. Botstein, 1998 : Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences* 95 p.14863-14868
- [flickr] <http://www.flickr.com/>

- [Flickner95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele et P. Yanker, 1995 : Query by image and video content : The QBIC system, in *IEEE Computer*
- [Frakes92] W.B. Frakes et R. Baeza-Yates, 1992 : Introduction to information storage and retrieval systems, *Information retrieval : data structures and algorithms* 1 p.1-12
- [ForsythFleck99] D.A. Forsyth et M.M Fleck, 1999 : Automatic detection of human nudes, in *Int'l J. Computer Vision* Vol.32, No.1, p.63-77
- [Fuhr92] N. Fuhr, 1992 : Probabilistic models in information retrieval, *The Computer Journal* Vol.35 No.3 p.243-255
- [GoogleImages] <http://images.google.com/>
- [GotliebKreyszig90] C.C. Gotlieb and H.E. Kreyszig, 1990 : Texture descriptors based on co-occurrence matrices, in *Computer Vision, Graphics and Image Processing* Vol.51 No.1 p.70-86
- [Harman92] D. Harman, 1992 : Relevance feedback and other query modification techniques, in *Information retrieval : Data structures and algorithms*, New York : Prentice-Hall
- [HarrisStephens88] C. Harris et M. Stephens, 1988 : Combined corner and edge detector, in *Fourth Alvey Vision Conference* p.147-151
- [Inoue04] M. Inoue, 2004 : On the need for annotation-based image retrieval, in *Workshop on Information Retrieval in Context* p.44-46
- [JonesWillett97] K.S. Jones et P. Willett 1997 : Readings in Information Retrieval, ed. Morgan Kaufmann Publishers
- [JainFarrokhnia90] A.K. Jain et F. Farrokhnia, 1990 : Unsupervised texture segmentation using Gabor filters, in *IEEE International Conference Proceedings on Systems, Man and Cybernetics* p.14-19
- [KrovetzCroft92] R. Krovetz et W.B. Croft, 1992 : Lexical Ambiguity and Information Retrieval, in *Information Systems* 10(2) p.115-141
- [LandauerDumais97] T.K. Landauer et S.T. Dumais, 1997 : A solution to Plato's problem : The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, in *Psychological Review* Vol.1 No.04 p.211-140
- [Landauer98] T.K. Landauer, P.W. Foltz et D. Laham, 1998 : Introduction to Latent Semantic Analysis, in *Discourse Processes* 25 259-284
- [LaineFan93] A. Laine et J. Fan, 1993 : Texture Classification by Wavelet Packet Signatures, *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol.15 No.11 p.1186-1191

BIBLIOGRAPHIE

- [LiWang03] J. Li et J.Z. Wang, 2003 : Automatic linguistic indexing of pictures by a statistical modeling approach, in *IEEE transactions on pattern analysis and machine intelligence* Vol.25 No.6 p.1075-1088
- [Lin97] H.C. Lin, L.L. Wang et S.N. Yang, 1997 : Color image retrieval based on hidden Markov models, in *IEEE Transactions on Image Processing* Vol.6 No.2 p.332-339
- [Lim01] J.H. Lim, 2001 : Building visual vocabulary for image indexation and query formulation, in *Pattern Analysis and Applications* Vol.4 No.2-3 p.125-139
- [Lindberg98] T. Lindeberg, 1998 : Feature Detection with Automatic Scale Selection, in *International Journal of Computer Vision* Vol.30 No.2 p.77-116
- [Lowe04] D.G. Lowe, 2004 : Distinctive Image Features from Scale-Invariant Keypoints, in *International Journal of Computer Vision* Vol.60 No.2 p.91-110
- [Lyman03] P. Lyman, H.R. Varian, K. Swearingen, P. Charles, N. Good, L.L. Jordan et J. Pal, 2003 : How much information 2003?, in <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>
- [McLachlanKrishnan97] G. McLachlan et T. Krishnan, 1997 : *The EM Algorithm and Extensions*
- [Maree05] R. Maree, P. Geurts, J. Piater et L. Wehenkel, 2005 : 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) Vol. 1 p.34-40
- [Mehrotra97] S. Mehrotra, Y. Rui, M. Ortega et T. Huang, 1997 : Supporting Content-based Queries over Images in MARS, in *icmcs* p.632
- [MikolajczykSchmid01] K. Mikolajczyk et C. Schmid, 2001 : Indexing based on scale invariant interest points, in *International Conference on Computer Vision* p.525-531
- [Miller99] D.H. Miller, T. Leek et R. Schwartz, 1999 : A hidden Markov model information retrieval system, in *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval* p.214-221
- [Mindru99] F. Mindru, T. Moons et L. van Gool, 1999 : Recognizing Color Patterns Irrespective of Viewpoint and Illumination, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Vol.1 p.368-373

- [MirmehdiPetrou00] M. Mirmehdi et M. Petrou, 2000 : Segmentation of Color Textures, *IEEE transactions on pattern analysis and machine intelligence* Vol.22 No.2 p.142-159
- [Netcraft06] Netcraft, May 2006 : Web Server Survey, in http://news.netcraft.com/archives/web_server_survey.html
- [Nie06] J.Y. Nie, 2006 : Cours de Recherche d'Information, <http://www.iro.umontreal.ca/nie/IFT6255/>
- [Porter80] M.F. Porter, 1980 : An algorithm for suffix stripping, in *Program*, Vol.14 No. 3 p.130-137
- [Praks03] P. Praks, J. Dvorsky, V. Snasel, 2003 : Latent Semantic Indexing for Image Retrieval Systems, in *Proceedings of the SIAM Conference on Applied Linear Algebra*
- [Pédaque03] R.T. Pédaque, STIC-CNRS, 2003 : Document : forme, signe et médium, les re-formulations du numérique
- [Pentland96] A. Pentland, R.W. Picard et S. Sclaroff, 1996 : in *International Journal of Computer Vision* Vol.18 No.3 p.233-254
- [Picard95] R.W. Picard, 1995 : Toward a Visual Thesaurus, *M.I.T Media Laboratory Perceptual Computing Section* Technical Report No. 358
- [PicardMinka95] R.W. Picard et T.P. Minka : Vision texture for annotation, *Multimedia Systems* Vol.3 p.3-14
- [PonteCroft98] J.M. Ponte et W.B. Croft, 1998 : A Language Modeling Approach to Information Retrieval, in *ACM-SIGIR '98*, p.275-281
- [Rehder97] B. Rehder, M.L. Littman, S. Dumais et T.K. Landauer, 1997 : Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing, in *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*
- [Robertson81] S.E. Robertson, C.J. van Rijsbergen et M.F. Porter, 1981 : Probabilistic models of indexing and searching, in Oddy R.N. et al. Eds. *Information Retrieval Research* Butterworths London 1981 p. 35-56.
- [SaltonBuckley88] G. Salton et C. Buckley, 1988 : Term-weighting approaches in automatic text retrieval, *Information Processing and Management* 24 p.513-523
- [SaltonBuckley90] G. Salton et C. Buckley, 1990 : Improving retrieval performance by relevance feedback, in *Journal of the American Society for Information Science*
- [Salton82] G. Salton, E.A. Fox et H. Wu, 1982 : Extended Boolean Information Retrieval, in *Communications of the ACM*, 26(12) p.1022-1036

BIBLIOGRAPHIE

- [SaltonMcGill83] G. Salton et M. McGill, 1983 : An introduction to modern Information Retrieval, New York : McGraw-Hill
- [Salton75] G. Salton, A. Wong et C.S. Yang, 1975 : A vector space model for automatic indexing, in *Commun. ACM*, Vol.18 No.11 p.613-620
- [Sanderson94] M. Sanderson, 1994 : Word Sense Disambiguation and Information Retrieval, in *SIGIR'94*
- [SavaresiBoley04] S. Savaresi et D. Boley, 2004 : A comparative analysis on the bisecting K-means and the PDDP clustering algorithms, in *Intelligent Data Analysis* Vol.8 No.4
- [Savoy93] J. Savoy, 1993 : Stemming of French Words Based on Grammatical Categories, in *Journal of the American Society for Information Science* 44 1-9
- [Sclaroff98] S. Sclaroff, M. la Cascia et S. Sethi, 1998 : Unifying textual and visual cues for content-based image retrieval on the World Wide Web, in *Computer vision and image understanding* Vol.75 Nos.1/2 p.86-98
- [Singhal96] A. Singhal, C. Buckley et M. Mitra, 1996 : Pivoted document length normalization, in *Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval* p.21-29
- [Sivic05] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman et W.T. Freeman, 2005 : in *IEEE International Conference on Computer Vision (ICCV'05)*
- [Sivic04] J. Sivic, F. Schaffalitzky et A. Zisserman, 2004 : Efficient object retrieval from videos, in *Proceedings of the 12th European Signal Processing Conference (EUSIPCO '04)*
- [Smeulders00] A.W.M Smeulders, M. Worring, A. Gupta et R. Jain, 2000 : Content-based image retrieval at the end of the early years, in *IEEE transactions on pattern analysis and machine intelligence* Vol.22 No.12 p.1349-1380
- [SmithChang96] J.R. Smith et S.F. Chang, 1996 : VisualSEEK : a fully automated content-based image query system, in *MULTIMEDIA '96 : Proceedings of the fourth ACM international conference on Multimedia* p.87-98
- [TrecEval] http://trec.nist.gov/trec_eval/
- [VanRijsbergen79] C.J. van Rijsbergen, 1979 : Information Retrieval, London : Butterworths
- [VanRijsbergen86] C.J. van Rijsbergen, 1986 : A Non-classical Logic for Information Retrieval, *The Computer Journal*, 29(6)

BIBLIOGRAPHIE

- [Voorhees93] E. Voorhees, 1993 : Using WordNet to disambiguate word senses for text retrieval, *ACM-SIGIR* 1993 p.171-180
- [Voorhees99] E. Voorhees, 1999 : Natural Language Processing and Information Retrieval, in M.T. Paziienza Ed., *Information Extraction : Towards Scalable, Adaptable Systems* p.32-48
- [Westerveld00] T. Westerveld, 2000 : Image retrieval : Content versus context, in *Proceedings Recherche d'Information Assistée par Ordinateur*
- [Won02] C.S. Won, D.K. Park et S.J. Park, 2002 : Efficient Use of MPEG-7 Edge Histogram Descriptor, in *ETRI Journal* Vol.24 No.1
- [WongYao95] S.K.M. Wong et Y.Y. Yao, 1995 : On modeling information retrieval with probabilistic inference, *ACM Transactions on Information Systems* 13(1) p.69-99
- [ZhaoGrosky02] R. Zhao et W. Grosky, 2002 : Narrowing the semantic gap - improved text-based web document retrieval using visual features, in *IEEE Transactions on Multimedia* Vol.4 No.2 p.189-200