



IRISA • Campus universitaire de Beaulieu • 35042 Rennes Cedex France • Tél. : +33 2 99 84 71 00 • Télécopie : +33 2 99 84 71 71 • Internet : www.irisa.fr

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

Projet TEXMEX

*Techniques d'exploitation des documents multimédias :
exploration, indexation et recherche dans de très grandes bases*

Rennes

_____ THÈME 3A _____



Table des matières

1	Composition de l'équipe	3
2	Présentation et objectifs généraux	3
3	Fondements scientifiques	6
3.1	Description des documents et méta-données	6
3.1.1	Les images fixes	7
3.1.2	La vidéo	9
3.1.3	Le texte	10
3.1.4	Multimédia et couplage entre médias	13
3.1.5	Évaluation des descripteurs	14
3.1.6	Méta-données	15
3.2	Exploitation efficace des descriptions	16
3.2.1	Stratégies de sélection et de calcul des descripteurs	17
3.2.2	Contrôle de la qualité et de la cohérence des données, descripteurs et méta-données	19
3.2.3	Statistiques pour les grands volumes de données	19
3.2.4	Indexation multidimensionnelle	20
3.2.5	Supports systèmes et matériels	21
4	Domaines d'applications	23
4.1	Gestion de bases d'images fixes	23
4.2	Gestion des bases de vidéos	23
4.3	Gestion des grandes bases textuelles	24
4.4	Robotique et asservissement visuel	24
5	Logiciels	25
6	Résultats nouveaux	25
6.1	Recherche dans de grandes bases d'images	25
6.1.1	Description des images fixes	25
6.1.2	Description des vidéos	28
6.1.3	Algorithmes d'indexation et de recherche	29
6.2	Recherche dans de grandes bases de textes	30
6.2.1	Traitement automatique des langues et apprentissage	30
6.2.2	Extraction et visualisation de connaissances à partir de corpus textuels	33
6.3	Méta-données et étude des usages	34
6.3.1	Dimension qualité	34
6.3.2	Dimension usage	35
6.3.3	Visualisation et web mining	35

7 Contrats industriels (nationaux, européens et internationaux)	36
7.1 Contrats industriels	36
7.1.1 Contrat Thalès Communications : analyse des caractéristiques d'un auditoire en vue de la conception d'un logiciel d'argumentation - Génération de lieux selon le type d'argument considéré	36
7.2 Contrats dans le cadre des réseaux nationaux de recherche technologique	36
7.2.1 Projet PRIAM Médiaworks	36
7.2.2 Projet RNRT Diphonet : Diffusion de photos par Internet	37
7.3 Contrats avec l'Union européenne	37
7.3.1 Projet européen IST BUSMAN : Bringing User Satisfaction to Media Access Networks	37
8 Actions régionales, nationales et internationales	38
8.1 Actions régionales	38
8.2 Actions nationales	38
8.2.1 ACI santé Neurobase	38
8.2.2 ACI Grid GénoGRID	38
8.2.3 Action Bio-info inter-EPST architecture parallèle et reconfigurable pour l'extraction de données génomiques	39
8.2.4 Action Bio-info inter-EPST Caderige-2 : catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques	39
8.2.5 Action nationale INRIA de R & D SYNTAX	40
8.2.6 Action JemSTIC TEXMEX	40
8.2.7 ACI jeunes chercheurs TEXMEX	40
8.2.8 Participation à des groupes de travail nationaux	40
8.3 Collaborations internationales	41
8.3.1 Groupe de travail Image Understanding d'ERCIM	41
8.3.2 Collaboration avec le NII au Japon	41
8.3.3 Collaboration bilatérale avec l'Islande	41
9 Diffusion de résultats	42
9.1 Organisation de conférences, workshops, séminaires	42
9.2 Animation de la communauté scientifique	42
9.3 Enseignement universitaire	42
9.4 Participations à des jurys	43
9.5 Participation à des colloques, séminaires, invitations	43
10 Bibliographie	43

TEXMEX est un projet commun avec le CNRS et l'université de Rennes 1. L'équipe a été créée le 1^{er} janvier 2002 et est devenue projet INRIA le 1^{er} novembre 2002.

1 Composition de l'équipe

Responsable scientifique

Patrick Gros [chargé de recherche CNRS]

Assistante de projet

Édith Blin-Guyot [TR, du 1/1/2002 au 30/4/2002, à temps partiel dans le projet]

Stéphanie Lemaile [CDD du 1/5/2002 au 31/8/2002, TR du 1/9/2002 au 15/11/2002, à temps partiel dans le projet]

Maryse Auffray [AA, du 15/11/2002 au 31/12/2002, à temps partiel dans le projet]

Personnel de l'université de Rennes 1

Laure Berti-Équille [maître de conférences]

Annie Morin [maître de conférences]

Pascale Sébillot [maître de conférences]

Personnel CNRS

Laurent Amsaleg [chargé de recherche]

Chercheurs doctorants

Sid-Ahmed Berrani [allocataire CIFRE avec Thomson]

Nicolas Bonnel [allocataire CIFRE avec France Télécom R&D, à 50 % dans l'équipe depuis le 15/11/2002]

Vincent Claveau [allocataire MJENR]

Ewa Kijak [allocataire CIFRE avec Thomson, à 20% dans l'équipe]

Anicet Kouomou-Choupo [boursier du gouvernement français avec un financement de la région Bretagne, depuis le 15/11/2002]

Rodolphe Priam [allocataire MJENR jusqu'au 30/9/2002, ATER à l'université de Nantes depuis le 1/10/2002]

Anthony Remazeilles [allocataire MJENR, en commun avec le projet VISTA]

Mathias Rossignol [allocataire MJENR]

François Tonnin [boursier INRIA – région Bretagne dans le cadre du projet RNRT Diphonet, en commun avec le projet TEMICS, depuis le 1/10/2002]

2 Présentation et objectifs généraux

Mots clés : accès au contenu de documents, exploration, indexation, recherche, bases de données, multimédia, traitement automatique des langues, reconnaissance d'images, apprentissage.

Résumé : *L'explosion de la quantité de documents numériques pose le problème de la gestion de ces documents. Au-delà des problèmes de stockage, nous nous intéressons aux problèmes de gestion des contenus : comment exploiter les grandes bases de documents, les classer, les indexer pour pouvoir y rechercher des documents, en visualiser le contenu ? Pour cela, nous proposons un travail pluridisciplinaire regroupant au sein*

d'une même équipe des spécialistes des médias : image, vidéo, texte, et des spécialistes des techniques d'exploitation des données et méta-données extraites de ces données : bases de données, statistiques, recherche d'information. Nos travaux se situent à l'intersection de ces champs disciplinaires et concernent plus particulièrement 3 points : recherche dans les grandes bases d'images, ajout de ressources linguistiques dans les moteurs de recherche, couplage entre médias pour la description des documents multimédias.

L'exploitation du contenu de grandes bases de documents multimédias numériques est un problème aux multiples facettes, et la construction d'un système exploitant une telle base fait appel à de nombreuses techniques : étude et description de documents, organisation des bases, algorithmes de recherche, de classification, de visualisation, mais aussi gestion adaptée des mémoires primaires et secondaires, interfaces et interaction avec l'utilisateur.

Les cinq défis majeurs du domaine nous paraissent être les suivants :

- il faut, tout d'abord, pouvoir **traiter de grands ensembles de documents** : il est important de mettre au point des techniques qui passent à l'échelle vis-à-vis de la quantité des documents pris en compte, et d'évaluer leurs résultats tant en qualité qu'en rapidité ;
- les documents multimédias ne sont pas qu'une juxtaposition de médias indépendants, et il est important de **mieux exploiter le couplage existant entre les différents médias** présents dans un même document ;
- **les bases de documents multimédias sont évolutives** : les collections de documents évoluent, mais les techniques de description des documents et les modes d'interrogation évoluent aussi, ce qui modifie en retour la manière dont sont utilisées les bases ;
- face à des requêtes pour la plupart d'ordre sémantique, les techniques de description des contenus n'ont accès qu'à la forme de ces documents ; il faut donc trouver des moyens pour **réduire cet écart entre des besoins sémantiques et des outils de description syntaxiques** ;
- la prise en compte de **l'interaction entre l'utilisateur et le système** est un point central : il faut permettre à l'utilisateur de traduire ses besoins de manière efficace et simple mais nuancée ; il faut lui permettre de guider le système ou d'évaluer les résultats ; il doit pouvoir piloter le système sans que ce soit ce dernier qui impose ses choix.

Nous avons adopté une organisation du travail de type matriciel. D'une part, nous disposons de compétences dans deux domaines principaux, la description automatique des documents et l'exploitation de ces descriptions, et d'autre part, nous avons défini trois sujets de recherche transversaux. L'idée sous-jacente est de nous concentrer sur les questions où la pluridisciplinarité de l'équipe nous paraît un atout pour obtenir des résultats originaux.

Notre premier domaine de compétence est donc la description des documents. Les documents ne sont généralement pas exploitables directement pour des tâches de recherche ou d'indexation : il est nécessaire de passer par des descriptions intermédiaires qui doivent être porteuses du maximum d'information sur la sémantique des documents, mais doivent aussi être calculables automatiquement. Aux documents et à leurs descripteurs, on peut ajouter des méta-données, que nous définissons ici comme l'ensemble de toutes les informations (autres que les descripteurs) qui ont la portée de renseigner, de compléter ou de qualifier les données (et les descripteurs) auxquelles elles sont associées.

Notre deuxième domaine de compétence concerne l'exploitation des descriptions. Il s'agit de définir les techniques qui permettent d'appréhender, manipuler et exploiter les volumes de données, méta-données et descripteurs, qui ont pu être extraits des documents : **organisation et gestion des bases**, mise en cohérence logique et temporelle, sélection et stratégies de calcul des descripteurs et méta-données ; **techniques statistiques** pour l'exploration de grands volumes de données ; **techniques d'indexation** visant à confiner au plus petit ensemble de données pertinent possible l'exploitation des données et à ainsi éviter un examen exhaustif dont le coût est certes maîtrisé mais rédhibitoire ; **problèmes système** liés à l'organisation physique de grands volumes de données, comme la gestion des accès disques ou la gestion de mémoires caches nécessitant de nouvelles techniques qui soient adaptées aux caractéristiques des descripteurs et à la façon dont ils sont utilisés.

Premier sujet de recherche : la recherche dans de grandes bases d'images

Passer de corpus de quelques milliers d'images à des corpus en contenant quelques millions reste un enjeu de recherche aujourd'hui. La solution ne peut venir des seuls descripteurs ou d'un nouvel algorithme d'indexation, mais ce sont tous les différents composants du système et leur articulation qui doivent être pris en compte simultanément. Nous proposons donc de travailler sur :

- la description des données, plus particulièrement les données compressées ou tatouées,
- les algorithmes d'indexation et de recherche,
- l'organisation des bases et l'utilisation des méta-données,
- les supports système et matériels,

et sur le couplage entre ces différentes techniques pour améliorer les performances des systèmes actuels, tant en vitesse qu'en qualité de reconnaissance.

Deuxième sujet de recherche : des moteurs de recherche plus sémantiques

Les moteurs de recherche sont des outils très utilisés mais souvent décevants par leur approche trop syntaxique des mots-clés utilisés pour les recherches. Des outils de traitement automatique des langues existent pourtant qui pourraient leur donner un fonctionnement plus sémantique, en leur permettant de désambiguïser un mot et surtout de reconnaître les différentes formulations d'un même concept. Il convient donc de marier ces deux techniques.

Cette union n'est toutefois pas simple, car elle oblige d'une part à ajouter aux moteurs de recherche des stratégies d'extension de mots, puis à traduire ces extensions en similarité et, d'autre part, à faire travailler les outils de traitement des langues dans des univers plus ouverts que ceux dans lesquels ils sont employés habituellement. L'apport d'une telle modification des moteurs doit aussi être démontré, ce qui demande un travail précis sur l'évaluation des résultats obtenus.

Troisième sujet de recherche : multimédia et couplage entre les médias

L'étude du couplage entre médias est entreprise de deux manières. Dans le cadre de la vidéo, nous nous intéressons aux descriptions qui font intervenir conjointement les deux versants son et image de la vidéo. Ceci concerne en particulier la structuration des vidéos, mais aussi l'amélioration des techniques de détection et de reconnaissance de personnes, que ce soit par leur visage ou leur voix.

Par ailleurs, nous étudions le couplage entre texte et image dans les documents où ces deux médias sont fortement couplés, ce qui est le cas des bases bibliographiques scientifiques, des journaux de presse, des livres d'art ou des documents techniques. Le but est de relier, dans un même document,

l'image et le texte qui s'y rapporte, ce qui doit permettre de fournir une description automatique et sémantique des images, puis de relier des documents entre eux, soit par la recherche d'images visuellement ressemblantes, soit par la recherche de textes traitant du même sujet, et d'ainsi améliorer la description des images et de supprimer les ambiguïtés éventuelles dans la compréhension du texte.

3 Fondements scientifiques

Le travail du projet s'appuie sur deux types de compétences : pour exploiter le contenu de documents, il faut tout d'abord être capable d'accéder à ce contenu, c'est-à-dire être capable de caractériser ce contenu d'une manière utilisable. Ainsi reconnaître qu'une image est bleue ou qu'un texte comporte le mot « bleu », c'est déjà accéder à une part de ce contenu. Il faut ensuite que cette description du contenu puisse être utilisable, et cela pose surtout de difficiles problèmes de complexité des algorithmes utilisés. Il faut enfin qu'algorithmes et descriptions puissent répondre aux besoins des utilisateurs : c'est là l'exigence la plus forte. Les utilisateurs ont le plus souvent des demandes d'ordre sémantique sur les documents, alors que les descriptions sont, pour la plupart, syntaxiques. On se trouve donc au centre d'un faisceau de contraintes et d'exigences avec le but de fournir le meilleur service à l'utilisateur.

Trouver une solution passe par la maîtrise des techniques de description des documents : textes, images et vidéos (l'aspect son et parole est traité par d'autres équipes dont METISS à Rennes). Il est aussi nécessaire de savoir exploiter le couplage entre les divers médias des documents qui en comprennent plusieurs, puis de pouvoir évaluer l'apport de chaque description particulière. C'est là le domaine des experts des divers médias qui fait l'objet de la section 3.1.

Il faut ensuite pouvoir exploiter ces descriptions pour satisfaire les attentes de l'utilisateur : sont alors nécessaires des algorithmes de tri, classification, indexation, recherche, visualisation... qui ont la double contrainte de fournir des résultats de qualité en un temps acceptable par l'utilisateur. Ces techniques font l'objet de la section 3.2.

Ces deux aspects ne sont pas indépendants. Les solutions n'intégrant qu'un des deux aspects ne fournissent que des réponses partielles ne constituant pas des systèmes utilisables par eux-mêmes. Mais la combinaison des deux dans le contexte de grandes collections de données pose des problèmes difficile dont la solution ne pourra venir que d'une étude conjointe et fortement couplée sur les deux aspects. C'est là notre programme de travail (cf. section 6).

3.1 Description des documents et méta-données

Mots clés : descripteur, méta-donnée.

Résumé : *Tous les documents multimédias présentent la caractéristique ambivalente d'être, d'un côté, très riches sémantiquement et, d'un autre côté, très pauvres, surtout lorsque l'on considère les composants élémentaires qui les constituent (suites de lettres ou de pixels). D'où la nécessité d'un travail sur les données pour obtenir des descriptions intermédiaires plus concises et plus informatives, descriptions qu'il faut pouvoir calculer automatiquement à partir des documents.*

3.1.1 Les images fixes

Mots clés : appariement d'images, reconnaissance d'images, indexation d'images, invariants.

Le calcul de descripteurs d'images fixes est un sujet étudié maintenant depuis une dizaine d'années. Le but est d'extraire des indices appelés descripteurs dont la distance reflète la ressemblance des images à partir desquelles ils ont été calculés. D'une certaine manière, on peut voir cela comme un problème de codage : comment faut-il coder les images pour que la distance entre les codes reflète la ressemblance entre les images ?

La difficulté de ce problème vient d'abord du fait que la ressemblance entre image n'est pas un concept univoque ni même clairement défini. Les images sont très polysémiques, et la ressemblance entre deux images va dépendre de l'utilisateur qui juge cette ressemblance, du problème qu'il a à résoudre lorsqu'il étudie cette ressemblance et de l'ensemble des autres images dont il peut disposer à cet instant. Autant dire qu'il est impossible de trouver un unique descripteur qui puisse résoudre le problème de manière générale.

On peut toutefois préciser le problème en catégorisant les utilisateurs, les bases et les besoins. Ainsi peut-on séparer les utilisateurs professionnels, qu'ils soient experts du contenu des images, experts en images et traitement des images, experts de la documentation et de la recherche de documents..., et les utilisateurs dits grand public, dont on ne peut supposer aucune expertise si ce n'est d'avoir éventuellement une idée précise de ce qu'ils recherchent. Au niveau des bases, il faut distinguer entre des bases généralistes et disparates contenant un grand nombre d'images sans caractéristique forte commune — ainsi en est-il des photos d'une agence de presse ou des photos prises par un particulier — et les bases d'images spécifiques, images présentant une forte unité ou une modalité unique et spécifique de prise de vue : bases d'images radiographiques, bases d'empreintes digitales, bases de visages en sont des exemples. Enfin, les bases d'images peuvent servir à des recherches lorsque l'utilisateur désire un document précis dans la base, ou à des recherches plus exploratoires lorsque l'utilisateur n'a pas d'idée précise sur ce qu'il cherche, mais souhaite soit affiner sa requête au cours de l'exploration ou simplement se faire une idée du contenu de la base ou d'une de ses parties.

Pour répondre aux divers scénarios que l'on peut envisager à partir de ces catégories, plusieurs types de descripteurs ont été proposés. Le cadre d'étude le plus fréquent est celui de la recherche dans des bases généralistes par des utilisateurs grand public. Les descripteurs proposés dans ce cas intègrent de l'information de l'ensemble de l'image : histogrammes de couleurs dans divers espaces de couleur, descripteurs de texture, descripteurs de forme (dont l'inconvénient majeur est de nécessiter une segmentation qui ne peut être automatique dans un grand nombre d'images). Ce champ de recherche est encore actif : les histogrammes de couleurs fournissent une information bien trop pauvre pour résoudre le moindre problème dès que la taille de la base d'images augmente^[SS94], et diverses propositions sont faites pour y remédier : corrélogrammes^[HKM⁺97], histogrammes pondérés^[BBV01]... Avec la texture, la difficulté est qu'aucun descripteur n'est pertinent pour l'ensemble des textures, mais chacun est spé-

[SS94] M. STRICKER, M. SWAIN, « The Capacity of Color Histogram Indexing », in : *Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, 1994.*

[HKM⁺97] J. HUANG, S. R. KUMAR, M. MITRA, W. ZHU, R. ZABIH, « Image Indexing Using Color Correlograms », in : *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA, p. 762–768, juin 1997.*

[BBV01] N. BOUJEMAA, S. BOUGHORBEL, C. VERTAN, « Soft Color Signatures for Image Retrieval by Content », in : *Eusflat'2001, 2, p. 394–401, 2001.*

cialisé pour un type de texture. Comme on ne sait pas reconnaître automatiquement ces types dans les images, cela limite fortement l'utilité pratique de ces descripteurs à l'heure actuelle. Quant aux formes, la difficulté déjà signalée est celle de l'extraction de ces formes.

De nombreux autres travaux ont été effectués pour des bases spécifiques. Ainsi plusieurs études concernent la détection automatique et la reconnaissance de visages, tâches souvent indispensables dans de nombreux contextes. D'autres recherches concernent les nombreuses modalités d'imagerie médicale ou les bases d'empreintes digitales.

Pour notre part, nous travaillons dans un paradigme différent où l'on utilise des descripteurs locaux : à une image n'est plus associé un seul descripteur mais un ensemble de descripteurs. Cette solution ouvre la voie à des techniques de reconnaissance dites partielles, reconnaissance d'objets indépendamment du fond de l'image, reconnaissance de portions de scènes^[SM97]...

Les principales étapes de cette méthode sont les suivantes. On commence par détecter dans l'image des primitives simples : des points dans notre cas, mais on peut aussi utiliser des régions ou des contours. Divers extracteurs ont été utilisés : le plus simple consiste à simplement sous-échantillonner l'image ; le plus répétable et probablement le plus utilisé est une version légèrement modifiée du détecteur de Harris^[HS88], mais qui ne présente pas une bonne précision. D'autres possibilités ont été étudiées, par exemple pour assurer une meilleure répartition des points dans l'image^[LSBJ00].

La notion de similarité entre portions d'images est alors traduite par la notion d'invariance : on cherche des grandeurs caractéristiques du signal qui sont invariantes à certaines transformations : rotations, translations, facteurs d'échelle, mais aussi transformations photométriques. Dans la pratique, cette notion d'invariance se révèle souvent trop forte, et l'on se ramène à des quasi-invariants^[BL93] ou à des propriétés établies expérimentalement seulement^[FDF94,FCF96].

Dans le cas des points, la technique classique consiste à caractériser le signal par quelques grandeurs, par exemple la convolution du signal avec une gaussienne et ses premières dérivées. On combine alors ces grandeurs pour obtenir les propriétés d'invariance traduisant le concept de similarité pertinent pour le problème. En cette matière, le travail initial de Florack^[FtHRKV94] sur l'invariance à la rotation a ensuite été complété pour le passage à l'échelle^[DSH00] ou pour des transformations affines^[MS02], puis

-
- [SM97] C. SCHMID, R. MOHR, « Local Grayvalue Invariants for Image Retrieval », *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 5, mai 1997, p. 530–534, http://www.inrialpes.fr/movi/people/Schmid/pub97_1.html.
 - [HS88] C. HARRIS, M. STEPHENS, « A Combined Corner and Edge Detector », in : *Proceedings of the 4th Alvey Vision Conference*, p. 147–151, 1988.
 - [LSBJ00] E. LOUPIAS, N. SEBE, S. BRES, J.-M. JOLION, « Wavelet-based Salient Points for Image Retrieval », in : *Proceedings of the IEEE International Conference on Image Processing, Vancouver, Canada, 2000*.
 - [BL93] T. BINFORD, T. LEVITT, « Quasi-Invariants: Theory and Exploitation », in : *Proceedings of DARPA Image Understanding Workshop*, p. 819–829, 1993.
 - [FDF94] G. FINLAYSON, M. DREW, B. FUNT, « Color Constancy: Generalized Diagonal Transforms Suffice », *Journal of the Optical Society of America A* 11, 11, novembre 1994, p. 3011–3019.
 - [FCF96] G. FINLAYSON, S. CHATTERJEE, B. FUNT, « Color Angular Indexing », in : *Proceedings of the 4th European Conference on Computer Vision, Cambridge, Angleterre*, p. 16–27, 1996.
 - [FtHRKV94] L. FLORACK, B. TER HAAR ROMENY, J. KOENDERINK, M. VIERGEVER, « General Intensity Transformation and Differential Invariants », *Journal of Mathematical Imaging and Vision* 4, 2, 1994, p. 171–187.
 - [DSH00] Y. DUFOURNAUD, C. SCHMID, R. HORAUD, « Appariement d'images à des échelles différentes », in : *Actes du 12e Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, Paris, France*, 2, p. 327–336, février 2000.
 - [MS02] K. MIKOLAJCZYK, C. SCHMID, « An Affine Invariant Interest Point Detector », in : *Proceedings of the 7th*

pour des transformations photométriques des niveaux de gris^[SM97] et des couleurs^[Gro00]. Il faut noter qu'une des difficultés est non seulement de calculer des grandeurs invariantes, mais que l'extracteur de primitives soit lui aussi invariant aux mêmes transformations.

Une des difficultés du domaine est l'évaluation et la comparaison des méthodes. Chacune répondant à des besoins un peu différents, les comparaisons ne sont pas aisées. Par ailleurs, les résultats affichés sont souvent très dépendants des bases utilisées : lorsque celles-ci sont de taille modeste, un simple critère syntaxique peut donner l'illusion de bonnes propriétés sémantiques : détecter des visages dans une base ne contenant que des visages et des images de forêt vierge ne pose guère de difficultés, mais cela ne valide pas la méthode pour autant.

3.1.2 La vidéo

Mots clés : indexation de la vidéo, structuration, événements-clés.

Les collections de vidéos, tant chez les particuliers que les professionnels, présentent le plus souvent un nombre d'images de l'ordre de mille fois supérieur à celui de leurs homologues dans le domaine de l'image fixe. Si la qualité de ces images est souvent plus faible (images floues, faibles résolutions, forts mouvements...), elles présentent une forte redondance temporelle.

Les travaux en description des vidéos sont orientés dans plusieurs axes : la structuration qui consiste à retrouver les unités temporelles (plans, scènes), la détection d'événements (buts, applaudissements...) et la caractérisation des unités temporelles extraites. Cette caractérisation peut utiliser différents types de descripteurs : il peut s'agir de descripteurs globaux du mouvement^[Fab01], de descripteurs basés « images fixes » tirant partie de la redondance temporelle des informations^[HM00].

Concernant la structuration temporelle de la vidéo, de nombreuses contributions ont été proposées pour réaliser la segmentation temporelle des vidéos en plans élémentaires à partir de la détection de transitions^[BGG99,GKA98]. Il faut néanmoins noter que le niveau du plan se révèle le plus souvent trop atomique pour permettre une exploitation directe de ce type de partitionnement. La segmentation en scènes fournit une première solution à partir du regroupement hiérarchique des plans élémentaires. Sur ce point, un axe de recherche important reste la combinaison des caractérisations de différentes formes de contenu (son, image, vidéo...).

European Conference on Computer Vision, Copenhagen, Danemark, 2002.

- [Gro00] P. GROS, « Experimental Evaluation of Color Illumination Models for Image Matching and Indexing », in : *Proceedings of the RIAO'2000 Conference on Content-Based Multimedia Information Access*, p. 567–574, avril 2000.
- [Fab01] R. FABLET, *Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images : application à l'indexation vidéo*, thèse de doctorat, Université de Rennes 1, juillet 2001.
- [HM00] R. HAMMOUD, R. MOHR, « Mixture Densities for Video Objects Recognition », in : *Proceedings of the 15th International Conference on Pattern Recognition, Barcelone, Espagne, 2*, IAPR, p. 71–75, septembre 2000.
- [BGG99] P. BOUTHEMY, M. GELGON, F. GANANSIA, « A Unified Approach to Shot Change Detection and Camera Motion Characterization », *IEEE Transactions on Circuits and Video Technology* 9, 7, octobre 1999, p. 1030–1044.
- [GKA98] U. GARGI, R. KASTURI, S. ANTANI, « Performance characterization and comparison of video indexing algorithms », in : *Proceedings of the Conference on Computer Vision and Pattern Recognition, Santa Barbara, Californie, États-Unis*, p. 559–565, juin 1998.

3.1.3 Le texte

Mots clés : traitement automatique des langues, sémantique lexicale, apprentissage automatique, acquisition de ressources linguistiques en corpus, analyse exploratoire des données.

L'indexation de documents textuels bénéficie, contrairement à certains autres médias, d'une longue expérience des archivistes et documentalistes. L'indexation manuelle permet, à l'aide de listes d'autorité (listes de mots-clés) ou de *thesaurus*, de représenter de manière quasi unifiée, au sein d'un système de recherche d'information, les concepts abordés dans un texte. Cependant, la quantité croissante de documents numériques a laissé place à une indexation automatique plein texte (*full text*) qui, outre le problème du choix des mots contenus dans les textes qui vont les représenter (mots simples ou complexes, mots suffisamment discriminants^[Sal75,Sal89], mots situés dans une certaine partie du texte...), pose de nouveaux problèmes liés à une indexation non plus au niveau des concepts mais des mots. Deux de ces problèmes d'ordre sémantique sont fondamentaux : celui de la formulation différente d'une même idée (comment apparier le même concept contenu dans une requête et un texte, mais exprimé différemment) et, problème dual, celui de la désambiguïsation (un même mot – même chaîne graphique – pouvant exprimer des concepts différents). À ces difficultés se combine le fait que le sens d'un mot dans un document portant sur un domaine, et donc les liens sémantiques que ce mot entretient avec d'autres mots, varient en fonction de ce domaine. Pour tenter de résoudre ces difficultés, le recours à des ressources linguistiques propres au domaine traité, permettant tant de désambiguïser les mots que d'étendre les requêtes par des synonymes, des hyperonymes ou mots entretenant un lien sémantique autre avec les éléments qu'elles contiennent, est une solution envisagée. Cependant, ces ressources n'existant pas *a priori* pour tout domaine, il convient de les apprendre en corpus.

Depuis une dizaine d'années, de nombreux travaux en acquisition d'éléments lexicaux sur corpus ont vu le jour^[Gre94,HNS97], souvent basés sur des méthodes statistiques, mais également depuis moins longtemps sur de l'apprentissage symbolique^[WRS96]. Nous nous intéressons à l'acquisition, par classification, de liens sémantiques entre mots (essentiellement Nom-Nom N-N), en implémentant des principes de la théorie de la sémantique différentielle de F. Rastier^[Ras96,RCA94] à l'aide d'une méthodologie originale. Nous mettons aussi l'accent sur les liens inter-catégoriels Nom-Verbe (N-V) permettant par exemple la reformulation de *magasin de disques* en *vendre des disques*, liens actuellement très peu

-
- [Sal75] G. SALTON, *A Theory of Indexing*, *Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphie, Pennsylvanie, États-Unis, 1975.
- [Sal89] G. SALTON, *Automatic Text Processing*, Addison-Wesley, 1989.
- [Gre94] G. GREFENSTETTE, *Explorations in Automatic Thesaurus Discovery*, Dordrecht: Kluwer Academic Publishers, 1994.
- [HNS97] B. HABERT, A. NAZARENKO, A. SALEM, *Les linguistiques de corpus*, Armand Collin/Masson, Paris, 1997.
- [WRS96] S. WERMTER, E. RILOFF, G. SCHELER (éditeurs), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, *Lecture Notes in Computer Science*, Vol. 1040, Springer Verlag, 1996.
- [Ras96] F. RASTIER, *Sémantique Interprétative*, édition Second, Presses universitaires de France, 1996.
- [RCA94] F. RASTIER, M. CAVAZZA, A. ABEILLÉ, *Sémantique pour l'analyse : de la linguistique à l'informatique*, Masson, 1994.

exploités malgré leur intérêt^[Gre97,FS99,BFSJ00] ; le lexique génératif de J. Pustejovsky^[Pus95,BB01] nous offre un cadre de définition des liens N-V pertinents, que nous acquérons par programmation logique inductive^[MDR94]. Nos travaux portent tant sur la méthode d'apprentissage que sur l'intérêt linguistique ou l'intérêt applicatif des liens acquis.

Acquisition de lexiques basés sur la sémantique différentielle de Rastier

La sémantique différentielle est une théorie linguistique dans laquelle l'accent est mis sur les relations entre les significations des mots au sein d'un lexique. La signification d'un mot est définie par les différences qu'elle entretient avec celles des autres mots. Un lexique est donc un réseau de mots structuré par des classes, où les différences entre les significations sont représentées par des sèmes (ou traits sémantiques). Au sein d'une même classe sémantique, correspondant à un groupe de mots pouvant être échangés dans certains contextes, les éléments possèdent des sèmes génériques marquant leurs points communs et permettant de constituer ces classes (par exemple /pour s'asseoir/ associé à la classe {pouf, chaise, fauteuil}), et des sèmes spécifiques, explicitant leurs différences (/a des bras/ différenciant fauteuil des deux autres). Pour F. Rastier, le sens d'un mot est déterminé par le co-texte qui l'entoure, et deux types de contextes linguistiques sont fondamentaux pour caractériser les relations de signification lexicales : le thème de l'unité de texte dans laquelle est située l'occurrence étudiée et son voisinage. La sémantique différentielle précise que c'est seulement au sein de thèmes homogènes qu'il est possible de définir des classes sémantiques valides dans lesquelles les sèmes spécifiques entre mots peuvent être définis. Un thème est reconnaissable dans un texte par la présence d'une isotopie, c'est-à-dire de la récurrence de sèmes dans les ensembles de sèmes (on parle de sémèmes) représentant les significations des mots de ce texte. Ainsi, l'occurrence dans un même texte de *soldat*, *char*, *offensive* et *général* est révélatrice d'une thématique guerrière, car tous les sémèmes de ces mots sont porteurs du sème /guerre/.

Nous avons mis au point une méthodologie d'acquisition de lexiques basée sur la sémantique différentielle^[PS00,PS99]. Dans un premier temps, nous apprenons automatiquement, sur corpus, les

-
- [Gre97] G. GREFENSTETTE, « SQLET: Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text », in : *Recherche d'Informations Assistée par Ordinateur, RIAO'97*, McGill-University (éditeur), Montréal, Québec, Canada, 1997.
- [FS99] C. FABRE, P. SÉBILLOT, « Semantic Interpretation of Binominal Sequences and Information Retrieval », in : *CIMA'99 (International ICSC Congress on Computational Intelligence: Methods and Applications, Symposium on Advances in Intelligent Data Analysis AIDA'99)*, Rochester, N.Y., États-Unis, juin 1999.
- [BFSJ00] P. BOUILLON, C. FABRE, P. SÉBILLOT, L. JACQMIN, « Apprentissage de ressources lexicales pour l'extension de requêtes », *Traitement automatique des langues, numéro spécial traitement automatique des langues pour la recherche d'information* 41, 2, 2000, p. 367–393.
- [Pus95] J. PUSTEJOVSKY, *The Generative Lexicon*, MIT Press, Cambridge, 1995.
- [BB01] P. BOUILLON, F. BUSA, *Generativity in the Lexicon*, Cambridge University Press, 2001.
- [MDR94] S. MUGGLETON, L. DE-RAEDT, « Inductive Logic Programming: Theory and Methods », *Journal of Logic Programming* 19-20, 1994, p. 629–679.
- [PS00] R. PICHON, P. SÉBILLOT, « From Corpus to Lexicon: from Contexts to Semantic Features », in : *PALC'99: Practical Applications in Language Corpora*, B. Lewandowska-Tomaszczyk et P. J. Melia (éditeurs), *Lodz studies in Language*, 1, Peter Lang, 2000.
- [PS99] R. PICHON, P. SÉBILLOT, « Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience », in : *TALN'99 (Traitement automatique des langues naturelles)*, Cargèse, France, juillet 1999.

classes de mots caractéristiques des thèmes abordés, à l'aide d'une technique de classification ascendante hiérarchique (analyse de la vraisemblance du lien, AVL^[Ler91]), réalisée sur la répartition des noms dans les paragraphes. Ces classes de mots sont ensuite utilisées pour découper le corpus initial en sous-corpus thématiquement homogènes, au sein desquels nous établissons des classes sémantiques de noms par AVL sur les contextes partagés. Enfin, nous cherchons à caractériser finement au sein de chaque classe les liens de similarité et dissemblance, automatisant des études menées manuellement dans^[FHL97] par exemple.

Comme signalé ci-dessus, notre travail porte à la fois sur l'algorithme d'apprentissage, par exemple, pour pouvoir traiter des matrices très creuses de très grande taille [23], sur l'optimisation de la qualité linguistique des éléments obtenus (pertinence des classes sémantiques et automatisation de la détection des sèmes), et sur la mise au point une méthode d'exploitation efficace de ces liens intra-catégoriels dans le cadre de la recherche d'information.

Acquisition d'éléments du lexique génératif de Pustejovsky

Dans les entrées lexicales du formalisme du lexique génératif de J. Pustejovsky, la structure des qualia définit les différentes facettes de la sémantique des noms par des formules prédicatives essentiellement verbales. Le rôle télique indique le but ou la fonction du nom (*couper* pour *couteau*), l'agentif son mode de création (*construire* pour *maison*), le constitutif ses parties (*anse* pour *tasse*), et le rôle formel l'articulation de ses catégories sémantiques (*contenir* (*de l'information*) pour *livre*).

Nous avons mis au point une méthode d'apprentissage sur corpus par programmation logique inductive de règles expliquant ce qui caractérise, en termes de contexte, les paires N-V dont les constituants sont liés par un des rôles qualia par rapport aux autres paires^[CSBF01][13]. Ces règles peuvent ensuite être utilisées pour acquérir des liens N-V sur tout le corpus et constituer une partie de lexique génératif.

Notre travail porte également sur trois aspects : l'accroissement de l'efficacité et de la portabilité à moindre coût de la méthode d'apprentissage, ainsi que de l'expressivité des règles qu'elle permet d'acquérir ; l'étude de la valeur linguistique des règles apprises et de la façon dont elles permettent de verbaliser la théorie ; l'exploitation des liens qualia N-V en recherche d'information, tant en reformulation qu'en désambiguïsation.

Caractérisation de vastes collections de textes portant sur un domaine donné

La recherche des mots-clés pertinents pour caractériser les documents d'un corpus portant sur une thématique donnée n'est pas triviale : il règne une grande incohérence dans les bases bibliographiques, même lorsqu'elles sont de petite taille comme l'est la base des rapports de recherche de l'INRIA. Il est aussi intéressant de pouvoir détecter des évolutions thématiques dans le corpus, afin de mettre à jour les listes d'autorité utilisées.

-
- [Ler91] I. LERMAN, « Foundations in the Likelihood Linkage Analysis Classification Method », *Applied Stochastic Models and Data Analysis* 7, 1991, p. 69–76.
- [FHL97] C. FABRE, B. HABERT, D. LABBÉ, « La polysémie dans la langue générale et les langages spécialisés », *Sémiotiques* 13, décembre 1997.
- [CSBF01] V. CLAVEAU, P. SÉBILLOT, P. BOUILLON, C. FABRE, « Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? », *Traitement automatique des langues, numéro spécial Lexiques sémantiques dans les applications du TAL*, à paraître 42, 3, 2001.

Les données textuelles sont généralement des données non structurées mais nous supposons que les fichiers traités ont soit une structure minimale, soit un thème général commun. Nous comptons poursuivre le travail de classification et d'indexation automatique que nous avons commencé^[Mor99, MKB00]. La méthode utilisée est entièrement basée sur l'analyse factorielle des correspondances suivie ou non d'une classification automatique. Elle permet de classer les documents ou textes dont on dispose, d'indexer ces documents et de visualiser les documents et leurs mots caractéristiques. Cependant, un certain nombre de procédures, dont la sélection des mots pour l'indexation, reste empirique bien que basé sur un ensemble de résultats statistiques. Notre objectif est d'étudier la qualité et la robustesse des procédures mises en œuvre.

Par ailleurs, l'automatisation de ces procédures permettra de traiter des fichiers de taille de plus en plus importante. D'autres méthodes de type carte auto-organisatrice de Kohonen ou des variantes sont en cours de développement pour accélérer l'étape d'indexation. Ce travail fait l'objet du travail de thèse de R. Priam.

3.1.4 Multimédia et couplage entre médias

Mots clés : multimédia, couplage entre médias, couplage vidéo–son, couplage texte–image.

Les techniques citées jusqu'à présent sont principalement monomédias, alors que de nombreux documents, comme les vidéos ou les documents accessibles sur le web, combinent non seulement plusieurs médias, mais ne peuvent être totalement compris qu'en faisant intervenir tous ces médias. Créer des systèmes basés sur l'utilisation conjointe et simultanée de plusieurs médias demande de trouver un formalisme qui permette de décrire ce couplage.

Un premier formalisme syntaxique est proposé par la norme MPEG'7 et a été expérimenté dans le cadre du projet RNRT AGIR. La norme permet, en effet, d'associer de manière quelconque des descripteurs ou des schémas de description relatifs à l'image ou au son pour créer de nouveaux schémas de description. Cette possibilité présente un intérêt particulier pour la structuration automatique des vidéos, où le son et l'image fournissent des segmentations différentes et complémentaires. L'image fournit un premier découpage naturel, celui lié aux plans. Au niveau supérieur, la détection des scènes et des séquences est plus difficile et pourrait bénéficier de l'apport de la composante sonore. Celle-ci est probablement plus apte à détecter les scènes par la recherche des ruptures musicales, mais il est nécessaire de savoir à quelle échelle temporelle de telles ruptures doivent être recherchées. Le rythme de montage des plans contient cette information.

D'autres formalismes sont envisageables pour combiner l'information extraites de chacun des médias, par exemple le formalisme bayésien ou les modèles de Markov cachés. Ces formalismes permettent d'accumuler des évidences issues de médias différents pour prendre une décision de reconnaissance globale. Cela exige toutefois que tous ces médias puissent fournir leurs résultats sous forme

[Mor99] A. MORIN, « Latent Semantic Analysis and Correspondence Analysis for Thematic Exploration in Texts », in : *Proceedings Applied Stochastic Models and Data Analysis, Lisbonne, Portugal*, H. Bacelar-Nicolau (éditeur), 1999.

[MKB00] A. MORIN, M. KERBAOL, J. BANSARD, « Étude des résumés en français des rapports de recherche d'un institut d'informatique publiés de 1989 à 1998 », in : *Actes des journées d'analyse statistiques des données textuelles, Lausanne, Suisse*, mars 2000.

probabiliste, ce qui est naturel dans le domaine du son, ou de la vidéo^[Fab01]. Dans le domaine de l'image fixe, on peut noter quelques travaux, dont ceux de R. Hammoud^[HM00] ou de S. Picard^[MPS97]. Pour le texte, certaines approches, dont celles utilisées par A. Morin ou J.-C. Chappelier à l'EPFL, se placent naturellement dans un cadre statistique.

3.1.5 Évaluation des descripteurs

Mots clés : évaluation, performance, pouvoir discriminant.

Sur ce sujet, la situation est très contrastée selon les médias. Dans le domaine du texte, du son et de la parole existent des bases de tests de référence ou des organismes qui organisent des campagnes de test régulières (NIST pour le son et la reconnaissance de la parole, TREC pour le texte en anglais, AMARYLLIS pour le texte en français, SENSEVAL ou ROMANSEVAL sur les aspects désambiguïsation)¹. Dans le domaine de l'image fixe et de la vidéo, le BENCHATLON a pour but de fournir des bases et des algorithmes d'évaluation des systèmes d'indexation d'images fixes, TREC développe des tests pour la vidéo, une base et un système d'évaluation des systèmes de détection des changements de plans a été mis au point par G. Quenot et P. Joly^[RJMMQ99].

Beaucoup reste pourtant à faire, à commencer dans le domaine de l'image fixe. Une fois un descripteur mis au point, il est nécessaire de le tester et d'en évaluer les performances vis-à-vis de la tâche pour laquelle il a été développé. Ce travail est souvent effectué par les concepteurs du descripteur, mais cette évaluation rencontre deux limites. Tout d'abord, elle est souvent biaisée par la base de données utilisée pour le test, base qui est souvent celle utilisée lors du développement du descripteur. D'autre part, la taille de ces bases est souvent très inférieure à celle nécessaire pour prouver l'utilité du descripteur dans des conditions réelles d'utilisation. Ainsi, la quasi-totalité des descripteurs d'images fixes est testée sur un à quelques milliers d'images, alors que le stock d'images d'une agence de photos se monte à quelques millions d'images.

Quels que soient les médias, la difficulté principale de l'évaluation des descripteurs sur de grands volumes de données est l'exploitation et l'interprétation des résultats fournis par cette évaluation. Pour contourner les problèmes soulevés, deux voies sont possibles. La première est la mesure du niveau de satisfaction de l'utilisateur, qui nécessite la mise en place d'un contexte réaliste de test, la deuxième

¹On peut d'ailleurs noter que, pour les documents textuels au sein de TREC, un glissement de la recherche d'information ramenant des documents pertinents par rapport à une requête vers des systèmes questions – réponses (SQR) a vu le jour ces dernières années. Dans ces SQR, plus que le document pertinent, c'est la réponse à la question qu'il contient qu'il convient de retourner : zone de texte la mentionnant, ou information extraite.

-
- [Fab01] R. FABLET, *Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images : application à l'indexation vidéo*, thèse de doctorat, Université de Rennes 1, juillet 2001.
- [HM00] R. HAMMOUD, R. MOHR, « Mixture Densities for Video Objects Recognition », in : *Proceedings of the 15th International Conference on Pattern Recognition, Barcelone, Espagne, 2*, IAPR, p. 71–75, septembre 2000.
- [MPS97] R. MOHR, S. PICARD, C. SCHMID, « Bayesian Decision Versus Voting for Image Retrieval », in : *Proceedings of the 7th International Conference on Computer Analysis of Images and Patterns, Kiel, Allemagne*, p. 376–383, 1997.
- [RJMMQ99] R. RUILOBA, P. JOLY, S. MARCHAND-MAILLET, G. QUENOT, « Towards a Standard Protocol for the Evaluation of Video-to-Shots Segmentation Algorithms », in : *Proceedings of the first European Workshop on Content Based Multimedia Indexing, Toulouse, France*, octobre 1999.

solution est la mise au point d'un protocole automatique d'évaluation qui vise au maximum à une certaine objectivité des résultats.

Suivant la première voie, nous avons débuté un travail d'évaluation de l'apport des paires nom – verbe dont les constituants sont liés pas un des rôles de la structure des qualia aux capacités de recherche d'un système de recherche d'information. Le critère d'évaluation dans ce travail est la satisfaction de l'utilisateur menant la recherche.

Dans l'optique de la seconde voie, nous avons commencé à développer une base de séquences d'images spécifiques². En mélangeant ces séquences dans des bases d'images disparates, on peut tester le degré de robustesse des descripteurs vis-à-vis de certains changements des paramètres de prise de vue. L'avantage de cette base est de fournir une vérité terrain sûre et graduée : chaque séquence contenant des images de plus en plus différentes, il est possible de mesurer un degré de robustesse plutôt qu'une simple réponse binaire. La dépendance vis-à-vis des autres images ne disparaît pas, mais devient plus faible, car ce n'est pas dans ces images que l'on cherche un résultat ou une image définie comme telle par une mesure approximative de l'évaluateur. Cette méthodologie est plus adaptée pour tester des algorithmes de reconnaissance d'objets que des algorithmes de reconnaissance plus généraux. Elle doit néanmoins pouvoir être étendue en ajoutant dans notre base de séquences des images appropriées pour de tels tests.

3.1.6 Méta-données

Mots clés : méta-données.

Pour améliorer l'organisation des données ou pour définir les stratégies de choix de certains descripteurs, il peut être judicieux d'utiliser des informations exogènes ou contextuelles, c'est-à-dire indirectement liées au contenu des données. Ces informations appelées méta-données (données sur les données) peuvent être liées à la production ou à l'usage des données. Des méta-données sur les requêtes, ou sur le comportement d'un utilisateur en recherche d'information, peuvent également être définies et, par exemple, décrire de façon explicite le contenu d'un document.

De manière comparable au problème de la sélection des descripteurs, le choix des méta-données à utiliser doit être considéré « au cas par cas » en tentant de réduire autant que possible l'ordre de grandeur du volume de données effectivement examinées.

V. Kashyap et A. Sheth^[KS98] ont proposé une classification des méta-données pour les données multimédias qui les répartit en deux grandes catégories : celles qui contiennent des informations exogènes (date, localisation...), et celles qui contiennent des informations endogènes liées au contenu ou à la représentation des documents, appelées descripteurs et qui peuvent être elles-mêmes réparties en deux groupes :

1. les méta-données qui sont calculées directement à partir du contenu : descripteurs d'images, indices *full text*, fichiers inversés et vecteurs d'indices d'un document textuel ou, pour l'audio, les différents descripteurs permettant la reconnaissance d'un locuteur ou les changements de sujets ;

²La base d'images est accessible en <http://www.irisa.fr/texmex/Images/>

[KS98] V. KASHYAP, A. SHETH, « Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies », in : *Cooperative Information Systems*, M. Papazoglou et G. Schlageter (éditeurs), Academic Press, San Diego, Californie, États-Unis, 1998, p. 139–178.

2. les méta-données descriptives de contenu : annotations décrivant le contenu d'une image par un ensemble de mots-clés tels les *Impression Vectors* proposés dans^[KKH94].

Chacun de ces types de méta-données peut être spécifique à un domaine d'application (comme le sont les descripteurs *Q-Features*^[JH94], *NDVI*^[AS94], *Content Classification Meta-data*^[BR94] pour les images et vidéos), auquel cas celles-ci renseignent davantage le contexte de production voire d'utilisation « prévue » de la donnée.

À cette première classification, on peut ajouter que la finalité des méta-données (l'accès aux données, la classification, le résumé, l'interopérabilité, la représentation du média ou de sa structure, la traçabilité...) et la manière dont celles-ci sont générées (par exemple, la méthode LSI utilisée pour l'indexation d'un texte, la technique de compression d'une image) ont une importance évidente sur le choix des données auxquelles sont rattachées ces méta-données.

Notre travail consiste à répertorier toutes les méta-données pertinentes pour les utilisations des documents multimédias que nous considérons et à compléter leur classification pour chaque type de média en ajoutant des aspects jusqu'alors non abordés dans la littérature : les méta-données relatives à la qualité d'une donnée monomédia, celles relatives au couplage entre médias ou à la qualité d'un descripteur (pouvoir discriminant, performance).

Les méta-données sont un outil privilégié pour maintenir diverses informations relatives aux données ou à leurs descripteurs et qui peuvent en faciliter l'exploitation future. C'est en particulier le cas des informations de cohérence qui permettent de limiter le renouvellement du calcul des descripteurs, ou des informations de fiabilité qui permettent de caractériser la confiance que l'on peut accorder à la réponse à une requête.

3.2 Exploitation efficace des descriptions

Mots clés : qualité, stratégies de calcul, statistiques, indexation.

Résumé : *La description des contenus, fût-elle automatique, ne suffit pas à réaliser un système utilisable. Il faut que ces descriptions puissent être exploitées dans des délais raisonnables pour satisfaire les attentes de l'utilisateur. Cette partie du rapport est ainsi consacrée aux outils d'exploitation des descripteurs et méta-données.*

On peut grossièrement diviser les différentes manières d'exploiter les données selon que cette exploitation se fait en-ligne ou hors-ligne. Le cas hors-ligne rassemble plutôt les procédures d'exploitation massive qui ont en général besoin d'analyser toutes les données, le temps de calcul n'étant pas alors un facteur prépondérant. À l'opposé, le cas de l'exploitation en-ligne rassemble de préférence des procédures qui doivent répondre

-
- [KKH94] Y. KIYOKI, T. KITAGAWA, T. HAYAMA, « A Meta-Database System for Semantic Image Search by a Mathematical Model of Meaning », *ACM SIGMOD Record, Special issue on Metadata for Digital Media* 23, 4, 1994.
- [JH94] R. JAIN, A. HAMPAPURAM, « Representations of Video Databases », *ACM SIGMOD Record, Special issue on Metadata for Digital Media* 23, 4, 1994.
- [AS94] J. ANDERSON, M. STONEBRAKER, « Sequoia 2000 Metadata Schema for Satellite Images », *ACM SIGMOD Record, Special issue on Metadata for Digital Media* 23, 4, 1994.
- [BR94] K. BÖHM, T. RAKOW, « Metadata for Multimedia Documents », *ACM SIGMOD Record Special Issue on Metadata for Digital Media* 23, 4, 1994, p. 21–26.

rapidement. Pour obtenir de telles performances, ces procédures s'appuient sur des résultats pré-calculés hors-ligne pour déterminer un sous-ensemble de la base qu'il suffit d'examiner pour réaliser la tâche. Le but de cette partie est donc de réfléchir aux techniques permettant de réduire les coûts des procédures d'exploitation de grands volumes de données multimédias.

3.2.1 Stratégies de sélection et de calcul des descripteurs

Mots clés : sélection, stratégies de calcul, calcul à la volée.

Le travail de spécification et de conception des descripteurs de données multimédias ressort de la compétence des spécialistes de chaque média, et peut être évalué en fonction des propriétés de discrimination de chaque descripteur et de la reconnaissance qu'il permet. De très nombreux descripteurs ont pu être proposés, correspondant à des médias et à des objectifs de reconnaissance variés.

Cependant, il n'est pas toujours clair de savoir à quel problème répond tel descripteur et quelles performances le caractérisent. Suivant le problème traité, il peut rester un important travail à faire pour trouver de bons descripteurs. Aussi, la caractérisation des performances reste un impératif. D'une manière générale, la plupart de ces descripteurs ont été spécifiés selon une pure approche ascendante sans prise en compte explicite, par exemple, de leur exploitation future par des algorithmes d'indexation.

Pour un même type de données, la palette des descripteurs potentiellement intéressants est souvent grande, et il est prohibitif, parfois inutile, de vouloir calculer tous les descripteurs pour toutes les données. Suivre l'évolution des techniques de description des données, mais aussi les « comportements d'interrogation » ou les modes d'usage, suggère de pouvoir enrichir la base de nouveaux traitements, de nouveaux descripteurs, d'en combiner certains ou, au contraire, d'en abandonner d'autres selon des stratégies spécifiques.

En interaction forte avec le travail de conception et d'évaluation des descripteurs de données multimédias, la première étape de notre travail s'attache ici à définir des critères et des stratégies permettant de préconiser le choix d'un ou d'une combinaison de descripteurs, selon la nature de la requête et l'usage des données. La démarche consiste d'abord en l'étude des propriétés des descripteurs de données multimédias avec leur pré- et post-conditions d'application pour garantir leur efficacité, puis de définir une algèbre minimale de descripteurs, formalisme permettant de représenter la sémantique des descripteurs de données multimédias et leurs dépendances éventuelles. Il s'agit de définir les notions d'équivalence et de congruence de descripteurs afin de proposer, par exemple, une substitution des descripteurs, ainsi que des opérateurs de composition pouvant être appliqués selon la nature et le volume de données à décrire. De même, il est intéressant d'évaluer la minimalité d'une combinaison de descripteurs pour procéder, par la suite, à une réécriture de requêtes (implémentant les opérateurs de manipulation des descripteurs).

Plusieurs travaux antérieurs (*Multimedia Presentation Algebra*^[ASS00], *Acoi Algebra*^[NK98]) ont permis d'identifier un ensemble minimal de caractéristiques sur les données multimédias, d'opérateurs topologiques et de structures d'indexation à utiliser pour la recherche d'images (ou de vidéos) par le

[ASS00] S. ADALI, M. SAPINO, V. SUBRAHMANIAN, « An Algebra for Creating and Querying Multimedia Presentations », *Multimedia Systems* 8, 3, 2000, p. 212–230.

[NK98] N. NES, M. KERSTEN, « The Acoi Algebra: a Query Algebra for Image Retrieval Systems », in : *Advances in Databases. 16th British National Conference on Databases*, p. 77–88, 1998.

contenu. La plupart de ces formalismes intègrent des opérateurs de type relationnel, topologique ou spatial, des opérateurs de tri et d'ordonnement, des contraintes gérant les aspects temporels basés sur des primitives algébriques relativement classiques en base de données. Ces propositions formelles ont été implantées dans de nombreux langages de requêtes qui peuvent se classer en trois catégories : 1) les langages de requêtes entièrement nouveaux et spécialisés^[ATS96], 2) les langages basés sur la programmation logique ou fonctionnelle^[MS96] et 3) les langages étendant SQL^[MHM96].

Plus rares sont les travaux qui intègrent des primitives permettant la manipulation des structures mathématiques (histogrammes, transformées de Fourier...) relatives à certaines classes de descripteurs^[GS00]. Ils n'ont, à notre connaissance, pas donné lieu à une implantation au sein d'un langage de requêtes. Pour notre part, nous proposons d'étendre l'algèbre d'histogrammes proposée dans la littérature^[GS00], de l'appliquer à d'autres structures mathématiques de descripteurs et à des types composés de plusieurs structures, en considérant notamment les possibilités de substitution de descripteurs, de congruence, de composition, comme phase préparatoire à la réécriture et à l'optimisation des requêtes sur ces structures de données complexes.

Dès lors qu'il est question du coût d'un descripteur ou d'une méta-donnée en termes de temps de calcul, notre objectif est de proposer un modèle de coût qui définit :

- le moment où il est le plus opportun de déclencher le calcul d'un descripteur ou d'une méta-donnée,
- le type de calcul (pré-calcul, calcul à la volée, calcul à la demande) qui doit être adopté selon l'application, le volume de données et la nature du descripteur,
- la taille du jeu de données ciblé par un descripteur : il s'agit de déterminer s'il est possible de ne calculer (ou pré-calculer) un descripteur que pour une partie des données,
- le pouvoir discriminant *a priori* et *a posteriori* d'un descripteur, c'est-à-dire évaluer si les résultats bâtis grâce à ce descripteur sont de bonne qualité et établir un retour sur la pertinence de ce descripteur (*relevance feedback*).

Cette analyse devra permettre, par la suite, la mise en œuvre de techniques telles que l'anticipation de chargement (*prefetching*) de données pour les pré-calculs.

De plus, le temps nécessaire au calcul d'un descripteur pour l'ensemble de la base, et le fait que les données puissent être stockées en mémoire tertiaire ou sur des supports externes créent des difficultés supplémentaires, accentuées par la variabilité des données et des requêtes. Ainsi, dans ce modèle de coût, il importe de prendre en compte ces contraintes matérielles pour préconiser un choix de calcul de descripteurs.

-
- [ATS96] H. ARISAWA, T. TOMII, K. SALEV, « Design of Multimedia Database and a Query Language for Video Image Data », in : *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, p. 462–467, 1996.
- [MS96] S. MARCUS, V. SUBRAHMANIAN, « Foundations of Multimedia Database Systems », *Journal of the ACM* 43, 3, 1996, p. 474–523.
- [MHM96] E. MEGALOU, T. HADZILACOS, N. MAMOULIS, « Modeling and Querying Very Large Interactive Multimedia Repositories », in : *Proceedings of the International Conference on Multimedia Modeling*, p. 323–338, 1996.
- [GS00] A. GUPTA, S. SANTINI, « Towards Feature Algebras in Visual Databases: The Case for a Histogram Algebra », in : *Proceedings of the IFIP Working Conference on Visual Databases (VDB5)*, 2000.

3.2.2 Contrôle de la qualité et de la cohérence des données, descripteurs et méta-données

Mots clés : qualité des données, cohérence des données.

Vu le temps considérable que peut prendre le calcul d'une famille de descripteurs sur une grande base de documents, il n'est pas possible de la recalculer à chaque modification de la base. La contrepartie est que, une fois les descripteurs calculés, ces derniers risquent d'être rapidement « décorrélés » du nouvel état de la base, de ne plus être pertinents ou encore, de perdre leur pouvoir discriminant. La cohérence entre données et méta-données conditionne, d'une part, la pertinence et la qualité des méta-données et descripteurs utilisés et, d'autre part, la validité des résultats de recherche sur un grand volume de données multimédias.

Si la création des méta-données et le calcul des descripteurs doivent être menés en cohérence avec le cycle de vie des données auxquelles ces éléments sont attachés (à la création, mise à jour, et suppression), il est nécessaire de trouver un compromis entre la fraîcheur des méta-données et des descripteurs, le coût de leur calcul et les contraintes de l'application.

Notre démarche consiste à étudier sur un domaine précis d'application (par exemple la vision en robotique) les contraintes diachroniques, c'est-à-dire les contraintes de cohérence entre les données et leurs méta-données à deux instants successifs, et les seuils de tolérance que l'on peut définir selon le contexte de l'application, la nature des descripteurs et méta-données, selon l'impact sur leur pertinence et leur efficacité et, enfin, selon la nature et le nombre des données modifiées.

De nombreux travaux s'intéressent aux processus de rafraîchissement des données et à la traçabilité des données dans un contexte d'entrepôt où les données sont issues de sources distribuées et où les vues matérialisées doivent être maintenues en cohérence avec les données^[BFG⁺99,GFS⁺01]. Dans ce contexte, l'utilisation des méta-données est encouragée mais le problème de leur rafraîchissement reste relativement peu évoqué sur de grands volumes de données.

Au-delà de la cohérence, il importe d'évaluer et de contrôler la qualité des données et des méta-données sur un ensemble de critères objectifs et quantifiables, qu'il nous faut définir tout en maîtrisant le coût. La thématique de recherche sur la qualité des données ne s'est véritablement révélée, en tant que thème de recherche à part entière au niveau international, que depuis environ sept ans^[WSF95,TB98,Red96].

3.2.3 Statistiques pour les grands volumes de données

Mots clés : analyse exploratoire des données, statistiques, échantillonnage.

Si le cas de données en petit nombre a déjà été étudié, la trop grande abondance de données pose d'autres problèmes : en particulier, l'utilisation des résultats de la statistique inférentielle classique

-
- [BFG⁺99] M. BOUZEGHOUB, F. FABRET, H. GALHARDAS, M. MATULOVIC, J. PEREIRA, E. SIMON, « Data Warehouse Refreshment », in : *Fundamentals in Data Warehouses*, Springer, 1999.
- [GFS⁺01] H. GALHARDAS, D. FLORESCU, D. SHASHA, E. SIMON, C.-A. SAITA, « Improving Data Cleaning Quality Using a Data Lineage Facility », in : *Workshop on Design and Management of Data Warehouses (DMDW)*, 2001.
- [WSF95] R. WANG, V. STOREY, C. FIRTH, « A Framework for Analysis of Data Quality Research », *IEEE Transactions on Knowledge and Data Engineering* 7, 4, 1995, p. 623–638.
- [TB98] G. TAYI, D. BALLOU, « Examining Data Quality », *Communication of the ACM* 41, 2, 1998, p. 54–57.
- [Red96] T. REDMAN, *Data Quality for the Information Age*, Artech House, 1996, ISBN 0-89006-8836.

dans le domaine des tests conduit à rejeter trop souvent l'hypothèse nulle. Les méthodes d'identification de modèles échouent trop souvent ou alors conduisent à une surestimation de la précision du modèle obtenu. Comment réaliser un échantillonnage représentatif dans de tels ensembles de données ? D'autre part, certains algorithmes de classification sont inutilisables sur de tels volumes compte tenu de leur complexité.

L'exploration de grands volumes de données pose des problèmes difficiles. Ces problèmes sont dûs tant à la complexité des calculs, qui est induite par les grands volumes à traiter, qu'au fait que ces données ne sont généralement pas connues avec précision : les modèles que l'on pourrait chercher à estimer sont alors rejetés.

Pour commencer, nous pourrions dire que la plupart des méthodes statistiques sont utilisables avec des précautions et à condition que les données soient de très grande qualité. Dans le cas de grands volumes de données où on recherche parfois des formes petites et très localisées ou alors des écarts à une règle, la non-qualité des données est un handicap et peut conduire à de grandes erreurs. La première étape consiste à filtrer, nettoyer les données et vérifier leur cohérence.

Nous distinguerons deux objectifs au traitement d'un grand volume de données : soit la construction d'un modèle global, soit la recherche de structures dans les données. Ces deux objectifs conduisent à deux classes générales de techniques. En statistique, on recherche souvent des modèles « globaux », qu'il s'agisse d'une partition d'un ensemble de données (classification), d'un modèle de régression, d'un arbre de segmentation. Dans ce cas, on peut travailler avec un échantillon de données et découvrir les caractéristiques importantes de l'ensemble de départ ; cette stratégie est la meilleure si on ne veut pas surestimer la précision du modèle. Dans le cas de recherche de structures, on souhaite découvrir des « niches » et on n'est pas certain de les trouver à partir d'un échantillon. Il faut utiliser d'autres heuristiques permettant de travailler avec de grands fichiers de données.

Le problème de l'échantillonnage dans de grandes bases de données n'est pas résolu. Que faire ? Constituer un panel de données en utilisant des techniques identiques à celles qu'utilisent les statisticiens lors du traitement des données de recensement ?

L'analyse de données exploratoire (EDA) est un outil indispensable pour le traitement de grands volumes de données. L'EDA explore les données de façon interactive sans hypothèses préconçues et fournit des représentations graphiques utiles. Les méthodes de visualisation pour des données qui sont dans un espace de dimension supérieur à trois sont tout aussi indispensables : citons les méthodes en coordonnées parallèles. Ces techniques utilisent les propriétés de l'œil pour explorer les données.

Il faut noter qu'actuellement les logiciels de *datamining*, très coûteux par ailleurs, proposent à leurs utilisateurs des outils qui sont des techniques statistiques adaptées pour l'occasion et pas toujours performantes dans ce contexte : statistique descriptive, analyses en composantes principales et des correspondances, arbres de segmentation et réseaux neuronaux la plupart du temps. Il y a un gros travail d'adaptation de ces méthodes et des méthodes statistiques en général à réaliser.

3.2.4 Indexation multidimensionnelle

Mots clés : indexation multidimensionnelle, malédiction de la dimension, complexité, espace de grandes dimensions.

Les algorithmes d'indexation sont destinés à limiter au maximum la quantité de données qui doit être accédée durant un processus de recherche. Ils sont particulièrement important pour des recherches

par similarités sur un contenu visuel car, sans eux, la quantité de données à analyser serait telle qu'il serait impossible d'examiner ces données de façon exhaustive en un temps réaliste. Ces techniques d'indexation, essentiellement développées par la communauté des bases de données, se classent en deux grandes catégories : celles qui regroupent les données en fonction de leur proximité (*data-partitionning index methods*), et celles qui découpent *a priori* l'espace multidimensionnel et qui ensuite stockent les données selon ce découpage (*space-partitionning index methods*).

Or, dans un contexte de description visuelle, les techniques d'indexation traditionnellement employées en bases de données sont toutes prises en défaut, essentiellement parce que les index sont inadaptes aux recherches de plus proches voisins sur des données de grande dimension et parce que leur complexité est exponentielle avec la dimension des données. R. Weber^[WSB98] démontre que le nombre de dimensions au-delà duquel *une recherche séquentielle devient plus performante que toute recherche au travers d'un index* se situe autour de 10. La cause de cet état de fait est que la recherche tend à explorer l'intégralité de l'espace pour identifier les voisins les plus proches, et selon une complexité fortement exponentielle avec la dimension des données.

Cette difficulté est exacerbée lorsque l'on utilise des techniques modernes de description du contenu, comme c'est le cas avec les descripteurs locaux pour la description d'images fixes. Ce type de descripteur est préféré aux descripteurs globaux, plus traditionnels, car ils sont plus souples, plus puissants, permettent des recherches plus fines et absorbent mieux certaines variations (illuminations différentes, recadrages, points de vues différents...), mais plus complexes à calculer et à gérer. La différence essentielle est qu'avec les descripteurs locaux, chaque image est décrite par beaucoup de descripteurs (entre 50 et 600 contre un seul pour une description globale). Ceci a pour conséquence de changer fondamentalement la manière dont la reconnaissance s'effectue : il est alors nécessaire d'interroger de nombreuses fois la base puis de synthétiser les résultats avant de pouvoir retourner une réponse, et non plus de se contenter d'une seule et unique interrogation comme dans le cas des descripteurs globaux. Ce nouveau mode d'interrogation disqualifie toutes les techniques connues en indexation de bases de données, déjà mises à mal dans des contextes de description plus traditionnels.

Nos travaux s'appuient sur les tendances actuelles de la recherche dans le domaine et sur notre expertise des descripteurs : prise en compte de la non-uniformité de données et de la multiplicité des interrogations effectuées pour répondre à une requête, utilisation de schémas d'approximation adaptés aux données, algorithmes de recherche basés sur des simplifications de la recherche séquentielle...

3.2.5 Supports systèmes et matériels

Mots clés : gestion de la mémoire, caches, gestion des disques, processeurs reconfigurables.

En complément des directions précédentes, nous voulons examiner les techniques de bas niveau permettant d'accéder aux données et de les traiter le plus rapidement possible. Ce problème est difficile car ces données sont stockées et traitées par des ressources partagées et critiques comme le sont les supports secondaires.

De nombreuses stratégies d'accès aux disques ont été proposées. Ces stratégies explorent certains

[WSB98] R. WEBER, H. SCHEK, S. BLOTT, « A Quantitative Analysis of Performance Study for Similarity-Search Methods in High-Dimensional Spaces », in : *Proceedings of the 24th International Conference on Very Large Data Bases, New York City, New York, États-Unis*, p. 194–205, août 1998, <http://www-dbs.ethz.ch/~weber/paper/VAFILE.ps.gz>.

aspects « systèmes » de l'exploitation de données : gestion de caches et éviction, ordonnancement des lectures (*scheduling*), anticipation de chargement (*prefetching*), mise en cache de résultats temporaires... Ces stratégies sont bien établies dans le contexte des systèmes de fichiers ou de bases de données classiques, mais peu ont été proposées dans la perspective de données multimédias.

Une seule publication aborde pour le moment ce point dans le contexte des index multidimensionnels^[BEKS00] : à partir de l'analyse de plusieurs requêtes soumises simultanément, il est déterminé un ordonnancement des lectures visant à ne charger en mémoire qu'une seule et unique fois les données et, plutôt que de traiter en séquence et intégralement chaque requête, toutes celles concernées par les données qui viennent d'être chargées peuvent progresser simultanément. Cette approche est un premier pas vers la prise en compte de la localité spatiale et temporelle des traitements multidimensionnels, clé de voûte des stratégies d'accès aux disques dans les contextes classiques.

De nouveau, les caractéristiques des descripteurs et la manière dont ils utilisent les données sont des éléments-clés pour fabriquer des politiques performantes d'accès aux disques. Il est peu probable que des politiques généralistes, incluant simplement un aspect multidimensionnel, apportent une amélioration conséquente. Parmi les pistes que nous désirons suivre, nous voulons exploiter le contenu des caches, réutiliser au maximum les données chargées, et regarder l'effet de l'anticipation de lectures (ces deux pistes semblent particulièrement intéressantes dans le cas des descripteurs locaux^[AG01]) et les risques encourus si les prévisions sont mauvaises. De plus, la connaissance des descripteurs permet éventuellement de collecter lors d'une recherche des informations non demandées, mais connexes à la question, qui faciliteraient des recherches ultérieures.

En complément, l'utilisation d'accélérateurs matériels semble très prometteuse. Dans les années 1980, des processeurs spécialisés dans des tâches BD étaient intégrés aux systèmes. Ceux-ci comportaient des algorithmes câblés (tris, jointures, filtrage) très efficaces, capables de traiter au vol les données. À l'époque, le coût de fabrication de ces processeurs était tel, les délais si importants et l'extensibilité si réduite que ces approches ont été abandonnées. L'évolution des techniques semble permettre une « réincarnation » des idées proposées à l'époque via l'utilisation de processeurs reconfigurables FPGA. En chargeant des filtres dans ces processeurs qui sont intercalés entre les disques et la mémoire, on peut filtrer (ou trier...) à la volée les données et ne laisser passer que les informations pertinentes ; la reconfigurabilité permet de changer instantanément la nature des filtres, du tri, ou la façon dont les différents étages mis en parallèle sont reliés. Ce travail, très prospectif et très amont, se fait dans le contexte d'une réflexion menée conjointement avec D. Lavenier (SYMBIOSE) et P. Quinton (R2D2), et vise à court-circuiter le cheminement typique des données au sein d'un système pour les traiter au plus vite le plus près de leur lieu de stockage.

[BEKS00] B. BRAUNMÜLLER, M. ESTER, H. KRIEGEL, J. SANDER, « Efficiently Supporting Multiple Similarity Queries for Mining in Metric Databases », in: *Proceedings of the IEEE International Conference on Data Engineering*, p. 256–267, San Diego, Californie, États-Unis, février 2000.

[AG01] L. AMSALEG, P. GROS, « Content-based Retrieval Using Local Descriptors: Problems and Issues from a Database Perspective », *Pattern Analysis and Applications 2001*, 4, 2001, p. 108–124.

4 Domaines d'applications

4.1 Gestion de bases d'images fixes

Mots clés : bases d'images, agences de photos, photographies numériques, imagerie médicale.

Nous nous intéressons plus particulièrement aux grandes bases d'images, comme celles gérées par les agences de photos. Ces agences possèdent entre cinq cent mille et douze millions d'images. L'agence Andia Presse en a un million, Sigma douze millions, l'agence Corbis qui regroupe l'ensemble des acquisitions de Microsoft en a trente six millions. Ces agences travaillent selon deux modes. Dans le premier, elles sélectionnent et envoient gratuitement à un client suite à sa demande un lot d'images, celui-ci ne payant ensuite, au moment de la diffusion, que les images qu'il a retenues. Dans le deuxième mode, les clients sont abonnés aux agences qui leur envoient systématiquement leurs nouvelles photos, le mode de paiement étant le même. Ce mode de travail est celui de l'AFP ou de Reuters.

Un des soucis des agences est bien sûr la gestion de la propriété des images, et le fait qu'elles ne soient pas indûment utilisées par des personnes ou institutions n'en ayant pas acquitté les droits. Le tatouage et l'indexation sont deux techniques envisagées pour contrôler la diffusion des images, soit en cherchant une marque de propriété dans les images pour le tatouage, soit en vérifiant par indexation que l'image n'est pas un fragment d'une image de la base de l'agence.

Un autre domaine important où la gestion des images acquiert une importance croissante est celui des images médicales. La difficulté est alors que l'accès au contenu médicalement intéressant est très ardu, alors que les contraintes en matière de qualité de la recherche sont très forte. Les applications des méthodes basées sur l'exploitation des contenus sont donc encore balbutiantes dans ce domaine.

4.2 Gestion des bases de vidéos

Mots clés : bases de vidéos, structuration de vidéos.

Les bases de vidéos existantes sont généralement peu numérisées. Le passage progressif de la télévision au numérique devrait rapidement changer ceci. Ainsi, TF1 est passé à la production entièrement numérisée, seules les caméras restant analogiques. Traitement, montage et diffusion sont numériques. Par ailleurs, les décodeurs numériques domestiques peuvent désormais être équipés de disques durs permettant un stockage d'abord modeste, d'une dizaine d'heures de vidéo, mais à terme important, d'un millier d'heures.

On peut alors distinguer deux types d'archives numériques. Tout d'abord celles des particuliers, comprenant des enregistrements de programmes diffusés et des films pris à l'aide de caméscopes numériques. Si l'effort de gestion de telles bases sera probablement faible, sans méthode rigoureuse, il y a un grand besoin d'outils pour aider l'utilisateur : création automatique de résumés et de sommaires pour permettre de retrouver aisément une information, ou de se faire en quelques minutes une idée générale d'un programme. Même si le service est rustique, il sera d'abord évalué en fonction de la plus-value qu'il apporte à un système (magnétoscope, décodeur), devra rester peu cher, mais profitera d'une grande diffusion.

D'un autre côté, on trouve les archives professionnelles : archives des chaînes de télévision, dépôt légal, cinémathèques, producteurs... Ces archives sont de taille beaucoup plus grande, mais bénéficient des soins attentifs de professionnels de la documentation et de l'archivage. Dans ce domaine, les sys-

tèmes peuvent être beaucoup plus chers et sont jugés en fonction des gains de productivité et de l'aide qu'ils apportent aux documentalistes, journalistes et utilisateurs.

4.3 Gestion des grandes bases textuelles

Mots clés : bibliographie, indexation.

La recherche dans les grands corpus textuels a fait l'objet de nombreuses recherches. Les enjeux actuels sont de pouvoir traiter de très grands volumes de données, de pouvoir faire des requêtes portant plus sur des concepts que sur la simple inclusion dans les textes de chaînes graphiques, et de pouvoir caractériser des ensembles de textes.

Nous travaillons ainsi sur l'exploitation de bases bibliographiques scientifiques. Il se trouve que l'explosion du nombre des publications scientifiques rend la recherche des données pertinentes pour un chercheur très difficile. La généralisation de leur indexation dans des banques de données n'a pas résolu le problème. La difficulté est de choisir les mots-clés qui vont cerner au mieux un domaine d'intérêt. La méthode statistique utilisée, l'analyse factorielle de correspondances, permet d'indexer les documents ou un ensemble de documents et fournit la liste des mots clés les plus discriminants pour ce ou ces documents. La validation de l'indexation est réalisée en faisant une recherche d'information dans des bases de données plus générales que celle qui a permis d'élaborer l'index et en étudiant les documents rapportés. Cela permet en général de réduire encore le sous-ensemble de mots caractérisant un domaine.

Une autre difficulté est de savoir correctement retrouver au sein même d'un document les parties qui abordent un sujet. Nous avons ainsi travaillé sur l'extraction automatique à partir de textes de bioinformatique provenant de bases telles Medline des zones de textes décrivant des interactions entre gènes et de modéliser l'interaction décrite. La modélisation nécessitant une analyse fine et coûteuse des phrases, elle ne doit être effectuée que sur des zones de textes susceptibles de contenir effectivement une interaction. Nos méthodologies d'apprentissage de liens sémantiques entre mots sont exploitées pour déterminer ces zones de textes pertinentes. Sur un corpus de résumés extraits de Medline, nous appliquons un apprentissage par PLI pour tenter d'apprendre ce qui distingue les phrases contenant des interactions des autres.

4.4 Robotique et asservissement visuel

Mots clés : robotique, asservissement visuel, mémoire visuelle, planification.

Si la collaboration entre robotique et vision est un sujet déjà ancien, elle a subi un changement important de paradigme dans les cinq dernières années. Jusqu'alors, la collaboration était envisagée au niveau de la planification : une caméra observait le monde autour d'un robot pour lui permettre de planifier ses déplacements. Les résultats n'ont pas été à la hauteur des attentes.

Le champ de la collaboration s'est alors déplacé vers l'asservissement : la vision ne sert plus alors à planifier un mouvement, mais à en assurer le suivi et la bonne exécution, en mettant en place une

boucle de contrôle fermée avec rebouclage par la vision justement^[ECR92,Cha98,MCB99]. Les résultats sont tout à fait prometteurs et de nombreuses applications industrielles existent déjà.

Quelques difficultés demeurent : les tâches à accomplir sont spécifiées à l'aide d'une image cible qu'il s'agit d'atteindre, mais cela suppose que le robot est capable d'établir un lien entre cette image et l'image courante fournie par la caméra. On retrouve là un problème d'appariement. Si ces deux images n'ont rien en commun, il va falloir se servir d'une collection d'images intermédiaires, qui vont définir autant de positions intermédiaires du robot pour arriver au but final.

On retrouve donc un problème de gestion de collections d'images, collections dynamiques pour suivre l'évolution de l'environnement du robot, avec des besoins d'accès rapide pour des reconnaissances. Cette application nous paraît importante car elle offre un élargissement notable des conditions expérimentales d'emploi des techniques d'asservissement visuel : une fois un environnement capté dans une base, le robot peut partir de n'importe quelle position pour aller vers n'importe quelle cible. Si ce genre d'approche présente peu d'intérêt pour un bras articulé pour lequel on peut lire directement les coordonnées articulaires, un véhicule autonome peut en bénéficier en environnement restreint tel un parking. Dans ce cas, les systèmes de positionnement comme les GPS n'offrent pas de précision relative suffisante et ne donnent pas d'information d'orientation.

5 Logiciels

L'équipe développe des logiciels pour tester les divers algorithmes mis au point, ou pour mener des campagne de test et d'évaluation. Mais ces logiciels ne sont pas diffusés ou mis à disposition à l'extérieur de l'équipe pour le moment.

6 Résultats nouveaux

6.1 Recherche dans de grandes bases d'images

6.1.1 Description des images fixes

Participants : Sid-Ahmed Berrani, Patrick Gros, Anthony Remazeilles, François Tonin.

Mots clés : base d'images, reconnaissance d'images, asservissement visuel, indexation d'images, compression d'images.

Notre travail concernant l'indexation et la recherche d'images fixes a suivi cette année deux orientations. La première est liée à la volonté affichée de l'équipe de travailler sur de très grandes bases d'images et en particulier d'évaluer diverses propositions faites dans la littérature dans ce contexte nouveau. Mais pour que de telles évaluations soient possibles, il est tout d'abord nécessaire de pouvoir

[ECR92] B. ESPIAU, F. CHAUMETTE, P. RIVES, « A New Approach to Visual Servoing in Robotics », *IEEE Transactions on Robotics and Automation* 8, 3, juin 1992, p. 313–326.

[Cha98] F. CHAUMETTE, *De la perception à l'action : l'asservissement visuel, de l'action à la perception : la vision active*, Habilitation à diriger des recherches, Université de Rennes 1, janvier 1998.

[MCB99] E. MALIS, F. CHAUMETTE, S. BOUDET, « 2 1/2 D Visual Servoing », *IEEE Transactions on Robotics and Automation* 15, 2, avril 1999, p. 238–250.

les mener en des temps raisonnables. Notre travail s'est donc concentré sur les algorithmes d'indexation et de recherche de faible complexité afin de disposer des moyens permettant de mener à bien des évaluations. Ce travail, relaté ci-dessous, nous ouvre maintenant le chemin vers de nouvelles expériences : vérification du pouvoir de discrimination des divers descripteurs, mesure de l'apport de la couleur, mesure de l'apport et de la sensibilité de certains descripteurs... L'intérêt de mener ces évaluations en grandeur nature est d'éliminer les biais liés à l'utilisation des trop petites bases d'images : avec de très grandes bases, il est peu probable que la base ne contienne que quelques catégories d'images et permette ainsi un accès facile à la sémantique des images à partir de simples critères syntaxiques.

L'autre orientation de nos travaux concerne l'utilisation de l'indexation dans des contextes applicatifs précis. La thèse de A. Remazeilles concerne l'utilisation de l'indexation et de la reconnaissance d'images pour doter les robots d'une mémoire visuelle et d'un moyen de planification de leurs mouvements proche des algorithmes de contrôle de ces mouvements par asservissement visuel. Enfin, la thèse de F. Tonin qui démarre étudie l'emploi des techniques de recherche d'images pour la détection de la copie d'images sur internet.

Couplage perception-action par indexation d'images et asservissement visuel

Ce travail est mené en commun avec le projet VISTA (F. Chaumette).

Nous nous intéressons à la réalisation de très grands déplacements par un système robotique muni d'une caméra embarquée. Ces mouvements sont définis par une image initiale (fournie par la caméra avant que le système robotique ne se déplace) et une image finale (que la caméra doit obtenir une fois le déplacement réalisé) totalement différentes, c'est-à-dire sans aucune primitive en commun. Cette absence de recouvrement rend impossible l'utilisation des techniques actuelles d'asservissement visuel.

Nous avons décomposé le problème en deux phases :

- La localisation du système robotique dans son environnement de navigation. Il est aussi nécessaire de localiser la position que l'on veut atteindre à la fin du déplacement, et de mettre en relation ces deux positions.
- La réalisation du déplacement, en contrôlant celui-ci par asservissement visuel.

Dans la première étape, nous utilisons une base d'images décrivant l'ensemble des images que peut obtenir la caméra du système robotique lors de son déplacement. Nous avons réalisé une application qui détermine tout d'abord les images de la base qui sont les plus proches des images requêtes (initiales et finales), et qui sélectionne ensuite dans cette même base une succession d'images décrivant le chemin à parcourir.

La détermination des images les plus proches de celles requêtes, ou reconnaissance d'images, est réalisée par indexation d'images. Des points d'intérêt sont extraits des images de la base grâce à un détecteur de coins équivalent à celui de Harris. En chaque point d'intérêt, le signal lumineux est caractérisé par un vecteur d'invariants photométriques, valeurs qui restent identiques lorsque l'image subit une transformation (déplacement rigide de la caméra ou changement d'illumination). Ces vecteurs caractéristiques sont les grandeurs indexées, et la comparaison des vecteurs calculés pour les points de l'image requête avec ceux obtenus pour les points des images de la base permet de déterminer les images les plus proches. L'intérêt de l'utilisation des techniques d'indexation d'images est que le temps de traitement est bien plus rapide que celui que fournirait une succession d'appariements image requête-image de la base.

Connaissant les images les plus proches des images requêtes, la définition du chemin à parcourir est obtenue en transposant notre problème en algorithmique des graphes. Chaque couple d'images successives du chemin doit permettre de décrire un asservissement visuel, c'est-à-dire qu'il doit y avoir suffisamment de primitives en commun entre ces deux images pour qu'un asservissement visuel permette de déplacer le système robotique de la première position (liée à la première image) à la deuxième. Un nœud du graphe représente ainsi une image requête ou une image de la base, et ces nœuds sont reliés par un arc dont le poids quantifie la « difficulté » qu'aurait le système robotique à effectuer le déplacement défini par ces deux images. L'algorithme du plus court chemin de Dijkstra, entre les nœuds des images initiale et finale, nous permet d'obtenir une succession d'images décrivant le chemin à parcourir.

Nous nous intéressons actuellement à la réalisation du déplacement défini par cette collection d'images. Nous désirons nous appuyer sur ces images et les primitives mise en correspondances entre celles-ci afin de planifier une trajectoire, permettant toujours d'aller de l'image initiale à l'image cible, mais sans forcément atteindre les positions liées aux images intermédiaires. Cependant, nous devons nous assurer que le système robotique reste dans un environnement connu, c'est-à-dire que des primitives mises en correspondances entre les images du chemin soient toujours visibles sur l'image fournie par la caméra lors du déplacement du système robotique.

Nous implémentons donc actuellement un algorithme de reconstruction, afin de posséder dans un même espace de représentation l'ensemble des primitives détectées dans les images du chemin. Nous effectuons plus précisément une reconstruction projective, afin d'être indépendants des paramètres de la caméra qui a pris les images de la base, et de celle embarquée sur le système robotique. Cette reconstruction projective nous permettra de savoir si une position de la caméra est valide, c'est-à-dire si assez de primitives connues sont visibles dans l'image fournie par la caméra. Pour cela nous effectuerons une projection des primitives 3D projectives sur le plan image correspondant à la position considérée de la caméra. Mais la caméra effectue des déplacements rigides dans l'espace euclidien, et notre reconstruction est elle projective, soit dans un espace plus vaste. Nous travaillons donc à la détermination de la transformation projective associée à un déplacement rigide de la caméra.

Enfin, parallèlement à ces travaux, nous avons implémenté un algorithme de suivi robuste de primitives, à cadence vidéo. Ce suivi nous est utile lors d'un déplacement par asservissement visuel, afin de pouvoir déterminer dans chaque nouvelle image fournie par la caméra où sont les primitives qui définissent l'asservissement en cours. Notre application s'appuie sur les travaux de Shi et Tomasi ; il est considéré que, entre deux images successives fournies par la caméra, le déplacement des primitives est très faible et peut donc être représenté sous la forme d'une translation. On cherche donc la translation permettant de superposer au mieux la fenêtre des pixels au voisinage du point d'intérêt dans l'image précédente avec celle dans l'image courante. L'intérêt de cet algorithme est qu'il détecte et prévient lorsqu'un point ne peut plus être suivi (les deux fenêtres n'ont pu être superposées correctement). L'intérêt est double pour l'asservissement visuel. D'une part, cela évite de persister à suivre un point qui est difficilement localisable, et donc potentiellement source d'échecs de la tâche robotique. D'autre part, cela permet de prendre en compte d'éventuelles occlusions qui peuvent se produire, et de pouvoir continuer la tâche robotique, si, bien-sûr, assez de primitives restent encore visibles.

Étude de l'interaction en indexation, compression et tatouage d'images

Ce travail est mené en commun avec le projet TEMICS (S. Pateux) et fait l'objet de la thèse de F. Tonin qui a débuté au 1^{er} octobre 2002.

Les agences de photos souhaitent pouvoir diffuser leurs images sur le web, en mettant à disposition des imageries. Mais leur souci est de ne pas se faire copier ces imageries par des personnes qui les diffuseraient à leur tour. Ces agences sont donc à la recherche de moyens leur permettant de détecter les fraudes et de prouver leur propriété en cas de conflit.

Plusieurs pistes vont être étudiées : le tatouage d'images est l'une d'elles, mais il ne peut fournir une preuve : il ne fournit qu'une présomption. La vérification finale appartient au juge. D'autres techniques sont envisageables : la stéganographie, par l'ajout de marques invisibles mais fragiles, peut fournir un moyen de montrer qu'un utilisateur a essayé de lessiver les images (*i.e.* d'en enlever les tatouages). Pour notre part, nous allons chercher à étudier la résistance des techniques de reconnaissance par le contenu dans ce contexte.

Pour cela, nous considérons les techniques de reconnaissance par descripteurs locaux qui sont les plus adaptées à ce type de reconnaissance. Le but du travail est d'étudier la robustesse de cette technique aux dégradations que peut subir l'image. Trois types de dégradations sont au programme : la compression des images, le tatouage et le lessivage des images.

Dans le cas de la compression, le problème est double : on étudiera, tout d'abord, l'influence des pertes occasionnées par les algorithmes de compression. Mais on cherchera aussi à calculer les descripteurs directement à partir des images compressées. Cela exige que l'algorithme de compression préserve le contenu et ne broie pas le signal, ce qui est, par exemple, le cas de JPEG'2000.

6.1.2 Description des vidéos

Participants : Patrick Gros, Ewa Kijak.

Mots clés : couplage image – son, structuration de la vidéo, sport, télévision.

Nos travaux en matière de description des vidéos sont menés en étroite collaboration avec les projets VISTA et METISS et la société Thomson où se déroule la majeure partie de la thèse d'E. Kijak.

Sous le terme d'indexation vidéo on trouve des travaux ayant rapport à la structuration de la vidéo (segmentation, macro-segmentation), à la caractérisation des unités ainsi détectées (mouvement de caméra, niveau d'activité) et à la détection d'événements clés. Pour notre part, nous nous intéressons à la structuration à plus haut niveau et au couplage entre son et image.

Pour ce faire, nous considérons un contexte précis, celui des retransmissions sportives. Nous travaillons pour le moment sur le tennis, où existe des règles de montage qui rejoignent celles de jeu : il y a des phases de jeu entrelacées avec des interruptions courtes et des interruptions longues avec par exemple de la publicité. Au-delà de la simple reconnaissance de ces phases, on peut les caractériser, ou les lier en phases de jeu de niveau supérieur (jeu, set...). Le but est de fournir une description de la retransmission qui permette une navigation aisée, et la possibilité de faire des résumés de durée variable, en n'ôtant pas les meilleurs extraits. Une telle possibilité intéresse les journalistes qui doivent faire des résumés très courts pour le journal du soir ou une émission spécialisée, ou le grand public qui voudrait voir en une heure un résumé de tous les matchs de la journée d'un grand tournoi par exemple.

La collaboration entre son et image peut se placer à divers niveaux : l'analyse de la bande son devrait permettre de détecter les frappés de balle, mais aussi des plages de silence, jeu, publicités, applaudissements... L'idée est d'utiliser de telles informations le plus tôt possible dans le processus de structuration de la vidéo.

Dans un premier temps, nous avons travaillé sur une classification et une structuration séparées de l'image d'un côté et du son de l'autre, puis sur la fusion des décisions dans un deuxième temps. Nous étudions maintenant l'intégration des informations le plus tôt possible. Nous utilisons pour cela un formalisme probabiliste (chaîne de Markov cachées), de manière similaire à ce qui est fait de façon classique pour le son dans le projet METISS.

6.1.3 Algorithmes d'indexation et de recherche

Participants : Laurent Amsaleg, Sid-Ahmed Berrani, Patrick Gros.

Mots clés : indexation multidimensionnelle, recherche de plus proches voisins.

Ce travail est mené en collaboration avec la société Thomson.

Nos travaux s'appuient sur les tendances actuelles de la recherche dans le domaine et sur notre expertise des descripteurs : prise en compte de la non-uniformité de données et de la multiplicité des interrogations de la base effectuées pour répondre à une requête, utilisation de schémas d'approximation adaptés aux données, algorithmes de recherche basés sur des simplifications de la recherche séquentielle...

Nous avons déjà commencé à explorer ces différentes pistes^[AGB01], notamment au travers d'une approche faisant le corps de la thèse de S.-A. Berrani. Cette approche est détaillée dans les paragraphes suivants, mais on peut toutefois en résumer le principe ainsi : nous essayons d'abord de partitionner l'ensemble des points en *clusters* (ou partitions) à l'aide d'un algorithme de *clustering*. Puis, la recherche des voisins les plus proches d'un descripteur-requête identifie les *clusters* susceptibles de contenir des descripteurs intéressants. Cette identification est faite de façon *approximative* car la diminution de la précision du résultat entraîne, dans ce cas, une forte diminution du temps de réponse. En plus de ses bonnes performances, notre technique contrôle la précision de la recherche en fonction de la probabilité maximale que l'on a de ne pas retrouver dans le résultat approché un descripteur qui serait présent dans le résultat exact. Sa conception a donné lieu au dépôt d'un brevet avec la société Thomson.

Notre recherche approximative de plus proches voisins présuppose que les vecteurs sont regroupés en *clusters*. Nous avons donc étudié les techniques de *clustering* existantes, mais à la lumière d'une grille d'analyse particulière. En effet, notre but n'est pas de chercher les raisons qui ont entraîné l'apparition des partitions finales, mais de réduire le nombre de points à considérer durant une recherche. Aussi, nous cherchons à obtenir un partitionnement qui ait les propriétés suivantes : forme hypersphérique et compacité des *clusters*, maîtrise du nombre de *clusters*, prise en compte du bruit, de la grande dimension et du volume important des données.

Ces propriétés ont orienté notre choix sur une méthode particulière dénommée Birch^[ZRL96] que nous avons adaptée aux contraintes spécifiques de notre application. Birch est un algorithme de *clustering* hiérarchique qui permet de regrouper les vecteurs d'une base en quatre étapes, dont la première

[AGB01] L. AMSALEG, P. GROS, S. BERRANI, « A Robust Technique to Recognize Objects in Images, and the DB Problems it Raises », in : *Proceedings of the Workshop on Multimedia Information Systems, Capri, Italie*, novembre 2001.

[ZRL96] T. ZHANG, R. RAMAKRISHNAN, M. LIVNY, « BIRCH: an Efficient Data Clustering Method for Very Large Databases », in : *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, p. 103–114, Montreal, Québec, Canada, juin 1996.

sert à construire un arbre hiérarchique de *clusters* en mémoire centrale. Les feuilles de l'arbre correspondent à de petits paquets de vecteurs appelés *micro-clusters*. Lors des phases suivantes, les centres de ces *micro-clusters* sont considérés comme des représentants des vecteurs de la base et sont utilisés pour calculer les centres des *clusters* finaux.

Comme le rôle du *clustering* dans notre contexte se réduit à une mise en petits paquets équilibrés des vecteurs de la base, nous avons utilisé et adapté la première phase de Birch, rejetant les 3 autres. Les principales modifications apportées concernent le calcul du rayon des *micro-clusters* et le mode d'insertion. Nous avons supprimé la hiérarchie car celle-ci introduit des erreurs d'insertion dues au fort chevauchement entre les *clusters*. En sortie, l'algorithme utilisé retourne un ensemble de *clusters* et un ensemble de points isolés. Pour déterminer l'ensemble des vecteurs isolés, nous utilisons la même heuristique que Birch : un *micro-cluster* n'est conservé que s'il contient plus de $\tau\%$ de vecteurs que la moyenne des populations de tous les *clusters*. Sinon, les vecteurs correspondants sont considérés comme des points isolés. τ est appelé le taux de bruit.

À l'issue de la phase de *clustering*, nous mémorisons pour chaque *cluster*, le centre de gravité, le nombre de vecteurs, le rayon exact et la distribution des distances des vecteurs par rapport au centre.

Pour permettre la recherche approximative, chaque *cluster* est ensuite analysé, et, pour chacun, en plus de son hypersphère englobante exacte, des hypersphères englobantes approximatives lui sont associées. Le rayon donné à chaque hypersphère approximative correspond à un niveau de précision prédéterminé. Cette précision est la probabilité maximale que l'on a de ne pas retrouver dans le résultat approché un vecteur qui serait présent dans le résultat exact. Lors de l'interrogation, la recherche considère les hypersphères englobantes approximatives ayant comme rayons ceux correspondant au niveau d'imprécision fourni par l'utilisateur. Prendre ces rayons en compte plutôt que les rayons des hypersphères englobantes exactes peut faire que certains *clusters* qui seraient analysés par la recherche exacte sont ignorés par la recherche approximative. Il est ainsi possible que les vrais plus proches voisins du point requête soient ignorés, mais notre approche permet de contrôler de façon probabiliste les chances pour que ceci ait lieu.

L'analyse des performances de notre technique montre, par exemple, que pour $1,5 \cdot 10^6$ vecteurs ayant 24 dimensions, la recherche approchée est jusqu'à 5 fois plus rapide qu'une recherche séquentielle, même lorsque la probabilité d'ignorer un des plus proches voisins exacts du point requête est inférieure à 0,01.

6.2 Recherche dans de grandes bases de textes

6.2.1 Traitement automatique des langues et apprentissage

Participants : Vincent Claveau, Mathias Rossignol, Pascale Sébillot.

Mots clés : traitement automatique des langues, apprentissage automatique, sémantique lexicale, acquisition de relations sémantiques en corpus, programmation logique inductive, classification hiérarchique.

Résumé : *Nos travaux portent sur le développement de méthodes d'apprentissage automatique sur corpus textuels de relations lexicales sémantiques permettant d'enrichir la description de noms, dans une double optique de désambiguïsation et de traitement de variantes sémantiques intra- et intercatégorielles, susceptibles d'être utilisées au sein*

d'applications d'accès au contenu de documents (recherche d'information, filtrage...). Des théories linguistiques nous servent de cadres pour déterminer les relations lexicales pertinentes, valider ce qui est acquis, voire mettre au point la méthode d'apprentissage nécessaire à l'acquisition. Nous nous intéressons plus particulièrement à deux familles de liens. D'une part, en nous positionnant dans le cadre de la sémantique différentielle de F. Rastier, nous cherchons à apprendre, par des méthodes statistiques (en particulier de classification ascendante hiérarchique), des liens intracatégoriels (synonymie..., mais aussi d'autres liens plus « fins »); dans le cadre d'applications de type recherche d'information, l'acquisition de ces liens vise à permettre l'appariement d'une requête contenant le mot automobile à un texte indexé par le terme voiture. D'autre part, en contrôlant leur pertinence grâce au formalisme du lexique génératif de J. Pustejovsky, nous acquérons par de l'apprentissage symbolique de type programmation logique inductive des liens transcatégoriels nomino-verbaux; en recherche d'information, ces liens conduisent également à des reformulations intéressantes de termes d'indexation nominaux – par exemple, la reformulation de jaugeur de carburant en mesurer du carburant sur la base de la relation téléique (but, fonction) liant jaugeur à mesurer – et permettent également de les désambiguïser.

Apprentissage de relations nomino-verbales basées sur le lexique génératif.

Le but de notre recherche est d'apprendre, sur un corpus étiqueté catégoriellement et sémantiquement, ce qui distingue les couples nom-verbe (N-V) liés par un rôle qualia des autres couples. Pour ce faire, des exemples positifs (E^+) et négatifs (E^-), constitués à l'aide des contextes d'apparition des N et V dans des phrases du corpus, sont générés et fournis en entrée d'Aleph, mise en œuvre de la programmation logique inductive (PLI) développée par Srinivasan, qui produit des hypothèses (clauses) par généralisation de certains E^+ .

Au cours de l'année 2002, nos travaux sur l'acquisition en corpus de couples N-V dont les constituants sont liés par un des rôles définis dans la structure des qualia du lexique génératif ont porté sur 4 points.

1. **Consolidation de la méthode d'apprentissage et de son évaluation.** Concernant l'apprentissage proprement dit, nous avons optimisé l'opérateur de raffinement bien adapté aux connaissances hiérarchisées que nous manipulons, afin de parcourir de manière efficace le treillis d'hypothèses organisé selon un ordre de généralité que nous avons précisé (dérivé de la θ -subsumption sous identité d'objet), et de produire des règles linguistiquement motivées et bien formées [13]. L'apprentissage, validé théoriquement à l'aide d'une méthode de validation croisée (*10-fold cross-validation*), a été mené sur un corpus technique et a permis d'obtenir 9 règles générales, qui, appliquées sur le corpus, permettent d'acquérir des paires N-V qualia. Nous avons validé empiriquement l'apprentissage réalisé en comparant les décisions des règles (caractère qualia ou non de couples N-V qu'elles extraient) à celles d'experts. Nous avons aussi étudié la pertinence linguistique des règles apprises. Il apparaît que les clauses donnent des indices de surface très généraux sur les structures qui, dans le corpus, favorisent l'expression de liens qualia, et sont suffisants pour donner accès à certains patrons spécifiques du corpus [17].
2. **Comparaison à des méthodes statistiques.** Nous avons comparé les performances de notre système fondé sur la PLI à celles obtenues à l'aide de 10 mesures statistiques connues (infor-

mation mutuelle au cube, χ^2 ...) permettant d'extraire des cooccurrences de noms et verbes, en testant le caractère qualia ou non des paires détectées par ces dernières. Seules quelques mesures statistiques donnent des résultats permettant de les utiliser dans une tâche d'extraction de paires qualia, sans toutefois atteindre la qualité globale des scores obtenus par notre méthode d'apprentissage de PLI.

3. **Insertion dans un système de recherche d'information.** Une des utilisations que nous visons pour les couples N-V qualia, acquis par application des règles apprises sur un corpus, est la reformulation de requêtes au sein d'un système de recherche d'information (SRI). Nous débutons actuellement l'insertion de mécanismes de prise en compte de paires N-V qualia dans un SRI pour évaluer de manière systématique leur apport sur les données de la seconde campagne AMARYLLIS (campagne d'évaluation des systèmes de recherche d'information en français). Nous présentons également en 6.3 d'autres modes d'évaluation des apports de ressources linguistiques aux SRI que nous avons commencé à mettre en place, pour ne pas nous restreindre à une évaluation dans des cadres présupposant une liste de réponses à découvrir et ne prenant pas en compte la pertinence des documents retournés pour un utilisateur effectif.
4. **Réduction du coût.** Nous avons également poursuivi nos efforts, débutés en 2001, visant à réduire le coût des tâches manuelles à effectuer pour mettre en place notre technique d'apprentissage supervisé. Nous avons, en particulier, testé, lors du stage de DEA d'E. Catz [25], la méthode de *co-training* proposée par Blum et Mitchell, qui permet d'utiliser deux méthodes d'apprentissage supervisé ensemble en limitant le nombre d'exemples étiquetés comme positifs ou négatifs en entrée. Nous avons ainsi implémenté l'algorithme qu'ils proposent, en l'appliquant, dans un premier temps, à un problème bien cerné : l'étiquetage catégoriel de corpus.

Par ailleurs, dans le cadre de l'action Bioinformatique Caderige (Catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques) (cf. 8.2), dont l'objectif global est d'extraire automatiquement, à partir de textes de bioinformatique, les zones de textes décrivant des interactions entre gènes et de modéliser l'interaction décrite, nous avons appliqué notre méthodologie d'apprentissage par PLI, pour tenter d'apprendre ce qui distingue les phrases contenant des interactions des autres. Ce travail, réalisé par P. Tchienehom dans le cadre de son stage de DEA [26] co-encadré par J. Nicolas du projet SYMBIOSE, a été effectué sur un corpus de 2209 résumés extraits de Medline, et nous avons extrait des règles permettant de caractériser les zones de textes pertinentes.

Apprentissage de relations intracatégorielles basées sur la sémantique différentielle.

Pour ce qui est de l'apprentissage en corpus de lexiques sémantiques basés sur la sémantique différentielle sans connaissances *a priori*, l'essentiel de nos efforts a porté cette année sur l'amélioration de la détection et caractérisation des thèmes présents dans notre corpus d'étude (archives de 1984 à 1998 du mensuel *le Monde diplomatique* - 11,4 millions de mots). La scission d'un corpus en sous-corpus thématiquement homogènes constitue en effet la première phase de notre méthode d'acquisition de liens N-N, puisque c'est au sein d'un domaine donné qu'il est possible de former des classes sémantiquement homogènes et de les structurer à l'aide de sèmes (traits sémantiques) dénotant les dissimilarités de sens entre leurs éléments. Dans notre travail, les thèmes sont caractérisés à l'aide de listes de mots automatiquement induites à partir du corpus d'apprentissage grâce à une classification hiérarchique fondée sur la similarité des distributions des mots dans les différents paragraphes du corpus. Ces listes de mots « symptomatiques » de thèmes sont ensuite utilisées pour scinder le corpus initial,

en utilisant la coprésence d'un certain nombre de leurs membres dans un paragraphe pour affecter un thème à celui-ci.

Nous avons proposé une algorithmique de remise en cause partielle des associations effectuées par l'algorithme de classification, en définissant une mesure de qualité des classes à l'aide d'une seconde classification fondée sur la cohésion lexicale des paragraphes. Plusieurs exécutions de la méthodologie, effectuées sur des ensembles de paragraphes tirés aléatoirement, conduisent à l'obtention de listes de mots-clés symptomatiques de thèmes quelque peu différentes. Nous avons regroupé ces listes au sein d'un graphe valué, dont les nœuds sont les N présents dans au moins une liste, et dont le poids d'un arc entre deux N correspond au nombre de fois où ces deux N ont été réunis dans une classe. Nous extrayons des composantes connexes de ce graphe et étendons ces noyaux par les mots dont la fréquence dans les paragraphes qu'ils reconnaissent (coprésence de certains mots de la liste-noyau) est particulièrement élevée par rapport à leur fréquence moyenne sur le corpus. Nous obtenons, sur le corpus du *Monde diplomatique*, une quarantaine de classes de 25-30 mots, toutes très satisfaisantes en termes de consistance thématique. En appliquant un critère de coprésence d'au moins trois mots d'une liste caractérisant un thème dans un paragraphe pour affecter le thème à ce segment textuel, et en prenant en compte la structuration en articles du corpus, nous répartissons une grande proportion du corpus initial en sous-corpus thématiquement homogènes avec une précision supérieure à 85%.

6.2.2 Extraction et visualisation de connaissances à partir de corpus textuels

Participants : Annie Morin, Rodolphe Priam.

Mots clés : Analyse exploratoire des données.

L'extraction de connaissances dans un ensemble de données textuelles n'est pas triviale. L'état de l'art utilise des méthodes d'analyse factorielle ou encore des méthodes issues des modèles neuronaux comme les cartes de Kohonen (SOM). Nos travaux ont en particulier porté sur l'utilisation plus efficace des techniques factorielles [20, 19] et sur l'élaboration de variantes des cartes de Kohonen. La thèse de R. Priam propose une adaptation des cartes de Kohonen à des données discrètes et au cas de volumes de données importants. L'objectif final est de visualiser l'information extraite de corpus textuels [22]. Pour cela, R. Priam applique l'algorithme du TPÉM (Topology Preserving EM) introduit en 1996 par C. Ambroise et G. Govaert, qui permet une cartographie associative de distributions gaussiennes, au cas d'un mélange de multinomiales. Le nouvel algorithme, appelé CASOM, s'avère être une extension non-linéaire de l'Analyse Factorielle des Correspondances et se caractérise par des propriétés prometteuses. Ces travaux ont fait l'objet d'une présentation au cours d'une conférence [15]. Un autre aspect des travaux de R. Priam consiste à utiliser des méthodes classiques de segmentation d'images comme les champs aléatoires de Markov dans le cadre des données textuelles [21].

6.3 Méta-données et étude des usages

Participants : Laure Berti-Équille, Vincent Claveau, Anicet Kouomou-Choupo, annie Morin, Pascale Sébillot.

Mots clés : Méta-donnée, usage.

Glossaire :

données géomédias : données multimédias relatives aux aspects culturels, économiques, politiques, historiques et architecturaux d'un lieu localisé sur une carte.

Résumé : *Étudier les critères de sélection des méta-données, proposer des techniques de génération automatique de méta-données et des algorithmes de sélection pour le précalcul (total ou partiel) des méta-données jugées a priori les plus adaptées selon une requête, un utilisateur ou une application multimédia sont les objectifs de cet axe de recherche. Sur ce sujet, plusieurs collaborations ont débutées cette année, notamment avec le Software Research Division du NII Tokyo ainsi qu'une thèse en fouille multimédia sur l'usage et la qualité des métadonnées, débutée en novembre 2002 par A. Kouomou Choupo sous la direction de L. Berti-Équille et A. Morin.*

6.3.1 Dimension qualité

Le travail qui a été initié cette année par L. Berti-Équille avec le National Institute of Informatics (NII, Prof. Andres) de Tokyo et l'UNESCO (Information Society Division, CI Sector (CI/INF), Dr. Kim) dans le cadre du projet M4 (Management of Metadata and MultiMedia data) se focalise sur l'exploitation des méta-données en tant que phase préalable à la recherche, à l'utilisation ou à l'analyse des données géomédias en masse (pour des données fournies par plusieurs producteurs d'information puis, intégrées, notamment les données de la Route de la Soie, cf. colloque ADTCARA'02, Bakou, Azerbaïdjan). Dans ce contexte, l'évaluation et le contrôle de la qualité de grands volumes de données multimédias est un axe de recherche important pour le projet M4 et il consiste à proposer des techniques de génération automatique de méta-données décrivant la qualité des données multimédias et détectant les problèmes de cohérence et de synchronisation entre données et méta-données.

Le projet M4 inclut une phase de prototypage et d'évaluation : d'une part, il concerne l'adaptation de l'outil d'extraction de méta-données GEOMEX développé au NII. GEOMEX est un générateur/optimizeur de plans d'extraction de méta-données. Un tel plan se compose de plusieurs méthodes d'extraction de méta-données. Le traitement dans GEOMEX se base sur la connaissance de ces méthodes : type de données en entrée, type de données en sortie, profil qualité/coût associé à la méthode. D'autre part, le projet se base sur un moteur de gestion de méta-données déjà existant au NII assurant le stockage et la manipulation des méta-données.

Les méta-données peuvent avoir de multiples finalités : améliorer l'accès aux données, permettre une classification, fournir un résumé, permettre l'interopérabilité des données, offrir une représentation du média ou de sa structure, la traçabilité... La manière dont elles sont générées a une importance évidente sur le choix *a posteriori* des données auxquelles peuvent être rattachées ces méta-données. Notre travail consiste dans le projet M4 à répertorier toutes les méta-données relatives aux données géomédias et compléter leur classification pour chaque type de média en ajoutant des aspects jusqu'alors peu abordés dans la littérature : les méta-données relatives à la qualité d'une donnée géomédia.

6.3.2 Dimension usage

Afin d'accroître les possibilités d'appariement entre les requêtes d'utilisateurs et les documents indexés et augmenter leurs performances (rappel, précision), les systèmes de recherche d'information (SRI) se tournent peu à peu vers l'intégration d'informations linguistiques, qu'elles soient de nature terminologique, morphologique, morpho-syntaxique, syntaxique ou sémantique (voire conceptuelle). Des bases de connaissances lexicales, génériques ou spécifiques à un domaine, préexistantes, construites manuellement ou acquises automatiquement sur corpus textuels, sont par exemple utilisées pour apparier un mot d'une requête et son synonyme présent dans la représentation d'un document. Cependant, à notre connaissance, aucune méthodologie d'évaluation adéquate de l'intérêt effectif de telles extensions n'a réellement été proposée, la plupart supposant par exemple la connaissance a priori des documents répondant "effectivement" à une requête, indépendamment de la notion d'intérêt qu'un usager réel attribuerait à ces documents ou à un des autres documents de la base interrogée.

Le projet initié par L. Berti-Équille, V. Claveau et P. Sébillot a pour objectif de mettre au point une méthodologie d'évaluation des apports linguistiques pour des moteurs de recherche, prenant en compte l'utilisateur lors de sa session interactive de recherche. Cette méthodologie d'évaluation est :

- 1- basée sur la satisfaction de l'utilisateur et son usage du SRI étendu,
- 2- générique pour être adaptable à des SRI portant sur un domaine précis ou bien des moteurs de recherche du Web,
- 3- capable de gérer des extensions de types divers.

La satisfaction d'un utilisateur est inférée à l'aide d'une technique d'apprentissage automatique s'appuyant sur les comportements d'utilisateurs « cobayes » lors d'un scénario de recherche. La correspondance entre la satisfaction d'utilisateurs et différents indicateurs (métriques) de leurs comportements d'interrogation et de navigation, d'une part lorsque les requêtes sont automatiquement étendues par les informations linguistiques et, d'autre part, lorsqu'elles ne le sont pas, est déterminée par apprentissage. Les modèles « utilisateurs satisfaits » appris servent ensuite à mesurer automatiquement la satisfaction de l'utilisateur et de manière non-intrusive (c'est-à-dire sans l'interroger). Plus précisément, ceci signifie qu'après l'apprentissage de ces modèles, il est possible, lors de requêtes ultérieures sur un moteur de recherche bénéficiant d'une extension linguistique donnée, en fonction du comportement de l'utilisateur, de regarder si les documents retournés possèdent ou non les caractéristiques du modèle de satisfaction et si l'extension amène donc ou non un apport réel au moteur. Les résultats attendus de ce projet sont à la fois une méthodologie d'évaluation des moteurs de recherches en situation réelle, et également des mesures de l'efficacité d'extensions linguistiques comme assistance à l'interrogation qui seront effectivement testées pour valider la méthodologie.

Ce projet a fait l'objet d'un stage d'été (Vincent Marqueton, DIIC2) et d'un projet de DIIC3 en cours (septembre 2002 à janvier 2003).

6.3.3 Visualisation et web mining

Participants : Nicolas Bonnel, Annie Morin.

Dans le cadre du web mining et de la visualisation des informations, Nicolas Bonnel commence une thèse CIFRE en collaboration avec France Télécom R & D sur la génération dynamique de présentations interactives en multimédia 2D et 3D, de données, pour des applications en ligne. En d'autres termes, il est nécessaire d'étudier des méthodes, des algorithmes et des langages :

- pour déterminer rapidement le contenu à visualiser,
- pour exprimer efficacement les métaphores de présentation 3D et multimédia,
- pour les lier aux sources de données pour former une présentation 3D et multimédia,
- pour introduire des outils « intelligents » permettant d'aider de façon efficace l'utilisateur dans sa tâche,
- pour permettre un travail « collaboratif » avec d'autres utilisateurs.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Contrats industriels

7.1.1 Contrat Thalès Communications : analyse des caractéristiques d'un auditoire en vue de la conception d'un logiciel d'argumentation - Génération de lieux selon le type d'argument considéré

Participant : Pascale Sébillot.

Mots clés : argumentation, auditoire.

Contrat avec Thalès Communications Colombes de 8 mois débuté en mars 2002, réalisé en collaboration avec P. Besnard (IRIT).

Un énoncé qui établit une conclusion sur la base de diverses hypothèses est un argument quand le destinataire du propos tend à être favorablement disposé vis-à-vis des hypothèses. Cet aspect de l'argumentation est qualifié d'accord avec l'auditoire. La recherche vise à produire des lieux, prémisses d'ordre très général permettant de fonder des valeurs ou hiérarchies ou de renforcer l'intensité de l'adhésion qu'elles suscitent, qui constituent l'un des meilleurs moyens de réaliser l'accord avec l'auditoire. Disposant des valeurs de l'auditoire (y compris sous forme de hiérarchies), il s'agit d'énoncer des lieux qui reflètent ces valeurs (confortant la relation entre l'auditoire et la personne qui argumente) et surtout préparent les arguments en mettant en lumière les valeurs sur lesquelles les arguments vont s'appuyer.

7.2 Contrats dans le cadre des réseaux nationaux de recherche technologique

7.2.1 Projet PRIAM Médiaworks

Participant : Patrick Gros.

Mots clés : archives télévisuelles, bases de vidéos.

Ce projet est mené conjointement avec le projet VISTA. Durée 36 mois, début septembre 2000. Participants : LIMSI – CNRS, AEGIS, INRIA (projet IMEDIA), TF1.

Le projet Médiaworks est un projet labellisé par le programme Priamm et le programme Société de l'information, financé par le Ministère de l'industrie, et qui a débuté au 1^{er} septembre 2000. Ce projet regroupe TF1, le LIMSI, la société AEGIS, l'INRIA (projets IMEDIA de l'INRIA Rocquencourt et VISTA), et traite de systèmes d'aide à l'indexation pour des documentalistes. Ses éléments principaux en sont la coopération entre les médias texte et image, et la mise au point d'un moteur de recherche

sémantique. TEXMEX est associé aux travaux du projet VISTA sur les outils de structuration automatique en plans et de représentation iconique de ces plans, ainsi que des descripteurs de niveaux d'activités au sein des plans.

7.2.2 Projet RNRT Diphonet : Diffusion de photos par Internet

Participants : Laurent Amsaleg, Patrick Gros.

Mots clés : bases d'images, reconnaissance d'images, tatouage, copyright.

Durée : 30 mois, début en janvier 2002. Participants : IRISA, Canon, L2S, Andia Presse, INRIA (projets CODES, TEMICS et TEXMEX).

Le but premier du projet est la conception du système d'indexation et de recherche destiné à identifier les images fixes piratées.

De l'avis des professionnels de la photographie, le problème principal lié à la mise à disposition de photos au travers de réseaux ouverts est le pillage illicite de leur patrimoine. Ce problème, exprimé dans un contexte informatique, demande de mettre en place un environnement permettant d'une part de détecter qu'une image fait illégalement partie d'une collection d'images mises à disposition par un tiers (traçage de copies) et d'autre part d'exhiber une preuve de propriété irréfutable (authentification). Pour atteindre ces deux objectifs, Diphonet propose de combiner les techniques de tatouage et d'indexation par le contenu. Cette combinaison, inédite, permet de couvrir les deux aspects que sont le traçage et l'authentification, mais aussi de renforcer la puissance et la robustesse de chaque technique prise individuellement car l'une et l'autre s'épaulent dans le processus de recherche et de vérification. De façon sommaire, nous abordons les problèmes de traçage via des recherches intelligentes au sein de bases d'images tierces d'images similaires à celles que l'on possède légalement et que l'on suppose avoir été volées. Les preuves de propriété sont abordées au travers de l'utilisation conjointe de techniques de tatouage et de protocoles cryptographiques.

Au cours de cette première année, nous avons rédigé une étude bibliographique sur les techniques de description d'images et sur les techniques d'indexation mises en œuvre dans les bases de données.

7.3 Contrats avec l'Union européenne

7.3.1 Projet européen IST BUSMAN : Bringing User Satisfaction to Media Access Networks

Participant : Laurent Amsaleg.

Mots clés : vidéo, serveur de vidéos, indexation de vidéos.

Durée : 30 mois. Participants : IRISA (projet TEMICS et TEXMEX), Motorola, Telefonica, Technical University Munich, Queen Mary University of London, BTextact Technologies, Heinrich-Hertz Institute Berlin, FramePOOL.

Le projet porte sur le développement de solutions d'indexation et de marquage des contenus vidéo pour la création de services multimédias à valeur ajoutée. Nos contributions portent sur la participation à la spécification du système d'indexation et de recherche au sein de bases de vidéos.

8 Actions régionales, nationales et internationales

8.1 Actions régionales

- L. Berti-Équille participe à une collaboration avec l'INSERM U522 sur la problématique de gestion et de nettoyage des données issues des expériences transcriptome hépatique.
- L. Berti-Équille participe à une collaboration avec l'INRA Génétique Animale sur la problématique de gestion et de partage des données sur les interactions entre gènes.
- A. Morin a participé à un contrat avec le Laboratoire de Biomatériaux en sites osseux de l'université de Rennes 1, visant à une étude statistique sur les effets de deux instruments sur les surfaces radiculaires. La procédure utilisée était une ANOVA à 1 facteur et nous a permis de tester l'hypothèse d'égalité des moyennes.

8.2 Actions nationales

8.2.1 ACI santé Neurobase

Participants : Laurent Amsaleg, Patrick Gros.

Mots clés : neuro-imagerie, imagerie médicale.

Ce projet est mené conjointement avec les projets VISTA et EPIDAURE.

Nous participons à l'action concertée incitative (ACI) du ministère de la recherche intitulée « NeuroBase : Système de gestion de données et de connaissances réparties en neuro-imagerie ». L'objectif est de faire coopérer, au travers de l'Internet, des bases d'informations en neuro-imagerie situées dans différents centres d'expérimentation, cliniques neurologiques ou établissements de recherche en neurosciences cognitives. Ce projet consistera à spécifier la façon de relier de telles bases et d'y accéder efficacement par la définition d'une architecture informatique permettant le partage de résultats d'expérimentations ou bien encore de méthodes de traitement des données au sein d'un même site ou entre sites différents. Cela permettra par exemple la recherche de résultats similaires, la recherche d'images contenant des singularités, ou encore des recherches transversales de type « fouille de données » pour mettre en évidence d'éventuelles régularités.

Ce projet, référencé sous le numéro TS 2001/23, a pour partenaires les projets VISTA (responsable de l'ACI) et TEXMEX de l'IRISA, l'IFR 49 « Neuroimagerie Fonctionnelle » (CEA SHFJ, INSERM U494, CHR Pitié Salpêtrière, Paris), le projet CARAVEL de l'INRIA Rocquencourt et le projet EPIDAURE de l'INRIA Sophia-Antipolis, l'unité INSERM U438 (Grenoble), l'équipe IDM de la Faculté de Médecine de Rennes 1, le laboratoire TIMC (Grenoble).

8.2.2 ACI Grid GénoGRID

Participant : Laurent Amsaleg.

Mots clés : architecture reconfigurable, FPGA, génomique.

Cette action est menée conjointement avec le projet SYMBIOSE et les actions ADEPT et R2D2.

Il s'agit d'une action concertée incitative nationale dans le cadre de l'appel d'offre Globalisation des Ressources Informatiques et des données (ACI GRID) lancé par le ministère de la recherche, dont

D. Lavenier est le coordonnateur. Cette ACI a pour objectif de mettre en place un portail par lequel des chercheurs en biologie peuvent accéder à des ressources de calcul réparties géographiquement. Les partenaires incluent divers laboratoires du Grand Ouest (INRA, IFREMER, Station Biologique de Roscoff), l'ABISS à Rouen, le LAMIH à Valenciennes, le LIH au Havre et le LIFL à Lille.

Notre équipe participe à la définition de l'architecture d'une machine spécialisée devant prendre part à la future grille.

8.2.3 Action Bio-info inter-EPST architecture parallèle et reconfigurable pour l'extraction de données génomiques

Participant : Laurent Amsaleg.

Mots clés : architecture reconfigurable, FPGA, génomique.

Cette action est menée conjointement avec le projet SYMBIOSE et l'action R2D2.

Ce contrat s'effectue dans le cadre de l'action inter-EPST Bioinformatique. Il concerne la mise au point d'architectures parallèles reconfigurables pour l'extraction des données génomiques. Le projet vise l'extraction rapide, et par le contenu, des données génomiques emmagasinées dans les banques et les bases de données. *Par le contenu* indique que la recherche porte sur l'information brute, et non sur les annotations pouvant y faire référence. Par exemple, extraire des séquences sur la base d'un alignement significatif, ou sur la base d'un motif exprimé à l'aide d'une expression régulière, est une tâche qui s'effectue essentiellement sur le texte des séquences, et non sur les annotations.

Notre équipe est impliquée dans la définition de l'architecture de la machine spécialisée reconfigurable.

8.2.4 Action Bio-info inter-EPST Caderige-2 : catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques

Participant : Pascale Sébillot.

Mots clés : Extraction d'information, bioinformatique, génomique.

Action inter-EPST Bio-informatique CNRS, INSERM, INRA, INRIA, Ministère de la Recherche regroupant, outre ceux des participants à TEXMEX, des membres de Symbiose, des laboratoires Leibniz de l'IMAG, du LIPN, du LRI, et de deux laboratoires INRA : MIG et INRA-ENSAR. Après un pré-contrat de 1 an obtenu en 2000, cette action a été renouvelée pour 2 ans en octobre 2001. Son objectif est de filtrer, dans des bases textuelles de bioinformatique telles que MedLine, les textes parlant spécifiquement d'interactions géniques, et d'extraire de ces textes des réseaux de telles interactions. La participation de notre équipe concerne la détection des zones de textes susceptibles de contenir une interaction.

8.2.5 Action nationale INRIA de R & D SYNTAX

Participante : Pascale Sébillot.

Mots clés : gestion de documents.

Cette action sur la recherche d'information dans les documents électroniques est en phase de mise en place depuis l'été 2002 et a pour coordonnateur L. Romary (INRIA Lorraine). Son but est de construire une plate-forme d'intégration et de gestion de documents en collaboration avec des industriels, en développant des outils à partir des résultats de recherche acquis dans différents projets INRIA.

8.2.6 Action JemSTIC TEXMEX

Participants : Laurent Amsaleg, Laure Berti-Équille, Patrick Gros.

Ce programme du département STIC du CNRS a pour but d'aider la création de nouvelles équipes par de jeunes chercheurs. Ce programme a accordé un financement de 13 k€ TTC pour 1 an pour soutenir le lancement de l'équipe TEXMEX.

8.2.7 ACI jeunes chercheurs TEXMEX

Participants : Laurent Amsaleg, Laure Berti-Équille, Patrick Gros.

Ce programme du ministère de la recherche a pour but d'aider et d'accompagner la création de nouvelles équipes par de jeunes chercheurs. Ce programme a accordé un financement de 103 k€ pour 3 ans pour soutenir le lancement de l'équipe TEXMEX.

8.2.8 Participation à des groupes de travail nationaux

- L. Berti-Équille participe au groupe GafQualité de l'AS GafDonnées du département STIC du CNRS.
- L. Berti-Équille participe aux groupes de travail Documents Multimédia et Médiation du GDR I3.
- L. Berti-Équille participe au groupe de travail de l'AS « Médiation d'informations via les méta-données » dans le cadre du RTP9 du département STIC du CNRS.
- P. Gros est membre des comités de pilotage des RTP 25 (Vision par ordinateur) et 33 (Documents et contenus : création, indexation, navigation) du département STIC du CNRS.
- P. Gros et A. Remazeilles participent à l'AS transmodalité du département STIC du CNRS.
- P. Gros participe à l'AS fouille d'images du département STIC du CNRS.
- P. Sébillot est membre du réseau thématique Information et connaissance : « découvrir et résumer » du CNRS, coordonné par P. Gallinari et A. Napoli.
- P. Sébillot est membre de l'action spécifique web sémantique du STIC-CNRS, coordonnée par J. Charlet, P. Laublet et C. Reynaud.
- P. Sébillot est membre du collège de l'AFIA Café (Collège apprentissage, fouille et extraction), animé par M. Sebag.

- P. Sébillot est membre du groupe de travail A3CTE : Application, Apprentissage, Acquisition de Connaissances à partir de Textes Électroniques du GdR-PRC I3, coordonné par A. Nazarenko et C. Nédellec.

8.3 Collaborations internationales

8.3.1 Groupe de travail Image Understanding d'ERCIM

Participant : Patrick Gros.

Participant également à ce groupe les projets VISTA, IMEDIA et ARIANA.

Ce groupe de travail a pour but de fédérer les activités en analyse et compréhension des images et de la vidéo des membres du consortium européen ERCIM. Son action a principalement consisté à répondre à l'appel à manifestation d'intérêt de l'Union européenne pour les réseaux d'excellence.

8.3.2 Collaboration avec le NII au Japon

Participante : Laure Berti-Équille.

Mots clés : méta-données, culture, géomédia, géographie.

Un MOU (Memorandum Of Understanding) a été signé entre l'IRISA (équipe TEXMEX) et le NII - National Institute of Informatics - de Tokyo (Japon) pour encadrer la collaboration et initier le projet M4 (Gestion des méta-données et données multimédias). Ce projet de recherche commun concerne la définition, la génération et le choix des méta-données les plus pertinentes pour gérer au mieux les données multimédias disponibles en très grands volumes dans le domaine culturel et géomédia. La problématique de la gestion de grands ensembles de documents multimédias incluant la dimension géographique, sociale, politique ou culturelle (tels que, par exemple, tous les documents relatifs à la route digitale de la Soie - Digital Silk Road) est au centre de cette collaboration. Des accords entre l'IRISA (TEXMEX) et l'UNESCO sur ce thème sont également à l'étude.

8.3.3 Collaboration bilatérale avec l'Islande

Participant : Laurent Amsaleg.

Mots clés : gestion de la mémoire, caches, disques.

L. Amsaleg a mis en place, avec le concours des relations internationales de l'INRIA Rocquencourt, une coopération franco-islandaise (Université de Reykjavik-IRISA). Un échange de chercheurs a déjà eu lieu début 2002. Cette coopération a pour but d'explorer les différents aspects systèmes (buffers, accès aux disques, . . .) des systèmes de recherche par le contenu.

Dans le cadre de cette collaboration, nous avons accueilli Björn Jónsson qui est *Associate Professor* à l'université de Reykjavík en Islande.

9 Diffusion de résultats

9.1 Organisation de conférences, workshops, séminaires

JADT'2002 A. Morin et P. Sébillot ont organisé (avec M. Kerbaol de l'INSERM) les sixièmes journées internationales d'analyse de données textuelles (JADT'2002) du 13 au 15 mars 2002 à Saint-Malo. Cette manifestation a lieu tous les deux ans et rassemble des chercheurs travaillant sur les traitements automatiques et statistiques de données textuelles. Il y a eu 177 participants et 70 communications.

JOBIM'2002 L. Berti-Équille a participé avec le projet Symbiose à l'organisation du colloque JOBIM'02 du 10 au 12 juin 2002 à Saint-Malo.

9.2 Animation de la communauté scientifique

- L. Amsaleg a fait partie du comité scientifique de BDA 2002 : 18^{es} Journées Bases de Données Avancées, Evry, octobre 2002.
- L. Amsaleg a fait partie du comité scientifique de MIR 2002 : *4th Intl Workshop on Multimedia Information Retrieval*, Juan-les-Pins, décembre 2002 (*workshop* satellite de *ACM Multimedia 2002*).
- L. Amsaleg fait partie du comité éditorial de la revue ISI : Ingénierie des Systèmes d'Information, numéro spécial « Bases de Données et Multimédia ». 2002.
- L. Berti-Équille est membre du conseil d'administration de l'association SPECIF.
- L. Berti-Équille a fait partie du comité de programme pour un numéro spécial Recherche d'information et filtrage d'information (volume 7, n° 1-2/2002) de la revue Ingénierie des systèmes d'information.
- L. Berti-Équille fait partie du comité de programme du congrès INFORSID 2003.
- L. Berti-Équille fait partie du comité de programme du SEMSOFT 2003.
- P. Gros est membre du comité éditorial de la revue « Traitement du signal ».
- P. Sébillot est membre du comité éditorial de la revue In Cognito.
- P. Sébillot est membre du comité éditorial du Jedai (Journal Électronique d'Intelligence Artificielle).
- P. Sébillot a été membre du comité de programme de CIFT'02 (Colloque international sur la fouille de texte), Hammamet, Tunisie, octobre 2002.
- P. Sébillot a été membre du comité de programme de JADT'2002 (Journées internationales d'analyse statistique des données textuelles), Saint-Malo, mars 2002.

9.3 Enseignement universitaire

- DESS Mitic, Ifsic, Rennes 1. L. Amsaleg, P. Gros et P. Sébillot : indexation des documents numériques. L. Berti-Équille : gestion des données.
- ENSTA Paris, 3^e année. P. Gros : géométrie projective et vision 3D.
- ANVAR, formation des chargés d'affaires. P. Gros. Indexation par le contenu. Quels éléments d'appréciation ?
- En octobre 2002, P. Sébillot a effectué un cours de 3h00 sur le thème « Traitement automatique des langues et recherche d'information » dans le cadre d'un cours INRIA intitulé « La

recherche d'information sur les réseaux » qui s'adresse aux professionnels de l'information, bibliothécaires, documentalistes et archivistes (cf. [16]).

- DEA d'Informatique à l'Université de Yaoundé I (Cameroun), L. Berti-Équille : Bases de Données Avancées.
- INSA de Rennes, 5^e année. L. Berti-Équille : bioinformatique - gestion des données biologiques.
- Formation continue à l'INRA et au CNRS pour des biologistes (chercheurs et techniciens), L. Berti-Équille.
- ENST Bretagne, 3^e année. L. Berti-Équille : entrepôts et fouille de données.
- IFSIC, journées pédagogiques à destination des enseignants-chercheurs. L. Berti-Équille : technologies XML, méta-données, XLink et XML Query.

9.4 Participations à des jurys

P. Gros a participé, en tant que rapporteur, au jury de thèse de Jérôme Fournier (Traitement du signal, université de Cergy Pontoise, 31 octobre 2002).

9.5 Participation à des colloques, séminaires, invitations

- L. Berti-Équille a été invitée au colloque ADTACARA - Regional Workshop on Advanced Digital Technology-Assisted Cultural Artwork Restoration and Archiving - organisé par l'UNESCO à Bakou (Azerbaïdjan) du 14 au 18 novembre 2002 et y a présenté un exposé intitulé « Metadata for Efficient Exploitation of Very Large Multimedia Databases ».
- L. Berti-Équille a présenté un exposé intitulé « Modéliser et contrôler la qualité des données biologiques » lors d'un séminaire au LORIA à Nancy le 6 novembre 2002.
- P. Gros a été invité lors du séminaire InTech de l'INRIA Rhône-Alpes du 21 mars 2002. Le titre de l'exposé était : « Indexation par le contenu. Quelques perspectives ».
- P. Gros a présenté un exposé intitulé « Techniques d'exploitation des documents multimédias » lors d'un séminaire à l'IRIN à Nantes.
- P. Gros a été invité au colloque « Invariants dans les systèmes complexes » organisé par l'université d'Amiens le 17 mai 2002 et y a présenté un exposé sur « Les invariants géométriques et photométriques en vision et robotique ».
- P. Sébillot a présenté un exposé intitulé « Automatic Topic Characterization and Detection for Text Classification » lors de la journée du CNRT TIM « Text Indexing for Multimedia », Rennes, avril 2002.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

- [1] L. AMSALEG, P. GROS, « Content-based Retrieval Using Local Descriptors: Problems and Issues from a Database Perspective », *Pattern Analysis and Applications 2001*, 4, 2001, p. 108–124.
- [2] J. ANDRÉ, A. MORIN, H. RICHY, « Comparison of Literary Texts using Biological Sequence Comparison and Structured Documents Capabilities », in : *Proceedings of the ICCLSDP, Calcutta, Inde*, février 1998.
- [3] L. BERTI-EQUILLE, « Annotation et recommandation collaboratives de documents selon leur qualité », *Revue ISI-NIS, Numéro spécial Recherche et filtrage d'information 7*, 1-2/2002, 2002, p. 125–156.

- [4] V. CLAVEAU, P. SÉBILLOT, P. BOUILLON, C. FABRE, « Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? », *Traitement automatique des langues, numéro spécial Lexiques sémantiques* 42, 3, 2001, p. 729–753.
- [5] B. LAMIROY, P. GROS, « Rapid Object Indexing and Recognition Using Enhanced Geometric Hashing », *in : Proceedings of the 4th European Conference on Computer Vision, Cambridge, Angleterre, 1*, p. 59–70, avril 1996.
- [6] R. PRIAM, A. MORIN, « Visualisation des corpus textuels par treillis de multinomiales auto-organisées - Généralisation de l'analyse factorielle des correspondances », *Revue Extraction des Connaissances et Apprentissage (Actes EGC'2002) 1*, 4, 2002, p. 407–412.
- [7] M. ROSSIGNOL, P. SÉBILLOT, « Automatic Generation of Sets of Keywords for Theme Characterization and Detection », *in : 6^{es} journées internationales d'analyse statistique des données textuelles*, A. Morin, P. Sébillot (éditeurs), Saint-Malo, France, 2002.

Thèses et habilitations à diriger des recherches

- [8] P. SÉBILLOT, *Apprentissage sur corpus de relations lexicales sémantiques - La linguistique et l'apprentissage au service d'applications du traitement automatique des langues*, Habilitation à diriger des recherches, Université de Rennes 1, France, décembre 2002.

Articles et chapitres de livre

- [9] L. AMSALEG, P. GROS, S.-A. BERRANI, « A Robust Technique to Recognize Objects in Images, and the DB Problems it Raises », *Special issue of the Journal of Multimedia Tools and Applications*, 2002, à paraître.
- [10] S.-A. BERRANI, L. AMSALEG, P. GROS, « Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation », *Ingénierie des systèmes d'information*, 2002, à paraître.
- [11] L. BERTI, F. MOUSSOUNI, A. ARCADE, « Integration of Biological Data on Transcriptome », *Revue ISI-NIS, Numéro spécial interopérabilité et intégration des systèmes d'information* 6, 3/2001, 2002, p. 61–86.
- [12] L. BERTI, « Annotation et recommandation collaboratives de documents selon leur qualité », *Revue ISI-NIS, Numéro spécial recherche et filtrage d'information* 7, 1-2/2002, 2002, p. 125–156.
- [13] V. CLAVEAU, P. SÉBILLOT, C. FABRE, P. BOUILLON, « Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming », *Journal of Machine Learning Research, special issue on Inductive Logic Programming*, 2002, à paraître.
- [14] B. DAILLE, C. FABRE, P. SÉBILLOT, « Applications of Computational Morphology », *in : Many Morphologies*, P. Boucher (éditeur), Cascadilla Press, Somerville, 2002, p. 210–234.
- [15] R. PRIAM, A. MORIN, « Visualisation des corpus textuels par treillis de multinomiales auto-organisées - Généralisation de l'analyse factorielle des correspondances », *Revue Extraction des Connaissances et Apprentissage (Actes EGC'2002) 1*, 4, 2002, p. 407–412.
- [16] P. SÉBILLOT, « Traitement automatique des langues et recherche d'information », *in : La recherche d'information sur les réseaux II (cours Inria)*, ADBS Éditions, 2002, p. 137–168.

Communications à des congrès, colloques, etc.

- [17] P. BOUILLON, V. CLAVEAU, C. FABRE, P. SÉBILLOT, « Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method », in : *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC'2002, Las Palmas, Iles Canaries, Espagne, 2002.*
- [18] Y. MEZOUAR, A. REMAZEILLES, P. GROS, F. CHAUMETTE, « Image Interpolation for Image-based Control Under Large Displacement. », in : *Proceedings of the IEEE International Conference on Robotics and Automation, Washington DC, États-Unis, mai 2002.*
- [19] A. MORIN, « Deux exemples d'analyse de données textuelles », in : *Premier colloque sur la statistique et l'analyse des données dans les sciences appliquées et économiques, Beyrouth, Liban, 2002.* à paraître.
- [20] A. MORIN, « Factorial Correspondence Analysis: a Dual Approach for Semantics and Indexing », in : *Proceedings of the Conference Compstat 2002 -Short communications and posters, 2002.*
- [21] R. PRIAM, A. MORIN, « Cartographie par Champ de Markov. Application aux données textuelles », in : *XXXIVèmes Journées de Statistique - JSBL'2002, Bruxelles, Belgique, p. 328–329, mai 2002.*
- [22] R. PRIAM, A. MORIN, « Visualisation des données textuelles », in : *6^{es} journée internationales d'analyse statistique des données textuelles, JADT'2002, Saint-Malo, France, A. Morin, P. Sébillot (éditeurs), p. 629–640, mars 2002.*
- [23] M. ROSSIGNOL, P. SÉBILLOT, « Automatic Generation of Sets of Keywords for Theme Characterization and Detection », in : *6^{es} Journées internationales d'analyse statistique des données textuelles, JADT'2002, A. Morin, P. Sébillot (éditeurs), Saint-Malo, France, 2002.*

Rapports de recherche et publications internes

- [24] S.-A. BERRANI, L. AMSALEG, P. GROS, « Controlling the Precision of Approximate Fast Nearest-Neighbor Searches », *rapport de recherche n°4423, INRIA, avril 2002.*
- [25] E. CATZ, « Apprentissage automatique de catégories de mots par co-training », *Rapport de DEA, IFSIC, Université de Rennes 1, France, juin 2002.*
- [26] P. TCHIENEHOM, « Sélection automatique d'unités de textes décrivant des interactions entre gènes », *Rapport de DEA, IFSIC, Université de Rennes 1, France, juin 2002.*