Indexation des données multimedia, utilisation dans le cadre d'un système de recherche d'informations

Catherine Berrut

▶ **To cite this version:**

Catherine Berrut. Indexation des données multimedia, utilisation dans le cadre d'un système de recherche d'informations. Autre [cs.OH]. Université Joseph-Fourier - Grenoble I, 1997. Français. ⟨NNT : ⟩. ⟨tel-00004920⟩

HAL Id: tel-00004920

https://theses.hal.science/tel-00004920v1

Submitted on 20 Feb 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Manuscrit présenté par

Catherine Berrut

pour obtenir le diplôme d'

Habilitation à Diriger des Recherches
spécialité Informatique

de l'Université Joseph Fourier - Grenoble I

Indexation des données multimédia, utilisation dans le cadre d'un système de recherche d'informations

date de soutenance :	le 17 octobre 1997	
composition du jury :	président	M. Michel Adiba
	rapporteurs	Mme Marion Créhange
		M. Arcot Desai Narasimhalu
		M. Alan Smeaton
	examinateurs	M. Patrick Bosc
		M. Yves Chiaramella

Travaux préparés au sein du laboratoire CLIPS
(Communication Langagière et Interaction Personne-Système)
de l'Institut IMAG

Je tiens à remercier

Monsieur Michel Adiba, Professeur à l'Université Joseph Fourier, d'avoir accepté de présider ce jury. Président de mon jury de thèse en 1988, puis responsable du projet scientifique Aristote, actuellement directeur de l'UFR Informatique et Mathématiques Appliquées, il suit avec intérêt mes travaux de recherche depuis de nombreuses années, et je le remercie du soutien scientifique et moral qu'il m'a toujours apporté ;

Monsieur Patrick Bosc, Professeur à l'Ecole Nationale Supérieure des Sciences Appliquées et de Technologies (ENSSAT) de Lannion, d'avoir accepté de participer à ce jury. Pour avoir participé à plusieurs jurys de thèse de l'équipe, avoir rapporté sur mes travaux de thèse en 1988, il suit mes travaux de recherche depuis plus de dix ans. Je souhaite que ce document lui en apporte une synthèse intéressante, et alimente ainsi nos prochaines discussions ;

Monsieur Yves Chiaramella, Professeur à l'Université Joseph Fourier, directeur du laboratoire Communication Langagière et Interaction Personne-Système, qui a dirigé mes travaux de recherche depuis mon arrivée à Grenoble en 1984. Directeur de ma thèse, puis co-directeur des thèses de Mourad Mechkour, François Paradis et Nathalie Denos, il a largement contribué à ce travail. Ses lectures de ce document en ont permis une nette amélioration. Je le remercie très sincèrement à la fois de la liberté d'action, du soutien, et des encouragements qu'il m'a toujours apportés. Je n'ai à ce jour qu'un regret dans notre longue collaboration : qu'il n'ait pu me transmettre son calme !

Madame Marion Créhange, Professeur à l'IUT A - Nancy II, responsable de l'équipe EXPRIM du Centre de Recherche en Informatique de Nancy, d'avoir accepté d'être rapporteur de mon travail. Nous avons beaucoup travaillé ensemble sur ce manuscrit, et je la remercie très chaleureusement de ses remarques, de ses questions, qui ont largement contribué à une nette amélioration de ce document. Sachant que Madame Marion Créhange prend très prochainement sa retraite, je suis très touchée qu'elle ait accepté ce travail de rapporteur ;

I really and sincerely thank Pr Alan Smeaton and Dr Arcot Desai Narasimhalu, for accepting to be referees of my work. I know this task was difficult for them, and I am really happy that they appreciated the document ;

Dr Arcot Desai Narasimhalu is Associate Director of the Institute of Systems Science (National University of Singapore), he is the responsible of the Business Development, and the New Initiatives departments. We met in 1995 in Grenoble during the MIRO workshop and had interesting discussions about research going on in ISS and Grenoble. Since then, Munkew Leong and Jian Kang Wu came in Grenoble in 1996, Yves Chiaramella and Philippe Mulhem went to Singapore. I hope that the MIX proposal will succeed and allow deeper research exchange between our groups ;

Pr Alan Smeaton is Professor in Computer Science, at Dublin City University. We first met in 1986, in Pisa, during the ACM-SIGIR Conference. Since then, we had several opportunities to work together : funding from the French Ministry of Research in 1995, MIRA (1996-1999), and, I hope again, MIX ;

Toutes les personnes que j'ai encadrées, et tout particulièrement René Cuzin, Pascal Bouchon, Mourad Mechkour, François Paradis, et Nathalie Denos. Ce travail est d'abord

et avant tout le leur. Je les remercie du travail qu'ils ont accompli, sans lequel je ne pourrais certainement pas présenter ce diplôme aujourd'hui ;

Jacques Demongeot, Professeur à l'Institut Universitaire de France, et Philippe Cinquin, Professeur à l'Université Joseph Fourier, pour être à l'initiative de RIME, et pour l'intérêt qu'ils ont toujours porté à ce travail ;

Claude Puech, Professeur à l'Université Joseph Fourier, Président de la Commission HDR, qui a assuré avec beaucoup de compréhension le bon déroulement de l'évaluation de ce travail ;

Marie-France Bruandet, Professeur à l'Université Joseph Fourier, responsable de l'équipe Modélisation et Recherche d'Informations Multimédia, pour son amitié, son soutien permanent et son implacable philosophie de la vie qui m'a souvent appris et guidé ;

Tous les membres de l'équipe Modélisation et Recherche d'Informations Multimédia pour leur disponibilité, leur serviabilité, et leur sympathie. Je voudrais rajouter un remerciement supplémentaire à Philippe Mulhem, dont j'use et j'abuse de la gentillesse, et aussi des chocolateries !

Tous les membres du laboratoire CLIPS, et plus particulièrement les "toujours présents" du deuxième étage du bâtiment B, grâce à qui les soirées et les week-ends deviennent moins moroses … et les pizzas du soir, un moment si agréablement partagé !

Tous mes amis et collègues de travail, notamment les membres de l'équipe STORM du LSR qui a également participé au projet Aristote, l'équipe MOVI du laboratoire GRAVIR, Pierre-Claude Scholl avec qui j'ai commencé à enseigner en 1985, les enseignants de LP1, l'équipe d'enseignants en Deug. J'ajouterais une pensée pour mes amis Hervé Martin et Jean-Pierre Peyrin, toujours là pour me rappeler que la vie est belle !

Ma famille et ma belle-famille ;

Et enfin, Michel et notre Vincent, … pour tout !

A Michel,

A Vincent

1. Les systèmes de recherche d'informations

Parmi les systèmes travaillant sur des corpus documentaires, le but des systèmes de recherche d'informations est de permettre l'accès aux documents par leur contenu sémantique ; ainsi l'utilisateur exprime son besoin d'informations en indiquant le contenu qu'il souhaite observer dans les documents retrouvés. Les systèmes de recherche d'informations sont généralement fondés sur un modèle formel de correspondance (modèle logique par exemple). L'efficacité et l'ergonomie avec lesquelles cette recherche s'effectue apparaissent comme une priorité lorsque l'on implémente cette théorie dans un système opérationnel. De façon plus générale, même si la définition de l'accès au contenu sémantique des documents constitue le problème central des systèmes de recherche d'informations, la recherche dans ce domaine couvre un spectre très large de travaux : depuis la signature de documents jusqu'à la conception d'interfaces homme-machine, depuis des travaux théoriques en logique jusqu'à l'étude des signaux des données multimédia.

1.1. Architecture générale

Classiquement, tout système de recherche d'informations se décompose en :

- un processus d'interrogation, permettant à l'utilisateur de formuler une requête et ainsi d'interroger le corpus. La requête et le corpus de documents sont représentés respectivement dans un modèle de requêtes et un modèle de documents (ou langage d'indexation). Un modèle de correspondance compare la requête aux documents : les documents répondant à la requête sont ainsi donnés en réponse. Il faut, à ce niveau, établir une comparaison sémantique (et non une égalité) entre les concepts figurant dans un document et ceux figurant dans la requête. Par exemple, dans le modèle booléen, les documents sont représentés sous la forme d'une liste conjonctive de mots-clés : $D = t_{d1} \wedge t_{d2} \wedge \ldots \wedge t_{dn}$. Les requêtes sont représentées sous la forme de listes de mots-clés reliés entre eux par les opérateurs et ($\wedge$), ou ($\vee$), sauf ($\neg$) : $Q = t_{q1} \wedge t_{q2} \vee \ldots \wedge \neg t_{qn}$. Dans ce modèle, un document D répond à une requête Q si et seulement si on peut démontrer que $D \rightarrow Q$, où $\rightarrow$ est l'implication de la logique des prédicats.

La comparaison entre requête et document aboutit rarement à des équivalences strictes, mais plutôt à des équivalences partielles : le document correspond à une partie seulement de la requête. Les documents peuvent ainsi être ordonnés selon une relation d'ordre permettant le classement des documents du plus pertinent au moins pertinent. On appelle pertinence-système la pertinence que le système accorde à chaque document pour une requête donnée ;

- un processus d'indexation, mis en œuvre afin d'extraire préalablement une représentation homogène du contenu sémantique, sous forme de termes d'indexation qui sont des éléments d'un langage d'indexation. Dans les systèmes classiques de recherche d'informations, l'indexation est organisée en trois étapes : extraction, sélection, pondération, dont l'objectif final est de définir pour chaque document ses termes d'indexation ;

- une modélisation de la connaissance. Classiquement cette modélisation apparaît sous la forme d'un thésaurus, et sert de référence aussi bien au processus d'interrogation qu'au processus d'indexation. Le processus d'interrogation l'utilise afin d'augmenter ou de restreindre les requêtes des utilisateurs, agissant ainsi directement sur le nombre et la qualité des réponses données par le système. Le processus d'indexation l'utilise essentiellement comme norme de définition des termes d'indexation.

1.2. Finalité d'un système de recherche d'informations

Satisfaire les besoins des utilisateurs constitue la finalité des systèmes de recherche d'informations que [Sal83] présente sous cinq critères fondamentaux :

- l'effort, intellectuel ou physique, nécessaire à l'utilisateur pour formuler les requêtes, conduire sa recherche, visualiser les documents donnés en réponse ;

- le temps entre l'envoi d'une requête et la présentation de la réponse ;

- la présentation des résultats qui influence l'utilisateur dans sa motivation à consulter les documents retrouvés ;

- le contenu du corpus, c'est-à-dire sa capacité à donner intrinséquement de bonnes réponses à une requête ;

- et surtout, la capacité du système à identifier uniquement les documents pertinents, et à éliminer les autres.

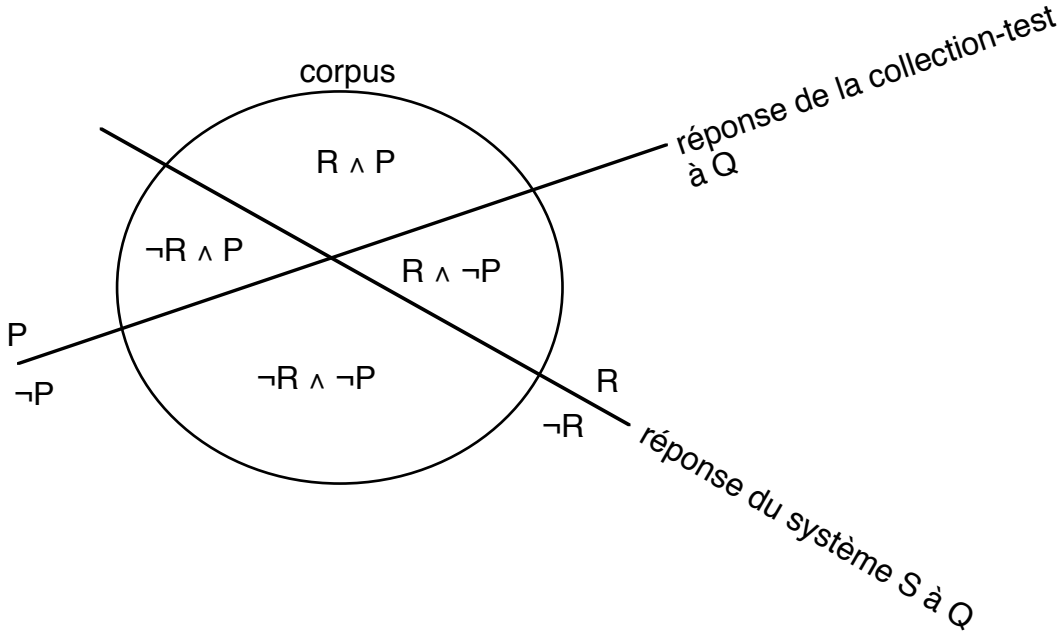
La plupart des systèmes insistent sur ce dernier point, qui constitue effectivement le point essentiel dans la qualité attendue d'un système. En effet, qu'un système fournisse de bonnes réponses à l'utilisateur est une condition sine qua none de son utilisation !

Au niveau opérationnel ce critère est testé en comparant les réponses données à un ensemble de requêtes par le système à celles souhaitées par l'utilisateur. Ces tests mettent en évidence la dualité entre la pertinence système (ce que le système retrouve) et la pertinence utilisateur (ce que l'utilisateur souhaite retrouver), en mesurant la distance entre ces deux pertinences.

En pratique, ces tests sont effectués en utilisant les collections-tests. Ces collections-tests proposent en effet un corpus et un ensemble de requêtes résolues. Ainsi pour toute requête Q d'une collection test, l'ensemble P des documents pertinents, c'est-à-dire répondant à la requête Q selon le point de vue de l'utilisateur, et l'ensemble ¬P, formé des documents non pertinents pour Q, sont définis en extension et partitionnent le corpus C de la collection-test.

La réponse d'un système de recherche d'informations S à la requête Q partitionne le corpus C en deux sous-ensembles : l'ensemble R des documents retrouvés par le système de recherche d'informations S, c'est-à-dire répondant à la requête Q selon ce système, et l'ensemble ¬R des documents non retrouvés.

Ainsi on obtient le schéma suivant :



Le système de recherche d'informations parfait est celui tel que R=P, pour toute requête Q de la collection-test, mais un tel système n'existe pas. Aussi existe-t'il un certain nombre de mesures permettant de caractériser chaque système, ces mesures sont présentées dans la partie suivante.

1.3. Mesures qualitatives d'un système de recherche d'informations

Pour mesurer les performances qualitatives des systèmes de recherche d'informations, on procède à la comparaison des ensembles P, ¬P, R, ¬R sur l'ensemble des requêtes. Il existe à cet effet de nombreuses mesures, chacune mettant en évidence telle ou telle propriété du système testé. Les mesures les plus classiques sont les mesures de rappel (recall) et de précision :

$$\text{Rappel} = \frac{R \wedge P}{P} \qquad \text{Précision} = \frac{R \wedge P}{R}$$

Le rappel mesure la capacité du système de recherche d'informations à trouver, pour une requête, tous les documents pertinents. Le rappel peut donc se définir comme la probabilité pour un document d'être retrouvé, sachant qu'il est pertinent.

La précision mesure la capacité du système à trouver, pour une requête, uniquement des documents pertinents. La précision est une mesure très intéressante pour mesurer la qualité des réponses d'un point de vue de l'utilisateur du système.

L'élimination (discrimination) et son complémentaire, l'hallucination (fallout), permettent de mesurer la capacité du système à éliminer les documents non pertinents :

$$\text{Elimination} = \frac{\neg R \wedge \neg P}{\neg P} \qquad \text{Hallucination} = \frac{R \wedge \neg P}{\neg P}$$

L'élimination mesure la capacité du système à éliminer tous les documents non pertinents. L'élimination peut donc se définir comme la probabilité pour un document d'être éliminé, sachant qu'il n'est pas pertinent.

D'autres mesures comme le bruit (le complémentaire de la précision), le silence (le complémentaire du rappel), la généralité (la proportion, pour une requête donnée, de documents retrouvés par rapport à tout le corpus, soit $R / (R \vee \neg R)$) existent. De par leur complémentarité, ces mesures peuvent s'exprimer entre elles : il suffit par exemple de connaître trois des mesures parmi le rappel, la précision, l'hallucination, la généralité, pour calculer la quatrième.

Majoritairement, les systèmes opérationnels affichent leurs tableaux de rappel et de précision. Pourtant certains travaux tels que [Rob 69] préconisent l'utilisation plutôt du rappel et de l'élimination ou bien du rappel et de l'hallucination. Leur argumentation se base sur le fait que, contrairement à l'hallucination ou à l'élimination, la précision est trop sensible aux variations de la généralité de chaque requête.

De notre point de vue, le choix entre rappel/précision et rappel/hallucination relève d'abord et avant tout de la mise en évidence de propriétés particulières du système. Le rappel indique la proportion des documents pertinents retrouvés, et la précision est une mesure de l'efficacité avec laquelle les documents pertinents sont retrouvés. En ce sens, le couple rappel/précision donnent une mesure orientée utilisateur des performances du système. Par ailleurs, l'hallucination mesure l'efficacité de l'élimination des documents non-pertinents. Ainsi le couple rappel/hallucination donne une mesure orientée système des performances du système.

1.4. Mise en cause de ces mesures de performance

Les performances des systèmes de recherche d'informations sont ainsi systématiquement calculées et publiées depuis plusieurs décennies. Pourtant [Bel80] fait un bilan pessimiste des performances des systèmes : 60% en rappel et 40% en précision en moyenne. De notre point de vue, ce résultat n'est pas un constat d'échec, mais le résultat d'un quiproquo sur l'interprétation des mesures que nous expliquons en deux points :

1) la valeur accordée à ces mesures doit être relativisée, et nous expliquons pourquoi dans la partie 1.4.1 ;

2) par l'utilisation de ces mesures, les performances des systèmes de recherche d'informations sont évaluées globalement, et ne permettent pas d'isoler les performances individuelles des différents processus. On constate ainsi que les systèmes mesurés sont généralement munis de langages à mots clés et que bien souvent la comparaison entre systèmes se ramène donc à une comparaison de leur fonction de correspondance. Nous revenons sur cet argument dans la partie 1.4.2.

1.4.1. Valeur relative des mesures qualitatives

Le problème posé par toutes ces mesures qualitatives réside dans le décalage entre leur sens et leur mise en œuvre. Comme nous le montrons dans le chapitre 2, deux utilisateurs différents formulant une même requête n'ont pas le même point de vue sur les réponses du même système. De plus si on confronte ces deux utilisateurs à deux systèmes différents, chaque utilisateur préfère l'un des deux systèmes, mais pas obligatoirement le même ! De notre point de vue, donner des mesures qualitatives d'un système n'a de sens que si l'on indique le contexte dans lequel le système est utilisé. Sinon, il existe sûrement un utilisateur à qui on donne entière satisfaction, un autre qui est très déçu, et il vaut mieux afficher uniquement les mesures de l'utilisateur satisfait !

De façon plus générale, cet argument remet en cause les comparaisons faites actuellement avec les collections-tests. Ces collections-tests donnent un corpus, des requêtes et leurs réponses. On suppose que le plus on s'approche de ces réponses, meilleur est le système. Malheureusement, ceci ne prouve pas que le système est bon, ceci prouve que le système fonctionne parfaitement pour la ou les personnes qui ont construit cette collection. La question peut alors être complètement déviée en choisissant la collection-test avec laquelle le système fonctionne le mieux.

a- L'hypothèse de l'existence de P, ¬P

Toutes les mesures prennent pour hypothèse l'existence d'une dichotomie du corpus en deux ensembles : les documents pertinents et les documents non-pertinents. L'existence de ces deux ensembles est une hypothèse peu réaliste pour des raisons pratiques et humaines :

- Qui peut définir de façon crédible l'ensemble P des documents pertinents d'une requête, … dans un corpus de plusieurs millions de documents ? Le rappel est donc une mesure intéressante, mais incalculable concrètement. Il en est de même pour l'élimination et l'hallucination qui nécessitent la connaissance de l'ensemble ¬P des documents non pertinents.

- Certes un utilisateur accordera à certains documents le fait d'être pertinent ou non pertinent. Mais tous les documents n'entreront pas dans cette catégorisation : certains seront, selon le point de vue de l'utilisateur, "relativement intéressants", d'autres encore auront le statut "à voir", … Par ailleurs, l'utilisateur peut déclarer des documents pertinents pour des raisons diverses : ce document est pertinent pour telle raison, cet autre document est pertinent pour telle autre raison. La sémantique de l'ensemble P est alors complexe.

b - Définition de R et ¬R dans un système pondéré

La pondération des documents retrouvés n'est pas introduite dans les mesures actuelles. Les calculs effectués montrent généralement l'évolution du rappel et de la précision en fonction du nombre de documents pris en compte. Mais afin de calculer le rappel et la précision, les documents sont catégorisés dans les ensembles R et ¬R sans tenir compte de la pondération accordée à chacun d'entre eux.

Ainsi ces mesures n'introduisent pas de statuts intermédiaires aux documents, et tout le "plus" de certains systèmes (faire découvrir de nouveaux documents, permettre des questions ouvertes, …) ne peut actuellement être valorisé.

c - Problématique de l'évaluation

Comparer statistiquement des systèmes de recherche d'informations élude complètement leur apport individuel : à taux de précision/rappel égaux, deux systèmes peuvent se compléter parfaitement pour satisfaire un utilisateur, en apportant chacun des réponses complémentaires. L'union de ces deux systèmes peut peut-être donner de meilleurs résultats qu'un système plus complexe.

Comme nous venons de le souligner, un des problèmes essentiels de l'évaluation vient du fait que la simulation de la satisfaction de l'utilisateur en deux ensembles (P et ¬P) et la visualisation binaire du système (R et ¬R) modélisent trop schématiquement la réalité de la recherche d'informations. Une première voie de meilleure modélisation serait sans doute d'introduire une plus grande discrétisation de la satisfaction de l'utilisateur et de la réponse du système, et d'introduire cette discrétisation dans les mesures actuelles.

Une seconde voie consisterait d'une part à introduire une dénotation de la requête et des jugements de pertinence de l'utilisateur, afin de prendre en compte une sémantique plus fine de ses besoins. D'autre part, le système doit alors restituer à l'utilisateur le même niveau de justifications de ses réponses : la réponse ne doit plus être une liste éventuellement pondérée de documents, mais une ou plusieurs listes de documents, chacune justifiée sémantiquement dans ses liens avec la requête. Utilisateur et système peuvent alors entrer dans un dialogue direct et explicite : les quiproquos introduits par l'opacité des systèmes classiques peuvent alors être évités.

Cette seconde voie nécessite cependant une remise en question des systèmes et du cloisonnement de leurs processus : notamment l'indexation ne peut plus uniquement produire des listes de termes, et ce de façon indépendante du reste du système. Interrogation et indexation doivent coopérer afin de restituer au mieux à l'utilisateur sa propre vision du corpus. C'est dans cette voie que se situent nos travaux de recherche, que nous introduisons ci-après.

Par ailleurs, cette seconde voie permettrait également d'envisager une typologie des systèmes en identifiant non seulement le nombre de documents satisfaisants, mais aussi en connaissant la sémantique de leur pertinence. Il serait alors intéressant de travailler sur de nouvelles mesures, ou bien dans un premier temps de voir sous un nouveau jour les mesures actuelles. Par exemple, l'hallucination pourrait être transformée en une mesure dite "découverte", montrant la capacité du système à montrer d'autres documents à l'utilisateur : la proximité plus ou moins grande de ces documents avec le besoin de l'utilisateur lui permettrait alors de découvrir dans le corpus des documents répondant moins bien à sa requête certes, mais lui donnant une satisfaction partielle. Ainsi soit sa question est une question ouverte et l'utilisateur est content de regarder dans d'autres directions, soit l'utilisateur ne trouve pas ou pas assez de bonnes réponses et il accepte de regarder d'autres documents.

Aussi l'évaluation des systèmes de recherche d'informations est-elle un des problèmes ouverts en recherche d'informations, et notamment le groupe de travail MIRA (Evaluation Framework for Interactive Multimedia Information retrieval Applications) financé par la CEE a commencé sur cette thématique ses travaux l'année dernière.

1.4.2. Pour une recherche en indexation

Quelle que soit la mesure choisie, l'expression des performances qualitatives d'un système en donne une vision globale. En ce sens, elle ne permet pas de justifier les performances d'un système par le rôle de chacun des éléments de son architecture, ou encore d'identifier dans un système un paramètre particulièrement performant.

Le constat que nous pouvons faire actuellement est que cette vision macroscopique des systèmes a pour conséquence de privilégier les travaux au niveau de l'interrogation (interface, fonction de correspondance, ...) : afin d'améliorer les performances, les chercheurs raisonnent généralement uniquement au niveau de la fonction de correspondance. On constate que la littérature croule sous la présentation de diverses fonctions ou modèles de correspondance, dont la plupart raisonne sur des langages à mots clés.

Pour des raisons que nous donnons plus loin, l'autre processus du système, l'indexation, a beaucoup moins été étudié : on constate que peu de travaux s'effectuent sur l'indexation, et qu'elle est encore trop souvent considérée comme une boîte noire relativement externe au système, et fournissant en sortie des listes de mots-clés.

Pourquoi ? La première des raisons est qu'il est vrai que pendant longtemps l'indexation a réellement été une boîte noire externe au système, et fournissant en sortie des listes de mots-clés. En effet, les documents n'étaient pas électroniques, et l'indexation était produite manuellement, sous forme de listes de mots-clés, par des documentalistes particulièrement entraînés. Lorsque les documents textuels électroniques sont apparus, les méthodes automatiques d'indexation sont apparues également. Leur seul objectif était de fournir une indexation similaire à celle produite manuellement, et donc des listes de mots-clés. Il est alors vrai et naturel que dans ce contexte, le seul paramètre avec lequel le système peut être amélioré est l'interrogation.

Depuis une dizaine d'années, le développement des données électroniques, le stockage des données multimédia, la mise à disposition des systèmes de recherche d'informations à un public, qui peut aller du néophyte au spécialiste, ont permis l'émergence d'une nouvelle vision de la recherche d'informations, et notamment de mettre au premier plan la nécessité d'indexer autrement : il n'est plus possible d'imaginer que l'indexation d'un document puisse être uniquement une liste fixe de mots clés. Flexibilité, adaptativité, richesse d'expression sont les conditions nécessaires à l'expression d'une indexation. Par ailleurs, son intégration dans le système, son utilisation paramétrée en font dorénavant une pièce maîtresse pour la satisfaction de l'utilisateur.

En ce sens, l'indexation d'un document n'est pas universelle, mais dépendante de la perception de chacun. C'est dans cette nouvelle vision de la recherche d'informations que se situent nos travaux de recherche, et nous détaillons par la suite comment nous avons décliné cette voie selon deux axes principaux :

- indexation complexe afin de restituer fidèlement le contenu des documents ;

- système dynamique d'indexation et de recherche d'informations. Interrogation et indexation coopèrent afin de produire pour chaque utilisateur une vision adaptée du corpus. L'indexation doit dorénavant justifier chaque terme d'indexation par le rôle effectif qu'il joue dans le document.

1.5. Présentation de nos travaux

L'expression de l'indexation et sa pleine utilisation dans le système sont parmi les éléments essentiels que devront posséder les systèmes. Pour montrer ceci, nous avons développé nos travaux selon deux axes : puissance de l'expression d'une part, adaptation de l'expression d'autre part. La puissance de l'expression nous montre le besoin d'indexer autrement dès lors que l'on aborde les nouveaux besoins en

recherche d'informations. Par ailleurs, l'adaptation de l'expression permet au système de disposer dynamiquement de plusieurs compréhensions des documents : cette approche remet en cause la trop grande rigidité des indexations classiques.

De façon plus générale, notre point de vue est de montrer que la nature et la variabilité du langage d'indexation d'une part et le processus d'indexation d'autre part sont des paramètres à part entière pour les systèmes. Nous pensons que l'association étroite entre l'indexation et la fonction de correspondance constitue la condition d'une amélioration des systèmes et de leur adaptation à des utilisateurs et à des recherches variés.

Pour introduire nos travaux, nous proposons dans le chapitre 2 de ce document de valoriser le rôle et l'impact de l'indexation dans un système de recherche d'informations. Pour cela, nous commençons par rappeler les objectifs de l'indexation, puis nous montrons son impact qualitatif dans un système. Le chapitre 3 montre les spécificités de l'indexation des données multimédia, et précise l'objectif de nos travaux. Les chapitres suivants développent nos deux axes de recherche, que nous résumons ci-dessous.

1.5.1. Indexation de données multimédia : textes et images

Telle que nous l'avons abordée, cette indexation nous a amenés à une réflexion sur la nature des langages d'indexation et la nécessité de la définition de processus d'indexation rigoureux et adaptés. Nous avons abordé cette problématique dans deux contextes : les données images d'une part, les données dédiées à des spécialistes d'autre part.

L'interprétation multiple et variée des données multimédia oblige à aller au delà d'une indexation par mots clés. Les langages d'indexation que nous proposons sont complexes et permettent la représentation des documents par des concepts reliés entre eux par des connecteurs sémantiques.

La nécessité de langages complexes est accrue dans certains systèmes dédiés à des utilisateurs particuliers : les langages complexes sont alors l'unique moyen d'atteindre le niveau de précision recherché. Nous avons ainsi défini la notion de système orienté précision. Ce sont des systèmes dédiés à des catégories d'utilisateurs : des spécialistes d'un domaine. Leur besoin en recherche d'information est clair en ce sens que la sémantique donnée à l'ensemble P des documents pertinents est plus tangible que dans un contexte ouvert. L'objectif pour ces systèmes est de ne retrouver que des documents pertinents, mais pas forcément tous.

Disposer d'un langage d'indexation ne constitue que la moitié de l'indexation, et il faut, pour deux raisons fondamentales, définir le processus d'indexation associé :

- sans processus d'indexation, un langage d'indexation n'est qu'une théorie qui ne sera pas utilisée ;

- un langage complexe est difficilement manipulable, même par des experts entraînés, et il est source de nombreuses erreurs d'indexation. Il faut donc lorsque c'est possible produire un processus automatique d'indexation : [Sme97] décrit les limites de cette automatisation dans le contexte de données textuelles et du traitement de la langue naturelle. Sinon, il faut construire une interface d'indexation dédiée aux indexeurs. Cette interface a alors un double rôle : aider l'indexeur et contrôler son travail en minimisant les erreurs d'indexation.

Aussi les chapitres 4 et 5 présentent-ils les travaux que nous avons effectués dans ce contexte :

- le chapitre 4 présente l'indexation de textes médicaux que nous avons proposée dans le système RIME ;

- le chapitre 5 présente l'indexation d'images que nous avons définie avec le langage EMIR2 et son interface d'indexation.

La validation de ces travaux a été effectuée grâce au prototype PRIME, dédié aux médecins, et qui permet la recherche de données médicales. Cette plateforme logicielle importante (40 000 lignes de code) est le fruit d'un long travail commencé en 1990 [Cuz92] que nous présentons dans le chapitre 6.

1.5.2. Systèmes de recherche d'informations dynamiques

La finalité des systèmes de recherche d'informations est de permettre à un utilisateur de retrouver dans un corpus des documents l'intéressant. La réalité est toute autre : le système propose à l'utilisateur de formuler une requête et de regarder sa réponse. En cas d'insatisfaction, l'utilisateur doit formuler une nouvelle requête jusqu'à trouver celle qui lui permet de retrouver les documents recherchés.

En les rendant dynamiques, nous proposons de rapprocher les systèmes de leur finalité. Si un utilisateur est insatisfait de la réponse du système, ce n'est pas sa requête qui est en cause et donc à modifier, mais c'est l'interprétation que le système en donne qui doit être modifiée. De notre point de vue, cette dynamicité résulte de la prise en compte dans le système de trois éléments fondamentaux :

- la variabilité de l'indexation : le contenu d'un document ne doit pas être une constante du système. Chaque utilisateur a sa propre interprétation du contenu des documents, et il faut que le système soit capable de la connaître ou de la déduire ;

- adaptation de la correspondance : deux utilisateurs différents peuvent formuler la même requête sans que la réponse attendue soit la même. De même que pour l'indexation, la correspondance doit être paramétrable ;

- l'association étroite entre indexation et correspondance devient alors une nécessité de leur paramétrisation.

Cependant la prise en compte d'une vision personnalisée de chaque document, et de chaque requête n'est pas sans difficulté : [Bar 93] a montré une vingtaine de critères pour déterminer la pertinence d'un document : l'information est récente ou ancienne, l'information est présentée clairement, … [Ing94] montre que modéliser et représenter ceci est une tâche difficile.

La dynamicité que nous proposons est décrite dans les chapitres 7 et 8 qui présentent respectivement une indexation dynamique de documents structurés et un feedback adaptatif appliqué à un corpus d'images.

2. L'indexation en recherche d'informations

Proposant un travail sur l'indexation, nous avons souhaité développer dans ce chapitre un état de l'art sur les recherches effectuées dans ce domaine. L'état de l'art que nous présentons ne propose pas une synthèse par catégories d'indexation, mais une double réflexion sur l'indexation : la définition de son rôle tout d'abord, puis une étude sur l'impact qualitatif qu'elle joue. Cette seconde partie a été construite à partir d'articles de documentalistes, dont le métier les rend particulièrement habiles dans la manipulation des systèmes. Ils nous montrent comment ils "jouent" avec le système et l'indexation afin d'obtenir les meilleures performances possibles. Ces travaux sont particulièrement intéressants car ils sont la preuve expérimentale de l'impact qualitatif de l'indexation sur les systèmes .

2.1. Le rôle de l'indexation

L'indexation a pour rôle de représenter de façon homogène le contenu sémantique des documents du corpus. L'homogénéité de l'indexation réfère à la conformité, de cette représentation, à un langage d'indexation définissant (en extension ou en intention) les termes d'indexation utilisables. La notion de termes d'indexation est à prendre dans notre contexte au sens large, et nous entendons par terme d'indexation toute forme produite par l'indexation d'un document, quelle que soit sa complexité.

Le terme "indexation" encapsule donc deux problèmes bien distincts :

- la définition d'un langage d'indexation, permettant la représentation des concepts des documents du corpus ;

- la mise en place d'un processus d'indexation permettant l'extraction, à partir des documents du corpus, de termes d'indexation, c'est-à-dire de leur représentation conforme au langage d'indexation.

Même si elle est très vague, cette première définition est cependant consensuelle. Par exemple Borko et Bernier [Bor 78] disent "indexing is the process of analysing the informational content of records of knowledge and expressing the informational content in the language of the indexing system", alors que Wellish [Wel 91] donne une définition plus technocrate se référant à la norme ISO 5127/1 : "an operation intended to represent the results of the analysis of a document by means of a controlled or natural indexing language". De même Rowley [Row 88] écrit "the indexing process creates a description of a document or information, usually in some recognized and accepted style or format". Salton [Sal 83] complète cette définition, en ajoutant trois objectifs à l'indexation :

"- to allow the location of documents dealing with topics of interest to the user ;

- to relate documents to each other, and thus relate the topic areas, by identifying distinct documents dealing with similar, or related, topic areas ;

- to predict the relevance of individual documents to specific information requirements through the use of index terms with well-defined scope and meaning."

Ces différentes définitions nous montrent la dualité de l'indexation : représenter le contenu des documents afin de permettre aux utilisateurs de les retrouver. Ces deux objectifs sont difficiles à réunir, et la recherche en indexation le montre bien. En effet dans la plupart des travaux, l'indexation est soit orientée document soit orientée requête. L'indexation orientée document a pour objectif de résumer ou de représenter le contenu de chaque document, c'est-à-dire son signifiant et son signifié. L'indexation orientée requête doit, pour chaque document, refléter les requêtes pour lesquelles il est pertinent : l'indexation d'un document doit alors représenter les raisons pour lesquelles un utilisateur consulte ce document.

2.1.1. Indexation orientée document

L'indexation orientée document consiste à définir, à partir du document seulement, son contenu, que l'on qualifie dans ce contexte, d' "à-propos". Lancaster [Lan 91] décrit cette indexation comme : "a conceptual analysis, which, first and foremost, involves deciding what a document is about - that is, what it covers". Indexer un document revient à définir le processus qui permet de passer de la forme ou du signal d'un document à son fond, en d'autres termes de son signifiant à son signifié. Déterminer le signifié d'un document est une démarche délicate et subjective, car beaucoup de paramètres interviennent dans cette identification : la qualité du signifiant, l'indexeur, la base de connaissances, …

Prenons deux exemples. Le premier est une image montrant le Président Kennedy serrant la main à Bill Clinton adolescent. Jusqu'à récemment cette image était indexée par "le Président Kennedy serrant la main à des adolescents", et était peu utilisée dans les média. Depuis l'élection de Bill Clinton, cette image est devenue particulièrement symbolique en montrant "la poignée de main entre Kennedy et Clinton". Ainsi le signifié de cette image a été modifié du fait de l'élection de Clinton. Le deuxième exemple est du Président Pompidou, qui, lors d'une interview, avait répondu nerveusement à un journaliste "Je ne suis pas Mme Soleil". Cette réponse n'avait jamais été ni indexée ni diffusée. Mais lors de la mort de Mme Soleil, cette réponse de Pompidou a été "ressortie" afin de montrer la notoriété de cette astrologue. Cette information n'ayant pas été indexée, les documentalistes de l'INA (Institut National de l'Audiovisuel) n'ont eu à leur disposition pour retrouver cette interview que la réécoute de toutes les interview du Président Pompidou. Et cette interview a maintenant été plus diffusée du fait de cette réponse, que de toutes les autres réponses politiques et économiques qui étaient indexées. Son signifié a donc évolué pour devenir l'interview dans laquelle "Pompidou parle de Mme Soleil".

Cette notion d'à-propos prend une dimension très complexe sur les données images et vidéo, où la partie subjective des documents est délicate à déterminer. En effet, et nous reviendrons sur ces arguments dans le chapitre 3, leur signifiant est constitué physiquement d'un signal numérique et leur signifié est perçu par un sens (la vue, l'ouïe). Contrairement aux données textuelles, la distance entre signifiant et signifié est importante. De plus la perception du signifié se modifie non seulement en fonction des connaissances, mais également d'une personne à une autre. Il suffit pour le prouver de montrer une même image à différentes personnes, et de leur demander ce qu'elles observent. Le passage du signifiant au signifié devient alors un problème particulièrement complexe. La littérature met souvent en évidence la difficulté des indexeurs (humains ou non) à mener cette tâche, et les tests de consistence d'indexation le montrent par les désaccords entre indexeurs.

2.1.2. Indexation orientée requête

La façon la plus classique de procéder à une indexation orientée requête est d'anticiper les requêtes et donc de confronter chaque document de la base à une liste de requêtes prédéfinies. La liste de requêtes forment alors le langage d'indexation.

Certains utilisent la méthode suivante pour déterminer un langage d'indexation : pour tout document du corpus, un groupe de documentalistes répond à la question "pourquoi un de nos utilisateurs serait-il intéressé par ce document ?" En répondant à cette question par une liste de termes, les documentalistes génèrent ainsi un langage d'indexation. Ensuite, le processus d'indexation procède à l'indexation grâce à un filtrage : l'indexeur vérifie chaque terme associé a priori à un document et se demande : "est-ce que l'un de nos utilisateurs intéressé par ce document utiliserait ce terme pour formuler sa requête ?".

Les problèmes majeurs posés par une indexation orientée requête résident dans son évolution face à de nouvelles requêtes, et surtout dans la difficulté à l'automatiser.

2.1.3. Quelle indexation choisir ?

Les indexations proposées dans les systèmes de recherche d'informations doivent être mixtes. En effet, les besoins des systèmes sont doubles :

- afin de servir au mieux un utilisateur, de ce point de vue, l'indexation doit être orientée requête ;

- afin de servir le spectre le plus large d'utilisateurs, le système doit disposer du maximum d'informations sur les documents. De ce point de vue, l'indexation doit donc être orientée document.

On peut malgré cela constater une prééminence des travaux de recherche en indexation orientée document. Ceci s'explique relativement facilement par le fait que d'une part, il est difficile de disposer d'une liste représentative de requêtes, alors que d'autre part le corpus de documents est disponible et donc utilisable.

Cependant, comme nous le disions dans le chapitre 1, ces approches peuvent générer un dysfonctionnement du système par rapport à ses utilisateurs en oubliant que l'indexation d'un document n'est pas unique. Cependant un certain nombre d'approches actuelles intègrent cette personnalisation du système. Ces approches définissent en fait l'indexation orientée documents comme une "couche basse" de l'indexation, à partir de laquelle ils proposent une indexation orientée requête. En effet la finalité de ces systèmes est de satisfaire pleinement le besoin des utilisateurs et donc de savoir parfaitement adapter (et non fixer) leur indexation à ces spécificités. On peut notamment citer les travaux sur l'apprentissage de M. Smail [Sma 94] ou ceux sur l'indexation dynamique de F. Paradis [Par 96a].

2.1.4. Rôle d'un terme d'indexation

Un document indexé est constitué d'un ensemble de termes d'indexation, dont le rôle est de refléter le contenu sémantique du document. Un bon terme d'indexation doit pour cela jouer une double fonction.

D'une part, il doit refléter tout ou partie du contenu du document, de façon à ce que le document soit retrouvé quand il est recherché. Cette propriété est à mettre en relation directe avec la mesure de rappel du système.

D'autre part, un bon terme d'indexation doit permettre de bien partitionner le corpus entre les documents qu'il indexe et les documents qu'il n'indexe pas. En ce sens, un bon terme d'indexation n'indexe pas tout le corpus. Cette propriété est à mettre en relation directe avec la mesure de précision du système. Ainsi un article sur les isotopes en hydrologie doit être indexé par hydrologie dans une base de données sur les isotopes, par isotope dans une base sur l'hydrologie.

2.2. Les paramètres de l'indexation

Montrer et comprendre le rôle qualitatif que joue l'indexation dans un système nécessite préalablement d'identifier de quels paramètres elle dispose. Ce chapitre a par conséquent pour objectif de les présenter.

Le processus et le langage d'indexation constituent les deux éléments de l'indexation : le langage définit la nature de l'indexation, le processus le moyen de la réaliser. La rigueur de l'extraction des termes d'indexation est le critère fondamental du processus. Nous montrons dans la partie 2.1 les différents aspects que prend cette rigueur. Les termes d'indexation sont définis par le langage d'indexation, et la partie 2.2 montre les différentes possibilités de leur définition.

2.2.1. Validation du processus d'indexation

Pour être valide, un processus d'indexation doit fournir une indexation des documents exacte. Une indexation d'un document est exacte si elle contient tous et uniquement les termes d'indexation corrects du document. Par ailleurs, l'indexation doit être consistente, et toujours fournir la même indexation pour des documents identiques.

a. Exactitude de l'indexation

Pour être exacte, une indexation doit éviter deux types d'erreurs : l'oubli d'un terme d'indexation correct, et l'indexation par un terme incorrect. Pour mesurer l'exactitude de l'indexation, on mesure la complétude de l'indexation et la pureté de l'indexation.

Complétude de l'indexation

Pour un document d, on définit la complétude de son indexation Complétude(d) par :

$$\text{Complétude}(d) = \frac{\text{nb (termes correctement affectés à d)}}{\text{nb (termes qui devraient être affectés à d)}}$$

Pour un terme t, on définit sa complétude Complétude(t) par :

$$\text{Complétude}(t) = \frac{\text{nb (documents correctement indexés par t)}}{\text{nb (documents qui devraient être indexés par t)}}$$

Une bonne indexation doit donc assurer une complétude proche de 1, pour tout document, pour tout terme. La complétude est très intéressante pour prédéterminer les performances du système, car, comme nous le montrons par la suite, elle est directement correllée à la mesure de rappel.

Pureté de l'indexation

Pour un document d, on définit sa pureté Pureté(d) par :

$$\text{Pureté}(d) = \frac{\text{nb (termes correctement rejetés pour d)}}{\text{nb (termes qui devraient être rejetés pour d)}}$$

Pour un terme t, on définit sa pureté Pureté(t) par :

$$\text{Pureté}(t) = \frac{\text{nb (documents correctement non indexés par t)}}{\text{nb (documents qui ne devraient pas être indexés par t)}}$$

Une bonne indexation doit donc assurer une pureté proche de 1, pour tout document, pour tout terme. La pureté est très intéressante pour prédéterminer les performances du SRI, car, comme nous le montrons par la suite, elle est directement correlée à l'élimination.

L'"impureté" d'un document ou d'un terme d'indexation peut également être calculée. Corollairement, l'"impureté" prédétermine la mesure d'hallucination du système.

Concrètement, l'exactitude donne une mesure interne à chaque système, mais ne permet pas de comparer les systèmes entre eux. Par exemple, un système disposant d'un langage d'indexation pauvre (avec peu de termes) peut être complet et pur à 100%, alors qu'un système disposant d'un langage riche peut être complet à 50%. En effet le

nombre de termes indexables par document augmentant, la probabilité d'inexactitudes grandit également. Mais la richesse d'un langage d'indexation peut compenser l'inexactitude ainsi générée.

b. Consistence

Assurer que des documents similaires aient toujours la même forme indexée est un problème particulièrement délicat, que le processus d'indexation soit manuel ou automatique. Lorsque le processus d'indexation est manuel, la consistence est posée sous la forme : est-ce que deux indexeurs différents indexent de la même façon le même document ? Lorsque le processus est automatique, le problème est vu sous un autre angle : est-ce que deux documents au contenu sémantique identique sont indexés de la même façon ?

La consistence peut être mesurée en comparant les réponses entre deux indexeurs différents (consistence inter-indexeurs).

Pour un document d, on définit sa consistence Consistence(d) par :

$$\text{Consistence}(d) = \frac{\text{nb (termes affectés à d par les indexeurs A et B)}}{\text{nb (termes affectés à d par les indexeurs A ou B)}}$$

Pour un terme t, on définit sa consistence Consistence(t) par :

$$\text{Consistence}(t) = \frac{\text{nb (documents indexés par t par les indexeurs A et B)}}{\text{nb (documents indexés par t par les indexeurs A ou B)}}$$

On peut également mesurer les réponses d'un même indexeur (consistence intra-indexeur) lors de sessions différentes ou bien sur des documents similaires.

La consistence est une condition nécessaire au bon fonctionnement d'un système, mais elle ne garantit pas, bien entendu, son exactitude. La consistence est un indicateur qui peut être utilisé de façon constructive, car elle constitue une prévision partielle de l'exactitude : une non-consistence peut aider à la mise en évidence de non-complétudes ou d'impuretés. Cependant, il faut l'utiliser avec vigilance : Cooper [Coo 69] a montré que la consistence ne pouvait être utilisée comme une mesure qualitative. Notamment, il montre un cas hypothétique, où une amélioration de la consistence pouvait amener à détériorer les performances du système. Ceci arrive par exemple lorsque tous les indexeurs excluent un terme correct.

Par ailleurs, Cooper [Coo 69] introduit la notion de consistence indexeur-utilisateur, permettant d'appréhender l'adéquation entre le système et l'utilisateur. Ainsi si les indexeurs indexent un document d avec un terme t, et que les utilisateurs utilisent ce même terme t dans leur requête afin de retrouver le même document d, alors la consistence indexeur-utilisateur est élevée. Par exemple, si 30% des indexeurs choisissent d'indexer un terme t à un document d, et qu'également 30% des requêtes pour ce document d contiennent ce terme t, alors la consistence indexeur-utilisateur est parfaite pour cette association entre le terme t et le document d. Cependant il n'existe pas de mesure précise permettant de calculer la consistence indexeur-utilisateur. En fait cette consistence indexeur-utilisateur est une façon de préconiser une indexation orientée requête des documents.

Les systèmes définissant une indexation dynamique tels que [Par 96a] généralisent cette notion de cohérence indexeur-utilisateur. En effet, ces systèmes typent les relations

d'indexation, et peuvent ainsi paramétrer le système selon les requêtes de l'utilisateur en adaptant la sélection des termes d'indexation potentiels.

2.2.2. Paramètres du langage d'indexation

Un langage d'indexation L peut être simple si L = {terme ∈ Vocabulaire} ou complexe si les termes d'indexation peuvent être sémantiquement reliés entre eux. Par exemple L = {terme, terme ∈ Vocabulaire ∪ (L x Opérateur x L)} définit un langage d'indexation complexe tel que nous le définissons dans les chapitres 4 et 5 de ce document. L'ensemble Vocabulaire est le vocabulaire de base d'expression des termes d'indexation, Opérateur est l'ensemble des connecteurs sémantiques entre termes d'indexation. Un langage d'indexation est dit fermé ou contrôlé si Vocabulaire et, le cas échéant, Opérateur sont fixés et ne peuvent être modifiés. Dans le cas contraire le langage d'indexation est dit ouvert. Cette première description fait donc apparaître quatre familles de langages d'indexation : les langages simples et ouverts, simples et fermés, complexes et ouverts, complexes et fermés.

Cette description permet cependant difficilement de comparer les effets qualitatifs des langages d'indexation. C'est pourquoi, à l'instar de [Soe94], nous présentons plutôt les notions d'exhaustivité et de spécificité des langages d'indexation. Les techniques de précoordination et postcoordination ainsi que l'utilisation de liens et de rôles sont également présentées.

a. Représentation du domaine : l'exhaustivité

L'exhaustivité d'une indexation définit la complétude de la relation entre les thèmes de chaque document et les termes d'indexation qui lui sont associés. Ainsi un langage d'indexation exhaustif contient les termes couvrant tous les thèmes mentionnés dans le corpus de documents. Pour être parfaitement utilisé, le processus d'indexation associé doit assurer que tous les thèmes des documents sont proprement reflétés dans les termes d'indexation associés.

L'exhaustivité se présente sous deux aspects : l'exhaustivité en terme des différents points de vue ou facettes exprimés dans le langage d'indexation, l'exhaustivité en terme d'importance des termes d'indexation retenus. Est-ce que toutes les facettes ou points de vue demandés par les utilisateurs sont représentés dans le langage d'indexation et disponibles pour la recherche ? Le degré du "oui" à cette réponse définit l'exhaustivité en termes de points de vue. L'exhaustivité en terme d'importance définit le seuil à partir duquel un concept d'un document est jugé suffisamment représentatif pour être indexé. Ainsi, pour obtenir un langage d'une haute exhaustivité, il faut nécessairement choisir un seuil bas. Beaucoup de systèmes utilisent un poids pour indiquer l'importance des termes d'indexation.

b. Représentation du domaine : la spécificité/généricité

Associer un concept d'un document à un terme d'indexation peut se faire avec un certain décalage lié à la généricité du terme d'indexation par rapport au concept. Un langage d'indexation est spécifique s'il n'indexe jamais avec un même terme d'indexation générique des concepts différents. Ainsi indexer un document décrivant le tri à bulle par "tri à bulle" est plus spécifique que l'indexer par "tri en n^2", ceci étant plus spécifique que l'indexer par "tri". La spécificité est généralement incompatible avec une utilisation du système par différentes catégories d'utilisateurs. Aussi un système utilise un langage spécifique afin de servir des utilisateurs généralement spécialistes du corpus, par exemple des médecins pour un corpus médical [Ber 88].

Certains systèmes parlent d'indexation profonde (deep indexing) ou superficielle (shallow indexing). Une indexation est profonde si elle est à la fois exhaustive et spécifique. Une indexation superficielle est une indexation non spécifique, et produit peu de termes (relativement génériques) par document indexé. Les performances

qualitatives du système à indexation superficielle peuvent être médiocres, notamment en termes de précision et d'élimination, mais son indexation est très rapide et son coût minimal.

c. Précoordination ou postcoordination du langage d'indexation

Un langage d'indexation est défini par des termes simples (des mots clés) ou bien par des termes complexes (des syntagmes, ou des termes reliés entre eux par des opérateurs sémantiques). On parle alors dans ce second cas de précoordination, dans le sens où les termes sont reliés entre eux avant l'interrogation.

Les langages d'indexation définissant des termes simples peuvent être postcoordonnés, c'est-à-dire que le système coordonne les termes d'indexation a posteriori lors de requêtes utilisateurs.

Par exemple, le livre de G. Salton et M.J. Mc Gill [Sal83] est indexé par "automatic information retrieval" dans un langage précoordonné. Bien évidemment, la réponse à la requête "automatic information retrieval" sera parfaite ! Il faut cependant que le système puisse déduire que ce document est une réponse à la requête "information retrieval", et donc définisse une stratégie de "cassure" des termes d'indexation telle que celle proposée par [Ker 84].

Par ailleurs, retrouver ce document par la requête "information and retrieval" alors qu'il est indexé par "automatic" "information" "retrieval" est une conséquence de la postcoordination du langage.

d. Palliatifs à une faible précoordination : les liens et rôles

De façon très adhoc, certains systèmes proposent d'associer par des liens les termes d'indexation. Par exemple, Medline propose des "links" permettant de contextualiser les termes d'indexation entre eux. Ainsi pour le document "the effects of alcohol dependence on experimentation with cocaine", plutôt que d'indexer simplement par les termes d'indexation "alcohol", "cocaine", "drug dependance" et "experimentation", l'indexeur peut les associer par des liens de la façon suivante :

drug dependance : alcohol

experimentation : cocaine

Lorsqu'ils existent, les liens sont en fait un palliatif à une faible précoordination du langage d'indexation, et permettent d'exprimer des relations entre des termes plus fortes que la simple co-occurrence de ces même termes dans un document.

De façon similaire aux liens, certains systèmes introduisent des rôles permettant de spécifier l'utilisation de certains termes d'indexation. C'est le cas des "subheadings" du système Medline. Ainsi ce système propose, entre autres, les "subheadings" "therapeutics uses" (utilisations thérapeutiques) et "adverse effects" (effet contraire) qui permettent de distinguer le rôle d'une substance chimique dans un document. Par exemple, pour un document décrivant le traitement du diabète, l'indexeur choisira :

diabetes -drug treatments

insulin- therapeutic use

triglycerin-adverse effects

Les rôles sont souvent utilisés pour donner à des termes larges un sens spécifique.

e. Base de connaissance

La disponibilité d'une base de connaissance est une ressource intéressante pour l'indexation. Le plus fréquemment, les bases utilisées sont des thésaurus représentant la hiérarchie des termes d'indexation.

Une base de connaissance fournit un canevas utile à l'indexation, car son utilisation est normalisatrice des termes d'indexation utilisés : on n'utilise comme termes d'indexation que des termes de la base, et leur choix se fait selon un mode défini dans le processus d'indexation.

2.2.3. Conclusion

Avant de regarder les effets de ces caractéristiques sur les performances du système, observons leurs effets sur l'indexation elle-même par le nombre de termes d'indexation que l'on peut avoir par document. Quatre facteurs interviennent pour fixer ce nombre : la nature du document, le degré de précoordination du langage d'indexation, l'exactitude de l'indexation, et les règles d'indexation de chaque système.

A même niveau d'exhaustivité, un document long a normalement besoin de plus de descripteurs qu'un document court. Cependant, toute règle possède au moins une exception : pour une exhaustivité moyenne, un journal économique d'un vingtaine de pages peut être indexé par une vingtaine de termes, alors qu'un livre d'économie, couvrant tous les domaines de l'économie, est indexé par "économie", "générale".

Pour un terme "les méthodes d'apprentissage de la lecture à l'école primaire" dans un langage précoordonné, un système postcoordonné contient trois concepts fondamentaux "méthodes d'apprentissage", "lecture", "école primaire". Il semble donc que plus le langage est précoordonné, moins le nombre de termes est élevé.

L'exactitude, ou plutôt l'inexactitude de l'indexation, bien entendu est source de termes corrects oubliés ou incorrects indexés.

Les règles d'indexation de chaque système peuvent ajouter des termes d'indexation. Par exemple, sans pour autant remettre en cause leur exhaustivité, certains systèmes ajoutent systématiquement les termes génériques de tous les termes d'indexation.

2.3. Effets de l'indexation sur les performances d'un système de recherche d'informations

Cette partie a pour objectif de montrer les effets qualitatifs de l'indexation dans un système binaire et non-interactif. Dans un système binaire, pour chaque requête, le corpus est partitionné en deux sous-ensembles P (les documents pertinents) et ¬P (les documents non pertinents). Il n'existe donc pas de document partiellement pertinent. Nous ne traitons pas non plus l'interaction entre le système et l'utilisateur : il n'existe pas de reformulation de requête (relevance feedback). Les systèmes binaires et non-interactifs sont très courants, et leurs utilisateurs sont des professionnels (des documentalistes ou des bibliothécaires) particulièrement habitués à leur rigidité.

Nous montrons que les différents paramètres de l'indexation interviennent dans les performances qualitatives du système. Cette partie montre pour chacun de ces paramètres ses apports bénéfiques, et aussi les déficiences qu'il peut générer.

2.3.1. Exactitude

C'est LE critère pour de bonnes performances qualitatives d'un système de recherche d'informations. Avec une indexation exacte, s'établit une confiance entre l'utilisateur et le système. Cette relation est fondamentale, et donne toute la plus value à un système de recherche d'informations par rapport à d'autres systèmes tels que, par exemple, les systèmes plein texte.

Imaginons une requête d'un seul terme, identifié parfaitement par un terme du langage d'indexation. Imaginons que l'importance de l'indexation corresponde au besoin de l'utilisateur. Alors, dans ce cas imaginaire, le rappel est égal à la complétude du terme, et l'élimination à la pureté de ce terme.

Par ailleurs, oublier des termes diminue le rappel, puisque les interrogations sur ce terme ne donneront pas tous les documents pertinents. Trop indexer avec un même terme peut détériorer l'élimination. Ceci arrive quand des documents sont indexés par un même terme, alors que les concepts des documents sont distincts.

2.3.2. Exhaustivité

Intuitivement, on pourrait penser que l'exhausitivité améliore le rappel aux dépens de l'élimination.

Prenons l'exemple de la requête " donnez-moi les documents parlant de X". Dans le cas d'un système S1 peu exhaustif (à seuil d'importance élevé), les documents parlant peu de X ne sont pas retrouvés. Ainsi le rappel et l'élimination sont faibles. Dans le cas d'un système exhaustif S2 (à seuil d'importance faible), tous les documents traitant de X sont donnés en réponse, même ceux pour qui ces termes sont peu importants. Ainsi le rappel est meilleur, mais l'élimination se détériore.

Cet argument n'est pas toujours vérifié. [Sal 83] ajoute que même si une forte exhaustivité peut assurer un fort rappel, cela se fait de toute façon aux dépens de la précision. En effet certains termes marginaux peuvent être retrouvés, ou bien certains termes d'indexation peuvent recouvrir différents concepts.

a. Rôle des facettes

Ajouter des facettes à une requête peut permettre d'améliorer l'élimination. En effet, si l'on développe beaucoup de facettes ou points de vue, on étend les types de requêtes du système. Cela permet à l'utilisateur d'atteindre une meilleure élimination, peut-être dans certains cas aux dépens du rappel.

Par exemple, si un système rajoute dans son indexation la facette allaitement. Lors d'une requête Les facteurs d'allergies chez les bébés élevés au lait maternel, l'utilisateur peut dorénavant ajouter un terme allaitement maternel à sa requête pour en renforcer l'élimination. Cependant la nouvelle requête peut ne pas retrouver certains documents, selon le seuil d'importance et/ou la complétude du terme allaitement maternel dans les documents. Les systèmes ont généralement des taux de complétude faibles lorsque les indexeurs ne sont pas suffisamment entrainés à reconnaître l'occurrence d'une nouvelle facette. Ainsi, l'ajout des facettes à une requête doit être considéré comme une option, qui bien qu'améliorant l'élimination peut détériorer le rappel.

b. Rôle de l'importance

Le rôle de l'importance est plus complexe. Comme nous le disions en introduction de cette partie, à formulation de requête constante, une plus grande exhaustivité produit un rappel au moins supérieur, une élimination au plus égale à ceux d'une exhaustivité moindre.

Considérons une requête "tumeur du cerveau" formulée par deux utilisateurs différents. L'utilisateur u1 est un médecin spécialiste qui veut connaître les derniers développements dans la recherche en "tumeur du cerveau". L'utilisateur u2 est un étudiant en médecine qui souhaite un état de l'art sur tous les aspects de la tumeur du cerveau. Ces deux utilisateurs formulent la même requête et obtiennent les mêmes résultats, mais selon l'exhaustivité du système seront satisfaits ou mécontents des réponses. Considérons deux systèmes : le système S1 à faible exhaustivité (importance forte), le système S2 à haute exhaustivité (importance faible). Pour une même requête, le système S2 retrouve les mêmes documents que le système S1 ainsi que les documents qui font plus ou moins allusion à la requête.

Dans S1, u1 retrouve tous les documents qui l'intéressent, avec éventuellement un peu de bruit. Le rappel et l'élimination sont bons. Les importances de S1 et u1 correspondent : quand l'utilisateur u1 demande une importance forte sur un terme,

l'indexeur a également eu une importance élevée sur ce terme. Toujours dans le cas du système S1, beaucoup de documents pertinents pour u2 ne sont pas retrouvés, mais par contre le système rejette la plupart des documents jugés non-pertinents : le rappel est mauvais, et l'élimination est bonne. L'importance de S1 est élevée, alors que celle de u2 est faible.

Dans S2, peu de ces documents intéressent u1. Le rappel est assez bon, mais l'élimination est mauvaise. L'importance du système S2 est basse, celle de u1 est élevée. Pour u2, le système S2 est idéal. Le rappel et la précision sont bons. L'importance de S2 et de u2 sont faibles.

La figure 2.1 nous résume cet exemple :

pour une même requête	S1 faible exhaustivité	S2 haute exhaustivité
u1 spécialiste	rappel bon, élimination bonne	rappel honnête, élimination mauvaise
u2 généraliste	rappel mauvais, élimination bonne	rappel bon, élimination bonne

Figure 2.1 : effet de l'exhaustivité sur le rappel et la précision pour une même requête

Par ailleurs, une haute exhaustivité peut être exploitée par l'utilisateur du système pour améliorer l'élimination. Prenons un exemple en considérant une requête dans laquelle on souhaite obtenir des documents qui traitent essentiellement de "tumeur du cerveau" et éventuellement de "métastase". Avec une faible exhaustivité, la requête doit être formulée "tumeur du cerveau", car ajouter "métastase" rejetterait beaucoup trop de documents pertinents. Avec un forte exhaustivité, la requête doit être "tumeur du cerveau (importance élevée) et métastase (quelle que soit la valeur de l'importance)". Ceci améliore grandement l'élimination sans pour autant toucher au rappel. Pour de telles requêtes, les utilisateurs des systèmes à faible exhaustivité ont souvent recours à une recherche plein-texte sur les documents retrouvés, afin de sélectionner ceux contenant la chaîne de caractères "métastase".

2.3.3. Spécificité/Généricité

Une haute spécificité est utilisée pour assurer une grande précision au système, puisque tous les documents retrouvés sont supposés être pertinents. En fait, comme nous le montrons ci-après, les effets de la spécificité dépendent de la spécificité de la requête. Des termes spécifiques peuvent améliorer l'élimination, mais pour une recherche large la spécificité pose des problèmes pour obtenir de bonnes performances.

a. La spécificité associée à une recherche spécifique

Son intérêt est évident, les résultats excellents, mais dépend directement de l'exactitude de l'indexation. Et une indexation spécifique est source d'erreurs d'indexation. Dans une indexation spécifique, en cas de doute, les indexeurs attribuent souvent des termes trop génériques aux documents. Si la requête initiale est modifiée avec un terme plus générique, tous les documents pertinents sont trouvés, mais également des documents non pertinents : cela augmente le rappel au détriment de l'élimination.

b. La spécificité associée à une recherche large

Si le système possède une base de connaissance correcte, tout se passe très bien : la recherche est transformée en une recherche des termes spécifiques. Ni les performances du système ni les efforts de formulation ne sont altérés.

Mais si ceci n'est pas possible, il est nécessaire de procéder manuellement à une expansion de la requête en y ajoutant tous les termes proches ou spécifiques. Et si la hiérarchie du langage d'indexation est incomplète voire incorrecte, les choses se compliquent encore plus. A partir du moment où l'utilisateur ne peut obtenir de listes de termes proches, le rappel diminue.

Par exemple, les termes "cuisine lyonnaise", "cuisine dombiste", "cuisine dauphinoise", "cuisine savoyarde" doivent tous être inclus dans une recherche sur "cuisine française". Si le système procède automatiquement à cette nouvelle recherche, ou si l'utilisateur compense manuellement les déficiences du système, en modifiant lui-même la requête, tout se passe bien. Sinon, faute de sélectionner tous les documents pertinents, le rappel diminue. L'élimination ne bouge pas.

2.3.4. Base de connaissance

La base de connaissance donne un canevas à l'indexeur, et de ce fait une certaine assurance à l'exactitude. Ce qui, comme nous l'avons montré, améliore les performances de la recherche.

La base de connaissance a un effet direct sur la recherche, en donnant à l'utilisateur la possibilité de formuler une requête en choisissant les termes les plus adéquats. Elle peut également assister l'utilisateur dans l'expression de son besoin, lui faire découvrir des ramifications ou des aspects nouveaux.

Un certain nombre de systèmes utilisent la hiérarchie de la base de connaissance pour faire des recherches inclusives, c'est-à-dire qu'un terme de la requête retrouve non seulement les documents indexés par lui-même mais également les documents indexés pas des termes plus spécifiques. Nous venons de le voir dans le cas d'une requête large sur un langage spécifique. La recherche inclusive applique donc la hiérarchie de la base pour permettre une recherche performante qui améliore considérablement le rappel.

2.3.5. Précoordination / postcoordination

Lorsqu'un terme précoordonné répond exactement à un terme de la requête, alors la précoordination joue un rôle prépondérant en améliorant efficacement l'élimination. Cependant, un document indexé par "methods of instruction", "reading" et "second grade" répond à un spectre plus large de réponses que s'il est indexé par "methods of instruction for reading in elementary schools".

De ce fait, la postcoordination est plus utilisée dans les systèmes à spectre "large" d'utilisateurs, et la précoordination est plus utilisée dans les systèmes dédiés. Cette conclusion n'est pas surprenante, étant donné qu'un système de recherche d'informations à spectre "large" d'utilisateurs doit supporter des formulations exprimées de façons très diverses. A partir du moment où les utilisateurs sont hétérogènes, les requêtes peuvent aller depuis la demande d'états de l'art sur un domaine jusqu'à des requêtes très détaillées. Dans de telles circonstances, une analyse précoordonnée excessive serait trop spécialisée pour beaucoup d'utilisateurs. Si le système possède une base de connaissance, un processus de "cassure" des termes d'indexation, alors le système peut améliorer ses performances (voir partie précédente).

2.3.6. Liens et rôles

Leur intérêt est de permettre la formulation de requêtes discriminantes. Par contre, leur utilisation pour d'autres raisons peut aussi créer un malentendu entre utilisateurs et indexeurs. Et alors, le rappel se détériore. Beaucoup de liens sont très clairs et faciles d'utilisation, alors que les rôles sont souvent d'une utilisation plus délicate. Aussi, afin d'éviter des inexactitudes d'indexation leurs règles d'utilisation sont souvent très strictes. Par exemple on impose un seul rôle par terme d'indexation alors que souvent un terme

a plusieurs rôles : un terme peut en effet avoir aussi bien le rôle "utilisation thérapeutique" que "effet contraire", comme on peut le trouver dans le système Medline.

2.4. Conclusion

Nous venons de montrer l'impact qualitatif de l'indexation dans un système, mais n'oublions pas que le système dispose d'autres paramètres (recherche inclusive, utilisation des indicateurs de rôles ou de bases de connaissance) pour améliorer ses réponses. Par ailleurs, il faut également rappeler, comme nous le disions dans l'introduction, la valeur toute relative qu'il faut accorder aux mesures des systèmes. Ainsi tout comme les mesures classiques telles que le rappel ou l'hallucination, les mesures de l'indexation ne sont pas calculables. Ainsi le nombre de termes qui devraient être affectés à un document n'existe pas, et la complétude est donc incalculable. De notre point de vue, ces mesures devraient également être redéfinies, et peut-être observées sous un autre angle. Notamment les mesures de consistence ou de complétude pourraient être reconsidérées de façon complètement différente. Une inconsistence n'est-elle pas la trace d'une autre perception d'un document ? Une non-pureté pourrait donner les prémisses d'une bonne mesure de "découverte" ? De ce point de vue, ce qui apparaissait comme un problème dans les systèmes, ne deviendrait-il pas une nécessité des applications actuelles ?

Les données telles que par exemple les images introduisent une notion à la fois plus générale, et plus incertaine de l'indexation. Plus générale, car l'indexation d'une image peut aller de l'expression d'éléments du signal (des couleurs par exemple), jusqu'à la sémantique des objets contenus. Plus incertaine, car une image peut par exemple être floue. Outre le fait que la qualité de l'image peut faire partie de son indexation, sa valeur introduit des alternatives dans l'indexation. Ainsi telle partie de l'image peut être difficilement identifiée: cela peut être telle chose ou telle autre chose, mais pas les deux. L'exactitude de l'indexation devient alors impossible à établir. Par contre l'inconsistence peut jouer un rôle déterminant en établissant deux identifications distinctes d'un même objet.

Par ailleurs, un long débat existe depuis longtemps sur la dualité entre processus automatiques ou manuels. Les partisans des premiers argumentent de leur faible coût, ceux des seconds de leur utilisation massive dans les systèmes. Par exemple, [Sal 83] pense que les méthodes d'indexation automatiques simples, telles que celle de Smart, sont rapides et peu coûteuses, et que les performances qu'elles donnent sont tout aussi bonnes que celles des indexations manuelles basées sur des langages contrôlés. Cependant, l'indexation manuelle est encore de nos jours très utilisée ! Une raison essentielle réside dans le fait que peu de données sont disponibles électroniquement dans les services qui les indexent (les bibliothèques, par exemple). Par ailleurs, chaque service a des habitudes d'indexation manuelle liées aux besoins spécifiques de ses utilisateurs, et essaie dans la mesure du possible de préparer le travail de recherche des utilisateurs. Ainsi ces systèmes proposent à leurs utilisateurs un mode de fonctionnement adapté à leurs besoins d'informations. Bien évidemment, leurs performances chutent dès que l'on sort des habitudes d'utilisation de ces systèmes. De notre point de vue, l'atout essentiel de l'indexation automatique est de pouvoir être dynamique et donc de s'adapter à chaque requête d'utilisateur. Ce qui est une plus-value importante, même par rapport aux processus orientés utilisateurs, car ils peuvent ajuster leur indexation aux besoins de chaque requête, aux besoins de chaque utilisateur. En devenant dynamique, l'indexation automatique devient complètement orientée utilisateurs, alors que l'indexation manuelle restera orientée pour un type d'utilisateurs.

Cependant ces réflexions n'apparaissent plus de nos jours en amont des systèmes. Actuellement, le compromis entre les applications visées et les traitements non seulement nécessaires mais surtout possibles sur les données multimédia dicte le mode

de l'indexation : réaliser une indexation automatique n'est pas possible, lorsque la sémantique nécessaire à l'application est profonde. Le chapitre 3 a pour objectif de présenter cette problématique apportée par l'indexation de données multimédia, et les grandes familles actuelles d'indexation.

3. Recherche d'informations et données multimédia

3.1. Problématique

Cette première présentation de la recherche d'informations est encore plus remise en question de nos jours, dès lors que l'on considère les corpus actuels constitués de données multimédia. D'une part ce sont des données volumineuses, et les corpus traités sont de ce fait de plusieurs téra-octets. Mais d'autre part les données multimédia apportent des difficultés nouvelles à la recherche d'informations :

- la distance entre le signal des données et leur sémantique est élevée. Pour comprendre ceci, regardons l'exemple contraire des documents textuels. Leur signal est une chaîne de caractères. Les mots, les phrases, briques de base nécessaires à l'expression de la sémantique des textes, s'extraient aisément. Dans le cas des images, le signal est une matrice de pixels, alors que la sémantique consiste à identifier symboliquement différents éléments de l'image. Par exemple, là où la matrice de pixels indique du vert et du marron, la sémantique ajoute, le cas échéant, le fait qu'il y a un arbre.

- diverses interprétations complémentaires. Toute donnée multimédia véhicule des informations diverses : ainsi une image contient des couleurs (du vert, du marron, …), du contraste, des contours d'objets (le tour de l'arbre, …), des objets symboliques (un arbre, un tronc, un feuillage, …), etc. Et toutes ces informations participent à la description de l'image.

- la dimension temporelle des données son et vidéo. Cette dimension peut être vue comme la structure séquentielle des données son et vidéo : telle information se situe avant telle autre, telle information apparait fréquemment, telle autre apparait peu mais longuement. Par ailleurs, les documents vidéo et son se visualisent séquentiellement et leur durée peut être longue. Les demandes des utilisateurs s'adressent alors surtout à des parties seulement, et non à tout le document : il faut dans ce cas identifier un découpage temporel du document.

Ces contraintes expliquent les difficultés rencontrées lors de l'indexation et de l'interrogation de données multimédia. Nous ne revenons pas sur l'impossibilité de mesurer le rappel, l'élimination ou l'hallucination : les arguments que nous avons avancés dans la partie précédente deviennent encore plus forts dans ce contexte.

3.2. Indexation des données multimédia

Au niveau de l'indexation, il est délicat d'exprimer le contenu sémantique des données multimédia, et surtout d'en fixer les limites de façon cohérente. En effet, pour un même document multimédia, des informations de nature très différente, correspondant à diverses interprétations du contenu du document, doivent être indexées. En outre, seulement une partie de ces informations peut être extraite du signal du document : les couleurs peuvent être extraites automatiquement de la matrice de pixels de l'image, mais l'identification d'un arbre et son association à une couleur sont généralement explicitées par un indexeur humain.

Il existe deux catégories d'approches pour indexer des données multimédia : les approches ascendantes et descendantes. La distinction entre ces deux approches se situe au niveau de la prise en compte du signifiant des documents. Les approches ascendantes traitent le signifiant en tant que signal numérique, et le manipulent directement. Dans ces approches, le signifié reste alors assez proche du signal lui-même : il s'agit de valeurs numériques. Les approches descendantes manipulent le signifiant à partir de sa perception par un humain. De plus l'expression du signifié (en langage d'indexation) est généralement complexe et représentative d'une application particulière. Nous détaillons ces deux types d'approches dans les paragraphes suivants.

3.2.1. Les approches ascendantes

Ces approches utilisent le signal des données. Les recherches de termes d'indexation sont réduites à des calculs issus des signaux (invariants géométriques des images par exemple). Ces approches intéressent généralement les applications grand public ou orientées rappel, c'est-à-dire les applications où très peu de connaissance sémantique de l'application est exigée. Elles offrent l'avantage indéniable d'être automatisées.

Parmi les approches ascendantes, citons les travaux de [Han 96]. Han et Myaeng proposent un système d'indexation et d'interrogation d'images. Le langage d'indexation est constitué de vecteurs associés à chaque image. En effet chaque objet d'une image est caractérisé par un vecteur f_{shape} à cinq dimensions : f_{form}, f_{rect}, $f_{ellipse}$, f_{ecc}, f_{bend}. Ces dimensions permettent de décrire l'objet : f_{form} indique la surface de l'objet, f_{rect} sa différence avec le rectangle le plus proche de ses dimensions, $f_{ellipse}$ sa différence avec l'ellipse la plus proche de ses dimensions, f_{ecc} son étendue, et f_{bend} ses irrégularités. Le processus d'indexation se décompose en deux étapes : l'extraction des objets des images, le calcul de la valeur de leurs vecteurs. Les objets des images sont extraits automatiquement en utilisant l'algorithme de "modified snake" [Kim 94], et leurs vecteurs sont initialisés. La seconde étape a pour objectif d'homogénéiser ces premiers calculs en tenant compte des ressemblances entre les images du corpus : l'objectif est alors d'assurer que des images similaires ont des indexations proches, et réciproquement des images distinctes ont des indexations différentes. Pour cela le système utilise un réseau de neurones de type SOM (Self Organizing Feature Map) qui regroupe les images par similarité et permet un apprentissage en modifiant en conséquence la pondération de chaque image. Une indexation est stable lorsqu'elle n'évolue plus d'un apprentissage à l'autre. Sur un corpus-test de 150 images, 2000 apprentissages ont été nécessaires avant que l'indexation se stabilise. Le système permet aux utilisateurs de naviguer dans les données (ainsi regroupées par le réseau de neurones), d'interroger soit en donnant une image en guise de requête, soit en indiquant directement le vecteur des objets cherchés. Enfin les auteurs ont mesuré la valeur de discriminance de chaque dimension : toutes sont nécessaires et $f_{ellipse}$ s'avère être la plus discriminante. Pour cela la similarité moyenne entre images est calculée, puis on en enlève une des cinq dimensions et on recalcule la nouvelle similarité moyenne. Une valeur positive de la différence entre la première et la seconde similarité indique que la dimension est discriminante, une valeur négative indique qu'avec cette dimension, les images se ressemblent plus, et cette dimension atténue donc les distinctions.

Ce système est particulièrement intéressant car c'est l'un des rares (le seul ?) systèmes ascendants à disposer d'un processus d'indexation global : les données sont tout d'abord indexées individuellement, et ensuite le système assure la consistence globale de son indexation en l'homogénéisant sur tout le corpus : les données similaires doivent avoir des indexations similaires, les données différentes ne doivent pas répondre, pour la même raison, à une même requête. Cette seconde étape est très délicate à mettre en œuvre dans les approches ascendantes, car les termes d'indexation sont des regroupements de valeurs numériques. Dans ce contexte, définir une stratégie permettant de décider lesquelles de ces valeurs doivent être modifiées afin d'homogénéiser l'indexation est très difficile : il est quasiment impossible dans ces systèmes d'établir une relation entre ces valeurs et le symbolisme de l'image.

Certaines approches ascendantes utilisent une certaine connaissance de l'application ou du corpus, et permettent une indexation plus spécifique des documents. C'est le cas notamment des travaux de Wu et Narasimhalu [Wu 94] sur la reconnaissance de visages, ou les travaux de Peter Schaüble [Sch 92] sur le son.

3.2.2. Les approches descendantes

Ces approches définissent, à partir des besoins de l'application cible, le langage d'interrogation et d'indexation nécessaires. Ceci permet d'ajuster le système vers, par exemple, des recherches précises. Les processus d'indexation sont généralement manuels.

Les travaux de [Hje 94] montrent un système de modélisation et d'interrogation de données vidéo. [Hje 94] modélise de façon très détaillée les informations vidéo : représentation logique/physique, décomposition de la séquence en sous-séquences, annotations des sous-séquences pouvant indiquer les personnages, le lieu, ou les événements apparaissant. La représentation des vidéos dans ce modèle est alors manuelle, afin de renseigner au mieux les données. Le système permet à l'utilisateur d'interroger les données vidéo selon cinq types de requêtes : la navigation structurelle (par une table des matières), la navigation par clip (les vidéos sont résumées par quelques images, à partir desquelles l'utilisateur peut les visionner), la navigation par résumé (les utilisateurs peuvent lire le résumé des vidéos et ainsi faire leur choix), les requêtes de contenu simple (par les descriptifs faits dans les vidéos ou leurs sous-séquences), et les requêtes thématiques (par des annotations générales sur chaque vidéo). Le système a été développé pour un corpus de journaux télévisés, et est donc adapté notamment à ces besoins spécifiques.

3.3. Pour quelle recherche d'informations ?

Les systèmes basés sur une indexation ascendante proposent aux utilisateurs de poser des requêtes par analogie avec des documents : l'utilisateur dispose d'une palette de documents-types à partir desquels il peut identifier, voire composer, sa propre requête. Cette démarche semble effectivement, dans ce contexte, la seule approche possible : l'indexation produit en effet des informations sous forme numérique.

Et les systèmes basés sur une indexation descendante proposent aux utilisateurs de formuler leur requête sous la forme d'une description externe, particulièrement précise.

Les fonctions de correspondance utilisées sont soit ad hoc soit des réutilisations des modèles classiques de recherche d'informations (modèle booléen ou vectoriel).

De même que l'indexation, l'interrogation de corpus multimédia est très délicate. Car il reste très difficile pour un utilisateur de formuler son besoin d'informations, et pour le système d'identifier ce besoin tant la diversité d'interprétations de la requête et des documents est grande.

3.4. Retour sur la présentation de nos travaux

Sans revenir sur notre argumentaire développé dans l'introduction, rappelons que nos travaux de recherche se situent dans un double contexte : indexations descendantes pour des systèmes orientés précision, et systèmes dynamiques adaptables à chaque situation de recherche.

3.4.1. Indexations descendantes pour des systèmes orientés précision

Dans ce contexte, nos travaux ont permis la définition d'approches descendantes pour des textes et des images. Dans le cadre du système RIME, nous avons défini un langage d'indexation complexe décrivant le contenu de textes médicaux, ainsi que le processus d'indexation automatique associé. RIME permet d'indexer automatiquement des comptes rendus médicaux (des textes) en dérivant de façon uniforme et très détaillée leur contenu sémantique. Les médecins utilisateurs peuvent, comme ils le souhaitaient dès le départ de ce projet, interroger le corpus en donnant des requêtes très précises. Dans EMIR2, nous avons défini un langage d'indexation pour un corpus d'images. Dans ce contexte, une image est considérée à la fois comme un objet complexe et multifacettes. Ainsi chaque image est décrite par l'association d'un ensemble de vues

dans une structure homogène regroupant les facettes jugées pertinentes pour son contenu. L'indexation d'une image est produite par des indexeurs humains, aidés et contrôlés dans leur tâche par une interface d'indexation que nous avons construite.

3.4.2. Le prototype PRIME

Les travaux que nous venons de présenter ont été validés expérimentalement dans le cadre du prototype PRIME, que nous avons développé sur un corpus médical et pour des médecins. PRIME permet le stockage, la manipulation, l'indexation et la recherche de données médicales (structurées et multimédia) par des médecins. Pour cela, PRIME offre un accès aux documents sur des critères externes (un nom ou un numéro de patient, ...), la navigation sur les données médicales, et une interrogation par le contenu des données médicales.

3.4.3. Vers un système de recherche d'informations dynamique

La plupart des systèmes instancient ce principe en proposant à l'utilisateur de formuler une requête et d'observer les réponses du système. En cas d'insatisfaction, le seul paramètre de l'utilisateur est de reformuler sa requête jusqu'à trouver celle par laquelle il trouvera la réponse à son besoin initial d'informations. Ainsi la plupart des systèmes se présentent comme des entités monolithiques proposant, au travers d'une fonction de correspondance, une et une seule liste de réponses à une même requête.

De notre point de vue, un système de recherche d'informations doit être dynamique. En ce sens, un système n'est plus associé à une correspondance, dans laquelle le seul paramètre est la requête : le système doit être orienté utilisateur et savoir trouver face à un besoin d'informations, aux réactions de l'utilisateur, la réponse adaptée à cette demande. Le système doit ainsi être capable d'adapter sa pertinence à celle de l'utilisateur. Nous montrons comment ceci peut être pris en compte au niveau de la reformulation (relevance feedback).

Parallèlement à la correspondance, l'indexation doit également être dynamique, en s'adaptant au besoin d'information de l'utilisateur. Dans ce contexte, nos travaux ont permis de généraliser la notion de langage d'indexation en l'encapsulant dans un langage de représentation générale des documents. Nous avons défini l'indexation comme un ensemble de règles de déduction sur ce langage de représentation. Ainsi le processus d'indexation n'est plus un processus préalable construisant les documents indexés. Le choix des règles d'indexation se fait dans une phase de recherche d'informations où le besoin de l'utilisateur induit les règles de déduction qui lui sont nécessaires. Adaptant l'indexation aux spécificités de chaque recherche, ces travaux sont un premier pas vers l'association dynamique entre indexation et correspondance.

Ce sont ces cinq points que les chapitres 4, 5, 6, 7 et 8 décrivent maintenant.

4. Indexation de textes : les comptes rendus médicaux dans le système RIME

RIME est un système de recherche d'informations médicales [Ber 88], [Ber 89], [Ber 90], [Nie 90]. Le corpus traité est constitué d'images et de comptes rendus médicaux. Chaque image est associée à un compte rendu, rédigé en langue naturelle et décrivant l'interprétation médicale qu'a le radiologue de l'image. Chaque compte rendu constitue ainsi un descriptif précis de l'image à laquelle il est associé. Indexer ces comptes rendus permet la recherche des images par leur contenu sémantique et doit permettre par exemple de retrouver des images dans lesquelles "on observe des métastases dans le poumon gauche".

Dans ce contexte, nous avons défini un langage d'indexation, que nous appelons dans ce contexte un modèle sémantique, des comptes rendus médicaux ainsi que leur processus d'indexation automatique. Le modèle sémantique assure la représentation spécifique et exhaustive des comptes rendus médicaux : il permet ainsi la prise en compte complète des informations rédigées par les radiologues. Automatiser le processus d'indexation est possible dans un tel contexte, car la langue d'expression des radiologues est une langue de spécialité, c'est-à-dire une langue relevant d'une langue naturelle (le français en l'occurrence) mais en y ajoutant des spécificités de vocabulaire (le vocabulaire utilisé est un vocabulaire médical de radiologistes), d'expressions syntaxiques et sémantiques (expressions et tournures médicales essentiellement).

4.1. Le modèle sémantique

4.1.1. Présentation

Les comptes rendus médicaux sont des documents courts, de moins d'une page, produits par des médecins radiologues. Ils comportent à la fois des attributs externes (le nom du patient, leur date de naissance, la date de l'examen, ...) et des attributs internes (des informations techniques sur l'examen lui-même, une description de l'image, et, dans quasiment tous les cas, un diagnostic médical). Alors que les attributs externes décrivent le contexte du compte rendu médical, les attributs internes en constituent le contenu sémantique. Les attributs externes sont décrits dans des formats prédéfinis, les attributs internes sont rédigés dans une langue de spécialité. Cette langue de spécialité se traduit par l'utilisation :

- d'une part d'un vocabulaire précis : ce sont des termes techniques liés à la technologie utilisée pour réaliser l'examen, ou bien des termes médicaux permettant la description des observations et du diagnostic médical ;

- d'autre part un style concis et direct : les textes sont courts, sémantiquement denses et se présentent déjà sous la forme d'un résumé. L'analyse de ces textes montre que peu d'ambiguïtés apparaissent. Aussi l'automatisation de leur analyse en vue d'une indexation est tout à fait envisageable. Par ailleurs, pour les radiologues, toute information contenue dans les comptes rendus doit nécessairement être représentée dans les documents indexés. En effet, les médecins souhaitent retrouver les documents médicaux à partir de toute information contenue initialement.

La recherche d'informations souhaitée est donc définie pour des spécialistes d'un domaine. Elle doit donc être orientée précision, en s'appuyant sur un corpus représenté de façon exhaustive et spécifique.

4.1.2. Principes généraux

En utilisant les propriétés linguistiques des comptes rendus médicaux, nous pouvons définir un processus d'indexation automatique capable de produire à partir des documents source des documents indexés respectant notre modèle sémantique.

La définition du modèle sémantique de RIME a été un long travail effectué en collaboration avec les médecins du CHU de Grenoble.

Lors de la définition d'un modèle sémantique, la première difficulté que l'on rencontre est de déterminer une limite au processus de compréhension des textes. Dans le contexte de RIME, ceci a été fixé par le niveau de précision requis par les médecins. Aussi le modèle sémantique représente-t-il les connaissances factuelles extraites des textes médicaux, et sa granularité (sa spécificité) a été définie par les médecins. Par exemple, le terme "poumon" est considéré comme un concept de base de notre modèle, sans aucune information indiquant qu'il s'agit d'un élément du système respiratoire. Ce type de connaissance, qui en l'occurrence caractérise sémantiquement le "poumon", est connu via un thésaurus lui-même accessible lors de la phase de recherche.

La seconde difficulté concerne la représentation de la connaissance dans ce contexte. Le modèle que nous avons défini dans RIME est issu des dépendances conceptuelles de Schank [Sch80] [Sch81]. Il a été enrichi par la définition d'une grammaire appelée langage conceptuel qui permet le contrôle des concepts indexés et d'utiliser complètement ces concepts lors de la phase de recherche. Ainsi le modèle se présente schématiquement de la façon suivante :

1) La précoordination est exprimée par des structures d'arbres binaires :

- les nœuds non-terminaux sont des opérateurs sémantiques explicitant la relation entre des concepts de plus bas niveaux, représentés par les deux sous-arbres. Par exemple, l'opérateur sémantique "dû-à" établit une relation causale entre ses deux opérandes ;

- les nœuds terminaux correspondent à des termes médicaux ou techniques, c'est-à-dire des mots simples ou composés du domaine médical couvert par le corpus. Ainsi ces nœuds définissent le domaine sémantique de notre application. Par exemple, les termes "foie", "poumon" ou "opacité" sont des nœuds terminaux.

2) Chaque compte rendu médical est traduit en un arbre binaire représentant son indexation, sa représentation conceptuelle. Par exemple le syntagme "l'opacité du poumon" est indexé par l'arbre binaire représenté par la forme préfixée [porte_sur, opacité, poumon].

3) Chaque arbre est construit en respectant un modèle formel défini par une grammaire. Le langage défini par cette grammaire est le langage conceptuel de RIME.

4.1.3. Le langage conceptuel

a. Présentation

Le contenu d'un compte rendu médical se présente de façon très structurée et organisée. Comme nous le disions précédemment, un compte rendu médical contient généralement des informations sur l'examen, les constatations faites, et un diagnostic médical. Chacun de ces éléments peut à son tour être décrit en méta-notions tels que des signes, des constatations, des lésions, jusqu'au niveau le plus fin de notre modèle conceptuel (les termes médicaux ou techniques). Ceci suggère une formalisation hiérarchique du modèle sémantique, représentée par une grammaire hors contexte. Comme toute grammaire de ce type, la grammaire du langage conceptuel est définie par un vocabulaire terminal (l'ensemble des symboles terminaux), un ensemble de symboles non-terminaux, un ensemble de règles, et un symbole initial.

Le vocabulaire terminal contient l'ensemble des concepts atomiques du modèle (les termes médicaux ou techniques, tels que "poumon") et les opérateurs sémantiques tels que "porte_sur". Dans la suite tout symbole terminal sera mis en lettres minuscules.

Le vocabulaire non-terminal contient l'ensemble des méta-notions du modèle tels que SIGNE ou DIAGNOSTIC, que nous écrivons en lettres majuscules.

Le symbole initial est CR pour Compte_Rendu, et il correspond au plus haut niveau de notre modèle.

Les règles de notre grammaire sont hors contexte : la partie gauche contient un seul méta-symbole du vocabulaire non-terminal. Une règle est présentée dans la suite selon le format BNF : partie_gauche ::= partie_droite. Chaque partie droite de règle peut contenir un seul méta-symbole, et dans ce cas cela induit l'équivalence entre le méta-symbole de la partie gauche de la règle et celui de la partie droite. La partie droite peut également contenir deux méta-symbole et un opérateur sémantique, et permet ainsi la dérivation d'une structure binaire. Par exemple la règle OBSERVATION ::= [montré_par, SIGNE, EXAMEN] | SIGNE exprime qu'une observation peut être uniquement un signe ou bien un signe montré par un examen. La grammaire n'est pas ambiguë : il n'y a pas de règles ayant une partie droite égale et une partie gauche distincte.

Les règles terminales de la grammaire sont les règles dont la partie droite n'est constituée que de symboles terminaux exprimant des termes techniques ou médicaux. Ces règles expriment le lien entre les méta-notions de la grammaire et le vocabulaire terminal. Par exemple, SIGNE ::= opacité | déviation | …

b. La grammaire du langage conceptuel

Cette grammaire a été définie avec les médecins du CHU de Grenoble, et ce travail nous a permis de disposer d'un langage contenant les éléments du domaine médical couvert par notre corpus, et une idée claire des besoins en informations des utilisateurs de RIME.

1) Le premier niveau de la grammaire donne une définition formelle des concepts de plus haut niveau, c'est à dire les constatations, le diagnostic, permettant ainsi de relier les différents composants d'un compte rendu. Par exemple :

```
CR ::= CONSTAT
CR ::= DIAGNOSTIC
CR ::= [permet_de_déduire, CONSTAT, DIAGNOSTIC]
```

Ces règles définissent un CR comme étant constitué d'un seul diagnostic, d'un seul constat ou bien des deux. Dans ce dernier cas, le constat permet la déduction du diagnostic, et ce lien sémantique entre constat et diagnostic se traduit par l'opérateur sémantique permet_de_déduire.

Par ailleurs, notons que nous n'avons introduit ici aucune notion de flou, et aucune négation dans notre langage.

2) Le second niveau définit les notions de constat, de diagnostic, ainsi que des sous-notions qui leur sont associées.

```
CONSTAT ::= [dû_à, CONSTAT, CONSTAT]
```

Ce qui exprime que des constats peuvent être interdépendants.

```
CONSTAT ::= [montre_par, SIGNE, EXAMEN]
CONSTAT ::= SIGNE
```

Ces règles définissent un constat à partir d'un signe seul, ou bien par un signe révélé par un examen, un signe est une entité observable, telle une "cavité".

```
DIAGNOSTIC ::= [et, DIAGNOSTIC, DIAGNOSTIC]
```

Cette règle exprime un diagnostic comme une combinaison de plusieurs diagnostics.

```
DIAGNOSTIC ::= LESION
```

Cette règle exprime un diagnostic comme une simple lésion, par exemple un "emphysème".

```
SIGNE ::= [a_pr_val, SIGNE, QUAL]
```

Cette règle définit un signe comme un signe modifié par un qualificatif, tel que "aérique", "bombé"

```
SIGNE ::= [p_sur, SIGNE, LOC]
```

Cette règle définit un signe portant sur une localisation particulière, par exemple une "opacité du poumon".

```
LESION ::= [p_sur, LESION, LOC]
```

Cette règle définit une lésion portant sur une partie de l'organisme, par exemple une "tumeur du poumon".

```
LESION ::= [EN_REL_TOPO_AVEC, LESION, LOC]
```

cette règle situe une lésion par rapport à une partie de l'organisme, par exemple une "tumeur derrière le gril costal"

```
LOC ::= [a_pr_val, LOC, POS]

LOC ::= CONST_ORG
```

Cette règle définit une localisation comme un constituant de l'organisme, ou bien comme une notion plus complexe combinant une localisation et une position particulière, par exemple "la partie supérieure du poumon droit".

3) Le troisième niveau contient les règles pré-terminales et les règles terminales de la grammaire, et correspond aux concepts de plus bas niveau dans notre modèle. L'ensemble des symboles terminaux de la grammaire, hormis les opérateurs sémantiques, est consigné dans un lexique. Une information sémantique est attribuée à chaque entrée du lexique, et dans cette information se trouve notamment une catégorie sémantique correspondant à un méta-symbole de la grammaire : par exemple, "augmentation" possède la catégorie sémantique "signe".

Par exemple, $SIGNE ::= \{t \in V_t$ / catégorie_sémantique(t) = "signe"$\}$

ce qui signifie que tout terme de catégorie sémantique "signe" est un SIGNE dans la grammaire. Ainsi les éléments de SIGNE ne sont pas définis en extension, mais en intension par rapport au contenu du lexique.

De même, $LESION ::= \{t \in V_t$ / catégorie_sémantique(t) = "lésion"$\}$, et les termes tumeur et cancer sont des LESION. $CONST\_ORG ::= \{t \in V_t$ / catégorie_sémantique(t) = "const_org"$\}$, et les termes poumon et cœur sont des CONST_ORG. $POS ::= \{t \in V_t$ / catégorie_sémantique(t) = "position"$\}$, et le terme côté est une POSITION.

Pour conclure, cette grammaire est composée :

- d'une soixantaine de règles ;

- d'une dizaine d'opérateurs sémantiques terminaux (a_pr_val, porte_sur, …) et d'un opérateur EN_REL_TOPO_AVEC qui représente différents opérateurs sémantiques terminaux, mais dont le rôle dans la grammaire est strictement identique (à gauche, à droite, en haut, en bas, …)

- d'un lexique qui contient tous les terminaux de la grammaire.

4.2. Le processus d'indexation

4.2.1. Principes

Le processus d'indexation de RIME consiste à définir une fonction capable à partir d'un compte rendu médical de générer sa traduction dans le modèle sémantique. D'un point de vue général, ce processus peut apparaître comme un processus classique de

traduction de la langue naturelle vers un langage formel. En réalité, dans le contexte de RIME, nous n'avons pas à traiter tous les problèmes linguistiques qui peuvent apparaître dans une application ouverte. En effet les textes traités sont écrits dans une langue de spécialité, utilisant un vocabulaire très fermé et peu ambigu, et également une syntaxe et une sémantique particulières.

Dans ce contexte, la définition d'un processus d'indexation passe par la mise en évidence des traitements et problèmes linguistiques que nous devons traiter dans RIME. A partir de cela, nous avons spécifié et défini un processus d'indexation automatique permettant la génération automatique de comptes rendus indexés. [Ber 88] décrit de façon détaillée les traitements et problèmes linguistiques qui apparaissent dans RIME et ceux qui n'apparaissent pas. Nous ne montrons ici que les éléments que l'on doit traiter dans RIME.

4.2.2. Les traitements linguistiques

a. Introduction

L'indexation dans RIME doit construire des arbres selon le modèle sémantique à partir de textes médicaux. Pour cela, nous devons :

- reconnaître les mots, identifier leurs attributs virtuels, en déduire leurs attributs actuels. Nous appelons cela le niveau infra-structurel ;

- regrouper les mots entre eux afin de construire des structures, de les nommer. Nous appelons cela le niveau intra-structural ;

- déduire des liens entre structures : c'est le niveau inter-structurel.

b. Le niveau infra-structurel

A ce niveau, les mots sont considérés comme des entités individuelles sans aucun lien entre elles. Le niveau infra-structurel doit :

1) identifier les mots en isolant chaque mot simple (une simple chaîne de caractères), et en déduisant à partir des mots simples les listes des mots des textes. Par exemple, la phrase "extension ganglionnaire médiastinale au niveau du groupe de la bifurcation de la chaîne para-trachéale droite" est composée de 15 mots simples, et de 13 mots du discours médical ("au niveau du" est un seul mot).

2) identifier les attributs virtuels de chaque mot, c'est-à-dire l'ensemble de ses attributs morphologiques, syntaxiques et sémantiques potentiels. Par exemple "ganglionnaire" a deux attributs morpho-syntaxiques virtuels (masculin singulier, féminin singulier), un attribut syntaxique (adjectif qualificatif), un attribut sémantique ([ganglion, const_org]). Cet attribut sémantique signifie que "ganglionnaire" est représenté sémantiquement par "ganglion" qui est un constituant del'organisme.

3) déduire les attributs actuels de chaque mot à partir de ses attributs virtuels et de son contexte. Par exemple, "ganglionnaire" est au (féminin singulier) dans le contexte "extension ganglionnaire".

La difficulté essentielle du niveau infra-structurel est l'extraction des attributs actuels, qui ne peut être systématiquement déterminée par le niveau infra-structurel seul. Par exemple "temps de coagulation dangereux" peut être aussi bien au singulier qu'au pluriel. Les problèmes sont en fait divisés en trois catégories :

1) les ambiguïtés morphologiques, comme par exemple "temps de coagulation"

2) les ambiguïtés syntaxiques (les homographies) qui sont très rares dans notre contexte. Notons enfin que les polysémies sont inexistantes dans RIME.

c. Le niveau intra-structurel

L'objectif du niveau intra-structurel est de regrouper les mots afin de construire et nommer des structures syntaxiques ou sémantiques. Ceci pose deux difficultés essentielles :

1) les problèmes d'attachement. Considérons une structure à trois composants X Y Z, où Z doit être relié soit à X soit à Y. Ce problème a dans l'absolu trois solutions (X (Y Z)) ou ((X Y) Z) ou (X (Y) (Z)). Par exemple, "confirmation d'une hypertrophie de densité tissulaire" peut être analysée comme "confirmation (d'une hypertrophie de densité tissulaire)" ou "confirmation d'une hypertrophie (de densité tissulaire)" ou "confirmation (d'une hypertrophie) (de densité tissulaire)". Ce problème est souvent présenté dans la littérature avec l'exemple "école de commerce de jeunes filles".

2) les problèmes de fermeture. Considérons une structure à trois composants X, Y, Z où Z est relié à Y et peut-être à X. Cette structure peut alors avoir deux représentations (X) (Y Z) ou (X Z) (Y Z). Par exemple, "le cancer et la tumeur du poumon" peut être représenté par "le cancer et (la tumeur du poumon)" ou "(le cancer du poumon) et (la tumeur du poumon)"

d. Le niveau inter-structurel

Détecter des dépendances entre structures afin d'en créer de nouvelles est l'objectif du niveau inter-structurel. Parmi les dépendances entre structures, se trouvent :

- l'effacement d'une partie de structure. Ceci arrive lors de coordination, d'utilisation d'adjectifs possessifs, de comparatives. Par exemple, "opacité pulmonaire en projection du lobe supérieur droit et d'aspect alvéolaire avec signe de nécrose" présente un effacement de coordination (utilisation de "et") de "opacité pulmonaire", "la lobaire supérieure droite présente une amputation au niveau de sa lumière" présente un effacement d'adjectif possessif, "la tumeur est plus grosse que sur la radio du 30.10" présente un effacement de comparative, …

- les anaphores nominales (par répétition ou inclusion) ou pronominales ;

- les portées de certains opérateurs, que l'on trouve essentiellement dans l'utilisation de la négation.

4.2.3. L'architecture du processus d'indexation

Compte tenu de ce cahier des charges, nous avons spécifié différents processus linguistiques capables de résoudre, dans le contexte de RIME, ces différentes tâches. Ces processus sont au nombre de trois : un processus morphologique, un processus syntaxique, et un processus sémantique. Chacun d'eux a été spécifié selon les tâches linguistiques qu'il doit prendre en charge (voir figure 4.1). Par exemple, le processus morphologique a pour tâche d'extraire les mots des textes et d'en déduire leurs attributs virtuels.

Ces trois processus doivent interagir afin de résoudre les problèmes linguistiques : tel processus doit signaler telle tâche, mais tel autre processus doit la confirmer ou la résoudre. Pour assurer le dialogue et surtout l'indépendance entre les processus, nous avons construit un processus de coopération (spécifié comme un "tableau noir"), le seul ayant la connaissance permettant de piloter les trois processus linguistiques.

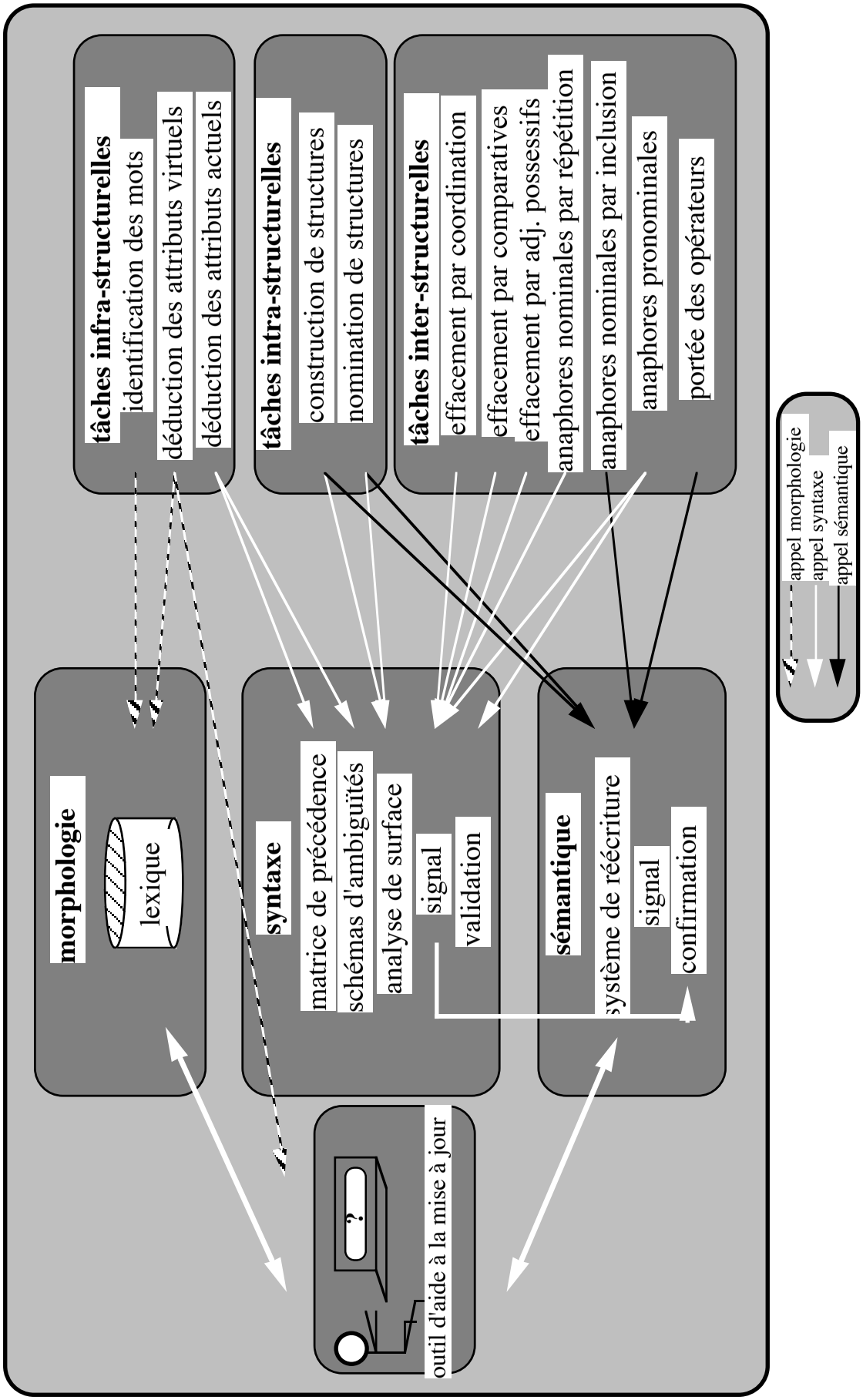


Figure 4.1 : la répartition des tâches dans RIME

a. Le processus morphologique

Le processus morphologique a pour tâche d'identifier les mots et d'extraire leurs attributs virtuels. Pour cela, le processus consulte un dictionnaire que nous décrivons succinctement ci-après.

Le dictionnaire

Chaque mot du corpus est stocké dans le dictionnaire de RIME, associé à ses attributs morphologiques, syntaxiques et sémantiques.

Les attributs sémantiques sont des couples (trait sémantique, catégorie sémantique). Le trait sémantique définit l'interprétation du mot dans le modèle. Si le mot appartient au vocabulaire de base, son trait sémantique est lui-même, sinon le mot est dit complexe et son trait sémantique est une expression du modèle. La catégorie sémantique correspond à un symbole de la grammaire. Par exemple poumon a pour attribut sémantique (poumon, const_org), hypertrophie ([a_pr_val, volume augmenté], signe).

Les attributs morphologiques et syntaxiques sont également des couples (catégorie grammaticale, valeurs grammaticales). La catégorie grammaticale donne la catégorie syntaxique du mot, les valeurs grammaticales donnent les informations telles que genre, nombre ou bien temps, personne, ... Par exemple, poumon a pour attributs (subc, (masc, sing)), où subc est une abréviation de substantif commun.

L'analyse morphologique

Nous effectuons une analyse morphologique classique de gauche à droite consistant en une segmentation d'une forme en racine + désinence. Les racines se trouvent dans un dictionnaire d'analyse [Pal 85] [Pal 90], les désinences dans de multiples tables spécifiques, le tout formant le lexique décrit précédemment. Pour une même forme, toutes les segmentations possibles sont mémorisées.

Nous utilisons un dictionnaire d'analyse où toutes les racines sont factorisées sous forme d'arbre lexicographique et rangées selon un critère de fréquence de consultation. Chaque fin de racine est matérialisée par le positionnement d'un indicateur. Au cours de l'analyse, la rencontre d'un tel indicateur valide un ensemble de tables de désinence. La reconnaissance est alors poursuivie parallèlement dans ces tables et dans le dictionnaire sans retour arrière.

b. Le processus syntaxique

Le processus syntaxique est chargé de certaines tâches, de signaler et de valider certaines autres :

- déduction des attributs actuels des mots homographes ;

- construction et nomination des structures syntaxiques (syntagmes nominaux, verbaux et prépositionnels) ;

- signal d'effacements et d'anaphores;

- validation de la proposition de résolution d'anaphores pronominales

La déduction des attributs actuels

Cette déduction se fait par filtrage syntaxique, et ce en deux étapes : la consultation d'une matrice de précédence, la consultation d'une liste de schémas d'ambiguïtés.

La matrice de précédence permet d'élaguer les impossibilités syntaxiques évidentes en regardant pour chaque mot ses liens avec le mot qui le précède. On procède à la simplification de certaines transitions en éliminant des successions morpho-syntaxiques interdites ou des accords grammaticaux non vérifiés. Ces transitions binaires sont stockées dans une matrice dite de précédence contenant pour tout couple de catégories

syntaxiques la valeur de leur utilisation successive (interdit, autorisé, autorisé si accord en genre et en nombre, …).

Un schéma d'ambiguïté peut succinctement se décrire comme une règle de transformation admettant en partie gauche un parcours ou chemin morpho-syntaxique ambigu et en partie droite un chemin résolu. Pour les utiliser lors du filtrage syntaxique, nous dressons une liste de schémas d'ambiguïtés. Lorsqu'une partie gauche d'un schéma est reconnue, on lui substitue sa partie droite.

La construction et nomination de structures syntaxiques

L'objectif de RIME est de construire des concepts médicaux connectés entre eux par des articulateurs. Nous avons observé que les concepts médicaux se trouvaient généralement dans les syntagmes nominaux, et les articulateurs dans les syntagmes verbaux ou les groupes prépositionnels. Aussi nous construisons des arbres syntaxiques à quatre niveaux :

- le niveau 4, le plus bas, contient les syntagmes nominaux ;

- au niveau 3 se trouvent les prépositions de type "de" ;

- au niveau 2 se trouvent les autres groupes prépositionnels ;

- enfin au niveau 1, le plus haut, apparaissent les groupes verbaux.

La génération de ces arbres se fait par une simple reconnaissance de catégories grammaticales : la liste des catégories appartenant à chaque type de syntagmes est connue.

Le signal d'effacements,

validation d'anaphores pronominales

Il s'agit des effacements par coordination, comparatives et adjectifs possessifs. Les effacements par coordination sont détectés par la succession d'une conjonction de coordination et d'une préposition, ou bien d'une virgule et d'une préposition. Les effacements par comparatives sont signalés par l'apparition de séquences "moins … que", "plus … que", … Les anaphores sont également facilement signalées.

Ces signaux sont directement intégrés dans l'arborescence syntaxique au niveau même de chaque syntagme présentant cette particularité linguistique.

La validation des anaphores pronominales consiste à vérifier si le syntagme proposé en remplacement du pronom est valide morpho-syntaxiquement.

c. Le processus sémantique

Le processus sémantique doit d'une part construire et nommer les structures sémantiques tout en respectant le modèle de RIME, d'autre part résoudre certaines problèmes inter-structurels (anaphores nominales, …).

Pour réaliser cela, le processus est constitué d'une enveloppe sémantique et d'un noyau sémantique. L'enveloppe sémantique élabore des solutions aux problèmes inter-structurels et des propositions pour le noyau sémantique, le noyau sémantique reçoit ces propositions et les traduit, lorsque cela s'avère possible, selon le modèle de RIME.

L'enveloppe sémantique reçoit les arbres à quatre niveaux construits par le processus syntaxique. Lorsque des problèmes inter-structurels apparaissent dans les arbres (résolution d'anaphores ou d'effacement), l'enveloppe sémantique élabore une première solution. Cette solution est proposée au noyau sémantique, et en cas de refus de la proposition, l'enveloppe sémantique élabore une nouvelle solution. En cas d'impasse, l'enveloppe sémantique demande une nouvelle proposition syntaxique.

Le noyau sémantique est basé sur un système de réécriture permettant la création des structures selon le modèle de RIME. Le système de réécriture possède les propriétés de terminaison et confluence afin d'une part d'être sûr d'obtenir des résultats mais aussi d'autre part d'assurer un résultat unique à une même proposition.

4.3. Conclusion

Cette présentation de l'indexation de RIME nous montre une approche d'indexation automatique sur un langage spécifique et exhaustif. Ces travaux nous ont permis de montrer que, dans un contexte spécialisé, un langage d'indexation complexe est une condition nécessaire à l'expression du signifié des documents. Contrairement à [Sal89], une solution de type mots-clés ne peut satisfaire de façon équivalente les utilisateurs médecins. Par ailleurs, l'automatisation du processus d'indexation constitue la condition à une indexation réaliste pour un langage complexe, et à sa validation.

Parallèlement à ce travail d'indexation, Nie [Nie88] [Nie90] a défini au niveau formel et opérationnel la correspondance de RIME. Pour cela, Nie a utilisé le modèle logique de Rijsbergen [Rij 86] et montré son utilisation dans le contexte du modèle sémantique de RIME et des utilisateurs médecins. Ainsi dans ce système, la réponse à une requête est le résultat d'une démonstration logique, et non pas, comme bien souvent, l'effet (quelquefois peu compréhensible) d'un algorithme complexe et adhoc.

5. Indexation d'images : le langage EMIR2 et son interface d'indexation

Prendre en compte la multiplicité des éléments dénotant le contenu d'une image est un problème délicat, que nous avons tenté de formaliser dans le langage EMIR2. Notre objectif est donc de proposer un langage permettant une indexation exhaustive d'images. En ce sens EMIR2 exprime les différents points de vue ou bien les différentes facettes d'une image [Mec 95a] [Mec 95b]. EMIR2 permet ainsi une indexation profonde d'images. Dans ce contexte, le processus d'indexation ne peut, bien entendu, être automatique. Aussi afin que la définition d'un tel langage soit réaliste, nous nous devons de construire une interface d'indexation capable d'aider l'indexeur dans sa tâche [Ber 95].

5.1. Présentation du problème

Le langage d'indexation EMIR2 doit tenir compte des caractéristiques suivantes des données images :

- Il faut pouvoir représenter les images physiquement et logiquement et permettre leur manipulation ;

- Il est nécessaire de prendre en compte la multiplicité des éléments dénotant le contenu d'une image ;

- L'interprétation du contenu des images est généralement propre à l'agent qui en effectue la description, et elle est effectuée dans un but précis qui peut ne pas être partagé par d'autres utilisateurs potentiels des mêmes images. Il est essentiel de prendre en compte l'aspect objectif tout en permettant la représentation de la subjectivité inhérente au processus d'interprétation. Le deuxième impératif est atteint essentiellement en permettant plusieurs descriptions de même type de la même image, chacune représentant une vision particulière de l'image.

La définition du modèle de données permet donc d'associer à chaque image une description multiforme de son contenu. Cette description est alors exploitée comme un index pour permettre une recherche d'informations effective dans un corpus d'images. La définition d'une fonction de comparaison des images doit être construite en se basant sur le modèle de données développé, et en respectant les paradigmes fondamentaux de la recherche d'images.

5.2. EMIR2 : un modèle de représentation des images

5.2.1. Introduction

Etant donné les différentes interprétations de leur contenu, les images sont dites multi-vues. La notion de vue correspond à la modélisation d'un type d'interprétation particulier qui peut être plus ou moins riche, selon par exemple la complexité des images :

a) l'image doit être représentée comme un objet physique, c'est-à-dire une matrice de pixels ;

b) l'image doit être représentée comme un ensemble d'objets géométriques associés à leur contour, et reliés entre eux par des relations spatiales ;

c) l'image doit être représentée par des descriptions symboliques correspondant à autant d'interprétations sémantiques de son contenu ;

d) l'image doit être représentée par l'aspect des objets qu'elle contient (couleur, texture, brillance) ;

e) l'image doit être vue comme un objet complexe permettant ainsi de lier les éléments intéressants de l'image entre eux par des relations de composition.

Un modèle de données général doit donc permettre la représentation d'une image comme un objet complexe, construit à partir d'éléments correspondant à des points de vue particuliers. Cet objet complexe peut être assimilé à une agrégation des différents points de vue estimés pertinents dans un contexte applicatif donné.

5.2.2. Le langage EMIR2

a. La notion d'objet image

Le contenu informatif utile d'une image physique est défini à partir d'ensembles de sous-images jugées pertinentes par rapport au contenu global de cette image. Autrement dit, une sous-image est un objet physique constituant un élément d'information utile pour la modélisation du contenu de l'image à laquelle elle appartient.

Dans notre modèle, chaque sous-image est décrite par une entité abstraite appelée "objet image" ayant pour référent physique cette sous-image elle-même. Les objets images peuvent se recouvrir, de même que la décomposition de l'image en objets images peut n'être que partielle ; elle dépend de l'interprétation retenue pour décrire les éléments jugés pertinents dans un contexte applicatif donné. Dans notre modèle, les objets images sont nommés par un identifiant qui permet de les référencer.

b. La notion de vue

On appelle vues les différentes interprétations des objets images que l'on peut avoir. Une vue correspond donc à un type de contenu informatif défini sur l'ensemble des objets images : à chacune d'elles est associé un modèle donnant une description des objets images, une description des relations qui, éventuellement, les lient et les opérations applicables sur ces descriptions.

Comme les interprétations qu'elles modélisent, les vues sont généralement complémentaires et chaque application en combine plusieurs pour capter un maximum d'informations pertinentes sur les images qu'elle manipule. Par exemple dans un système de cartographie, il ne suffit pas d'associer des descriptions géométriques aux différents objets dans une carte (route, ville, etc), encore faut-il leur associer également une description symbolique pour les identifier et les lier au monde réel, ce qui revient à leur attribuer une sémantique. A partir de l'objet image représentant une ville, une figure géométrique décrivant son contour est alors représentée, et on lui associe des données symboliques telles que son type (ville), le nom de la ville, la population, etc. Ainsi donc, un objet image est décrit suivant différentes vues : il en va de même des relations liant ces différents objets dans le contexte particulier de chaque vue.

EMIR2 modélise cinq vues regroupées en deux catégories : la vue physique qui considère l'image telle que perçue par l'œil dans sa représentation plane, bi-dimensionnelle, les vues logiques (structurelle, spatiale, symbolique et perceptive) qui regroupent les interprétations de l'image et de son contenu. La décomposition d'une image en un ensemble d'objets images est déterminée par la vue structurelle : c'est elle qui exprime la décomposition jugée pertinente de l'image en composants, et les trois autres vues logiques sont définies à partir de ce même ensemble.

c. La vue physique

La vue physique d'une image correspond aux données brutes constituées par les images numérisées. Elle est décrite par la donnée des caractéristiques générales de l'image, telles que ses dimensions, la résolution (nombres de bits par pixel), le format de représentation utilisé intégrant ou non les formats de compression comme les standards GIF ou JPEG, la table des couleurs, la matrice de pixels. Cette image peut être visualisée sur un écran, sauvée dans un fichier, et peut subir des transformations à l'aide de fonctions de traitement d'images comme le zoom, le changement de luminosité, la rotation, …

Le modèle de la vue physique est défini de la façon suivante :

$M_{ph} = (I_{ph}, POINT, EC, TYPE, h, l, tc, pixels, type)$

avec

- I_{ph} l'ensemble des identifiants des vues physiques de EMIR2

- POINT est l'ensemble des paires d'entiers positifs représentant les coordonnées de tous les points possibles $POINT = N^+ \times N^+$

- EC est l'ensemble des couleurs définies dans un espace de couleurs particulier,

$EC = \{O, 1…255\} \times \{O, 1…255\} \times \{O, 1…255\}$

- TYPE est l'ensemble des types des vues physiques. Pour le moment il est formé par quatre éléments, TYPE = {NB, NG, PC, CR} où NB = Noir_Blanc, NG = Niveauxdegris, PC = PaletteCouleur, CR = CouleurRéelle

- $h : I_{ph} \rightarrow N^+$, h est une fonction qui associe à chaque identifiant d'une vue physique un entier positif représentant la hauteur de l'image correspondante

- respectivement l identifie leur largeur

- $tc : I_{ph} \rightarrow P(EC)$, tc est une fonction qui associe à chaque identifiant d'une vue physique l'ensemble des couleurs de l'image physique correspondante. P(EC) est l'ensemble des parties de EC.

- pixels : $I_{ph} \rightarrow P(POINT \times EC)$, qui associe à chaque identifiant d'une vue physique l'ensemble des pixels de l'image physique correspondante

- type : $I_{ph} \rightarrow$ TYPE, qui associe à chaque identifiant d'une vue physique le type de l'image physique correspondante

Nous ne détaillons dans ce document aucune contrainte de cohérence du modèle EMIR2, elles sont exprimées formellement dans [Mec 95a].

d. La vue structurelle

La vue structurelle représente la décomposition d'une image en objets images. Elle permet également d'exprimer la décomposition d'objets images en objets images composants. A cette vue ne correspond qu'un seul type de relation : la relation de composition "CONTIENT" (CONT). La vue structurelle est donc représentée par un graphe ayant pour nœuds des objets images et pour arcs des instances de ce lien de composition.

Un objet image est complexe s'il est lui-même composé d'un ensemble d'objets images. Un objet image peut être partagé par plusieurs objets images. Ce cas se produit, par exemple, dans une image contenant deux maisons qui partagent un mur.

Nous définissons le modèle de la vue structurelle de la façon suivante :

$M_{st} = (I_{oi}, CONT)$

- I_{oi} est l'ensemble des identifiants de la vue structurelle

- CONT est la relation de composition entre les objets images, $CONT \subseteq I_{oi} \times I_{oi}$

e. La vue spatiale

La vue spatiale comporte des informations géométriques décrivant les objets spatiaux associés aux objets images et des relations spatiales décrivant leurs positions relatives. Par exemple la relation Proche est une relation spatiale. La forme des objets est représentée par une combinaison d'éléments géométriques de base qui sont les points, les segments et les polygones. Cette vue a un double usage : d'abord dans l'opération de recherche d'informations, pour décrire avec plus de précision les objets désirés en

considérant leur forme et leur position relative, et enfin pour effectuer des opérations spatiales, telles que les calculs de distance, de surface, de longueur, …

Le modèle formel de la vue spatiale est défini comme suit :

$M_{sp} = (I_{sp}, POINT, OS, RSPA, forme, R_{sp})$

- I_{sp} est l'ensemble des identifiants des objets spatiaux

- POINT est l'ensemble des paires d'entiers positifs représentant les coordonnées de tous les points, $POINT = N^+ \times N^+$

- OS est l'ensemble des objets géométriques de base définis dans EMIR2 et susceptibles d'être utilisés pour représenter la forme des objets. Trois types de base sont considérés pour le moment : le point, le segment, le polygone.

- RSPA désigne l'ensemble des relations spatiales : RSPA = {loin, près, est, ouest, nord, sud, dans, disjoint, touche, couvre, coupe}

- forme est la fonction qui lie chaque identifiant d'objet spatial aux objets géométriques le formant : forme : $I_{sp} \rightarrow P(OS)$

- R_{sp} est une relation qui définit l'ensemble des relations spatiales qui lient les objets spatiaux dans l'image, $R_{sp} \subseteq RSPA \times I_{sp} \times I_{sp}$

f. La vue symbolique

La vue symbolique correspond à la représentation du contenu sémantique d'une image. Elle est définie par des objets symboliques associés aux objets images de référence, ainsi que par des relations entre objets symboliques. A chaque objet symbolique correspond directement un attribut contenant des éléments de contenu décrivant l'objet en question.

Les relations symboliques correspondent à la description de scènes, ou d'actions faisant intervenir des objets symboliques.

Le modèle de la vue symbolique se définit formellement par :

$M_{sy} = (M_{app}, I_{sy}, cl, RI, PI)$

- I_{sy} est l'ensemble des identifiants des objets symboliques

- cl est une fonction qui associe à chaque objet symbolique sa classe, cl : I_{sy} -> ID_{cl} (ID_{cl} est défini dans M_{app})

- $RI \subseteq ID_{rs} \times I_{sy} \times I_{sy}$ est la relation qui permet de lier les objets symboliques entre eux par les relations de ID_{rs} de M_{app} .

- $PI \subseteq ID_{pr} \times I_{sy} \times VAL\_PROP$ est la relation qui permet de lier les objets images à leurs propriétés (ID_{pr} est défini dans M_{app}).

Toute vue symbolique se définit dans un contexte applicatif modélisé par le modèle d'application M_{app} :

$M_{app} = (ID_{cl}, ID_{pr}, ID_{rs}, VAL\_PROP, PROP, RSYMB, COMP, domaine)$

- ID_{cl} est l'ensemble des identificateurs des classes d'objets. Cet ensemble est muni d'une relation d'ordre partiel noté $\leq$ et est augmenté d'un élément minimal $\perp$ et d'un élément maximal T

- ID_{pr} est l'ensemble des identificateurs de propriétés

- ID_{rs} est l'ensemble des identificateurs des relations symboliques

- VAL_PROP est l'ensemble des valeurs possibles que peut prendre une propriété : VAL_PROP = Réel ∪ Entier ∪ Chaîne ∪ Booléen

- PROP est l'ensemble des définitions de propriétés : PROP ⊆ ID_{pr} x ID_{cl} x P(VAL_PROP)

- RSYMB est l'ensemble des définitions des relations symboliques : RSYMB ⊆ ID_{rs} x ID_{cl} x ID_{cl}

- COMP ⊆ ID_{cl} x ID_{cl} est la relation de composition entre les classes d'objets images

- domaine est la fonction qui définit pour chaque type de valeurs de propriétés l'ensemble des valeurs possibles, domaine : ID_{pr} → P(VAL_PROP)

g. La vue perceptive

La perception des images est fonction non seulement des objets qu'elle contient, mais aussi de l'aspect de ces objets dans l'image qui est un élément de leur description. Ainsi une image contenant un objet correspondant à une voiture, est incomplètement décrite si on ignore dans sa représentation la couleur de la voiture, ou toute autre caractéristique visuelle importante de l'image.

La vue regroupe donc l'ensemble des attributs visuels objectifs des objets images, qui en constituent les modalités de représentation. Nous incluons, dans EMIR2, trois modalités de représentation correspondant aux trois attributs visuels de base résumant la présentation visuelle des composants de l'image, et qui sont la couleur, la texture, et la brillance. Les valeurs associées à chacun de ces attributs sont relatives au contexte de l'application.

Le modèle formel de la vue perceptive se définit comme suit :

M_{pe} = (I_{pe}, TX, BR, CL, tx, br, cl)

- I_{pe} est l'ensemble des identifiants des objets perceptifs possibles

- TX est l'ensemble des modèles de texture

- BR est l'ensemble des valeurs de brillance

- CL est l'ensemble de valeurs de couleur

- tx : I_{pe} → TX associe aux identifiants d'objets perceptifs une texture

- br, et cl de la même façon leur associe une brillance, et une couleur.

h. Le modèle d'image

Une image s'exprime dans le modèle de la façon suivante :

M_{im} = (I_{im}, M_{ph}, M_{st}, M_{pe}, M_{sp}, M_{sy}, L_{sp}, L_{sy}, L_{pe})

- I_{im} est l'ensemble des identifiants de images dans EMIR2

- M_{ph}, M_{st}, M_{pe}, M_{sp}, M_{sy} les modèles des différentes vues

- L_{sp} ⊆ I_{oi} x I_{sp} associe les objets spatiaux aux objets images

- L_{pe}, L_{sy} associent les objets perceptifs et symboliques aux objets images

5.2.3. Un exemple

La figure 5.1 nous montre une indexation selon EMIR2 d'une image. Cette image montre une maison proche d'un arbre, cette maison appartient à un grand-père, l'arbre a été planté par Oncle Jules, … La vue structurelle contient deux objets images oi1, oi2, dont on trouve un correspondant dans toutes les autres vues logiques.

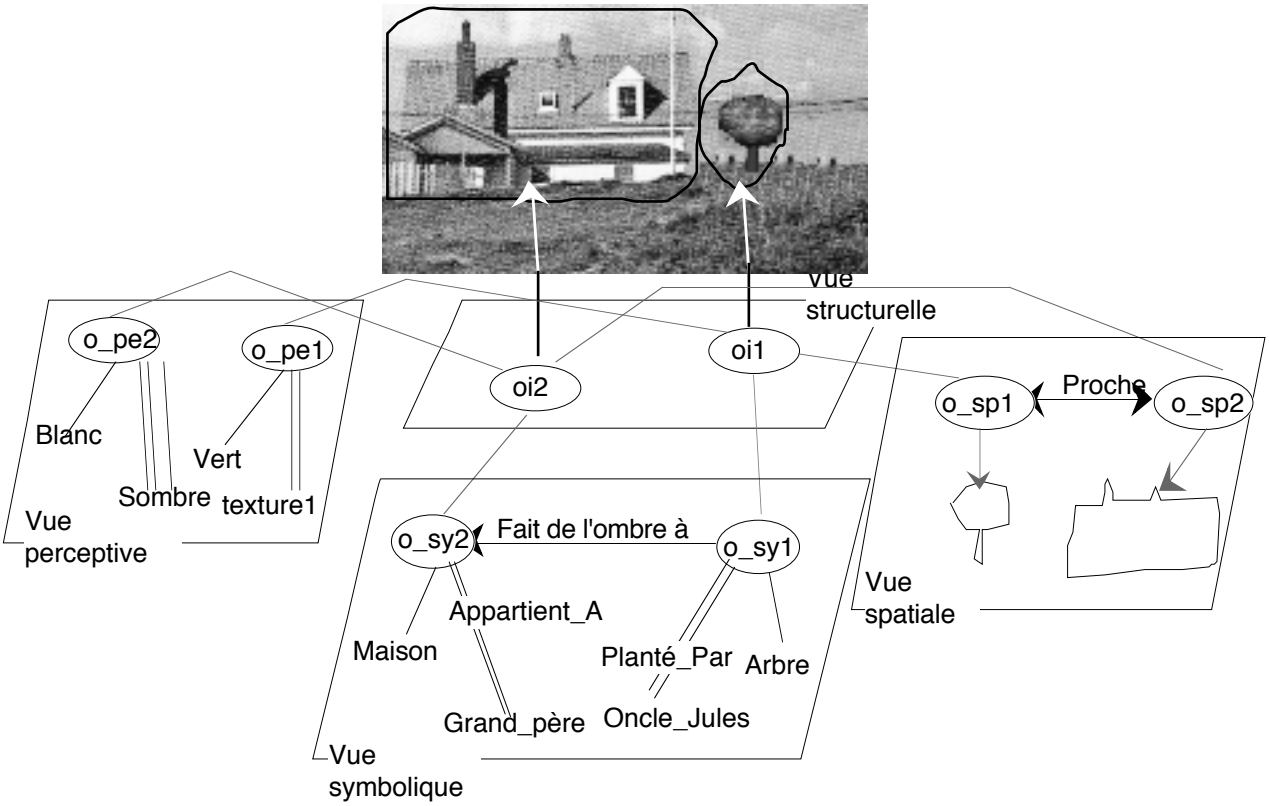


Figure 5.1 : un exemple d'image indexée selon EMIR2

5.3. Vers un modèle opérationnel : utilisation des graphes conceptuels

Le langage d'indexation EMIR2 est défini au niveau opérationnel en terme de graphes conceptuels [Sow 84] et l'opérateur de projection sur ces graphes nous permet de disposer d'une correspondance. Ce choix des graphes conceptuels nous permet de réaliser un système opérationnel respectant le modèle logique de recherche d'informations [Che 92].

Le modèle logique de recherche d'informations consiste à démontrer qu'un document D est une réponse à une requête Q si et seulement si D→Q, où → est une implication logique. Le prédicat de projection d'un graphe conceptuel a dans un graphe conceptuel b, notée $\Pi_a(b)$, est un prédicat indiquant si a se projette dans b, i.e. si b est un graphe conceptuel spécifique de a. Sowa a montré l'équivalence entre la projection et l'implication de la logique des prédicats, lorsque les concepts des graphes conceptuels sont simples. Ainsi, $D \rightarrow Q \Leftrightarrow \Pi_{gc(Q)}gc(D)$, où → est l'implication de la logique des prédicats, gc(Q) et gc(D) la représentation en graphes conceptuels de, respectivement, Q et D.

Ainsi, l'équivalence entre projection de graphes conceptuels et implication dans la logique des prédicats nous permet de disposer d'une correspondance à la fois formelle et opérationnelle. [Mec 95a] décrit de façon détaillée la transduction de EMIR2 en graphes conceptuels.

5.4. L'interface d'indexation définie pour EMIR2

Afin de permettre l'utilisation de EMIR2, nous avons défini une interface d'indexation [Bou 95], [Ber 95]. Le rôle de cette interface est double : fournir à l'indexeur un environnement ergonomique lui facilitant son travail, et ce tout en surveillant l'indexation réalisée. C'est cette interface que nous décrivons maintenant. Le corpus sur lequel nous

l'avons testé est le corpus médical de RIME. Pour présenter cette interface, nous procédons en deux étapes : nous montrons tout d'abord sa partie statique (les fenêtres, leur contenu), puis sa partie dynamique (la réaction de l'interface à des actions de l'indexeur). Nous terminons cette partie en montrant sa conception logicielle.

5.4.1. La partie statique de l'interface

L'interface d'indexation développée est composée de quatre fenêtres indépendantes : "Indexeur_Images", " image à indexer", "compte rendu", "image d'origine". Les fenêtres "compte rendu" et "image d'origine" affichent les données disponibles avant l'indexation : l'image sur laquelle il faut travailler, et le texte médical du dossier de cette image. La fenêtre "image à indexer" contient initialement l'image d'origine, s'ajouteront ensuite les éléments spatiaux de son indexation. Enfin la fenêtre Indexeur _Images fournit l'interface de l'indexation, comprenant entre autres des outils de traitements graphiques et les fenêtres de visualisation des différentes facettes de EMIR2.

a. La fenêtre "Indexeur_Images"

Comme le montre la figure 5.2, cette fenêtre met à la disposition de l'utilisateur indexeur les moyens nécessaires à son travail. Elle comprend en fait les outils d'indexation et est découpée en sept zones distinctes. Ces zones peuvent être regroupées en deux ensembles, l'un correspondant à des fonctions généralistes (1) puisant leur justification dans les principes de base que doit suivre une interface, l'autre (2) fournissant un support au modèle d'images à générer et donc orienté application.

Une fenêtre à zones multiples montre les différentes vues logiques de l'image (vues symbolique, spatiale et structurelle). Une image peut donc être vue à un instant donné sous le prisme de la vue structurelle, les deux autres vues étant non considérées et donc pouvant être invisibles. Cette idée de configuration à volonté des informations visibles selon la (ou les) vue(s) considérée(s) nous a poussée à choisir une fenêtre découpée en plusieurs zones. Chaque zone correspond à une des facettes du langage d'indexation et sa taille peut être modifiée par l'utilisateur selon son besoin.

Nous appliquons une règle de cohérence d'ensemble en choisissant les dénominations Détruire-Construire pour les boutons-poussoirs des diverses zones. La cohérence est donc sémantique et visuelle car ces divers boutons sont alignés les uns sous les autres. Elle est aussi comportementale car les boutons Détruire, par exemple, entraînent l'ouverture de fenêtres de dialogue demandant la confirmation de l'option sélectionnée. Des raccourcis clavier sont également proposés.

Les fonctions généralistes proposées à l'utilisateur

La barre de menu qui contient les choix généraux tels que SAUVER, ANNULER, AUTOMATIQUE et AIDE. L'aide proposée est une aide en ligne ce qui signifie que les informations listées sont relatives à la fonctionnalité active. L'utilisateur peut également, à partir d'une fenêtre d'aide affichée, sélectionner le sujet qui l'intéresse. Le choix AUTOMATIQUE permet d'assister le radiologue dans sa tâche d'indexation.

La zone ETATS qui, à travers l'emploi de "status man" [Bas 91], délivre à l'utilisateur l'état de son travail vis-à-vis du modèle d'images à renseigner. Pour chacune des vues du modèle, sémantique, structurelle et spatiale, une icône est disponible. Un clic souris sur l'icône d'un "status man" entraîne l'affichage d'une fenêtre de dialogue dans laquelle sont consignés les éventuels problèmes ou lacunes introduits par l'utilisateur lors de son travail d'indexation. L'utilisateur a la possibilité de sélectionner un objet image en défaut par la sélection de la ligne correspondante dans la liste des défauts présents dans la zone de dialogue. En effet l'utilisateur étant maître de l'interaction, ce principe permet d'installer quelques gardes-fous quant à la qualité et l'exhaustivité des termes d'indexation générés. Plus généralement, ils servent d'aide-mémoire par la vue synthétique qu'ils offrent de l'état du système et du travail effectué et permettent dans notre cas la réduction de la sollicitation de la mémoire à court terme. C'est un retour