Un modèle logique général pour les systèmes de recherche d'informations : application au prototype RIME

Jianyun Nie

THESE

présentée par

Jianyun NIE

pour obtenir le titre de
docteur de l'Université Joseph Fourier - Grenoble I
(arrêté ministériel du 5 juillet 1984)

spécialité ***Informatique***

Un modèle logique général pour les Systèmes de Recherche d'Informations

Application au prototype RIME

date de soutenance: 13 juillet 1990

composition du jury:

président:	M. Michel ADIBA
rapporteurs:	M. Joseph SIFAKIS
	M. Keith van RIJSBERGEN
examinateurs:	M. Richard BOUCHE
	M. Yves CHIARAMELLA
	M. Philippe CINQUIN

Thèse préparée au sein du Laboratoire de Génie Informatique - IMAG
à l'Université Joseph Fourier - Grenoble I

UNIVERSITE Joseph FOURIER (GRENOBLE I)

Président de l'Université :
M. NEMOZ Alain

Année Universitaire 1988 - 1989

MEMBRES DU CORPS ENSEIGNANT DE SCIENCES ET DE GEOGRAPHIE

PROFESSEURS DE 1ère Classe

ADIBA Michel	Informatique
ANTOINE Pierre	Géologie I.R.I.G.M.
ARNAUD Paul	Chimie Organique
ARVIEU Robert	Physique Nucléaire I.S.N.
AUBERT Guy	Physique C.N.R.S
AURIAULT Jean-Louis	Mécanique
AYANT Yves	Physique Approfondie
BARBIER Marie-Jeanne	Electrochimie
BARJON Robert	Physique Nucléaire ISN
BARNOUD Fernand	Biochimie Macromoléculaire Végétale
BARRA Jean-René	Statistiques-Mathématiques Appliquées
BECKER Pierre	Physique
BEGUIN Claude	Chimie Organique
BELORISKY Elie	Physique
BENZAKEN Claude	Mathématiques Pures
BERARD Pierre	Mathématiques Pures
BERNARD Alain	Mathématiques Pures
BERTRANDIAS Françoise	Mathématiques Pures
BERTRANDIAS Jean-Paul	Mathématiques Pures
BILLET Jean	Géographie
BOELHER Jean-Paul	Mécanique
BRAVARD Yves	Géographie
CARLIER Georges	Biologie Végétale
CASTAING Bernard	Physique
CAUQUIS Georges	Chimie Organique
CHARDON Michel	Géographie
CHIBON Pierre	Biologie Animale
COHEN ADDAD Jean-Pierre	Physique
COLIN DE VERDIERE Yves	Mathématiques Pures
CYROT Michel	Physique du Solide
DEBELMAS Jacques	Géologie Générale
DEGRANGE Charles	Zoologie
DEMAILLY Jean-Pierre	Mathématiques Pures
DENEUVILLE Alain	Physique
DEPORTES Charles	Chimie Minérale
DOLIQUE Jean-Michel	Physique des Plasmas
DOUCE Roland	Physiologie Végétale
DUCROS Pierre	Cristallographie
FINKE Gerde	Informatique
GAGNAIRE Didier	Chimie Physique
GAUTRON René	Chimie
GENIES Eugène	Chimie
GERMAIN Jean-Pierre	Mécanique,
GIDON Maurice	Géologie
GUITTON Jacques	Chimie
HICTER Pierre	Chimie
IDELMAN Simon	Physiologie Animale
JANIN Bernard	Géographie
JOLY Jean René	Mathématiques Pures

JOSELEAU Jean Paul	Biochimie
KAHANE André, détaché	Physique
KAHANE Josette	Physique
KRAKOWIAK Sacha	Mathématiques Appliquées
LAJZEROWICZ Jeanine	Physique
LAJZEROWICZ Joseph	Physique
LAURENT Pierre-Jean	Mathématiques Appliquées
LEBRETON Alain	Mathématiques Appliquées
DE LEIRIS Joël	Biologie
LHOMME Jean	Chimie
LLIBOUTRY Louis	Géophysique
LOISEAUX Jean-Marie	Sciences Nucléaires I.S.N.
LONGEQUEUE Nicole	Physique
LUNA Domingo	Mathématiques Pures
MACHE Régis	Physiologie Végétale
MASCLE Georges	Géologie
MAYNARD Roger	Physique du Solide
OMONT Alain	Astrophysique
OZENDA Paul	Botanique (Biologie Végétale)
PANNETIER Jean	Chimie
PAYAN Jean-Jacques	Mathématiques Pures
PEBAY-PEYROULA Jean-Claude	Physique
PERRIER Guy	Géophysique
PIERRE Jean Louis	Chimie Organique
RENARD Michel	Thermodynamique
RIEDTMANN Christine	Mathématiques
RINAUDO Marguerite	Chimie CERMAV
ROSSI André	Biologie
SAXOD Raymond	Biologie Animale
SENGEL Philippe	Biologie Animale
SERGERAERT Francis	Mathématiques Pures
SOUCHIER Bernard	Biologie
SOUTIF Michel	Physique
STUTZ Pierre	Mécanique
TRILLING Laurent	Mathématiques Appliquées
VAN CUTSEM Bernard	Mathématiques Appliquées
VIALON Pierre	Géologie

PROFESSEURS de 2ème Classe

ARMAND Gilbert	Géographie
ATTANE Pierre	Mécanique
BARET Paul	Chimie
BERTIN José	Mathématiques
BLANCHI J.Pierre	STAPS
BLOCK Marc	Biologie
BLUM Jacques	Mathématiques Appliquées
BOITET Christian	Mathématiques Appliquées
BORNAREL Jean	Physique
BORRIONE Dominique	Automatique informatique
BOUVET Jean	Biologie
BROSSARD Jean	Mathématiques
BRUANDET J.François	Physique
BRUGAL Gérard	Biologie
BRUN Gilbert	Biologie
CASTAING Bernard	Physique
CERFF Rudiger	Biologie
CHIARAMELLA Yves	Mathématiques Appliquées
CHOLLET Jean Pierre	Mécanique
COLOMBEAU Jean François	Mathématiques (ENSL)
COURT Jean	Chimie
CUNIN Pierre Yves	Informatique
DAVID Jean	Géographie

DHOUAILLY Danielle	Biologie
DUFRESNOY Alain	Mathématiques Pures
GASPARD François	Physique
GIDON Maurice	Géologie
GIGNOUX Claude	Sciences Nucléaires
GILLARD Roland	Mathématiques Pures
GIORNI Alain	Sciences Nucléaires
GONZALEZ SPRINBERG Gérardo	Mathématiques Pures
GUIGO Maryse	Géographie
GUMUCHAIN Hervé	Géographie
HACQUES Gérard	Mathématiques Appliquées
HERBIN Jacky	Géographie
HERAULT Jeanny	Physique
HERINO Roland	Physique
JARDON Pierre	Chimie
KERCKHOVE Claude	Géologie
MANDARON Paul	Biologie
MARTINEZ Francis	Mathématiques Appliquées
MOREL Alain	Géographie
NEMOZ Alain	Thermodymamique CNRS - CRTBT
NGUYEN HUY Xuong	Informatique
OUDET Bruno	Mathématiques Appliquées
PAUTOU Guy	Biologie
PECHER Arnaud	Géologie
PELMONT Jean	Biochimie
PELLETIER Guy	Astrophysique
PERRIN Claude	Sciences Nucléaires I.S.N.
PIBOULE Michel	Géologie
RAYNAUD Hervé	Mathématiques Appliquées
REGNARD Jean René	Physique
RICHARD Jean-Marc	Physique
RIEDTMANN Christine	Mathématiques Pures
ROBERT Danielle	Chimie
ROBERT Gilles	Mathématiques Pures
ROBERT Jean-Bernard	Chimie Physique
SARROT-REYNAULD Jean	Géologi:
SAYETAT Françoise	Physique
SERVE Denis	Chimie
STOECKEL Frédéric	Physique
SCHOLL Pierre-Claude	Mathématuques Appliquées
SUBRA Robert	Chimie
VALLADE Marcel	Physique
VIDAL Michel	Chimie Organique
VINCENT Gilbert	Physique
VIVIAN Robert	Géographie
VOTTERO Philippe	Chimie

MEMBRES DU CORPS ENSEIGNANT DE L' IUT 1

PROFESSEURS de 1ère Classe

BUISSON Roger	Physique IUT 1
CHEHIKIAN Alain	E.E.A. I.U.T.1
DODU Jacques	Mécanique Appliquée IUT 1
NEGRE Robert	Génie Civil IUT 1
NOUGARET Marcel	Automatique IUT 1
PERARD Jacques	EEA. IUT 1

PROFESSEURS de 2ème classe

BEE Marc	Physique IUT 1
BOUTHINON Michel	EEA. IUT 1
CHAMBON René	Génie Mécanique IUT 1
CHENAVAS Jean	Physique IUT 1

CHILO Jean	Physique IUT 1
CHOUTEAU Gérard	Physique IUT 1
CONTE René	Physique IUT 1
FOSTER Panayotis	Chimie IUT 1
GOSSE Jean-Pierre	EEA.IUT 1
GROS Yves	Physique IUT 1
HAMAR Roger	Chimie IUT 1
KUHN Gérard, (Détaché)	Physique IUT 1
LEVIEL Jean Louis	Physique IUT 1
MAZUER Jean	Physique IUT 1
MICHOULIER Jean	Physique IUT 1
MONLLOR Christian	EEA.IUT 1
PERRAUD Robert	Chimie IUT 1
PIERRE Gérard	Chimie IUT 1
TERRIEZ Jean-Michel	Génie Mécanique IUT 1
TOUZAIN Philippe	Chimie IUT 1
TURGEMAN Sylvain	Génie civil
VINCENDON Marc	Chimie IUT 1
ZIGONE Michel	Physique IUT 1

PROFESSEURS DE PHARMACIE

AGNIUS-DELORD Claudine	Physique	Faculté La Tronche
ALARY Josette	Chimie Analytique	Faculté La Tronche
BERIEL Hélène	Physiologie et Pharmacologie	Faculté La Tronche
CUSSAC Max	Chimie Therapeutique	Faculté La Tronche
DEMENGE Pierre	Pharmacodynamie	Faculté La Tronche
FAVIER Alain	Biochimie	C.H.R.G.
JEANNIN Charles	Pharmacie Galénique	Faculté Meylan
LATURAZE Jean	Biochimie	Faculté La Tronche
LUU DUC Cuong	Chimie Générale	Faculté La Tronche
MARIOTTE Anne-Marie	Pharmacognosie	Faculté La Tronche
MARZIN Daniel	Toxicologie	Faculté Meylan
RENAUDET Jacqueline	Bactériologie	Faculté La Tronche
ROCHAT Jacques	Hygiéne et Hydrologie	Faculté La Tronche
SEIGLE-MURANDI Françoise	Botanique et Cryptogamie	Faculté Meylan
VERAIN Alice	Pharmacie Galénique	Faculté Meylan

MEMBRES DU CORPS ENSEIGNANT DE MEDECINE

PROFESSEURS CLASSE EXEPTIONNELLE ET 1ère CLASSE

AMBLARD Pierre	Dermatologie	C.H.R.G.
AMBROISE-THOMAS Pierre	Parasitologie	C.H.R.G.
BEAUDOING André	Pédiatrie-Puericulture	C.H.R.G.
BEZEZ Henri	Orthopédie-Traumatologie	Hopital SUD
BONNET Jean-Louis	Ophtalmologie	C.H.R.G.
BOUCHET Yves	Anatomie	Faculté La Merci
	Chirurgie Générale et Digestive	C.H.R.G.
BUTEL Jean	Orthopédie-Traumatologie	C.H.R.G.
CHAMBAZ Edmond	Biochimie	C.H.R.G.
CHAMPETIER Jean	Anatomie-Topographique et Appliquée	C.H.R.G.
CHARACHON Robert	O.R.L.	C.H.R.G.
COLOMB Maurice	Immunologie	Hopital sud
COUDERC Pierre	Anatomie-Pathologique	C.H.R.G.
DELORMAS Pierre	Pneumophtisiologie	C.H.R.G.
DENIS Bernard	Cardiologie	C.H.R.G.
GAVEND Michel	Pharmacologie	Faculté La Merci

HOLLARD Daniel	Hématologie	C.H.R.G.
LATREILLE René	Chirurgie Thoracique et Cardiovasculaire	C.H.R.G.
LE NOC Pierre	Bactériologie-Virologie	C.H.R.G.
MALINAS Yves	Gynécologie et Obstétrique	C.H.R.G.
MALLION Jean-Michel	Médecine du Travail	C.H.R.G.
MICOUD Max	Clinique Médicale et Maladies Infectieuses	C.H.R.G.
MOURIQUAND Claude	Histologie	Faculté La Merci
PARAMELLE Bernard	Pneumologie	C.H.R.G.
PERRET Jean	Neurologie	C.H.R.G.
RACHAIL Michel	Hépato-Gastro-Entérologie	C.H.R.G.
DE ROUGEMONT Jacques	Neurochirurgie	C.H.R.G.
SARRAZIN Roger	Clinique Chirurgicale	C.H.R.G.
STIEGLITZ Paul	Anestésiologie	C.H.R.G.
TANCHE Maurice	Physiologie	Faculté La Merci
VIGNAIS Pierre	Biochimie	Faculté La Merci

PROFESSEURS 2ème CLASSE

BACHELOT Yvan	Endocrinologie	C.H.R.G.
BARGE Michel	Neurochirurgie	C.H.R.G.
BENABID Alim Louis	Biophysique	Faculté La Merci
BENSA Jean-Claude	Immunologie	Hopital Sud
BERNARD Pierre	Gynécologie-Obstétrique	C.H.R.G.
BESSARD Germain	Pharmacologie	ABIDJAN
BOLLA Michel	Radiothérapie	C.H.R.G.
BOST Michel	Pédiatrie	C.H.R.G.
BOUCHARLAT Jacques	Psychiatrie Adultes	Hopital Sud
BRAMBILLA Christian	Pneumologie	C.H.R.G.
CHIROSSEL Jean-Paul	Anatomie-Neurochirurgie	C.H.R.G.
COMET Michel	Biophysique	Faculté La Merci
CONTAMIN Charles	Chirurgie Thoracique et Cardiovasculaire	C.H.R.G.
CORDONNIER Daniel	Néphrologie	C.H.R.G.
COULOMB Max	Radiologie	C.H.R.G.
CROUZET Guy	Radiologie	C.H.R.G.
DEBRU Jean-Luc	Médecine Interne et Toxicologie	C.H.R.G.
DEMONGEOT Jacques	Biostatistiques et Informatique Médicale	Faculté La Merci
DUPRE Alain	Chirurgie Générale	C.H.R.G.
DYON Jean-François	Chirurgie Infantile	C.H.R.G.
ETERRADOSSI Jacqueline	Physiologie	Faculté La Merci
FAURE Claude	Anatomie et Organogénèse	C.H.R.G.
FAURE Gilbert	Urologie	C.H.R.G.
FOURNET Jacques	Hépato-Gastro-Entérologie	C.H.R.G.
FRANCO Alain	Médecine Interne	C.H.R.G.
GIRARDET Pierre	Anesthésiologie	C.H.R.G.
GUIDICELLI Henri	Chirurgie Générale et Vasculaire	C.H.R.G.
GUIGNIER Michel	Thérapeutique et Réanimation Médicale	C.H.R.G.
HADJIAN Arthur	Biochimie	Faculté La Merci
HALIMI Serge	Endocrinologie et Maladies Métaboliques	C.H.R.G.
HOSTEIN Jean	Hépato-Gastro-Entérologie	C.H.R.G.
HUGONOT Robert	Médecine Interne	C.H.R.G.
JALBERT Pierre	Histologie-Cytogénétique	C.H.R.G.
JUNIEN-LAVILLAUROY Claude	O.R.L.	C.H.R.G.
KOLODIE Lucien	Hématologie Biologique	C.H.R.G.
LETOUBLON Christian	Chirurgie Générale	C.H.R.G.
MACHECOURT Jacques	Cardiologie et Maladies Vasculaires	C.H.R.G.
MAGNIN Robert	Hygiène	C.H.R.G.
MASSOT Christian	Médecine Interne	C.H.R.G.

MOUILLON Michel	Ophtalmologie	C.H.R.G.
PELLAT Jacques	Neurologie	C.H.R.G.
PHELIP Xavier	Rhumatologie	C.H.R.G.
RACINET Claude	Gynécologie-Obstétrique	Hopital Sud
RAMBAUD Pierre	Pédiatrie	C.H.R.G.
RAPHAEL Bernard	Stomatologie	C.H.R.G.
SCHAERER René	Cancérologie	C.H.R.G.
SEIGNEURIN Jean-Marie	Bactériologie-Virologie	Faculté La Merci
SELE Bernard	Cytogénétique	Faculté La Merci
SOTTO Jean-Jacques	Hématologie	C.H.R.G.
STOEBNER Pierre	Anatomie Pathologique	C.H.R.G.
VROUSOS Constantin	Radiothérapie	C.H.R.G.

Je tiens à remercier,

Monsieur Michel Adiba, Professeur à l'Université Joseph Fourier, qui m'a fait l'honneur de présider ce jury;

Monsieur Joseph Sifakis, Directeur de recherche au C.N.R.S, qui a bien voulu être rapporteur de ce travail et qui, grâce à ses critiques constructives, a beaucoup contribué à la forme finale de ce document. Qu'il trouve ici toute ma reconnaissance;

Monsieur Keith Van Rijsbergen, Professeur à l'Université Glasgow, pour avoir accepté d'être rapporteur de ce travail malgré les problèmes de langue, et surtout pour ses apports théoriques dans le domaine de la Recherche d'Informations qui sont à la base de cette thèse. Qu'il trouve ici l'expression de mes sincères remerciements pour l'intérêt qu'il a manifesté pour ce travail;

Monsieur Richard Bouché, Professeur à l'Ecole Nationale Supérieure Bibliothécaire de Lyon, pour son aimable participation à ce jury;

Monsieur Philippe Cinquin, Professeur à l'Université Joseph Fourier, dont les compétences dans le domaine d'application ont été fort utiles dans la réalisation de ce travail;

Monsieur Yves Chiaramella, Professeur à l'Université Joseph Fourier, Directeur du Laboratoire de Génie Informatique et Responsable de l'équipe Systèmes Intelligents de Recherche d'Informations, qui m'a accueilli dans son équipe et qui a consacré un temps important pour diriger ces travaux. Il a, par ses critiques constructives et pertinentes, fait progresser ce travail, ses

relectures ont permis d'améliorer la qualité de ce document, sa patience et son amitié m'ont encouragé pendant toutes ces années;

Mademoiselle Annie Culet, Monsieur Patrick Palmer et Mademoiselle Dalila Kerkouba, pour l'amitié, la gentillesse et l'aide qu'ils m'ont toujours accordées, grâce auxquelles j'ai encore plus apprécié mon séjour en France et échappé à la nostalgie ...;

Monsieur Michel Simonet, pour les améliorations qu'il a apportées à ce document;

Madame Marie-France Bruandet, pour les encouragements qu'elle m'a prodigués;

Tous les membres de l'équipe Système Intelligents de Recherche d'Informations, les Catherines, Bruno, Jean-Pierre, Sahar, Jean, ... et les collègues du Laboratoire de Génie Informatique pour leur amitié et leur collaboration;

Yan et Hugo, pour leurs travaux qui ont contribué au prototype RIME;

Et surtout, mon arlésienne et mon dragon, qui m'ont toujours soutenu et encouragé, ainsi que mes parents, qui se sont beaucoup inquiété pour mon travail.

Table des matières

PARTIE 0

INTRODUCTION

1. DÉFINITION D'UN SYSTEME DE RECHERCHE D'INFORMATIONS

Un *système de recherche d'informations* (SRI) est un système qui stocke, gère et manipule un ensemble de *documents*, de façon à permettre aux utilisateurs de retrouver ceux qui *correspondent* à leur *besoin* d'information.

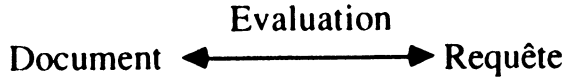
Dans ce contexte, la notion de *document* a beaucoup évolué. Elle était dans un premier temps identifiée à des textes dont le contenu était représenté par des mots-clés. Plus tard, cette notion a été étendue aux textes originaux. Actuellement, elle est en train de s'étendre à tout type d'informations: les textes, les images, les sons ... De plus, les documents peuvent être structurés ou non.

Les *besoins* d'un utilisateur conduisent à la sélection de documents satisfaisant certaines conditions. Ils sont exprimés à travers une *requête*, pouvant être considérée comme un ensemble de spécifications d'attributs (ex: auteur, date, contenu ...) relatifs aux documents que l'utilisateur veut obtenir. En parallèle de l'évolution de la notion de document, la notion de besoin a aussi évolué. Non seulement les utilisateurs peuvent exprimer leurs besoins en spécifiant les *attributs externes* (les attributs autres que le contenu), mais également les *attributs internes*, relatifs au contenu sémantique des documents.

La réponse à une requête est constituée des documents (ou des références aux documents) que le SRI estime *correspondre* à la requête. La définition de ce *critère* de *correspondance* varie beaucoup selon les systèmes: selon la solution choisie, un document peut être sélectionné par un SRI et pas par un autre. Le problème clé pour un SRI est donc de définir correctement le critère de correspondance. L'*évaluation* d'une réponse consiste alors à juger si un document correspond à une requête ou non selon le critère défini.

La définition du critère d'évaluation a aussi beaucoup évolué. Elle était fondée au départ sur des comparaisons superficielles de données: il s'agissait simplement de déterminer si deux données étaient identiques ou non. De plus en plus, elle s'étend vers des comparaisons sémantiques de données.

Les trois notions spécifiées ci-dessus (document, besoin, correspondance) impliquent chacune un composant essentiel des SRI: les documents, la requête, et l'évaluation. La relation entre ces trois composants peut être représentée comme suit:

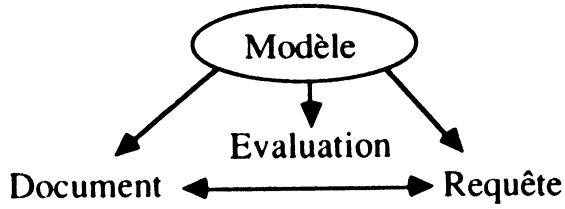


2. MODELES DE SRI

Définition:

> Un *modèle* de SRI définit le modèle des documents et des requêtes, ainsi que la *méthodologie* d'évaluation de la correspondance entre requête et documents.

Ainsi, un modèle de SRI peut être encore représenté comme suit:



Les modèles de SRI proposés sont très variés, notamment en ce qui concerne les deux aspects ci-dessous:

- Le critère d'évaluation d'un modèle peut être fondé sur une comparaison stricte des documents et de la requête (c'est-à-dire sur l'identification des termes de la requête avec ceux des documents), ou bien sur une comparaison plus souple, prenant en compte la sémantique de la requête et celle des documents.

- Un modèle peut être défini avec un degré d'abstraction plus ou moins élevé quant à la représentation des données.

Relativement à ces deux points, on peut regrouper les modèles en trois catégories. Apparus successivement, ils correspondent à des niveaux d'abstraction de plus en plus poussés:

- les modèles classiques
- les modèles intelligents
- les modèles abstraits ou formels

2.1. Modèles classiques

Par *modèles classiques*, on fait référence aux premiers modèles de SRI développés dans les années 70, tels que le modèle booléen, le modèle vectoriel et le modèle probabiliste indépendant.

Ces modèles classiques possèdent deux caractéristiques communes:

1. la dépendance par rapport à une représentation spécifique des données
2. la non prise en compte de la sémantique

A chacun de ces modèles correspond un mode particulier de représentation des données: un document et une requête sont considérés dans le modèle booléen comme une *expression booléenne* et dans le modèle vectoriel comme un *vecteur*, ... Un modèle classique ne s'applique donc pas à une représentation autre que celle qui lui est propre.

La notion de sémantique est absente dans ces modèles. Un document et une requête sont considérés comme bâtis sur un ensemble de termes indépendants, ces termes étant des mots-clés. Chaque mot-clé possède une sémantique déterminée différente de celle de tout autre mot-clé. Aucune relation n'est considérée entre deux termes différents.

2.2. Modèles intelligents

Les *modèles intelligents* sont ceux qui permettent une évaluation de correspondance fondée (plus ou moins) sur des comparaisons sémantiques entre document et requête.

Une évaluation sémantique se traduit par l'utilisation d'un ensemble de relations sémantiques entre les termes pour effectuer certains *raisonnements* (*déductions*). Par rapport aux modèles classiques, deux problèmes supplémentaires s'ajoutent:

(1). Quelles sont les relations entre les données à prendre en compte?
(2). Comment la déduction peut-elle s'effectuer?

Actuellement, on peut répondre à la première question par une contrainte plutôt que par une vraie réponse, car les relations qui *doivent* être prises en compte dans une application sont souvent tellement nombreuses qu'il est impossible de toutes les considérer. La première question doit donc être modifiée en une autre question:

(1 bis). Quelles sont les relations que le système est capable de prendre

en compte de manière efficace?

Les relations qu'un SRI peut prendre en compte, sont souvent stockées dans une base appelée un *thésaurus*. Les types de relations les plus souvent considérés dans un thésaurus sont les suivantes ([Croft85, Croft88, ...]):

la relation spécifique (la relation générique)
la relation de synonymie
la relation contextuelle (la relation de synonymie dans un contexte donné)
la relation de voisinage sémantique
la relation voir-aussi
...

On peut aussi trouver d'autres représentations possibles des relations, par exemple en utilisant des règles de production, ...

En ce qui concerne l'utilisation des relations pour effectuer des déductions, les techniques développées dans le domaine de l'intelligence artificielle (IA) figurent aussi souvent dans les SRI intelligents. En effet, le problème de déduction dans les SRI est tout à fait comparable à celui en IA. Ce dernier est présenté dans beaucoup d'ouvrages ([Davis77, Duda77, Laurière81]). Nous ne le détaillons donc pas ici.

Parmi les modèles existants présentés en I.2, le modèle sémantique, le modèle linguistique et le modèle logique sont considérés comme des modèles intelligents.

2.3. Modèles formels

On appelle les *modèles formels* ceux qui sont très indépendants d'une représentation spécifique de données (les autres sont appelés par opposition les *modèles spécifiques*).

Les modèles classiques sont des modèles spécifiques (cf.2.1). Dans certains modèles intelligents (ex: le modèle sémantique et le modèle linguistique (cf.I.2)), la dépendance entre le modèle et une représentation spécifique existe, mais est souvent beaucoup moins marquée. En effet, ces modèles s'appliquent souvent à un *ensemble* de représentations de données plutôt qu'à une représentation spécifique.

Par exemple, dans le modèle sémantique de Croft ([Croft85, Croft88]), l'opération d'évaluation s'applique sur plusieurs modèles classiques (le modèle booléen, le modèle vectoriel, le modèle probabiliste ...).

Ces modèles ne sont donc pas tout à fait des modèles généraux, mais ils ne sont pas non plus des modèles spécifiques. On les considère comme des modèles intermédiaires.

Le modèle logique de van Rijsbergen ([vanRijsbergen86]) (cf.I.2.5) est indépendant de toute représentation spécifique de données. Les documents et les requêtes sont supposés interprétés dans une représentation quelconque. C'est donc un modèle général, purement formel.

3. SRI ET MODELES LINGUISTIQUES

3.1. Les SRI fondés sur des approches linguistiques

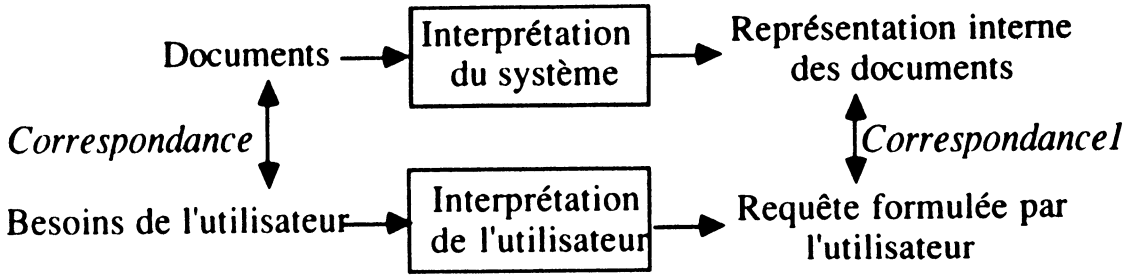
Les SRI sont destinés, dans la plupart des cas, à des utilisateurs non spécialistes ou occasionnels qui n'ont pas de connaissances à priori sur le système. Les SRI doivent donc fournir des moyens d'utilisation simples du système. Les SRI anciens possèdent souvent un langage spécifique pour l'interrogation, du type booléen, vectoriel .... Pour les interroger, soit les utilisateurs ont besoin d'une formation préalable concernant le langage d'interrogation, soit les requêtes des utilisateurs sont formulées par un spécialiste jouant le rôle d'intermédiaire. Ces deux solutions ne sont pas conviviales. Pour faciliter leur accès, les SRI développés actuellement, tendent à adopter de plus en plus une interface "évoluée". Un des moyens consiste à utiliser un langage d'interrogation plus "naturel", dans lequel l'utilisateur peut formuler des requêtes librement sans formation préalable.

Une raison supplémentaire en faveur de l'interrogation en langue naturelle est qu'il est plus facile pour un utilisateur occasionnel de formuler une requête en langue naturelle exprimant correctement ses besoins qu'en un langage d'interrogation artificiel.

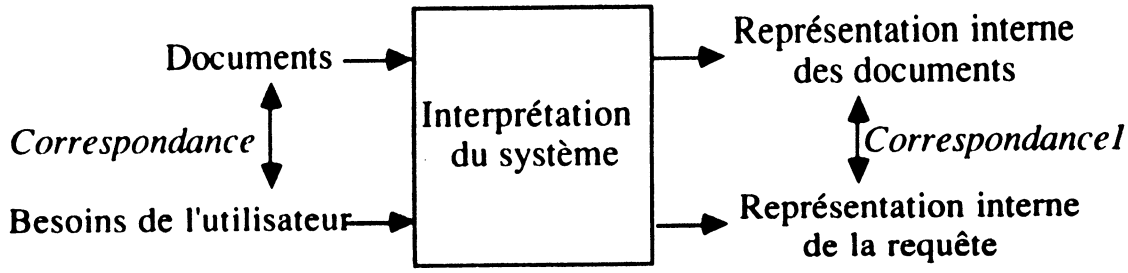
Le processus d'indexation visant à représenter dans un format interne le contenu des documents, eux-mêmes souvent en langue naturelle, impose une certaine interprétation des documents. Si un langage artificiel est utilisé, les utilisateurs doivent exprimer leurs besoins dans ce langage pour l'interrogation. La formulation choisie reflète leur propre interprétation de la requête (Fig.0.1.a). En effet, celle-ci doit tenir compte de la représentation interne choisie pour les documents. Pour un utilisateur occasionnel, un écart peut facilement se produire entre les deux représentations.

Si le système possède un langage d'interrogation naturel (ou proche), c'est au système d'interpréter les besoins de l'utilisateur dans une représentation interne (Fig.0.1.b). En interprétant la requête avec un

processus "identique" à celui utilisé lors de l'indexation des documents, l'écart peut être diminué.



a. *Interrogation avec un langage artificiel: correspondance ≠ correspondance1*



b. *Interrogation en langue naturelle: correspondance ≈ correspondance1*

Fig.0.1. *La comparaison entre une interrogation avec un langage artificiel et une interrogation en langue naturelle*

Diverses techniques sont proposées pour effectuer une interrogation en langue naturelle, mais toutes les techniques sont constituées par trois (ou éventuellement moins) phases conceptuelles: *l'indexation des documents, l'interprétation des requêtes* et *l'évaluation de la correspondance.*

L'*indexation des documents* vise à identifier le contenu sémantique des documents du corpus et représenter celui-ci dans le modèle interne du système.

L'*interprétation des requêtes* est le dual de l'indexation. Elle vise à reconnaître les requêtes de l'utilisateur et à les représenter dans le format interne du système.

L'*évaluation de correspondance* vise à comparer la représentation interne d'une requête avec celle des documents afin d'estimer si un document est une bonne réponse à la requête.

L'*évaluation de la correspondance* entre une requête en langue naturelle et un document est donc remplacée par l'évaluation de la correspondance entre les deux représentations internes correspondantes.

L'indexation et l'interprétation s'appuient beaucoup sur l'analyse linguistique des documents et des requêtes. Beaucoup de moyens existent. Selon les outils utilisés dans l'analyse et l'objectif visé, on peut les regrouper en trois catégories (cf. Annexe): analyse syntaxique, analyse sémantique et analyse syntaxico-sémantique.

Une *analyse syntaxique* opère sur la structure des phrases, et permet de reconnaître l'ordonnacement des syntagmes à l'intérieur de celle-ci.

Une *analyse sémantique* vise à reconnaître le sens véhiculé par une phrase (pour aboutir à une représentation de la sémantique de la phrase).

Une *analyse syntaxico-sémantique* fait collaborer des informations de nature syntaxique et sémantique pour aboutir à une reconnaissance syntaxique et/ou sémantique de la phrase.

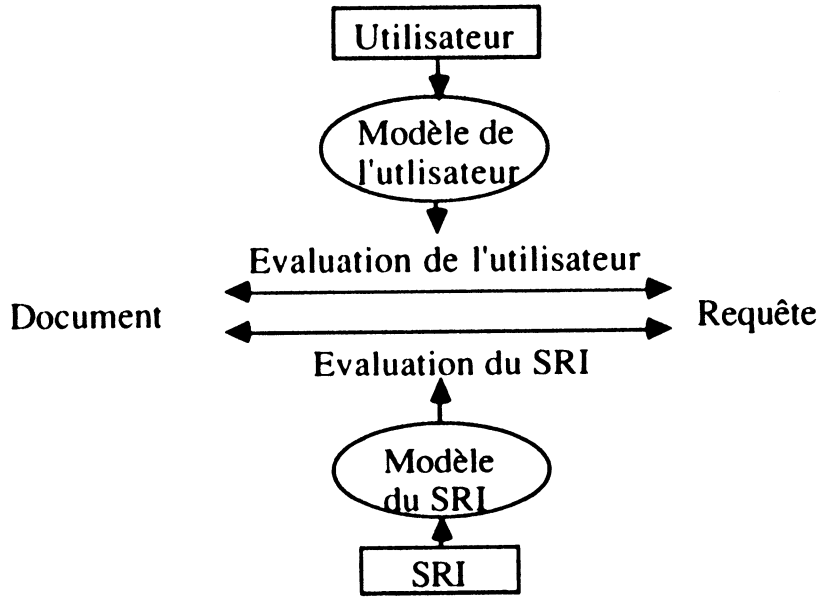
Une évaluation de SRI est théoriquement fondée sur des comparaisons sémantiques. Cela implique que les outils purement syntaxiques ne sont pas suffisants pour réaliser l'indexation des documents et l'interprétation des requêtes. Mais les analyses purement sémantiques s'avèrent souvent très coûteuses (cf. Annexe). Ainsi, les analyses syntaxico-sémantiques apparaissent comme une optique raisonnable.

3.1. Evaluation qualitative d'un SRI

Lorsqu'un document est fourni comme réponse à une requête, l'utilisateur évalue lui aussi la correspondance du document par rapport à son besoin. Son évaluation n'est pas nécessairement celle du SRI. Autrement dit il existe souvent un écart entre la correspondance évaluée par le système et celle évaluée par l'utilisateur. Il y a en fait deux modèles de correspondance: un modèle au niveau du système et un modèle au niveau de l'utilisateur.

Le modèle au niveau du système correspond à ce que le SRI réalise. Tandis que le modèle au niveau des utilisateurs correspond à ce que les utilisateurs souhaitent.

De ce fait, l'ensemble d'un SRI et de ses utilisateurs peut être considéré de la façon suivante:



Pour différencier ces deux évaluations, on appelle le résultat de l'évaluation des utilisateurs la *pertinence* (terme classique) et celle du système la *correspondance*. Cette différence sera discutée plus en détail dans la partie I.

Dans un SRI idéal, l'évaluation de la correspondance devrait être identique à celle de la pertinence. Mais cela est peu réaliste. Pour qualifier la proximité entre les deux évaluations, plusieurs paramètres sont souvent utilisés ([Salton83a]): *la précision, le rappel, le bruit* et *le silence.*

La *précision* mesure le taux de documents pertinents par rapport à l'ensemble des documents extraits:

$$\text{précision} = \frac{\text{nb(documents pertinents extraits)}}{\text{nb(documents extraits)}}$$

(où "nb" est une fonction mesurant le nombre)

Le *rappel* mesure le taux de documents pertinents extraits par rapport à l'ensemble des documents pertinents existants dans le corpus:

$$\text{rappel} = \frac{\text{nb(documents pertinents extraits)}}{\text{nb(documents pertinents)}}$$

Ces paramètres mesurent en effet la similarité existant entre le modèle de l'utilisateur et le modèle du SRI. Plus la valeur de ces deux paramètres est grande, plus le SRI est considéré proche de celui espéré par l'utilisateur et donc plus le SRI est considéré comme performant.

Le *bruit* mesure le taux de documents non pertinents extraits par rapport à l'ensemble des documents extraits:

$$\text{bruit} = \frac{\text{nb(documents non pertinents extraits)}}{\text{nb(documents extraits)}}$$

Le *silence* mesure le taux de documents pertinents non extraits par rapport à l'ensemble des documents pertinents existants dans le corpus:

$$\text{silence} = \frac{\text{nb(documents pertinents non extraits)}}{\text{nb(documents pertinents)}}$$

Ces deux paramètres mesurent la différence existant entre le modèle de l'utilisateur et le modèle du SRI.

Les relations entre ces paramètres sont les suivantes:

précision + bruit = 1
rappel + silence = 1

Un SRI de bonne qualité doit avoir un taux de précision et de rappel élevé et donc un taux de bruit et de silence bas.

5. ORGANISATION DE LA THESE

L'exposé précédent a souligné deux problèmes dans le domaine des SRI: l'établissement d'un modèle SRI général et l'intérêt d'un SRI doté d'une interface en langue naturelle.

1). Il apparaît que les développements actuels dans le domaine des SRI sont arrivés à un stade où l'établissement d'un modèle général devient crucial.

• *l'établissement d'un modèle général est un moyen pour comparer différents modèles spécifiques.*

Nous avons vu qu'il existe de nombreux modèles spécifiques dans le domaine des SRI. Etant donnée une application, il est possible de comparer deux modèles différents avec les paramètres précédents (rappel, précision, ...). Mais ces paramètres ne mesurent que les *caractéristiques extérieures* (le résultat) d'un SRI. Celui-ci est alors considéré comme une boîte noire. Aucun paramètre permettant de comparer les *caractéristiques internes* (telles que la définition du critère de l'évaluation, le processus d'évaluation) n'existe. Autrement dit, on peut conclure qu'un SRI est meilleur qu'un autre, mais on

ne peut pas savoir pourquoi.

L'établissement d'un modèle général permet de décrire différents modèles spécifiques dans un même cadre. Etant donnée une application, il est alors possible de comparer ces modèles spécifiques en fonction de leurs caractéristiques internes, c'est-à-dire de savoir pourquoi un modèle est mieux adapté à cette application qu'un autre.

• *l'établissement d'un modèle général est la base du développement ultérieur des SRI*

Dans les développements relatifs à n'importe quel domaine scientifique, il y a toujours une phase d'*abstraction*. Celle-ci permet de dégager les problèmes clés du domaine et de créer une base théorique solide pour les études ultérieures.

De nombreuses études ont été effectuées sur les modèles spécifiques des SRI. Celles-ci concernent souvent des problèmes particuliers liés à la réalisation. Mais une phase d'abstraction est indispensable pour fonder une base solide et commune à tous les systèmes. Les études sur des modèles généraux ont déjà contribué à cette abstraction.

2). Les SRI fondés sur des modèles linguistiques font l'objet de nombreuses études. Cela est fortement lié à leur extension, car ceux-ci sont de plus en plus utilisés dans des contextes où les utilisateurs ne sont pas des spécialistes. Une interface en langue naturelle favorise largement l'accès à un SRI. Mais sa réalisation pose des problèmes liés au traitement de la langue naturelle. Ces problèmes pratiques sont non moins intéressants que le problème du modèle. La résolution de ces problèmes nécessite le développement de certaines techniques particulières qui sont indispensables pour l'extension du domaine des SRI.

Notre travail s'inscrit dans ces deux axes, et la suite de cette thèse est organisée en deux parties:

- Dans la première partie, une étude sur un modèle général permettra de dégager les critères de jugement des utilisateurs dans un contexte libre. Nous proposerons ensuite un modèle général de SRI fondé sur ces critères. Ce modèle général sera comparé avec les modèles existants pour démontrer sa généralité. Nous étudierons aussi le problème de la réalisation (au niveau général) des SRI fondés sur un tel modèle avec l'aide de différents outils informatiques développés dans d'autres domaines, notamment celui des bases de données.

- Dans la seconde partie, nous appliquerons le modèle général proposé à un SRI particulier ayant une forte composante linguistique (RIME), réalisé dans le contexte du domaine médical (radiologique). Parmi les trois processus fondamentaux d'un SRI, l'indexation des documents, l'interprétation des requêtes et l'évaluation de la correspondance, notre travail s'est concentré sur les deux derniers (fonction d'interrogation). La partie indexation a fait l'objet d'une autre thèse ([Berrut88]).

- Nous présentons ensuite les réalisations et évaluations effectuées autour d'un système prototype réalisé en Prolog, pour conclure finalement sur les principaux acquis de cette étude et leurs prolongements possibles.

PARTIE I

UN MODELE GÉNÉRAL POUR LES SYSTEMES DE RECHERCHE D'INFORMATIONS

1. INTRODUCTION

La plupart des systèmes de recherche d'informations (SRI) actuels sont fondés sur les modèles de recherche développés à partir de la fin des années 70, tels que le modèle booléen, le modèle vectoriel, le modèle probabiliste, etc... ([vanRijsbergen79, Salton83a]). Trois caractéristiques sont communes à ces modèles:

1. La spécificité à un domaine d'application particulier
2. La dépendance de la représentation des données
3. L'insuffisance de la prise en compte de la sémantique.

1). Les SRI sont utilisés dans des domaines très variés, par exemple, la recherche documentaire dans des bibliothèques, la recherche d'informations cliniques, la recherche de documents en bureautique, ... voire dans les systèmes de question-réponse. Les besoins des utilisateurs varient d'une application à une autre. L'objectif d'un modèle de SRI est de modéliser l'opération de recherche de sorte que le système puisse donner la même (ou presque la même) estimation de la pertinence que l'utilisateur. Pour satisfaire les différents besoins particuliers, de nombreux modèles sont proposés, chacun s'adaptant à un cas d'application. Cela implique qu'un tel modèle ne peut être appliqué correctement que dans le contexte pour lequel il a été conçu, car autrement, une incohérence sera engendrée entre le jugement du système et le jugement de l'utilisateur. A cause de cette diversité, il n'est pas possible d'établir un modèle uniforme à partir des modèles existants qui permette d'élaborer les fonctionnalités de la recherche d'informations d'une façon générale.

2). La dépendance d'une représentation spécifique de données existe dans la plupart des modèles existants. Etant donné un modèle, les données (les documents et les requêtes) doivent être représentées d'une façon prédéfinie. Par exemple, le modèle booléen n'est applicable que lorsque les données sont des expressions booléennes, le modèle vectoriel exige que les données soient des vecteurs, ... Cette dépendance a été peu remarquée dans les applications classiques, car les données y sont considérées comme étant sémantiquement indépendantes les unes des autres. Ainsi, une représentation de données peut facilement être transcrite dans une autre représentation. Dans les applications plus récentes, par contre, cette caractéristique est souvent un inconvénient, car elle présume une équivalence entre la sémantique et une représentation spécifique. Or, dans la recherche actuelle, la sémantique est souvent étudiée d'une façon générale, indépendamment d'une représentation spécifique. La

dépendance des modèles à une représentation spécifique des données ne correspond évidemment pas à cette tendance unificatrice des approches sémantiques.

En effet, un modèle SRI a pour but de modéliser le jugement de correspondance du système afin d'approcher le jugement de l'utilisateur. Celui-ci étant indépendant de la façon dont les documents et les requêtes sont représentés, il n'est donc pas cohérent que le modèle, pour satisfaire l'usager, soit fonction d'une représentation spécifique.

3). La discussion sur la deuxième caractéristique souligne naturellement l'insuffisance des modèles existants quant à l'aspect sémantique. Dans les SRI, les données sont souvent représentées à l'aide d'un ensemble de *termes*. Dans les applications classiques, les termes sont considérés comme des *mots-clés* indépendants: aucune relation sémantique entre les mots-clés n'est prise en compte.

Ici, nous associons au terme "mots-clés" une signification bien spécifique: un mot-clé est la *seule* représentation reconnue du contenu sémantique de documents. La présence ou l'absence d'un mot-clé dans la représentation d'un document signifie la présence ou l'absence de l'information sémantique correspondante dans le document. Ceci est très différent de la notion récente de *terme*, dans laquelle un terme est considéré comme *une des* représentations possibles d'une certaine sémantique. Autrement dit, la présence d'un terme implique la présence d'une certaine sémantique, mais cette implication n'est pas nécessairement vérifiée en cas d'absence, car sa sémantique peut aussi être impliquée par d'autres termes. Nous adoptons le mot *descripteur* (cf.[Deogun88]) pour exprimer cette dernière signification de "terme". Sauf dans les cas ambigus où sa signification sera indiquée, le mot *terme* signifiera donc *descripteur*.

Les modèles existants fondés sur les mots-clés indépendants ne sont pas capables de prendre en compte cette multi-représentation de la sémantique. Dans les applications plus récentes, on est souvent amené à surajouter à un SRI classique un élément (un thésaurus) pour prendre en compte les relations sémantiques entre les descripteurs ([Croft85, Giger88]). Bien que cette approche permette une certaine prise en compte de la sémantique dans les SRI, une hétérogénéité est introduite: d'une part, on a des données indépendantes les unes des autres et l'évaluation est fondée sur cette indépendance, d'autre part, on a un ensemble de relations sémantiques sur les données qui modifient l'évaluation précédente en réintroduisant une dépendance. Cette hétérogénéité est évidemment préjudiciable à une approche plus fondamentale. En réalité, l'utilisation des relations sémantiques n'est pas intégrée dans la définition des modèles classiques.

Ces trois caractéristiques ne correspondent pas à la tendance de la recherche actuelle sur les SRI, ce qui justifie notre intérêt pour un modèle plus général, susceptible de les intégrer.

Depuis quelques années, de nombreuses études en recherche d'informations s'orientent vers la définition de modèles plus généraux. Comme dans d'autres domaines, cette activité se déroule selon deux tendances: la première consiste à étendre un modèle existant, afin de recouvrir un plus grand nombre de cas, la deuxième vise à développer un modèle à partir d'une nouvelle base théorique. Le modèle booléen étendu de Salton ([Salton83b]) est un exemple typique de la première approche. Ce modèle est capable de recouvrir une série de modèles déjà existants, du modèle booléen standard jusqu'au modèle vectoriel. Parmi les modèles développés selon la seconde approche, on trouve le modèle de Dabrowski ([Dabrowski75]) fondé sur la distribution des données et le modèle logique récemment développé par Van Rijsbergen ([vanRijsbergen86]). Notre approche a été surtout inspirée par le modèle de Van Rijsbergen, car les trois caractéristiques précédentes des modèles classiques y sont plus ou moins résolues: il est assez général pour recouvrir la plupart des modèles classiques, il est indépendant d'une représentation particulière des données et il a introduit une certaine prise en compte de la sémantique.

Selon nous, une bonne approche pour développer un SRI doit être fondée sur un modèle *général* pour assurer une bonne définition des fonctionnalités de base, pour ensuite adapter ce modèle aux cas particuliers d'application. Une fois que le modèle général est suffisamment développé, l'adaptation à un cas particulier n'est qu'un problème secondaire. Pour situer le contexte de cette étude, nous présentons quelques modèles représentatifs des SRI existants.

2. PRÉSENTATION DE QUELQUES MODELES EXISTANTS

2.1. Le modèle booléen

Le terme "modèle booléen" recouvre en fait de multiples formalismes particuliers: le modèle booléen standard, les modèles booléens modifiés, voire certains modèles flous. Pour des raisons de simplicité, nous ne considérerons ici que le modèle booléen standard.

Dans ce modèle, un document (D) est représenté par une conjonction de termes (ou de mots-clés) indépendants, que l'on représente sous forme d'un ensemble: $\{t_1, t_2, \ldots t_n\}$. Une requête (Q) est une expression logique composée de termes connectés par les opérateurs logiques $\vee$, $\wedge$ et $\neg$. Un document sera

jugé "pertinent" si l'expression logique de la requête est satisfaite par l'ensemble des termes du document (ou autrement dit, par l'expression de la conjonction des termes du document). Sinon, le document sera jugé "non-pertinent" pour la requête. Le modèle booléen est donc caractérisé par une valuation de pertinence binaire ("vrai" ou "faux").

Définissons une fonction P(D,Q) pour évaluer la pertinence d'un document D pour une requête Q. La valuation booléenne peut être plus précisément définie ainsi:

- $P(D,t_i)=1$ si $t_i \in D$ (où t_i est un terme)
- $P(D,Q_1 \wedge Q_2)=1$ si $P(D,Q_1)=1$ et $P(D,Q_2)=1$.
- $P(D,Q_1 \vee Q_2)=1$ si $P(D,Q_1)=1$ ou $P(D,Q_2)=1$.
- $P(D,\neg Q)=1$ si $P(D,Q)=0$.
- $P(D,Q)=0$, autrement.

2.2. Le modèle vectoriel

Dans ce modèle, un document D_i est représenté par un vecteur de termes d'indexation (ou de mots-clés) à n-dimensions:

$$D_i = (a_{i1}, a_{i2}, \ldots, a_{in})$$

où n est le nombre de termes d'indexation t_j définis dans le système, et $a_{ij} \in [0,1]$ $(1 \leq i \leq n)$ représente la pondération du terme t_j dans D_i. Ainsi, un ensemble de m documents peut être représenté par une matrice:

	t_1	t_2		t_n
D_1	a_{11},	a_{12},	...	a_{1n}
D_2	a_{21},	a_{22},	...	a_{2n}
...		...		
D_m	a_{m1},	a_{m2},	...	a_{mn}

Une requête Q est aussi formalisée par un vecteur de termes pondérés: $(b_1\ b_2\ \ldots, b_n)$, b_j étant le poids du terme t_j dans la requête.

Durant le processus d'évaluation, le système sélectionne les documents D_i dont la similarité avec la requête est la plus grande. Il existe diverses façons de calculer cette similarité. La plus utilisée ([Salton83a]) correspond au cosinus de l'angle des deux vecteurs Q et D_i dans l'espace à n dimensions:

$$Sim(Q, D_i) = \frac{\sum_j (a_{ij} \cdot b_j)}{[\sum_j (a_{ij})^2 \sum_j (b_j)^2]^{1/2}}$$

Deux autres mesures de similarité sont aussi souvent utilisées:

$$Sim(Q, D_i) = \frac{2\sum_j (a_{ij} \cdot b_j)}{\sum_j (a_{ij})^2 + \sum_j (b_j)^2}$$

$$Sim(Q, D_i) = \frac{\sum_j (a_{ij} \cdot b_j)}{\sum_j (a_{ij})^2 + \sum_j (b_j)^2 - \sum_j (a_{ij} \cdot b_j)}$$

2.3. Le modèle probabiliste

Etant donnée une requête Q, la probabilité de pertinence pour un document D est exprimée sous une forme conditionnelle - P(rel | D): c'est la probabilité que la réponse soit pertinente relativement à la requête, étant donné D comme réponse. Mais cette mesure n'est pas directement calculable. Elle est estimée selon la formule suivante ([vanRijsbergen79]) d'après le théorème de Bayes:

$$P(rel|D) = \frac{P_r(D)\ p(rel)}{P_r(D)\ p(rel) + P_n(D)\ p(nrel)}$$

où - $P_r(D)$ et $P_n(D)$ représentent repectivement la probabilité pour D d'être un document pertinent ou non-pertinent,

- p(rel) et p(nrel) sont respectivement la probabilité de pertinence et de non-pertinence d'un document quelconque.

Pour un corpus donné, p(rel) et p(nrel) sont supposés fixés, et correspondent à une sorte de connaissance *a priori*. L'estimation de P(rel|D) est donc ramenée à celle de $P_r(D)$ et $P_n(D)$. Mais ces dernières restent elles-mêmes difficiles à calculer. Un moyen de les estimer ([Salton83a]) est de choisir d'abord deux ensembles de documents considérés comme étant "pertinents" et "non pertinents", puis de calculer les probabilités d'apparition de chaque terme dans l'ensemble pertinent et non pertinent à partir du nombre d'occurrences des termes dans ces ensembles. Ces probabilités sont notées: $P_r(t_i)$ et $P_n(t_i)$. On peut alors établir les probabilités de pertinence et de non-pertinence d'un document en fonction des probabilités des termes contenus.

Dans le modèle probabiliste indépendant, par exemple, ces probabilités sont évaluées de la façon suivante:

$$P_r(D) = \prod_{t_i \in D} P_r(t_i),$$

$$P_n(D) = \prod_{t_i \in D} P_n(t_i)$$

Les probabilités ainsi établies peuvent difficilement être "correctes", car les probabilités d'un document quelconque ne sont en réalité pas fonction du choix initial des documents pertinents et non-pertinents et l'hypothèse d'indépendance des termes est trop simplificatrice. La révision des probabilités est donc nécessaire et cruciale au long de la vie du système pour ajuster leurs valeurs.

Si l'on suppose que les termes sont dépendants, le modèle probabiliste devient beaucoup plus compliqué. En général, les probabilités $P_r(D)$ et $P_n(D)$ doivent être calculées par les formules suivantes:

$$P_r(D) = P_r(t_1)\, P_r(t_2|t_1)\, P_r(t_3|t_1,t_2) \ldots P_r(t_n|t_1,t_2,\ldots,t_{n-1})$$

$$P_n(D) = P_n(t_1)\, P_n(t_2|t_1)\, P_n(t_3|t_1,t_2) \ldots P_n(t_n|t_1,t_2,\ldots,t_{n-1})$$

dans lesquelles, $P_r(t_i|t_1,t_2, \ldots,t_{i-1})$ (ou $P_n(t_i|t_1,t_2, \ldots,t_{i-1})$) représente la probabilité de pertinence (de non-pertinence) de t_i relativement à l'ensemble $t_1, t_2,\ldots,t_{i-1}$. Cette dépendance est d'ordre (i-1).

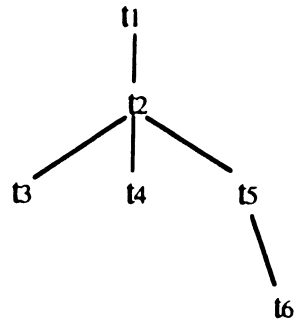
Supposons qu'il n'existe que des dépendances d'ordre 1 (i.e. un terme ne dépend que d'un autre terme), ces probabilités peuvent être alors calculées par les formules suivantes:

$$P_r(D) = \prod_{i=1}^{n} P_r(t_{m_i} \mid t_{m_{j(i)}}),$$

$$P_n(D) = \prod_{i=1}^{n} P_n(t_{m_i} \mid t_{m_{j(i)}})$$

où
- $(m_1,m_2,\ldots m_n)$ est la permutation de $1,2,\ldots,n$;
- $j(i)$ est une fonction donnant les entiers inférieurs à i;
- et en particulier, $P_r(t_i|t_0)$ et $P_n(t_i|t_0)$ signifient $P_r(t_i)$ et $P_n(t_i)$.

Par exemple, soit D=(t_1,t_2,t_3,t_4,t_5,t_6), et la dépendance suivante (cf.[vanRijsbergen79]):



Les probabilités $P_r(D)$ et $P_r(D)$ sont calculées par les formules suivantes:

$$P_r(D) = P_r(t_1)\, P_r(t_2|t_1)\, P_r(t_3|t_2)\, P_r(t_4|t_2)\, P_r(t_5|t_2)\, P_r(t_6|t_5)$$

$$P_n(D) = P_n(t_1)\, P_n(t_2|t_1)\, P_n(t_3|t_2)\, P_n(t_4|t_2)\, P_n(t_5|t_2)\, P_n(t_6|t_5)$$

2.4. Les modèles sémantico-linguistiques

Les modèles sémantico-linguistiques visent d'une part à fonder l'opération de recherche sur la sémantique, d'autre part, à considérer les documents (et éventuellement les requêtes) directement en langue naturelle.

Du côté linguistique, étant donné que la comparaison entre un document et une requête en langue naturelle ne peut pas s'effectuer directement, un intermédiaire est nécessaire: une représentation formelle du document et de la requête. Ces représentations formelles du document et de la requête sont communément désignées par *indexation de document* et *interprétation de requête*.

La comparaison entre le document initial et la requête initiale est donc ramenée à la comparaison entre le document indexé et la requête interprétée, qui se traduit par une mesure de correspondance. Cela peut être illustré comme suit:

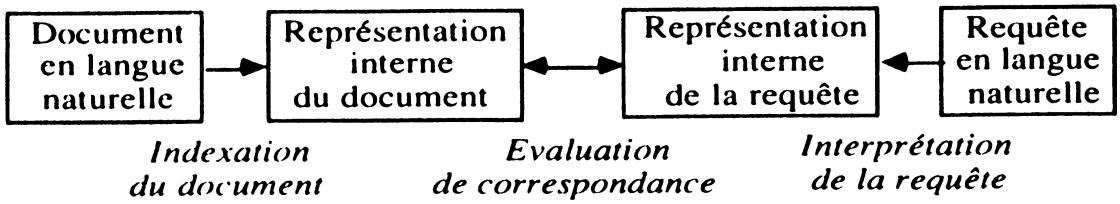


Fig.I.1. *Processus du modèle linguistique*

Beaucoup de méthodes d'indexation des documents et d'interprétation des requêtes sont proposées. En général, les représentations internes issues de ces indexations sont relativement similaires et consistent en un ensemble de *termes d'indexation* jugés représentatifs du contenu du document (cf.[Kerkouba85]). Les représentations formelles issues des différentes interprétations des requêtes sont souvent constituées d'un ensemble de termes d'indexation mis en relation (le plus souvent booléenne) les uns avec les autres.

Une grande variété de mesures de correspondance entre le document indexé et la requête interprétée est proposée. La plupart sont fondées sur la logique booléenne et/ou la théorie des ensembles flous ([Bookstein80, Salton83b, Waller79, ...]). Les documents fournis en réponse sont ordonnés selon la valeur de la mesure de correspondance avec la requête.

L'indexation des documents et l'interprétation des requêtes sont deux processus étroitement liés à la langue naturelle. A cause de la complexité et des ambiguïtés de celle-ci, ces deux processus ne peuvent jamais représenter *exactement* le document et la requête. Une certaine perte de précision est toujours introduite. Dans cette situation, il est évident que la mesure de correspondance fournie par le système n'est pas tout à fait celle qu'un utilisateur pourrait estimer entre le document initial et la requête initiale. Quelques mesures sont alors particulièrement utiles pour qualifier la performance d'un tel SRI: les taux de *rappel* et de *précision*, les taux de *silence* et de *bruit* (cf.Partie Introduction et [Salton83a]).

En ce qui concerne l'aspect sémantique, l'hypothèse de base ([Croft85]) est qu'une opération de recherche est dépendante non seulement de la requête et des documents acquis, mais aussi du domaine d'application, de l'utilisateur et d'une stratégie de recherche. Le modèle sémantique vise à fonder une opération de recherche spécifique sur ces données. Pour ce faire, il est nécessaire d'organiser et de représenter différentes *connaissances* dans le système pour spécifier respectivement le domaine et l'utilisateur, afin de choisir ensuite une stratégie de recherche appropriée.

La structure suivante est typique d'un SRI fondé sur le modèle sémantique ([Croft85]):

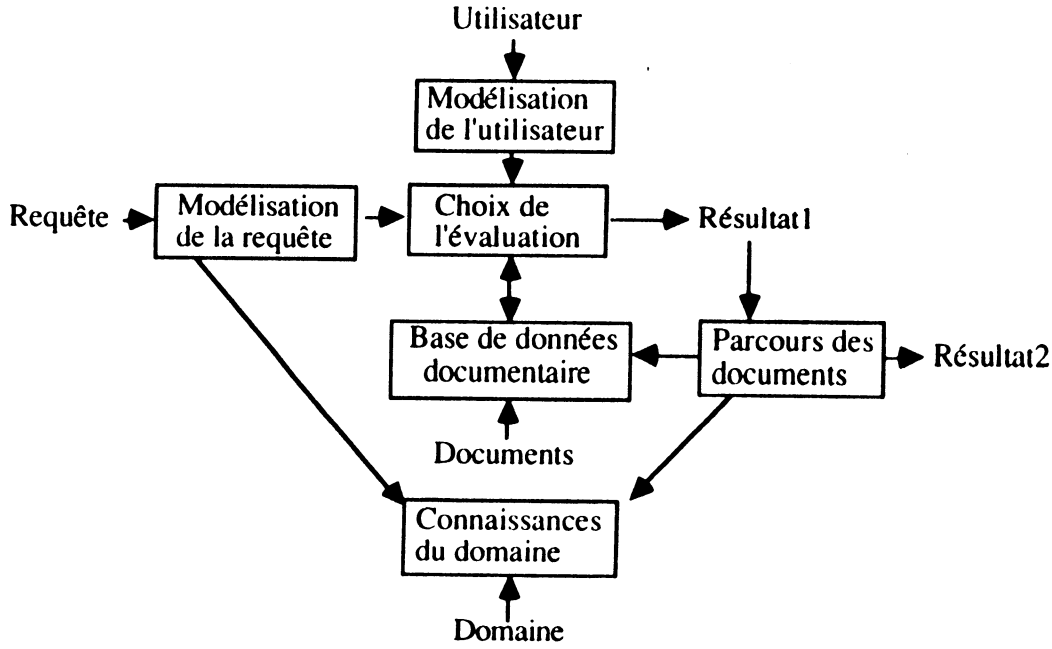


Fig.I.2. *Le schéma d'évaluation d'un modèle sémantique*

A titre d'exemple, nous présentons par la suite un modèle sémantico-linguistique qui a été développé par le groupe "Systèmes intelligents de recherche d'informations" du LGI (cf.[Chiaramella86, Defude86]): le modèle IOTA, qui intègre l'aspect linguistique et l'aspect sémantique.

Comme dans un modèle linguistique, une opération de recherche dans IOTA repose sur trois processus principaux: l'indexation des documents, l'interprétation de la requête et l'évaluation de la requête sur un ensemble de documents (ou de références aux documents). L'aspect sémantique est représenté par un système expert de recherche qui gère certaines connaissances sur l'utilisateur et sur la stratégie de recherche, correspondant à des connaissances expertes.

Le processus d'indexation ([Kerkouba85]) est fondé sur une analyse linguistique qui vise à représenter les documents par un ensemble de *termes d'indexation*, auquel sont associés deux mesures de représentativité. La structure logique des documents est prise en compte. Un document est hiérarchiquement divisé en un ensemble d'*unités textuelles* à différents niveaux structurels: chapitre, sous-chapitre, paragraphe et sous-paragraphe. A une unité (d) et un terme isolé (t), deux mesures statistiques sont associées: la *représentativité* de l'unité pour le terme REP(d,t), qui mesure le développement du thème (t) dans l'unité (d) relativement au reste du corpus et la *représentativité* du terme pour l'unité REP(t,d), qui mesure l'importance du thème (t) dans l'unité (d). Ces mesures sont respectivement définies de la façon suivante:

$$REP(d,t) = \frac{\text{fréquence d'apparition de t dans d}}{\text{fréquence d'apparition de t dans le corpus}}$$

$$= \frac{FLOC(t,d)}{FTOT(t)} \in [0,1]$$

$$REP(t,d) = \frac{\text{fréquence d'apparition de t dans d}}{\text{fréquence d'apparition de tous les termes dans d}}$$

$$= \frac{FLOC(t,d)}{TAILL(d)} \in [0,1]$$

A l'issue de l'indexation, à chaque terme d'indexation t est associé un ensemble de triplets:

$$t \leftrightarrow \{\langle d, REP(d,t), REP(t,d)\rangle : t \in d\}$$

Le processus d'interprétation des requêtes en "langue naturelle" ([Nie87]) est aussi fondé sur une analyse linguistique. Ce processus permet d'identifier les différents attributs d'une requête (tels que l'auteur, la date de publication ... et le contenu) et les transformer en une forme canonique, qui est une arborescence dont les feuilles sont des termes d'indexation (attributs de contenu), l'auteur, la date, ... (attributs externes), et dont les noeuds sont des opérateurs de recherche (du type booléen).

Le système expert de recherche ([Defude86]) permet ensuite de comparer la requête interprétée avec les documents indexés, en évaluant une mesure de pertinence pour chaque document retrouvé.

Schématiquement, le processus d'évaluation se décompose en plusieurs phases à partir d'une requête interprétée:

1. estimation de la typologie de l'utilisateur
2. évaluation de la requête (recherche d'unités de documents pertinents)
3. estimation de la qualité de la réponse
4. reformulation automatique éventuelle de la requête en fonction de la qualité de la réponse et de la typologie de l'utilisateur

1). La première phase est une modélisation très simple de l'utilisateur. On considère qu'un utilisateur *spécialiste* espère avoir des réponses précises relativement à sa requête, tandis qu'un utilisateur *débutant* peut être satisfait par des réponses moins précises. L'estimation de la typologie d'utilisateur est fondée essentiellement sur la spécialisation des termes employés dans la requête. L'utilisation de termes spécialisés est considérée comme caractéristique d'un utilisateur spécialiste.

2.) La requête étant ramenée à une expression booléenne, l'évaluation est composée de deux phases: une sélection des documents susceptibles de répondre à la requête et une estimation de la "pertinence" pour chaque document sélectionné.

Pour une requête Q, l'évaluation consiste d'abord à mettre en correspondance les termes fournis par l'usager et ceux du système (termes d'indexation). Cette opération est réalisée via un pattern-matching syntaxique et sémantique ([Defude86]). Une fois cette transformation réalisée, on aboutit à une expression booléenne de termes d'indexation permettant la sélection des unités textuelles. Cette sélection utilise l'association (4) établie durant l'indexation et peut être vue comme la valuation d'une fonction f(Q) (où D est un document, t un terme d'indexation, Q_1 et Q_2 des sous-requêtes, $Q_i = t_i \mid bool(t_1, \ldots, t_n)$):

$$f(t) = \{D \in corpus : t \in D\}$$
$$f(Q_1 \wedge Q_2) = f(Q_1) \cap f(Q_2)$$
$$f(Q_1 \vee Q_2) = f(Q_1) \cup f(Q_2)$$
$$f(Q_1 - Q_2) = f(Q_1) - f(Q_2)$$
$$f(\neg Q_1) = corpus - f(Q_1)$$

Comme dans le modèle booléen, tout document sélectionné doit satisfaire totalement l'expression logique de la requête.

3). La mesure de la pertinence associée à tout document appartenant à l'ensemble f(Q) est fonction des représentativités entre le document et les termes. Elle est définie d'une façon empirique comme ci-dessous:

Soient REP(t,D) et REP(D,t) les deux mesures de représentativité associées à un terme quelconque (t) par rapport à un document donné (D). Une fonction F est définie pour mesurer la correspondance entre une requête (Q) et un document (D):

$$F_1(t) = REP(t,D), \qquad F_2(t) = REP(D,t))$$

$$F_i(Q_1 \vee Q_2) = MAX[F_i(Q_1), F_i(Q_2)] \quad (i \in \{1,2\})$$

$$F_i(Q_1 \wedge Q_2) = \frac{F_i(Q_1) * F_i(Q_2)}{1 - F_i(Q_1) - F_i(Q_2) + 2 * F_i(Q_1) * F_i(Q_2)}$$

$$F_i(Q_1 - Q_2) = F_i(Q_1)$$

La mesure de la pertinence finale est définie par:

$$R(D,Q) = \frac{F_1(Q) + F_2(Q)}{2}$$

Les références ainsi pondérées sont ensuite soumises à un filtrage pour sélectionner les plus pertinentes. Ce filtrage est défini par des seuils sur les représentativités, différents selon la typologie de l'utilisateur. Il est fondé sur l'hypothèse qu'un spécialiste exige une réponse précise et de bonne qualité (mesure de pertinence élevée), alors qu'un débutant cherche plutôt un bon *rappel* et peut tolérer dans la réponse une pertinence moindre.

Une estimation globale portant sur la qualité et la quantité des références obtenues par rapport à la typologie de l'utilisateur permet de déduire un degré de satisfaction de l'utilisateur, celui-ci pouvant, s'il désire, provoquer une reformulation automatique de la requête.

4). Le processus de reformulation automatique est déclenché par un diagnostic qui indique (par rapport à une typologie d'utilisateur donnée) que le nombre de référence est trop faible, trop grand, ou que la qualité des réponses (la mesure de pertinence) est trop faible. Cette reformulation consiste à modifier soit les termes (utilisation des relations sémantiques du thésaurus), soit les relations logiques entre les termes dans l'expression booléenne de la requête. Cette nouvelle requête est alors reprise par le processus d'évaluation. Ce processus est réitéré jusqu'à satisfaction de l'usager ou obtention d'un degré de dégradation intolérable de la requête initiale (évalué par le système).

2.5. Le modèle logique

Il est généralement admis (dans les modèles classiques) qu'une interrogation en recherche d'informations est fondée sur une comparaison entre une requête de l'utilisateur et l'ensemble des documents disponibles dans le système. Dans le modèle logique de Van Rijsbergen ([vanRijsbergen86]), cette comparaison ensembliste est divisée en un ensemble de comparaisons élémentaires, chacune comparant la requête avec un document.

Un document est considéré comme un ensemble de "phrases" interprétées dans une certaine représentation sémantique prédéfinie. Il en est de même pour une requête, qui est généralement formée d'une seule phrase.

L'idée de base dans ce modèle est la suivante: étant donnés une requête Q et un document D, D répond bien à Q si D *implique* Q, ce qui est noté $D \rightarrow Q$. Ainsi, la mesure de la pertinence est transformée en une mesure de la *certitude de l'implication* de la requête par le document, notée $P(D \rightarrow Q)$. Remarquons que l'implication "$\rightarrow$" n'est pas une implication classique ($\supset$) comme dans la logique du premier ordre. Un contre-exemple simple est qu'un

document vide ne peut répondre à aucune requête non vide. Tandis qu'avec l'implication classique, "faux" $\supset x$ est toujours évalué à "vrai" (un document vide est équivalent à "faux", car aucune proposition peut y être satisfaite). A la différence des logiques classiques où l'implication est évaluée à une valeur stricte {vrai, faux}, l'implication dans les SRI est, dans la plupart du temps, "plausible". Ainsi, $P(D \rightarrow Q)$ est évaluée dans un intervalle continu [0,1] dont les valeurs extrêmes représentent "faux" et "vrai".

Dans le modèle logique, une valuation de la logique conditionnelle est proposée pour mesurer la certitude de l'implication, qui est représentée par la formule suivante:

$$P(D \rightarrow Q) = P(Q|D) = \frac{P(D \cap Q)}{P(D)} \qquad (1)$$

La partie droite de cette équation représente la probabilité que Q soit vraie étant donné que D est vrai, ou autrement dit, la probabilité que Q soit satisfaite étant donné D comme réponse. Nous donnons ci-dessous un exemple d'une telle valuation ([vanRijsbergen86]):

Soit A l'évènement "un nombre inférieur à 3 sera sélectionné" en jetant un dé et B l'évènement "un nombre pair sera sélectionné". P(B|A) est évalué à:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{2/6} = \frac{1}{2}$$

Tandis que $P(A \supset B) = P(\neg A \vee B) = \frac{5}{6}$

Lorsque la requête et le document sont représentés par des ensembles Q et D de termes indépendants, si $n(\mathcal{E})$ désigne la cardinalité d'un ensemble $\mathcal{E}$, cette formule devient:

$$P(D \rightarrow Q) = \frac{n(D \cap Q)}{n(D)} \qquad (1')$$

Cette expression représente la proportion de la partie de document concernant la requête $(D \cap Q)$ par rapport au document entier (D).

En généralisant cette approche, Van Rijsbergen propose un principe d'incertitude pour la valuation de la certitude de l'implication $D \rightarrow Q$:

Principe d'incertitude 1 :

> Etant données deux phrases x et y, une mesure d'incertitude de x→y relative à un ensemble d'informations donné, est déterminée par l'extension minimale de cet ensemble que l'on doit effectuer pour établir la vérité de x→y.

Etant donné qu'un document est représenté par un ensemble de phrases et une requête par une phrase, ce principe exprime dans sa forme la plus générale la mesure de la correspondance entre le document et la requête. Si une ou plusieurs phrases du document impliquent (d'une certaine façon) la phrase de la requête, la requête est alors considérée comme satisfaite. L'ensemble d'informations correspond à l'ensemble de connaissances dont on dispose au moment de la valuation de l'implication.

Si l'implication x→y ne peut pas être satisfaite par rapport à cet ensemble d'information, on doit ajouter de nouvelles informations de façon que l'implication devienne satisfaite par rapport à l'ensemble d'informations étendu. L'*extension minimale* signifie qu'au cours de la mesure d'incertitude, seules les informations nécessaires à la satisfaction de x→y sont *ajoutées*.

Une information ajoutée peut être de deux types différents: elle peut être la conséquence logique déduite des informations déjà existantes dans l'ensemble, elle peut aussi être totalement indépendante de celles-ci. L'ajout d'une information du premier type n'affecte pas la certitude de x→y, tandis qu'un ajout du second type diminue cette certitude. Nous pouvons aussi exprimer cela avec la notion d'état de l'ensemble d'informations: l'ajout d'une conséquence logique ne change pas l'état de l'ensemble d'informations, tandis que l'ajout d'une nouvelle information provoque un changement d'état. La certitude de x→y est déterminée d'après la relation entre deux états.

Nous pouvons appliquer ce principe pour l'exemple précédent de la manière suivante:

Soient D un document contenant "un nombre inférieur à 3" et Q une requête demandant "un nombre pair". L'information dont on dispose permet de dire que D contient 1 ou 2. Cette information ne suffit pas pour satisfaire totalement la requête. Pour satisfaire celle-ci, une information supplémentaire doit s'ajoute - "D contient un nombre supérieur à 1". La requête Q est alors totalement satisfaite par le document D par rapport à ce nouvel ensemble d'informations. L'extension effectuée sur l'ensemble d'informations réduit de moitié de l'ensemble des valeurs possibles contenues dans D. L'incertitude de l'application par rapport à l'ensemble initial d'informations est donc déterminée par l'extension effectuée, évaluée à 1/2.

La comparaison de ce modèle avec les modèles existants a montré sa généralité et sa puissance descriptive: beaucoup de modèles existants peuvent être décrits dans ce modèle ([vanRijsbergen86]). A notre avis, la généralité de ce modèle provient de l'utilisation de l'implication $D \rightarrow Q$ pour représenter la correspondance entre une requête et un document. En effet, l'implication de la requête par le document (cf. $D \rightarrow Q$) est un critère très important dans l'évaluation des SRI. Bien qu'il existe actuellement diverses façons de mesurer la correspondance entre document et requête, la plupart de ces mesures reflètent la valuation de $P(D \rightarrow Q)$ et peuvent donc être exprimées sous forme de $P(D \rightarrow Q)$.

Le modèle logique n'a spécifié aucun mode de représentation sémantique. Ce modèle, en effet, s'appuie surtout sur une formalisation des fonctionnalités de base d'un SRI. La particularité d'une représentation spécifique de données ne doit pas influencer le modèle si le modèle est suffisamment général et bien développé. Le problème de la définition de structures de données adaptées à la mise en oeuvre d'un modèle est secondaire et ne modifie pas la base philosophique du modèle (ceci rejoint évidemment un principe rencontré dans beaucoup d'autres domaines, et nous montrerons des applications relatives aux SRI dans les chapitres qui suivent).

La formule de valuation proposée dans le modèle logique (1 et 1') a néanmoins une forme particulière qui ne permet pas de retrouver beaucoup de valuations pratiquées dans les modèles existants. Par exemple, la valuation dans le modèle booléen est fondée uniquement sur la satisfaction de la requête par le document (qui n'est pas exprimée dans (4')), la proportion de la partie de document concernant la requête (4') n'ayant pas d'importance. Ainsi, on peut conclure que la formule (4') n'est pas une formule de valuation générale.

Le principe d'incertitude montre un aspect très intéressant: la mesure de l'implication est transformée en une comparaison entre deux ensembles d'informations. En effet, diverses significations sont associées à une implication dans la littérature. On peut la considérer comme une implication dans la logique du premier ordre, ou bien comme une implication dans la théorie des ensembles flous .... Van Rijsbergen a généralisé ces diverses considérations dans le principe d'incertitude. Pour obtenir l'interprétation de la logique du premier ordre par exemple, la règle suivante doit être établie: si une proposition n'est pas satisfaite par rapport à un ensemble de connaissances, une extension des connaissances doit s'effectuer. Cette extension (quelle qu'elle soit) correspond à l'incertitude de la satisfaction de la proposition par rapport à l'ensemble initial des connaissances. En d'autres termes, si une proposition n'est pas satisfaite par rapport à un ensemble de connaissances et que l'on ne peut pas étendre cet ensemble pour qu'elle soit satisfaite, la proposition est évaluée à "faux".

Un autre aspect intéressant dans ce principe est sa vision dynamique de la valuation: au lieu de mesurer une implication localement à un état donné, Van Rijsbergen propose de le faire en tenant compte des évolutions possibles d'états. Ainsi, ce principe introduit une vision plus large de la valuation.

En considérant l'évolution d'états sous-jacente à ce principe, on peut s'apercevoir qu'elle est similaire à la notion de transition entre un monde initial et des mondes possibles dans la logique modale ([Hughes68]). On peut donc approfondir l'idée du principe d'incertitude au moyen de la logique modale.

3. DÉFINITION D'UN MODELE GÉNÉRAL

3.1. Introduction

Pour obtenir un modèle d'évaluation des requêtes applicable à tous les cas (ou presque), il est indispensable de mettre en évidence les aspects importants de l'évaluation d'une requête. Nous allons donc analyser d'abord quelques exemples typiques pour dégager intuitivement ces aspects fondamentaux et les exprimer ensuite à travers une expression formelle générale.

La question posée est la suivante: comment peut-on caractériser un document répondant "parfaitement" à une requête?

Une des conditions nécessaires exprimée dans la plupart des modèles existants est que le document doit satisfaire *toute* la requête. En logique, ceci signifie: étant donné un document D, la requête Q doit être totalement satisfaite, et l'implication "document→requête" doit être évaluée à vrai, *i.e.* $P(D \rightarrow Q)=1$.

Prenons le cas du modèle booléen: une requête comportant une conjonction de deux termes $Q=t_1 \wedge t_2$ ne peut être satisfaite par un document D que si les deux termes sont mentionnés dans ce document: $t_1 \in D$ et $t_2 \in D$. La requête doit donc être totalement satisfaite.

Le fait que l'implication soit évaluée à "vrai" ou "faux" constitue un cas idéal pour la valuation. Dans la plupart des cas, cette implication ne peut être strictement vraie ou fausse, mais seulement "plausible". Deux cas de plausibilité peuvent être distingués: l'*implication incertaine* et l'*implication partielle*. Une implication est partielle quand seulement une partie de la requête est mentionnée dans le document. Nous illustrons ces deux cas par les exemples suivants. Supposons que les documents et les requêtes soient représentés par des termes (dépendants) non pondérés.

cas idéal: D={système de recherche d'informations, système expert}
Q={système expert}

cas non idéaux: D'={système de recherche d'informations, système expert}
Q'={système d'information}

D"={interrogation de base de données, représentation des connaissances}
Q"={interrogation de base de données, stockage de données}

Dans le cas idéal, la requête est totalement impliquée par le document. L'implication D→Q est donc évaluée à 1. Dans le premier cas non-idéal, l'implication est discutable. En effet, certains admettent qu'un "système de recherche d'informations" est un "système d'information" particulier ([Salton83a]). Mais d'autres les considèrent comme deux types de systèmes tout à fait différents. Par conséquent, l'implication D'→Q' n'est pas sûre. Dans le deuxième cas non-idéal, le document implique seulement une partie de la requête. L'implication D"→Q" est donc partielle.

La distinction entre ces deux cas non-idéaux apporte peu de différence quant au jugement sur la pertinence du document par rapport à la requête; elle est donc ignorée par la suite, et nous désignons ces deux cas non-idéaux par le terme unique d'*implication incertaine.*

Supposons maintenant que l'on ait deux documents D_1 et D_2 qui impliquent tous les deux *totalement* une requête Q et que les documents et la requête soient représentés par des listes de termes non pondérés. Cette situation peut être illustrée par l'exemple suivant:

Q = {base de données}
D_1 = {base de données}
D_2 = {intelligence artificielle, base de données, ...
système d'exploitation, système d'information, ...}

Intuitivement, il est bien évident que le document D_1 correspond "mieux" à la requête Q que le document D_2, parce que le thème "base de données" est plus important pour D_1 que pour D_2, et donc probablement mieux développé. La mesure de "correspondance" pour D_1 doit donc être plus élevée que pour D_2. La différence entre les deux documents ne porte pas sur la satisfaction du critère précédent, car tous les termes de Q (ici un seul - "base de données") sont inclus dans les deux documents, mais elle réside dans le développement du thème de la requête dans les deux documents. Sans

prendre en compte les tailles des documents, cette différence implique une différence d'importance du thème de la requête pour les documents. Cet aspect peut être qualifié par l'implication "thème→document". Etant donné que Q représente ce thème, cette implication peut être également représentée par Q→D, dont la valuation sera notée P'(Q→D).

3.2. Fondement du modèle

On peut donc conclure, à partir des discussions précédentes, qu'il existe deux critères pour évaluer la pertinence d'une réponse par rapport à une requête. Si on nomme le premier critère (D→Q) l'*exhaustivité* du document pour la requête, on peut nommer le deuxième (Q→D) la *spécificité* du document pour la requête. Ainsi la proposition suivante peut être établie pour une valuation de la correspondance:

Proposition :

Etant donnés un document D et une requête Q, la *correspondance* R entre D et Q est déterminée à la fois par l'exhaustivité du document pour la requête et par la spécificité du document pour la requête:

$$R(D,Q) = F[\ P(D \rightarrow Q),\ P'(Q \rightarrow D)\ ] \qquad (2)$$

où P et P' sont des fonctions mesurant la force d'implication, et F dénote une fonction combinant ces deux implications.

Quelques remarques sont nécessaires sur la terminologie employée. La *correspondance* que l'on a définie n'est pas exactement ce qui est classiquement défini par *pertinence* (ou *relevance*) ([Salton83a]). La *pertinence* est définie par rapport aux jugements des utilisateurs (ce sont les utilisateurs qui jugent si un document est "pertinent" relativement à une requête), tandis que la *correspondance* mesure la relation estimée *par le système* entre un document et une requête.

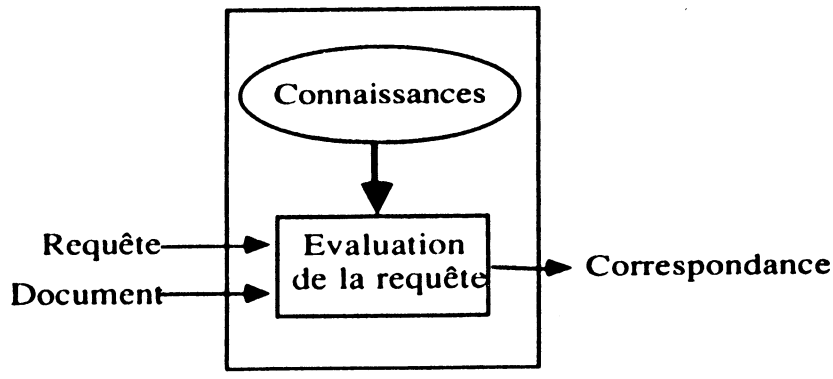


Fig.I.3. *Evaluation d'une requête*

En effet, un système d'évaluation de requêtes peut être vu, ainsi que le montre la figure I.1, comme dépendant d'un ensemble de connaissances. S'il existe une connaissance exprimant que "système expert" est un système particulier dans le domaine de l'"intelligence artificielle", un document sur "système expert" peut répondre à une requête portant sur l'"intelligence artificielle". Si cette connaissance est absente, le document ne peut pas être considéré comme "correspondant" à la requête.

En appliquant la notion classique de pertinence, c'est l'utilisateur qui joue le rôle du système pour juger si un document est "pertinent" pour une requête. Les connaissances dans Fig.I.1 seront donc celles de l'utilisateur.

La relation entre les deux notions de correspondance et de pertinence est donc la suivante: plus les connaissances du système s'approchent de celles de l'utilisateur, plus la mesure de correspondance se rapproche de la mesure de pertinence. Ainsi, ce rapprochement est souvent utilisé pour estimer le bon fonctionnement d'un système ([Salton83b]): un système donnant des mesures de correspondance équivalentes aux mesures de pertinence est donc considéré comme un bon système.

Cette différence de notions peut conduire à la conception de plusieurs niveaux de mesures. Dans un SRI, la valuation de la relation entre un document et une requête se fait à deux niveaux, celui du système et celui de l'utilisateur. La correspondance se fait au premier niveau et la pertinence au second niveau.

Pour conclure ces discussions, nous définissons une formule de valuation de relation (R) entre un document et une requête qui permet d'unifier les deux notions:

$$R(D,Q) = F\ [\ P_K(D \rightarrow Q),\ P'_K(Q \rightarrow D)\ ] \qquad (2')$$

où la valuation de la vérification des deux critères s'effectue par rapport à un ensemble de connaissances K (notée par P_K et P'_K), qui peuvent être celles du système ou celles de l'utilisateur.

Par la suite, on ne considérera que l'estimation du système, et on supposera que les connaissances du système (K) sont déjà déterminées quand on évaluera une requête.

3.3. Valuation des implications

Le principe d'incertitude de Van Rijsbergen exprime un moyen général pour la valuation d'une implication. L'implication est considérée comme une proposition à vérifier relativement à un ensemble de connaissances. Si la proposition est satisfaite relativement à l'ensemble initial de connaissances, elle est assignée à la valeur "vrai", sinon on doit élargir l'ensemble des connaissances successivement, pour que la proposition soit satisfaite par rapport à l'ensemble élargi des connaissances. Dans ce deuxième cas, la proposition par rapport à l'ensemble initial des connaissances est "vraie d'une façon incertaine". L'incertitude est déterminée par l'élargissement de l'ensemble des connaissances.

Comme nous l'avons déjà évoqué, l'idée de base de ce principe ressemble beaucoup aux transitions entre monde initial et "mondes possibles" dans la logique modale ([Hughes68, Zeman75]): si une proposition est satisfaite dans le monde initial, elle est assignée à "vrai". Sinon, le monde initial doit être transformé en des mondes possibles successifs jusqu'à ce que la proposition soit vérifiée dans un de ces mondes dérivés. Dans ce cas, la certitude de la satisfaction de la proposition par le monde initial dépend des transformations effectuées entre le monde initial et le monde possible dans lequel la proposition est satisfaite.

La notion d'ensemble d'informations évoquée dans le principe d'incertitude est comparable avec la notion de monde de la logique modale, et l'implication $x \rightarrow y$ peut être considérée comme une proposition à valuer.

Par rapport à la logique modale, on peut également proposer une autre façon de valuer la certitude de $D \rightarrow Q$:

On considère le document D comme le monde initial et la requête Q comme une formule à valuer. Si D peut directement satisfaire Q, alors $D \rightarrow Q$ est valuée à 1. Sinon, D doit être transformé en d'autres formes pour satisfaire Q. Ces formes transformées correspondent aux mondes possibles. La certitude de l'implication $D \rightarrow Q$ est déterminée selon la modification nécessaire effectuée entre le monde initial et le monde possible dans lequel la requête devient satisfaite.

Cette méthode de valuation de $D \rightarrow Q$ correspond tout à fait à une valuation de formule dans la logique modale. C'est cette idée que l'on va approfondir par la suite, au moyen de la logique modale. Nous rappelons brièvement les principes de la logique modale dans la section suivante.

3.3.1. La logique modale

Informellement, une logique modale peut être considérée comme une logique classique du premier ordre à laquelle sont ajoutées deux modalités: la *possibilité* ($\Diamond$) et la *nécessité* ($\Box$). La logique du premier ordre, ou logique propositionnelle, permet d'exprimer des propositions comme: "la propriété p est vérifiée par l'objet a". Une logique modale peut exprimer, en outre, qu'une propriété est *possiblement* vérifiée ou *nécessairement* vérifiée par un objet (NB: pour être cohérent avec la théorie de la logique modale, nous utilisons le mot possiblement, bien qu'il n'existe pas en français). Une propriété *possible* ($\Diamond p$) est celle qui est vraie dans au moins un cas, une propriété *nécessaire* ($\Box p$) est celle qui est vraie dans tous les cas. Par définition, on a:

$$\Diamond = \neg\Box\neg \qquad \Box = \neg\Diamond\neg$$

Les deux modalités sont souvent obtenues à partir de la notion de "monde possible".

On peut considérer un *monde* comme étant constitué d'un ensemble de propriétés vérifiées. Si à partir d'un monde initial (w), un autre monde (w') peut être obtenu en faisant des inférences, ce dernier monde est considéré en relation avec le premier (wRw') et est appelé un *monde possible* du monde initial. En général, à partir d'un monde initial, on peut avoir un ensemble de mondes possibles. Une proposition est *possiblement* vraie pour un monde initial si elle est vérifiée dans au moins un de ses mondes possibles; une proposition est *nécessairement* vraie si elle est vérifiée dans tous ses mondes possibles.

Plus formellement, une logique modale est composée d'un doublet:

1. un "frame" F = (W, R) où W est l'ensemble des mondes, et $R \subset W \times W$ est une relation binaire (qui peut être intuitivement comprise comme une relation de possibilité relative).

2. une application V: $\mathcal{P} \rightarrow 2^W$ (où $\mathcal{P}$ est l'ensemble de propositions atomiques) qui est une fonction d'une proposition atomique ($P \in \mathcal{P}$) vers un sous-ensemble de W. Intuitivement, V est une fonction assignant à chaque proposition atomique P ($P \in \mathcal{P}$), l'ensemble des mondes dans lesquels P est vraie. La fonction V est ensuite étendue sur toute formule bien formée qui est définie de la manière suivante:

$$f ::= P \in \mathcal{P} \mid f_1 \vee f_2 \mid \neg f \mid \Diamond f \mid \text{true}$$

La fonction V est définie comme suit:

$$V(P) = \{w \in W : w \models P\}$$
$$V(f_1 \vee f_2) = V(f_1) \cup V(f_2)$$
$$V(\neg f) = W - V(f)$$
$$V(\Diamond f) = \{w \in W : wRw' \wedge w' \in V(f)\}$$
$$V(true) = W$$

Selon les propriétés définies sur la relation R, on peut obtenir des systèmes modaux différents. Si la relation est transitive (wRw', w'Rw" $\Rightarrow$ wRw"), on obtient le système K4 tel que $V_w(\Diamond\Diamond p)=1 \Rightarrow V_w(\Diamond p)=1$. Le système S4 correspond à une relation R réflexive et transitive.

De façon générale, on peut représenter la vérification d'une formule f dans un monde w par rapport à un système S donné de manière uniforme:

$$w \models_S f$$

Selon les propriétés définies dans un système S donné, on a une valuation particulière de cette vérification.

En comparaison avec la logique du premier ordre, la logique modale est évidemment mieux adaptée pour modéliser les SRI, car on a introduit une notion de plausibilité (possible et nécessaire) qui permet partiellement de modéliser la notion d'incertitude dans les SRI (cf.II.3.1).

Malgré cela, la logique modale est encore insuffisante:

1). La transition d'un monde initial (w) à un monde possible (w') ne peut s'effectuer que si on est *sûr* de la relation wRw'. Comme dans beaucoup d'applications de la logique modale, il est souvent difficile dans les SRI de définir de manière fiable une relation R entre deux ensembles d'informations. Par exemple, la faisabilité de transformer {système de recherche d'informations} en {système d'information} n'est pas toujours certaine.

2). Les modalités introduites sont issues des valuations classiques où le résultat est la valeur "vrai" ou "faux". Autrement dit, pour qu'une proposition soit "possiblement" satisfaite dans un monde, elle doit être *totalement* satisfaite dans un de ses mondes possibles. Or la notion de la satisfaction totale est difficile à établir dans les SRI. Par exemple, on ne peut jamais garantir qu'un terme t décrit parfaitement le contenu d'un document.

En conclusion, la logique modale considérée telle qu'elle, paraît encore trop restreinte pour modéliser complètement les SRI.

Dans le domaine de la logique modale, des études ont aussi porté sur la logique modale floue ([Schotch75]), dans lesquelles sont ajoutées la notion de relation incertaine entre deux mondes et la notion de satisfaction incertaine d'une proposition dans un monde. Ces deux notions d'incertitude nous paraissent très intéressantes pour compléter la modélisation des SRI.

3.3.2. La logique modale floue

Dans la logique modale floue, la relation de possibilité relative R de la logique modale classique est remplacée par une fonction δ (δ: $W \times W \rightarrow [0,1]$) qui mesure la *certitude* de cette relation. En particulier, $\delta(w,w')=1 \Rightarrow wRw'$. De même, une fonction C est définie pour remplacer la notion de satisfaction totale d'une proposition atomique: $Cp(w)$ mesure le *degré* de satisfaction de la proposition atomique $P \in \mathcal{P}$ dans le monde w. Correspondant à ces deux cas non-classiques, un nouvel opérateur que l'on note par μ, est introduit. Il signifie qu'une formule (f) *peut être* possible (dans [Schotch75], μ est noté par M):

$$V(\mu f) = \{w \in W : \delta(w,w') \neq 0 \wedge C_p(w') \neq 0\}$$

Remarquons la différence entre μf et $\Diamond\Diamond f$: la modalité μ est issue d'une relation incertaine et/ou d'une satisfaction incertaine, la modalité $\Diamond$ est issue d'une relation et d'une satisfaction certaines. La relation entre les deux modalités est: $V_w(\Diamond f)=1 \Rightarrow V_w(\mu f)=1$. En particulier, si on accepte la condition suivante sur δ: $\delta(w,w') \neq 0 \wedge \delta(w',w'') \neq 0 \Rightarrow \delta(w,w'') \neq 0$, on obtient alors: $V_w(\Diamond\Diamond \ldots \Diamond f)=1 \Rightarrow V_w(\mu f)$.

A part les formules existantes dans la logique modale classique, on peut considérer dans ce nouveau modèle qu'une formule (f) "peut être vraie" (μf), et sa valuation est fondée sur des contraintes plus réduites que pour $\Diamond f$.

Pour considérer plus formellement les valuations non-classiques, Schotch propose un nouveau modèle, baptisé *modèle Baroque*, qui est défini par le n-uplet suivant:

$$\langle (W, W_1, \ldots, W_k, \delta), V, V_{W_1}, \ldots, V_{W_k}, \mathcal{V} \rangle$$

où W est un ensemble de mondes classiques définis comme précédemment,
W_i est un ensemble de mondes non-classiques, $W_i \cap W_j = \emptyset$ si $i \neq j$.
δ est une fonction mesurant la certitude de la relation entre ces mondes. $\delta: W \cup W_1 \cup \ldots \cup W_k \rightarrow [0,1]$. Cette fonction est associée avec les propriétés suivantes:

$\delta(w,w') = 0$ si $w \in W_i$ et $w' \in W$

$\delta(w,w') < 1$ si $w \in W$ et $w' \in W_i$

et la notion classique wRw' $(w,w' \in W)$ correspond à: $\delta(w,w') = 1$

V est une valuation de la logique modale classique

V_{W_i} est une valuation de Zadeh: $\mathcal{P} \rightarrow [0,1]^{W_i}$, mesurant la certitude $(\in [0,1])$ de la vérification d'une proposition atomique $(\in \mathcal{P})$ dans un monde non-classique $(\in W_i)$

$\mathcal{V}$ est une valuation Baroque d'une formule quelconque, définie comme suit:

$$\mathcal{V}(p) = \langle V(p), V_{W_1}(p), V_{W_2}(p), \ldots, V_{W_k}(p) \rangle$$

Cette valuation est définie par un ensemble de propriétés[note] .

Intuitivement, ce que le modèle Baroque propose, peut être exprimé de la façon suivante:

L'univers du modèle est d'abord défini par un ensemble de mondes classiques (W) possédant une valuation comme dans la logique modale classique, et un certain nombre d'ensembles de mondes non-classiques (W_i), possédant chacun une valuation floue. Une valuation Baroque $(\mathcal{V})$ consiste à assigner à une proposition donnée, l'ensemble des mondes classiques dans lesquels elle est vraie et les *ensembles* de mondes non-classiques dans lesquels la proposition *peut être* vraie. A un monde non-classique de ces ensembles est associé une mesure de degré de vérité.

---

note

Dans le cas d'un seul ensemble (W) de mondes possibles non-classiques, les propriétés sont les suivantes:

$\mathcal{V}(f) = \langle V(f), V_W(f) \rangle$

$V(P) = \{w \in W: C_P(w)=1\}$

$V(\neg f) = W - V(f)$

$V(f_1 \vee f_2) = V(f_1) \cup V(f_2)$

$V(\Diamond f) = \{w \in W: \delta(w,w')=1 \wedge w' \in V(f)\}$

$V(\mu f) = \{w \in W: \delta(w,w') \neq 0 \wedge (w \in W \Rightarrow w' \in V(f)) \wedge (w' \in W \Rightarrow C_f(w') \neq 0)\}$

$V_W(f)$ est une fonction qui assigne à chaque proposition f une fonction caractéristique floue: $C_f(w) \in [0,1]$, mesurant le degré de la satisfaction de p dans le monde non-classique $w \in W$. $V_W(f)$ est défini avec les propriétés suivantes:

$V_W(\neg f) = 1 - V_W(f)$

$V_W(f_1 \vee f_2) = MAX(V_W(f_1), V_W(f_2))$

$V_W(\Diamond f)$ = une fonction caractéristique binaire sur:

$$\{w \in W: \delta(w,w')=1 \wedge C_f(w') \neq 0\}$$

c'est-à-dire que $V_W(\Diamond f)=1$ pour tous les mondes de cet ensemble, et $V_W(\Diamond f)=0$ pour les autres.

$V_W(\mu f)$ = une fonction binaire sur $\{w \in W: \delta(w,w') \neq 0 \wedge C_f(w') \neq 0\}$

Par rapport à la logique modale classique, le modèle Baroque a introduit deux notions d'incertitude très intéressantes: l'incertitude de la relation entre deux mondes et l'incertitude de la satisfaction d'une formule dans un monde. Cette logique présente donc des aspects très intéressants pour modéliser les SRI, car elle permet de modéliser le traitement des informations *imprécises* (correspondant à une relation incertaine) ou *incomplètes* (correspondant à une satisfaction partielle), comme dans ce type de système.

Sa formalisation, par contre, n'est pas tout à fait adéquate pour cette modélisation:

a). La distinction de k ensembles de mondes non-classiques ($W_1 \ldots W_k$) est superflue par rapport à la valuation Baroque définie. La distinction entre deux ensembles de mondes non-classiques n'apporte pas beaucoup d'intérêt. En effet, les valuations (V_{Wi}) définies dans tous les ensembles de mondes non-classiques sont analogues, car elles possèdent les mêmes propriétés (cf.note). Ces ensembles peuvent donc être facilement regroupés en un seul ensemble auquel est associé une valuation fondée sur ces propriétés. L'ancienne distinction entre deux ensembles non-classiques peut être reflétée, dans ce nouvel ensemble, par la relation δ définie entre deux mondes non-classiques.

b). Il existe deux sortes de valuations dans ce modèle Baroque: une valuation floue ($\in [0,1]$) et une autre binaire ($\in \{0,1\}$). Cela introduit d'une part une hétérogénéité, mais on pert également l'intérêt de la notion d'incertitude dans le cas de valuation binaire, notamment pour la valuation de $\Diamond f$ et μf. Si on appliquait ce modèle directement aux SRI, cela signifierait que l'on ne peut pas qualifier un document comme une réponse "possible", "possiblement possible" ... pour une requête. Or, des valeurs plus fines sont souvent nécessaires pour ordonner de façon plus précise les documents en réponse.

c). Dans l'ensemble de mondes classiques (W), une formule est évaluée par un sous-ensemble de mondes classiques. Cette valuation peut être également considérée comme une valuation binaire:

$$V_w(f) = 1, \text{ si } w \in V(f)$$
$$V_w(f) = 0, \text{ si } w \notin V(f)$$

Exprimés comme ceci, les mondes classiques et les mondes non-classiques peuvent être vus d'une même façon, ce qui nous permet d'établir une valuation uniforme.

d). Si la notion d'incertitude est intégrée dans le modèle, non seulement cette notion peut apparaître sur les descriptions dans les documents, mais aussi, elle peut apparaître dans les requêtes, c'est-à-dire que les utilisateurs peuvent demander des documents vérifiant une proposition avec une valeur de certitude donnée. Ainsi, il est aussi nécessaire de créer un opérateur

supplémentaire pour exprimer la valeur de certitude associée à chaque formule. Nous notons cet opérateur par "$=_v$". "$=_v f$" signifie que la formule f doit être vérifiée avec la certitude v ($v \in]0,1[$).

En conclusion, nous proposons de considérer tous les mondes non-classiques dans un seul ensemble ($\mathcal{W} = W \cup (\bigcup_{i=1}^{k} W_i)$) et d'associer à chaque monde (classique ou non) une valuation du degré de satisfaction pour une formule donnée.

Nous proposons donc ci-dessous un modèle de logique modale généralisé pour les besoins des SRI, prenant en compte ces objectifs.

3.3.3. Le modèle de logique modale pour les SRI

L'idée de base est la suivante:

Etant donnée une implication $w \rightarrow f$, la valuation de sa certitude peut être transformée en la valuation de la vérification de la conclusion de l'implication (f) dans un monde de logique modale constitué par la prémisse de l'implication (w), c'est-à-dire: $w \models f$.

En tenant compte des remarques énoncées dans la dernière section, nous proposons de définir un modèle logique de la manière suivante:

- $\mathcal{P}$ est l'ensemble de propositions atomiques. Une formule bien formée (f) est définie comme suit:

$$f ::= P \in \mathcal{P} \mid f_1 \wedge f_2 \mid \neg f \mid \Diamond f \mid =_v f \mid true$$

$\mathcal{F}$ est constitué par toutes les formules bien définies.

- $\mathcal{W}$ est l'univers du modèle, constitué d'un ensemble de mondes.

- $C_P(w)$ est une fonction issue du modèle Baroque, qui mesure la certitude de la vérification d'une proposition atomique ($P \in \mathcal{P}$) dans un monde (w).

- δ: $\mathcal{W} \times \mathcal{W} \rightarrow [0,1]$ est une valuation de la relation entre deux mondes. Elle vérifie les propriétés suivantes:

$$\delta(w,w)=0$$

$$\delta(w,w')>0 \wedge \delta(w',w'')>0 \Rightarrow \delta(w,w'')>0 \quad w,w',w'' \in \mathcal{W}$$

- Δ: $[0,1]^2 \rightarrow [0,1]$ est une fonction qui combine une valeur de certitude d'une relation et une valeur de certitude d'une vérification pour former une valeur de certitude de vérification. Cette fonction vérifie les propriétés suivantes (où c, c1, c2 sont des valeurs de certitude quelconque $\in [0,1]$):

$$\Delta(c1,c2) \leq c1, \qquad \Delta(c1,c2) \leq c2$$
$$c1>0 \wedge c2>0 \Rightarrow \Delta(c1,c2)>0 \qquad \text{(idempotence)}$$

Un cas particulier de Δ peut être, par exemple, la multiplication.

Les deux propriétés suivantes peuvent être déduites des deux premières:

$$\Delta(1,c) = c, \qquad \Delta(c,1) = c$$
$$\Delta(c,0) = 0, \qquad \Delta(0,c) = 0$$

- V: $\mathcal{F} \rightarrow [0,1]^{\mathcal{W}}$ est une fonction assignée à chaque monde ($w \in \mathcal{W}$), qui calcule la certitude de vérification d'une proposition ($f \in \mathcal{F}$) dans ce monde

$V_w(f)$ est défini de la façon suivante:

- $V_w(P) = MAX\ [\ C_P(w),\ V_w(\Diamond P)]\ ,\ P \in \mathcal{P}$
- $V_w(f_1 \wedge f_2) = M(V_w(f_1), V_w(f_2))$
- $V_w(\neg f) = 1 - V_w(f)$
- $V_w(\Diamond f) = \underset{w'}{MAX}(\Delta[\delta(w,w'), V_{w'}(f)])$, avec $w' \in \mathcal{W}$
- $V_w(=_v f) = 1 - |V_w(f) - v|$
- $V_w(true) = 1$

La fonction M est une fonction qui, sous sa forme générale, est définie pour la valuation d'une conjonction. La définition exacte de cette fonction dépend fortement de l'application. Cette fonction sera définie comme MAX et MIN dans deux cas particuliers étudiés ultérieurement (cf.3.3.4).

Implicitement, la valuation d'une disjonction est définie de façon suivante:

$$V_w(f_1 \vee f_2) = V_w(\neg(\neg f_1 \wedge \neg f_2))$$
$$= 1 - M\ [(1-V_w(f_1)),\ (1-V_w(f_2))]$$

Intuitivement, comme dans la logique modale classique, une relation entre deux mondes signifie aussi une relation de possibilité relative (cf.3.3.1), mais la relation unique R de la logique modale classique est élargie en un ensemble de relations non-classiques. En particulier, la valeur de $\delta(w,w')=1$

correspond à une relation certaine entre w et w'. Cette relation est similaire à R dans la logique modale classique. La valeur de $\delta(w,w')=0$ signifie qu'aucune relation n'existe entre w et w'. Les valeurs intermédiaires $\delta(w,w') \in]0,1[$ dénotent une relation incertaine.

La valuation $V_w(=_v f)$ est définie comme le rapprochement de la certitude de la vérification de la formule f dans le monde w à la certitude demandée. Si les deux certitudes sont identiques, $V_w(=_v f)$ est alors valué à 1.

Si l'on avait admi la valeur de v dans l'intervalle [0,1], on aurait deux équations suivantes:

$$V_w(=_0 f) = 1 - V_w(f) = V_w(\neg f)$$
$$V_w(=_1 f) = V_w(f)$$

Ces équations montrent que l'opérateur $=_0$ est équivalent à $\neg$, et $=_1$ est un opérateur implicite. Pour ne pas introduire de la redondance, nous avons donc défini la valeur de v dans l'intervalle]0,1[.

La comparaison de l'opérateur $=_0$ avec l'opérateur implicite et la négation permet aussi d'étudier la propriété de la distributivité de $=_v$ sur les opérateurs booléens.

On peut considérer que la propriété de distributivité sur les opérateurs booléens $\wedge$ et $\vee$ est vérifiée par l'opérateur implicite. Ainsi, on a:

$$=_1(f_1 \wedge f_2) = (=_1 f_1) \wedge (=_1 f_2))$$
$$=_1(f_1 \vee f_2) = (=_1 f_1) \vee (=_1 f_2))$$

Mais cette propriété n'est pas vérifiée par l'autre cas extrême correspondant à la négation. Ainsi, on a:

$$=_0(f_1 \wedge f_2) \neq (=_0 f_1) \wedge (=_0 f_2))$$
$$=_0(f_1 \vee f_2) \neq (=_0 f_1) \vee (=_0 f_2))$$

mais plutôt:

$$=_0(f_1 \wedge f_2) = (=_0 f_1) \vee (=_0 f_2))$$
$$=_0(f_1 \vee f_2) = (=_0 f_1) \wedge (=_0 f_2))$$

Etant donné que l'opérateur $=_v$ est défini entre ces deux cas extrêmes ($v \in]0,1[$), il ne vérifie donc généralement pas cette distributivité:

$$=_v(f_1 \wedge f_2) \neq (=_v f_1) \wedge (=_v f_2)$$

$$=_V(f_1 \vee f_2) \neq (=_V f_1) \vee (=_V f_2)$$

Nous avons associé la signification suivante à l'opérateur $\Diamond$:

- La vérification de $\Diamond f$ (i.e.$V_w(\Diamond f)=1$) signifie que f est vérifié dans au moins un monde possible w' (i.e.$V_{w'}(f)=1$) et w' a une relation certaine avec w (i.e.$\delta(w,w')=1$);

- $V_w(\Diamond f)>0$ signifie que dans au moins un monde possible w' ayant une relation $\delta(w,w')>0$ avec w, on a $V_{w'}(f)>0$.

Par rapport au modèle Baroque de Schotch, le premier cas - $V_w(\Diamond f)=1$ correspond à la vérification de $\Diamond f$ dans un monde "classique", et le second cas - $V_w(\Diamond f)>0$ correspond à $V_w(\mu f)=1$. Il n'est donc plus nécessaire de définir l'opérateur μ dans notre modèle.

Avec la valuation d'une proposition atomique $V_w(P)$, nous avons associé la propriété suivante à l'opérateur $\Diamond$:

$$V_w(P) \geq V_w(\Diamond P)$$

Cette propriété implique la propriété de transitivité pour l'opérateur $\Diamond$:

$$V_w(\Diamond P) \geq V_w(\Diamond\Diamond P)$$

Cette propriété est comparable avec la transitivité de $\Diamond$ définie dans le modèle K4 de la logique modale.

Avec la valuation de $\Diamond f$ et de P, nous pouvons développer la formule de valuation de $V_{w_0}(P)$ (dans un monde w_0) de la façon suivante:

$$V_{w_0}(P) = MAX\{C_P(w_0), V_w(\Diamond P)\}$$
$$= MAX\{C_P(w_0), \underset{w_1}{MAX}(\Delta[\delta(w_0,w_1), V_{w_1}(P)])\} \quad (\text{avec } w_1 \in \mathcal{W})$$

En utilisant $V_{w_1}(P) = MAX(C_P(w_1), V_{w_1}(\Diamond P))$
et $V_{w_1}(\Diamond P) = \underset{w_2}{MAX}(\Delta[\delta(w_1,w_2), V_{w_2}(P)])$,
on a:

$$V_{w_0}(P) = MAX\{C_P(w_0), \underset{w_1}{MAX}(\Delta[\delta(w_0,w_1), \underset{w_1}{MAX}(C_P(w_1), V_{w_1}(\Diamond P))])\}$$

$$= MAX\{\ C_P(w_0),\ \underset{w_1}{MAX}(\Delta[\delta(w_0,w_1),\ \underset{w_1}{MAX}(C_P(w_1),\underset{w_2}{MAX}(\Delta[\delta(w_1,w_2),V_{w_2}(P)])\ ))\}$$

$$= MAX\{\ C_P(w_0),\ \underset{w_1}{MAX}(\Delta[\delta(w_0,w_1),C_P(w_1)]),\ \underset{(w_1,w_2)}{MAX}(\Delta[\delta(w_0,w_1),\Delta[\delta(w_1,w_2),V_{w_2}(P)]\ ])\ \}$$

$$= MAX\{\ C_P(w_0),\ \underset{w_1}{MAX}(\Delta[\delta(w_0,w_1),C_P(w_1)]),\ \underset{(w_1,w_2)}{MAX}(\Delta[\delta(w_0,w_1),\Delta[\delta(w_1,w_2),C_P(w_2)]\ ])\ \underset{(w_1,w_2,w_3)}{MAX}(\Delta[\delta(w_0,w_1),\Delta[\delta(w_1,w_2),\Delta[\delta(w_2,w_3),V_{w_2}(P)]]\ ]\ )\ \}$$

$$= \ldots$$

Le développement sur le dernier élément de cette formule $V_{w_i}(P)$ peut continuer jusqu'à ce qu'il n'existe plus de monde possible pour w_i. Supposons que l'on a au maximum n transformations possibles. Cette formule peut alors transcrite en:

$$V_{w_0}(P) = MAX\{\ C_P(w_0),\ \underset{w_1}{MAX}(\Delta[\delta(w_0,w_1),C_P(w_1)]),\ \underset{(w_1,w_2)}{MAX}(\Delta[\delta(w_0,w_1),\Delta[\delta(w_1,w_2),C_P(w_2)]\ ])\ \ldots\ \underset{(w_1,\ldots,w_n)}{MAX}(\Delta[\delta(w_0,w_1),\Delta[\ldots,\Delta[\delta(w_{n-1},w_n),C_P(w_n)]\ldots]])\}$$

Les éléments dans l'opération MAX représentent respectivement:

- la valuation directe de $V_{w_0}(P)$
- la valuation obtenue avec une transformation,
- la valuation obtenue avec deux transformations,
- ...
- la valuation obtenue avec n transformations.

Pour simplifier cette formule, nous définissons la fonction suivante:

$$V^{i}_{w_0}(P) = \underset{(w_1,\ldots,w_i)}{MAX}\Delta[\delta(w_0,w_1),\Delta[\ldots,\Delta[\delta(w_{i-1},w_i),C_P(w_i)]\ldots]]$$

en particulier, $V^{0}_{w_0}(P) = C_P(w_0)$

La valuation de $V_{w_0}(P)$ peut être alors exprimée par:

$$V_{w_0}(P) = \underset{i=0}{\overset{n}{MAX}}\{V^{i}_{w_0}(P)\} \qquad (3)$$

C'est-à-dire que la valuation de $V_{w_0}(P)$ est la valeur maximale des valuations obtenues avec toutes les transformations possibles.

Etudions maintenant la valuation de $V^{i}_{w_0}(P)$. Cette fonction est valuée via l'opérateur Δ. Si l'on peut définir un autre opérateur $\otimes$ qui vérifie l'équation suivante:

$$\Delta(\delta_1, \Delta(\delta_2, c)) = \Delta(\otimes(\delta_1,\delta_2), c)$$

alors la valuation de $V^{i}_{w_0}(P)$ peut être exprimée par:

$$\underset{(w_1,\cdots,w_i)}{MAX}\Delta\{\otimes[\otimes[\ldots\otimes[\delta(w_0,w_1),\delta(w_1,w_2)],\ldots],\delta(w_{i-1},w_i)], C_P(w_i)\}$$

On abrège la première opérande de Δ en $\delta^i(w_0,w_i)$. Cette formule devient:

$$V^{i}_{w_0}(P) = \underset{(w_1,\cdots,w_i)}{MAX}\Delta[\delta^i(w_0,w_i), C_P(w_i)]$$

ce qui signifie que la valuation de $V_{w_0}(P)$ obtenue avec i transformations (i.e.$V^{i}_{w_0}(P)$) est déterminée par la relation entre w_0 et les w_i obtenus après i transformations (i.e.$\delta^i(w_0,w_i)$) et par la validité de P dans w_i.

La valuation de $V_{w_0}(P)$ correspondante est donc la suivante:

$$V_{w_0}(P) = \underset{i=0}{\overset{n}{MAX}}\{\underset{(w_1,\cdots,w_i)}{MAX}\Delta[\delta^i(w_0,w_i), C_P(w_i)]\} \qquad (4)$$

A partir de cette nouvelle formule, on peut donner une nouvelle expression à la valuation de $V_{w_0}(P)$ ($P \in \mathcal{P}$) de la façon suivante:

Durant la valuation, P doit être non seulement valué par rapport à w_0, mais il doit aussi être valué par rapport aux mondes possibles de w_0, ceux-ci doivent aussi être transformés en leurs mondes possibles, ... Cette génération

des mondes possibles continue tant qu'il existe un monde w_i déjà généré et un monde w_{i+1} non généré tel que $\delta(w_i,w_{i+1})>0$.

Pour chaque monde donné, il peut exister plusieurs transformations possibles. En traçant toutes les transformations possibles, on obtient une arborescence dont les noeuds représentent un monde et les liens correspondent à une transformation apportant une relation >0.

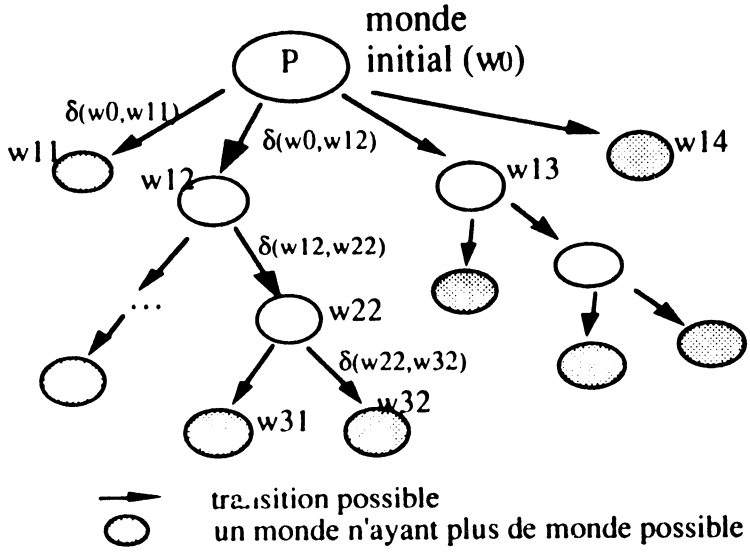


Fig.I.4. *L'arborescence de transformation pour la valuation d'une proposition atomique*

On appelle les transformations du monde initial à un monde possible un *chemin de transformation*. Dans la figure ci-dessus, $w_0 \rightarrow w_{12} \rightarrow w_{22}$ est un chemin de transformation. Un tel chemin peut être caractérisé par son dernier monde, appelé le *monde final*.

Relativement à un chemin de transition $w_0 \rightarrow w_1 \rightarrow \ldots \rightarrow w_i$ (caractérisé par w_i), on peut calculer une valeur de certitude pour $V_{w_0}(P)$ par:

$$V_{w_0}^{w_i}(P) = \Delta\{\delta^i(w_0,w_i), C_P(w_i)\} \qquad (5)$$

et la valeur finale assignée à $V_{w_0}(P)$ est la valeur maximale de toutes les valuations ci-dessus:

$$V_{w_0}(P) = \underset{w_i \in \text{arborescence}}{\text{MAX}} V_{w_0}^{w_i}(P) \qquad (5')$$

Comme l'opérateur $\otimes$ est défini à partir de Δ, pour que Δ vérifie les propriétés définies, il faut que $\otimes$ vérifie les mêmes propriétés:

$$\otimes(c1,c2) \leq c1, \quad \otimes(c1,c2) \leq c2$$
$$c1>0 \wedge c2>0 \Rightarrow \otimes(c1,c2)>0$$

Dans le principe de Van Rijsbergen, la notion de certitude de satisfaction d'une proposition dans un monde et celle de la certitude de relation entre deux mondes sont unifiées en la notion de certitude de la relation. La valuation de la satisfaction $C_P(w)$ est devenue binaire. En effet, dans le cas où une proposition peut être évaluée à une valeur incertaine dans un monde, Van Rijsbergen considère qu'il faut continuer la transformation du monde pour l'étendre en un monde final dans lequel la proposition est *totalement* satisfaite. La certitude de la satisfaction dans le monde initial est fonction des transformations effectuées pour obtenir ce monde final.

Cette vision est plus restreinte que le modèle que nous proposons, car il suffit de continuer à transformer le monde final d'un chemin de transformation w_i de notre modèle à un monde possible w_{i+1} (le nouveau monde final), dans la cas où $C_P(w_i)<1$, pour que $C_P(w_{i+1})=1$. Pour obtenir la valuation de $V_{w0}^{w_i}(P)$ dans la vision de Van Rijsbergen, il suffit de définir la relation entre w_i et w_{i+1} de la manière suivante:

$$\delta(w_i, w_{i+1}) = C_P(w_i) \text{ de notre cas.}$$

3.3.4. Expression du modèle initial en fonction du modèle de logique modale

Dans la section II.3.2, nous avons proposé comme extension du modèle logique de Van Rijsbergen l'expression suivante de la fonction de correspondance:

$$\mathbf{R}(D,Q) = F[P_K(D \rightarrow Q), P'_K(Q \rightarrow D)].$$

Nous montrons ici comment le modèle de logique modale recouvre cette spécification.

Comme dans la présentation de la logique modale classique, on peut aussi représenter de la façon suivante la vérification d'une formule (f) par rapport à un monde initial (w) dans un modèle particulier S:

$$w \models_S f$$

Si l'on définit une fonction P pour mesurer la certitude de la vérification de l'expression ci-dessus, on peut exprimer de la façon suivante la

certitude de satisfaction de la formule f dans le monde w par rapport au système S:

$$P(w \models_S p)$$

En effet, $P_K(D \rightarrow Q)$ peut être représenté dans la forme $P(w \models_S f)$ de trois façons différentes:

1. $P(D \models_{\mathcal{K}} Q)$, avec w=D, p=Q et S=$\mathcal{K}$
2. $P(Q \models_{\mathcal{K}} D)$, avec w=Q, p=D et S=$\mathcal{K}$
3. $P(K \models_{\mathcal{T}} (D \models_{\mathcal{K}} Q))$, avec w=K , p=$(D \models_{\mathcal{K}} Q)$, et S=$\mathcal{T}$

Dans la première représentation, on value la vérification d'une proposition Q (requête) dans un monde D (document) par rapport au modèle $\mathcal{K}$ défini comme précédemment. C'est l'approche fondée sur la modification de la prémisse de l'implication initiale.

Dans la seconde représentation, on value la vérification d'une proposition D (document) dans un monde Q (requête) par rapport au modèle $\mathcal{K}$. C'est l'approche fondée sur la modification de la conclusion de l'implication initiale.

Dans la dernière représentation, on considère $D \models_{\mathcal{K}} Q$ comme une *proposition* à valuer par rapport à un ensemble de connaissances K qui constitue un *monde*. Le modèle $\mathcal{K}$ dans $D \models_{\mathcal{K}} Q$ correspond à l'ensemble de connaissances K. Par l'expression $P(K \models_{\mathcal{T}} (D \models_{\mathcal{K}} Q))$, il s'agit, en réalité, de modifier le modèle $\mathcal{K}$ par rapport auquel $D \models_{\mathcal{K}} Q$ est valué. Le modèle $\mathcal{T}$ est donc un méta-modèle, qui gère la modification d'un modèle ordinaire ($\mathcal{K}$). C'est donc une approche fondée sur la modification du modèle.

C'est cette valuation que Van Rijsbergen suggère dans le principe d'incertitude. Au niveau théorique, c'est une méthode aussi valable que les autres. En pratique, la définition d'un méta-modèle est un problème extrêmement compliqué. Pour la raison de simplicité, nous ne nous intéresserons, par la suite, qu'aux deux premières valuations.

Pour chaque valuation, il est supposé que l'ensemble des connaissances du système (ou le modèle de valuation) K est déjà déterminé. Ainsi, dans la suite d'étude, nous ne précisons plus le paramètre K dans les expressions, et nous abrégeons $P_K(Q \rightarrow D)$ en $P(Q \rightarrow D)$, ...

Les études vont porter uniquement sur la valuation de P(D→Q), dont la prémisse est le document et la conclusion et la requête. L'implication P'(Q→D) pourra être évaluée d'une façon duale, car la description d'une requête peut être généralement considérée comme une description particulière de document.

3.3.4.1. Valuation par modification de la prémisse de l'implication

Dans cette approche de valuation de P(D→Q), la prémisse (document D) est considérée comme un monde appartenant à l'univers du modèle ($D=w_0 \in \mathcal{W}$). La conclusion de l'implication (requête Q) est considérée comme une formule bien formée ($Q=f \in \mathcal{F}$). P(D→Q) est donc valuée par:

$$P(D \rightarrow Q) = V_{w_0}(f), \text{ avec } w_0=D \text{ et } f=Q.$$

La fonction $V_{w_0}(f)$ est définie dans le modèle suivant:

- $\mathcal{P}$ est l'ensemble des formules atomiques. Une formule bien formée (f) est définie comme suit:

$$f ::= P \in \mathcal{P} \mid f_1 \wedge f_2 \mid \neg f \mid \Diamond f \mid =_v f \mid \text{true}$$

$\mathcal{F}$ est constitué par toutes les formules bien définies. Une requête Q est définie comme une formule bien formée $f \in \mathcal{F}$.

- $\mathcal{W}$ est l'univers du modèle, constitué de l'ensemble des documents initiaux et leurs formes transformées. Un document D est donc considéré comme un monde initial $w_0 \in \mathcal{W}$.

- $C_P(w)$ est une fonction identique à celle du le modèle Baroque: elle mesure la certitude de la vérification d'une formule atomique ($P \in \mathcal{P}$) dans un monde (w).

- δ: $\mathcal{W} \times \mathcal{W} \rightarrow [0,1]$ est la valuation de la relation entre deux mondes, définie en 3.3.3.

- Δ: $[0,1]^2 \rightarrow [0,1]$ est la fonction définie en 3.3.3..

- V: $\mathcal{F} \rightarrow [0,1]^{\mathcal{W}}$ est une fonction mesurant la vérification d'une formule dans un monde. Elle est définie de la façon suivante:

- $V_w(P) = MAX[\ C_P(w),\ V_w(\Diamond P)]\ ,\ P \in \mathcal{P}$
- $V_w(f_1 \wedge f_2) = MIN(V_w(f_1), V_w(f_2))$
- $V_w(\neg f) = 1 - V_w(f)$
- $V_w(\Diamond f) = \underset{w'}{MAX}(\Delta[\delta(w,w'), V_{w'}(f)])$, avec $w' \in \mathcal{W}$
- $V_w(=_v f) = 1 - |V_w(f) - v|$
- $V_w(true) = 1$

Dans la définition de la valuation d'une conjonction, nous avons instancié l'opérateur général M défini précédemment par l'opérateur MIN. En effet, la valuation définie sur $f_1 \wedge f_2$ et sur $\neg f$ est la valuation de la logique floue de Zadeh ([Zadeh68]).

Ce modèle suggère la valuation suivante de $P(D \rightarrow Q)$:

Un document D est considéré comme un monde initial (noté $D=w_0$), qui est constitué d'un ensemble de descriptions. Dans un SRI, il existe un ensemble de connaissances (éventuellement vide) qui permettent d'obtenir des nouvelles descriptions à partir des descriptions existantes.

Correspondant à cet ensemble de connaissances, $\mathcal{W}$ est construit de la façon suivante:

1. Tout document existant dans le système constitue un monde dans $\mathcal{W}$.
2. Etant donné un monde quelconque $w \in \mathcal{W}$, si les connaissances du système permettent de dériver une description dans w vers une nouvelle description, alors le remplacement de l'ancienne description par la nouvelle description constitue un nouveau monde w'. Ce nouveau monde fait partie aussi de $\mathcal{W}$.

En effet, $\mathcal{W}$ est constitué par la fermeture de l'ensemble des documents, en utilisant les connaissances du système.

Les documents (les mondes initiaux) existants dans le système sont supposés consistants, ainsi que les connaissances du système. En appliquant ces connaissances, on ne peut obtenir que des mondes consistants à partir des mondes initiaux.

La définition de la relation entre deux monde dépend aussi des connaissances. Si à partir d'un monde, on peut dériver un autre monde en utilisant les connaissances, alors il existe une relation entre les deux mondes. La certitude de cette relation est déterminée par l'importance de la

modification apportée par la dérivation. Plus la modification est grande, moins la relation est certaine.

Selon la valuation définie dans ce modèle, la valuation d'une proposition atomique (P) dans un monde initial (w_0) nécessite le rattachement progressif des mondes possibles au monde initial w_0, comme nous l'avons montré dans la description du modèle proposé. On peut obtenir ainsi une arborescence dont la racine correspond au monde initial, un noeud à un monde possible et un lien à une dérivation possible. Un lien entre w_i et w_{i+1} dans l'arborescence implique une relation entre les deux mondes: $\delta(w_i, w_{i+1})>0$.

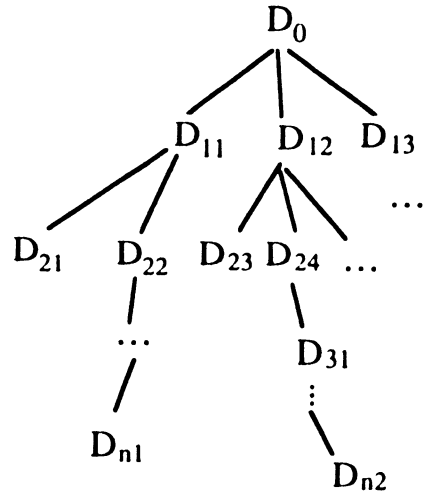


Fig.I.5. *Transformations du document*

En utilisant les formules (5) et (5'), la valuation d'une formule atomique (f=P) dans le monde initial (w_0=D) devient:

$$V_{w_0}(P) = \underset{w_i \in \text{arborescence}}{MAX} V_{w_0}^{w_i}(P)$$

$$= \underset{w_i \in \text{arborescence}}{MAX} \Delta[\delta^i(w_0,w_i), C_P(w_i)]$$

Cette approche de la valuation peut être également exprimée à travers un principe d'incertitude:

Principe d'incertitude 2:

> Etant donnés deux ensembles d'informations x et y, la mesure d'incertitude de x→y relative à un certain ensemble K de connaissances donné est déterminée par l'extension nécessaire de x en x' et la certitude de la vérification de x'→y.

La valuation définie sur les opérateurs $\vee$ et $\neg$ (et $\wedge$) est la valuation floue de Zadeh ([Zadeh65]). La propriété d'idempotence est vérifiée:

$$V_w(f \vee f) = V_w(f), \qquad V_w(f \wedge f) = V_w(f)$$

Par contre, la propriété de complétude n'est pas toujours vérifiée:

$$V_w(f \vee \neg f) \neq 1 \text{ et } V_w(f \wedge \neg f) \neq 0, \text{ si } V_w(f) \in]0,1[$$

Cette propriété est vérifiée quand $V_w(f)$ prend les valeurs extrêmes de $\{0,1\}$. Dans ce cas, $V_w(f \vee \neg f) = 1$ et $V_w(f \wedge \neg f) = 0$.

Cela signifie que $f \vee \neg f$ n'est pas toujours équivalent à la formule "true", et $f \wedge \neg f$ n'est pas toujours équivalent à la formule "false".

3.3.6.2. Valuation par modification de la conclusion de l'implication

Contrairement à la valuation précédente, la valuation par modification de $P(D \rightarrow Q)$ considère la conclusion de l'implication (requête Q) comme un monde, et la prémisse de l'implication (document D) comme une formule à valuer par rapport à ce monde: $P(D \rightarrow Q) = P(Q \models D)$.

Pour cette approche, on définit le modèle suivant, légèrement modifié par rapport au précédent:

- $\mathcal{P}$ est l'ensemble des formules atomiques. Une formule bien formée (f) est définie comme une assertion de la section précédente:

$$f ::= P \in \mathcal{P} \mid f_1 \wedge f_2 \mid \neg f \mid \Diamond f \mid =_v f \mid \text{true}$$

$\mathcal{F}$ est constitué par toutes les formules bien définies et $D = f \in \mathcal{F}$

- $\mathcal{W}$ est l'univers du modèle. Il est construit de la façon suivante:

1. Une requête Q donnée par l'utilisateur constitue un monde initial: $Q = w_0 \in \mathcal{W}$.

2. Si à partir d'un monde dans $\mathcal{W}$, on arrive à dériver un nouveau monde w', en utilisant les connaissances du systèmes, alors $w' \in \mathcal{W}$.

- $C_P(w)$ est une fonction qui mesure la certitude de la validité d'une formule atomique ($P \in \mathcal{P}$) dans un monde (w).

- δ: $\mathcal{W} \times \mathcal{W} \rightarrow [0,1]$ est la valuation de la relation entre deux mondes, définie en 3.3.3.

- Δ: $[0,1]^2 \rightarrow [0,1]$ est la fonction définie en 3.3.3.

- V: $\mathcal{F} \rightarrow [0,1]^{\mathcal{W}}$ est une fonction qui mesure la certitude de la vérification d'une formule dans un monde. Elle est définie de la façon suivante:

  - $V_w(P) = MAX[\ C_P(w),\ V_w(\Diamond P)]\ ,\ P \in \mathcal{P}$
  - $V_w(f_1 \wedge f_2) = MAX(V_w(f_1), V_w(f_2))$
  - $V_w(\neg f) = 1 - V_w(f)$
  - $V_w(\Diamond f) = \underset{w'}{MAX}(\Delta[\delta(w,w'), V_{w'}(f)])$, avec $w' \in \mathcal{W}$
  - $V_w(=_v f) = 1 - |V_w(f) - v|$
  - $V_w(true) = 1$

Par rapport au modèle précédent, la valuation de $V_w(f_1 \wedge f_2)$ est définie différemment. La définition ci-dessus signifie que, si un document est constitué de deux assertions en conjonction, il suffit que l'une des deux assertions puisse satisfaire la requête pour que celle-ci soit satisfaite par le document. Dans la section précédente, $V_w(f_1 \wedge f_2)$ est défini par $MIN(V_w(f_1), V_w(f_2))$, ce qui signifie qu'une requête constituée de deux formules en conjonction peut être satisfaite si les deux formules sont vérifiées. Ces deux interprétations sont tout à fait similaires et les approches ne sont donc pas contradictoires.

La formule "true" correspond ici à un document pouvant satisfaire toute requête. C'est donc un document maximal. Dans la section précédente, la formule "true" correspond à une requête vide.

La dernière différence à noter porte sur la définition de $C_P(w)$:

Notons D_w l'ensemble des descriptions vérifiées dans un monde w. Etant donnée une proposition atomique P, la définition de $C_P(w)=1$ de la précédente section est fondée sur la valuation de l'implication classique $D_w \Rightarrow P$, c'est-à-dire que P doit être une conséquence logique des descriptions de w.

Or, dans cette section, la valuation $C_P(w)=1$ est fondée sur la valuation de l'implication logique $P \Rightarrow D_w$, c'est-à-dire que $C_P(w) = 1$ si les descriptions de w peuvent être déduites à partir de la proposition P.

Schématiquement, cette seconde approche suggère la méthode de valuation suivante pour une requête constituée d'une proposition atomique:

Si le monde initial constitué de la requête initiale ne peut pas assigner $C_P(w_0)$ à 1, le monde w_0 (la requête) doit être transformé en d'autres formes qui constituent les mondes possibles w_i, en utilisant les connaissances du système. Ces transformations doivent continuer tant que cela est possible. Comme précédemment, ces transformations constituent une arborescence de transformation dont la racine correspond au monde initial (la requête initiale), les noeuds aux mondes possibles (requêtes transformées) et les liens aux transformations. Une valeur, mesurant la certitude de la relation, est associée à chaque lien.

Pour la racine et chaque noeud (w_i) de l'arborescence, on peut obtenir une valuation de la validité de la proposition atomique: $C_P(w_i)$. En combinant la relation entre la racine et w_i avec $C_P(w_i)$, on obtient une valuation possible de $V_{w_0}^{w_i}(P)$. Parmi toutes ces valuations, la valeur maximale est assignée à $V_{w_0}(P)$:

$$V_{w_0}(P) = \underset{w_i \in \text{arborescence}}{\text{MAX}} V_{w_0}^{w_i}(P)$$

$$= \underset{w_i \in \text{arborescence}}{\text{MAX}} \Delta[\delta^i(w_0,w_i), C_P(w_i)]$$

dans laquelle nous avons défini en particulier: $\delta^0(w_0,w_0) = 1$.

Pour comparer cette approche avec celle de la section précédente, on applique les deux approches dans l'exemple suivant. Cette comparaison montre aussi que les deux approches sont équivalentes.

Soient Q={base de données déductive} et D={base de données}. La valuation de $P_K(D \rightarrow Q)$ s'effectue de la manière suivante:

- Dans le cas de valuation par modification du document, si la vérification de la formule f="base de données déductive" n'est pas directement validée dans D={base de données}, il faut *modifier* le document de "base de données" en "base de données déductive" pour obtenir un monde possible w_1={base de données déductive} dans lequel f est validée. Cette modification consiste à *rajouter* des descriptions dans le document pour qu'il devient plus spécifique. La certitude de $V_{w_0}(f)$ est déterminée par rapport à la relation

entre w_0 et w_1 et la vérification de la proposition dans w_1. Plus on considère que la modification est importante, moins la relation est certaine.

- Dans le cas de valuation par modification de la requête, si la formule f="base de données" n'est pas directement vérifiée dans w_0={base de données déductive}, il faut *modifier* w_0 de "base de données déductive" en "base de données". Cette modification consiste à *restreindre* la requête en une requête plus générale.

On remarque que ces deux approches utilisent la relation entre "base de données" et "base de données déductive" dans des sens inverses. Dans le premier cas, on utilise la relation "base de données"→"base de données déductive", dans la seconde approche, on utilise "base de données déductive"→"base de données", ceci devant aboutir au même résultat. Il existe donc une différence au niveau de l'utilisation des connaissances dans les deux approches. Cela sera discuté en 3.4.

Correspondant à la méthode de valuation par modification de la conclusion, on peut également établir un autre principe d'incertitude.

Principe d'incertitude 3:

> Etant donnés deux ensembles d'informations x et y, la mesure d'incertitude de x→y relative à un certain ensemble K de connaissances donné est déterminée par la restriction nécessaire de y en y' et la certitude de la vérification de x→y'.

Comme dans la section précédente, la propriété d'idempotence est vérifiée par la valuation définie:

$$V_w(f \wedge f) = V_w(f), \qquad V_w(f \vee f) = V_w(f)$$

Mais la propriété de complétude n'est pas toujours vérifiée:

$$V_w(f \wedge \neg f) \neq 0 \text{ et } V_w(f \vee \neg f) \neq 1, \text{ si } V_w(f) \in]0,1[,$$
$$V_w(f \wedge \neg f) = 0 \text{ et } V_w(f \vee \neg f) = 1, \text{ si } V_w(f) \in \{0,1\}.$$

3.3.6.3. Conclusion

Jusqu'ici, on a développé deux approches pour la valuation de P(D→Q). Etant donné qu'une requête peut être considérée comme une description (spécification) particulière de document, la valuation de P'(Q→D) est strictement similaire à celle de P(D→Q). Cela implique que la démarche suivie pour la valuation de P(D→Q) est également applicable pour P'(Q→D).

On peut donc proposer (sans détailler) les deux approches suivantes pour la valuation de P'(Q→D):

1. Valuation par modification de la requête (prémisse).

La requête Q est considérée comme un monde initial (w_0=Q). Le document D est une formule à valuer. Durant la valuation, une arborescence de transformation est établie à partir de w_0. La certitude est déterminée par la valeur maximale des combinaisons entre les transformations effectuées et les vérifications de D dans les mondes obtenus.

Cette valuation est fondée sur la modification de la prémisse de l'implication Q→D, la définition du modèle est donc similaire à celle de la section 3.3.4.1.

2. Valuation par modification du document (conclusion).

La requête Q est considérée comme une formule à valuer et le document D comme un monde initial (w_0=D). Durant la valuation, une arborescence de transformation est établie à partir de w_0. La certitude est déterminée par la valeur maximale des combinaisons entre les transformations effectuées et les vérifications de Q dans les mondes obtenus.

Cette valuation est fondée sur la modification de la conclusion de l'implication Q→D, la définition du modèle est donc similaire à celle de la section 3.3.4.2.

3.4. Mise en œuvre du modèle

On a présenté trois approches possibles pour évaluer la certitude des implications. La question ne s'était pas encore posée quant à la nature de la transformation dérivant un monde vers un monde possible. Cette transformation est dépendante d'un ensemble de connaissances sémantiques du système. Par exemple, selon qu'une relation de synonymie entre deux termes existe dans les connaissances du système ou non, la transformation changeant un terme en l'autre peut ne pas apporter (quand la relation existe) ou au contraire apporter une dégradation de la certitude de l'implication.

Dans la plupart des modèles, aucune relation n'est prise en compte entre les termes (nous les appelons "modèles indépendants"). L'évaluation des requêtes est un processus fondé uniquement sur la présence de termes. La recherche d'informations développée à partir d'un tel modèle simpliste, manque souvent de finesse dans un domaine où la sémantique joue un rôle capital. Il est prévisible que le traitement de la sémantique sera à la base de

tous les futurs SRI; c'est ce que montrent les tendances dans les recherches actuelles ([Croft85, Rijsbergen86]).

Pour concrétiser les valuations proposées précédemment, nous allons examiner, à un niveau très général, quelques aspects concernant les transformations sémantiques pendant l'évaluation des requêtes. Il est à remarquer que nous ne donnons pas une présentation exhaustive, mais seulement des cas particuliers, jugés très représentatifs dans les SRI.

3.4.1. Les connaissances gérées par les SRI

Dans les SRI, les connaissances les plus souvent traitées sont les relations entre termes. La notion de *terme* est définie de la façon suivante:

> Un mot jugé significatif par un SRI est un terme; la connexion de termes par un opérateur sémantique forme un autre terme plus complexe.

Par exemple: "informatique" est un terme, "informatique" et "gestion" connectés par l'opérateur "utilisé-dans" forment un terme plus spécifique: "informatique utilisé-dans gestion".

Ce que nous appelons *connaissances d'un SRI* (noté K précédemment) correspond donc à un ensemble de relations sémantiques entre les termes d'un SRI. Les relations existant entre les termes ne sont souvent pas des implications logiques, mais des *relations de dérivation* qui sont reconnues dans un système particulier. Les implications logiques peuvent être considérées comme des cas extrêmes des relations de dérivation. L'utilisation de ces connaissances sur un monde conduit le monde vers son monde possible. C'est ce second type de connaissances que l'on va étudier en particulier.

Dans les systèmes existants, on trouve les connaissances bien formalisées, représentées sous forme des règles d'inférence, des règles de production, ... ([Croft88, Davis77, Duda77, Laurière81]). Un exemple typique concerne les règles de diagnostic dans des systèmes médicaux qui, à partir d'un ensemble de signes, permettent de déduire un constat, avec éventuellement une valeur de certitude. Les connaissances ainsi formalisées peuvent être utilisées directement dans la valuation.

Dans beaucoup d'autres cas, les connaissances sont moins formalisées. La représentation la plus souvent utilisée est le thésaurus. Un thésaurus contient, le plus souvent, les relations suivantes: les relations d'équivalence, les relations spécifiques et les relations génériques.

Les relations d'équivalence sont des relations entre des termes de même *niveau*, comme la relation de synonymie:

synonymie: accroissement ↔ augmentation,
banque de données ↔ base de données
...

Les relations spécifiques sont des relations entre un terme et un terme plus spécifique. Ces relations peuvent être une relation d'instanciation, une relation de décomposition, etc...:

relation d'instanciation: ordinateur → micro-ordinateur
document → livre
système d'exploitation → système Unix
...

relation de décomposition: ordinateur → processeur
texte → paragraphe
...

Les relations génériques sont les relations inverses des relations spécifiques. Elles peuvent être donc des relations de généralisation, des relations de composition, etc.:

relation de généralisation: micro-ordinateur → ordinateur
...
relation de composition: processeur → ordinateur
...

Il est évident que plus les relations sont finement identifiées, plus les raisonnements seront subtils. On ne peut malheureusement pas les préciser plus dans un cadre général.

Pour ces relations, le problème clé est de les modéliser en les relations de dérivation que l'on peut directement appliquer dans les transformation d'un monde. Or, cela est souvent difficile. Par exemple, la relation de composition "processeur" - "ordinateur" ne peut souvent pas être modélisée comme une relation de dérivation parfaite, car la dérivation de "processeur" vers "ordinateur" représente une modification assez importante. Mais cette relation n'est pas non plus impossible, car, par exemple, un document sur "processeur" peut aussi répondre à une requête sur "ordinateur" dans certains cas. Ainsi, la certitude de l'implication entre les deux termes doit être estimée dans l'intervalle]0,1[.

La modélisation des relations existant entre termes en des relations de dérivation dépend non seulement de l'application, mais aussi du modèle utilisé. Par la suite, nous analysons cette modélisation par rapport aux modèles définis: le modèle fondé sur la modification du document et le modèle fondé sur la modification de la requête. Nous ne considérons, dans un contexte général, que quelques relations sémantiques typiques (la synonymie, la relation générique et la relation spécifique).

Par la suite, nous adoptons la notation suivante pour une relation de dérivation:

$$a \Rightarrow_v b$$

ce qui signifie que a peut dériver b avec une certitude de valeur v. La valeur de certitude est définie dans l'intervalle]0,1] et en particulier, $a \Rightarrow b$ signifie $a \Rightarrow_1 b$.

3.4.2. Utilisation des connaissances

3.4.2.1. Utilisation dans la modification du document

Soit un document sur "base de données". Il peut répondre parfaitement à une requête demandant "banque de données", s'il l'on considère qu'il existe une relation de synonymie entre les deux termes. Ainsi, une relation de synonymie est une relation de dérivation certaine.

$$\text{synonymie}(a,b) \rightarrow (a \Rightarrow b) \text{ et } (b \Rightarrow a)$$

Soit un document sur "base de données déductive". Il est clair que ce document peut répondre aux requêtes sur "base de données", "informatique", ... Cela signifie qu'il est possible, dans ce cas, de dériver "base de données déductive" vers "base de données" et ensuite vers "informatique", avec une grande certitude. En considérant la relation générique entre "base de données déductive", "base de données" et "informatique", on peut conclure, dans ce cas, qu'une relation générique est une relation de dérivation de grande certitude.

$$\text{générique}(a,b) \rightarrow (a \Rightarrow_v b) \qquad \text{avec } v \sim 1$$

Soit un document sur "base de données". Il peut difficilement répondre à une requête sur "base de données déductive". On peut donc dire que la relation entre ces deux termes, qui est une relation spécifique, peut être modélisée en une relation de dérivation avec faible certitude.

$$\text{spécifique}(a,b) \rightarrow (a \Rightarrow_v b) \qquad \text{avec } v \sim 0$$

3.4.2.2. Utilisation dans la modification de la requête

Soit une requête sur "base de données". Elle peut être parfaitement satisfaite par un document sur "banque de données", si l'on considère qu'il y a une relation de synonymie entre ces termes. Cela signifie que le terme "base de données" peut être dérivé vers "banque de données". La relation de synonymie est également une relation de dérivation dans ce cas.

$$\text{synonymie}(a,b) \rightarrow (a \Rightarrow b) \text{ et } (b \Rightarrow a)$$

Soit une requête sur "informatique". Cette requête peut être satisfaite par des documents sur "base de données", "base de données déductive", ... c'est-à-dire que le terme "informatique" peut être, dans ce cas, dérivé vers "base de données" et ensuite vers "base de données déductive", ... Entre ces termes, il existe une relation spécifique. Cela suggère donc que la relation spécifique est une relation de dérivation avec grande certitude.

$$\text{spécifique}(a,b) \rightarrow (a \Rightarrow_v b) \qquad \text{avec } v \sim 1$$

Soit une requête sur "base de données déductive". Cette requête peut être difficilement satisfaite par un document sur "base de données". Cela signifie que la relation entre eux - la relation générique - ne peut être modélisée en une relation de dérivation qu'avec une faible certitude.

$$\text{générique}(a,b) \rightarrow (a \Rightarrow_v b) \qquad \text{avec } v \sim 0$$

En comparant ces analyses avec celles de la section précédente, on voit que la relation générique est une relation de dérivation avec grande certitude dans la section précédente, mais avec faible certitude dans cette section, et l'inverse pour la relation spécifique. Cela signifie que, pour avoir une relation de grande certitude entre deux mondes, il faut appliquer les relations dans le sens générique dans l'approche par modification de document, et il faut appliquer les relations dans le sens spécifique dans l'approche par modification de la requête. Ceci est résumé dans la figure suivante:

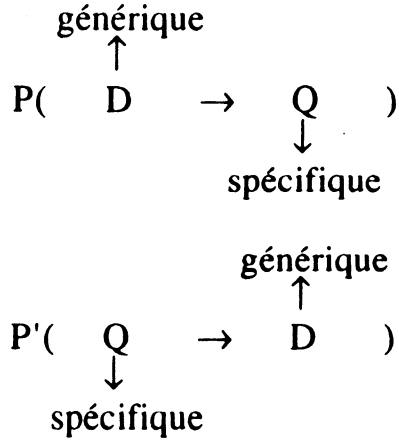


Fig.I.6. *Utilisations des relations apportant une grande certitude*

En généralisant cette idée, nous pouvons suggérer le critère de modélisation suivant:

Soit R(a,b) une relation existante. Si R(a,b) est modélisée en une relation de dérivation ($a \Rightarrow_V b$) dans un modèle fondé sur la modification du document, alors cette relation doit être modélisée en ($b \Rightarrow_V a$) dans le second modèle pour aboutir à un résultat identique (et vise versa).

4. COMPARAISON AVEC LES MODELES EXISTANTS

4.1. Le Modèle vectoriel

Rappelons que la pertinence dans le modèle vectoriel est mesurée par la similarité entre l'expression de la requête et celle du document. Une des formules les plus utilisées est la suivante (cf. 2.2):

$$\text{Sim}(D,Q) = \frac{\sum_i (a_i b_i)}{[\sum_i (a_i)^2 \sum_i (b_i)^2]^{1/2}} \qquad 1 \le i \le n$$

Pour décrire le modèle vectoriel, le modèle proposé doit être largement réduit:

Prenons l'approche par modification du document pour la valuation de $P(D \rightarrow Q)$. Le document D est considéré comme le monde initial w_0. La requête est considérée comme une formule à valuer. L'univers du modèle $\mathcal{W}$ est constitué d'un ensemble de mondes, chacun correspondant à un vecteur.

1. Dans le modèle vectoriel, nulle relation n'est prise en compte entre les différents termes. Un terme est considéré comme étant totalement indépendant des autres. Ainsi, il n'existe pas de transformation d'un monde vers un autre. Entre deux mondes différents, la relation est toujours valuée à 0.

2. On considère qu'un vecteur est une proposition atomique. Aucun opérateur n'existe dans le modèle vectoriel. Une formule bien définie est réduite en une simple proposition atomique (ou "true"):

$$f ::= P \in \mathcal{P} \mid \text{true}$$

Les fonctions δ et Δ n'ont plus besoin d'être définies dans ce contexte.

La valuation de $P(D \rightarrow Q)$ est déterminée par: $V_{w_0}(f) = V_{w_0}(P)$.

Pour obtenir le même résultat que dans le modèle vectoriel, nous définissons:

$$V_{w_0}(P) = Sim(D,Q) = \frac{\sum_i (a_i b_i)}{[\sum_i (a_i)^2 \sum_i (b_i)^2]^{1/2}} \quad 1 \leq i \leq n$$

et $V_{w_0}(\text{true}) = 1$.

Ainsi, on démontre que le modèle vectoriel est une instance très simplifiée de notre modèle.

Nous pouvons également effectuer cette comparaison en adoptant la vision d'évaluation plus stricte de Van Rijsbergen: si une proposition n'est pas parfaitement satisfaite (valuée à 1) dans un monde, ce monde doit être encore dérivé vers ses mondes possibles pour obtenir un monde final dans lequel la proposition est parfaitement satisfaite.

En effet, la similarité dans cette formule est définie par la valeur de $\cos(\varphi)$, φ étant l'angle entre les vecteurs D et Q. Dans le cas le nombre de dimensions est égal à 2 (correspondant à deux termes x et y), cela peut être illustré par la figure I.7.a:

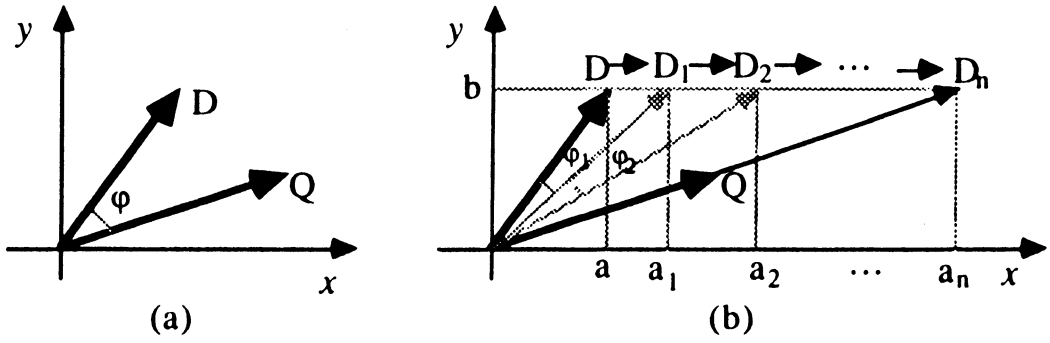


Fig.I.7. *Dérivations du document dans le modèle vectoriel*

Nous pouvons décrire comme suit la mesure de l'exhaustivité P(D→Q):

On choisit de modifier le document pour que la requête soit satisfaite dans le document modifié. On considère qu'une requête (Q) est totalament satisfaite dans un document (D) quand la requête et le document sont des vecteurs colinéaires. Etant donné qu'aucune relation sémantique n'existe dans ce contexte, si la requête n'est pas satisfaite dans un document, il faut donc réaliser un *ajout* d'information dans le document. Ceci se traduit par une augmentation du poids d'un terme dans le vecteur D.

Dans le cas de deux attributs (termes), si D et Q sont colinéaires, aucune modification n'est nécessaire sur D. L'implication P(D→Q) est directement évaluée à "1".

Si cela n'est pas le cas, comme dans la Fig.I.7.a, l'attribut *x* du vecteur D doit être modifié progressivement vers a_1, a_2, ... jusqu'à a_n, comme le montre la Fig.I.7.b. Durant cette modification, le vecteur D est progressivement transformé en D_1, D_2, ... D_n qui est colinéaire à Q.

On définit la certitude liée à une transformation d'un monde (représenté par un vecteur) vers un monde possible par la similarité existant entre eux:

$$\delta(D_{i-1},D_i) = Sim(D_{i-1},D_i) = \cos(\varphi_i)$$

où φ_i est l'angle formé par D_{i-1} et D_i.

Selon la valuation définie dans le modèle proposé, P(D→Q) est donc mesuré par:

$$P(D\rightarrow Q) = \Delta[\delta(D_0,D_1), \Delta[\ldots, \Delta[\delta(D_{n-2},D_{n-1}), \delta(D_{n-1},D_n)]\ldots]]$$

La fonction Δ est définie comme suit:

$$\Delta(x,y) = x \cdot y - [(1-x^2)(1-y^2)]^{1/2}$$

c'est-à-dire, avec $x=\cos(\varphi_1)$ et $y=\cos(\varphi_2)$:

$$\begin{aligned}&\Delta(\cos(\varphi_1),\cos(\varphi_2))\\ &= \cos(\varphi_1)\cdot\cos(\varphi_2) - [(1-\cos^2(\varphi_1))(1-\cos^2(\varphi_2))]^{1/2}\\ &= \cos(\varphi_1+\varphi_2)\end{aligned}$$

Si une séquence de dérivations $D \rightarrow D_1 \rightarrow D_2 \rightarrow \ldots \rightarrow D_n$ conduit le vecteur D à un vecteur D_n colinéaire à Q, cette séquence donne la valeur de certitude suivante:

$$\begin{aligned}P(D\rightarrow Q) &= \Delta[\delta(D_0,D_1), \Delta[\ldots, \Delta[\delta(D_{n-2},D_{n-1}), \delta(D_{n-1},D_n)]\ldots]]\\ &= \cos(\sum_{i=1}^{n} \varphi_i) = \cos(\varphi)\\ &\qquad /^*\text{où } \varphi \text{ est l'angle entre les vecteurs D et Q}^*/\\ &= \mathrm{Sim}(D,Q).\end{aligned}$$

On a donc démontré que le modèle booléen peut être également considéré sous la vision de Van Rijsbergen.

On peut également voir dans cette démonstration, que toute séquence de dérivations rapprochant (sans retour) le vecteur de document au vecteur Q est une séquence optimale (par exemple, les dérivations illustrées dans Fig.7.c), car une telle séquence donne la valeur maximale à la certitude de l'implication $D \rightarrow Q$, qui est indépendante aux étapes intermédiaires des dérivations.

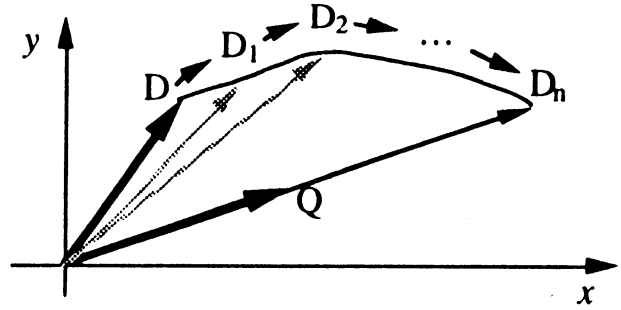


Fig. 7.c. *Une séquence de dérivations optimale*

D'une façon similaire, on peut également valuer $P'(Q \rightarrow D)$ par $\mathrm{Sim}(D,Q)$. Cela nous conduit à la remarque suivante:

Dans le modèle vectoriel, le document et la requête ne présentent aucune différence tant au niveau de la représentation qu'au niveau du rôle joué durant la valuation de leur correspondance (cf.R(D,Q)). Cela traduit, dans une certaine mesure, une neutralisation des deux critères: un document

est exhaustif si et seulement s'il est spécifique pour une requête. Cette optique est très restrictrice et n'est adoptée que dans de rares applications. Il nous semble que c'est cette restriction que Salton *et al.* ont voulu lever dans un modèle étendu ([Salton83b]) par l'introduction d'un paramètre permettant de différencier les deux critères et de faire varier leur importance relative.

Un autre inconvénient provient de l'hypothèse d'indépendance entre les termes. Cela implique qu'un terme présent dans un vecteur est un terme considéré comme un représentant d'une certaine famille. Toute la sémantique doit être véhiculée par les termes choisis. Etant donnée la pauvreté de cette représentation, cette approche ne peut donner un résultat satisfaisant que dans les cas très simples. Quand le domaine d'application devient complexe, cette approche devient rapidement une approche approximative. Ce modèle implique donc une prise en compte simpliste de la sémantique, réduisant largement la finesse qu'il peut atteindre.

Dans les études plus récentes, ce modèle a été plus ou moins étendu. Les études ont portée sur l'introduction des relations entre termes, l'introduction des opérateurs booléens, ... ([Salton83b]). Ces extensions démontrent, dans une certaine mesure, la nécessité du développement d'un modèle général.

4.2. Le modèle booléen

Dans le modèle booléen standard, les documents et les requêtes sont représentés par une expression booléenne (cf.2.1). Comme dans le modèle vectoriel, aucune relation entre les termes n'est prise en compte. L'inclusion de ce modèle dans le modèle général est évidente. Il suffit d'instancier notre modèle dans un cas beaucoup plus restreint. Par exemple, la fonction V ($\in \{0,1\}$) sera définie comme suit:

$$V_w(P) = C_P(w).$$
$$V_w(f_1 \wedge f_2) = MIN(V_w(f_1), V_w(f_2))$$
$$V_w(\neg f) = 1 - V_w(f)$$
$$V_w(true) = 1.$$

où $C_P(w)$ est définie comme suit ("|=" signifie une déduction logique):

$$C_P(w) = 1, \text{ si } w \models P$$
$$C_P(w) = 0, \text{ sinon.}$$

Par rapport aux deux critères évoqués dans la valuation de la correspondance entre un document et une requête, on remarque qu'un seul des deux critères (l'exhaustivité) est pris en compte dans le modèle booléen.

L'absence du critère sur la spécificité signifie que l'importance des termes de la requête dans le document, ou le développement du thème de la requête dans le document n'est pas important. Cette position est souvent prise dans des applications où l'on cherche une réponse précise, ainsi que dans des systèmes de question-réponse. L'hypothèse sous-jacente est que, si un terme est présent dans un document, il a une importance maximale; s'il est absent, le document ne concerne pas du tout le thème correspondant.

En général, ce modèle standard paraît trop strict pour les SRI: un document satisfaisant la plus grande partie d'une requête est souvent jugé aussi mauvais qu'un document ne la satisfaisant pas du tout, et un document développant très en détail les problèmes posés par la requête n'est pas meilleur qu'un document survolant ces problèmes. Par conséquent, les documents obtenus en réponse ne sont pas ordonnés (ce qui peut poser des problèmes dans le cas où la requête est satisfaite par beaucoup de documents), et un taux de silence élevé peut se produire (surtout dans le cas où peu de documents satisfont la requête).

Pour remédier à ces défauts, beaucoup de modèles ont assoupli leur mesure de correspondance en pondérant les termes des documents et/ou les termes des requêtes, et en définissant une mesure fondée sur la théorie des ensembles flous ou les logiques multi-valuées ([Bookstein80, Salton83b]). On peut remarquer que la pondération d'un terme peut être interpréter de deux façons: l'*importance* du terme pour un document (ou pour une requête) ou l'*implication* du terme dans un document (ou une requête). La première correspond à notre définition de spécificité et la deuxième à l'exhaustivité. Ainsi, nous pouvons conclure que toutes ces modifications sur le modèle booléen standard ont pour but de différencier la mesure de correspondance en prenant en compte l'inclusion de la requête dans le document et/ou l'importance du thème de la requête dans le document. Les valuations modifiées peuvent donc aussi être décrites par notre modèle.

Si l'on veut encore étendre ce modèle en introduisant les relations non classiques entre termes (autres que les relations d'implications logiques), le modèle général sera aussi très adéquat pour décrire la valuation dans le modèle étendu, en considérant les relations non classiques comme des relations de dérivation du modèle général.

4.3. Le modèle probabiliste

Soient une requête Q et un document D. Pour valuer $P(D \rightarrow Q)$, on utilise l'approche de la valuation par modification du document. Un document est considéré comme un monde initial. Une requête est considérée comme une

formule à valuer. Une requête dans le modèle probabiliste est toujours considérée comme une proposition atomique P.

Dans le modèle probabiliste indépendant, il n'existe aucune relation entre deux termes différents. Ainsi, la notion de transformation entre mondes n'existe pas. Les connaissances du système K sont constituées d'un ensemble de probabilités sur les termes, par rapport à une proposition atomique P:

$$\{P_r(t_1), P_n(t_1|nrel), P_r(t_2), P_n(t_2), \ldots P_r(t_n), P_n(t_n)\}$$

Pour décrire ce modèle, nous instancions notre modèle de la façon suivante:

- Une formule ne peut être qu'une proposition atomique.

- L'univers du modèle $\mathcal{W}$ est constitué d'un ensemble de mondes, chaque monde (w) étant constitué d'un ensemble de termes: $w=(t_1, t_2, \ldots t_n)$. Entre deux mondes différents, la relation est toujours valuée à 0.

- On a pas besoin des fonctions δ et Δ.

- La fonction V est définie comme suit:

$$V_w(P) = C_P(w).$$
$$V_w(true) = 1.$$

$$\text{avec } C_P(w) = \frac{p(rel) \prod_{i=1}^{m} P_r(t_i)}{p(rel) \prod_{i=1}^{m} P_r(t_i) + p(nrel) \prod_{i=1}^{m} P_n(t_i)}$$

et $t_i \in w$ $(1 \leq i \leq m)$.

Le modèle probabiliste indépendant est donc totalement inclus dans notre modèle.

Il est intéressant de montrer comment notre modèle peut décrire assez facilement le modèle probabiliste dépendant.

Par rapport au modèle indépendant, on possède en plus un ensemble de probabilités de dépendance: $p(t_i|t_1,t_2,\ldots,t_{i-1})$.

Pour décrire ce modèle par le modèle proposé, on considère cette probabilité de dépendance comme une relation de dérivation:

$$t_i \Rightarrow_v (t_1,t_2,\ldots,t_{i-1})$$

$$\text{et } v = \frac{p(rel)P_r(t_i|(t_1,t_2,\ldots,t_{i-1}))}{p(rel)P_r(t_i|(t_1,t_2,\ldots,t_{i-1}))+p(nrel)P_n(t_i|(t_1,t_2,\ldots,t_{i-1}))}$$

c'est-à-dire que t_i peut être dérivé en $(t_1,t_2,\ldots,t_{i-1})$ avec la certitude v.

Entre deux mondes dans $\mathcal{W}$, la relation est définie comme suit:

$$\delta(w,w') = \frac{p(rel)P_r(t_i|w')}{p(rel)P_r(t_i|w')+p(nrel)P_n(t_i|w')}$$

si $w= w'\cup\{t_i\}$

Dans ce cas, on doit introduire la notion de transformation d'un monde vers un monde possible. Une transformation consiste donc à retirer un terme à partir du monde.

La fonction $C_P(w)$ est définie par:

$$C_P(w) = \frac{p(rel)P_r(t_i)}{p(rel)P_r(t_i)+p(nrel)P_n(t_i)} \quad \text{si } w=t_i$$

$C_P(w) = 0$, autrement.

Une formule est définie par la syntaxe suivante:

$f::= P\in \mathcal{P} \mid \Diamond f \mid true$

La valuation $V_w(f)$ est définie comme suit:

$$V_w(P) = MAX(C_P(w), V_w(\Diamond P))$$
$$V_w(\Diamond f) = \underset{w'}{MAX}(\Delta(\delta(w,w'),V_{w'}(P))$$
$$V_w(true) = 1$$

où la fonction Δ est définie de façon suivante:

$$\Delta(x,y) = \frac{p(nrel)\ x\ y}{p(nrel)\ x\ y + p(rel)\ (1-x)\ (1-y)}$$

c'est-à-dire: soient $w=w'\cup\{t_i\}$,

$$\delta(w,w') = \frac{p(rel)P_r(t_i|w')}{p(rel)P_r(t_i|w')+p(nrel)P_n(t_i|w')}$$

$$V_{w'}(P) = \frac{p(rel)P_r(w')}{p(rel)P_r(w')+p(nrel)P_n(w')},$$

on a:

$$\Delta[\delta(w,w'),V_{w'}(P)]= \frac{p(rel)P_r(t_i|w')P_r(w')}{p(rel)P_r(t_i|w')P_r(w')+p(nrel)P_n(t_i|w')P_n(w')}$$

Dans le modèle probabiliste dépendant, l'ordre dans lequel les termes sont pris en compte n'a pas d'importance, c'est-à-dire:

$$\begin{aligned} P_r(t_1,t_2,t_3,\ldots) &= P_r(t_1|(t_2,t_3,t_4,\ldots))\ P_r(t_2,t_3,t_4,\ldots) \\ &= P_r(t_2|(t_1,t_3,t_4,\ldots))\ P_r(t_1,t_3,t_4,\ldots) \\ &= \ldots \end{aligned}$$

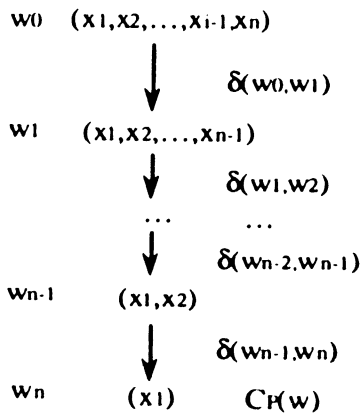
En général, soit w un ensemble de termes dont t_i:

$$P_r(w) = P_r(t_i|\ (w-\{t_i\}))\ P_r(w-\{t_i\}).$$
$$P_n(w) = P_n(t_i|\ (w-\{t_i\}))\ P_n(w-\{t_i\}).$$

Ainsi, tout chemin de transformation d'un monde initial vers un monde final (contenant un seul terme) est un chemin de transformation qui donne la valeur maximale à $V_w(\Diamond f)$. Ainsi, $V_w(\Diamond f)$ peut être défini simplement par:

$$V_w(\Diamond f) = \underset{w'}{MAX}[\Delta(\delta(w,w'),\ V_{w'}(f))] = \Delta(\delta(w,w'),\ V_{w'}(f)).$$

Soit un document D, correspondant à un monde initial $w_0=(t_1,t_2,\ldots,t_n)$. Avec les fonctions définies ci-dessus, on peut valuer $V_{w_0}(P)$ en faisant les transformations suivantes:



$$\text{où } \delta(w_0,w_1) = \frac{p(rel)P_r(t_n|w_1)}{p(rel)P_r(t_n|w_1)+p(nrel)P_n(t_n|w_1)}$$

$$\delta(w_1,w_2) = \frac{p(rel)P_r(t_{n-1}|w_2)}{p(rel)P_r(t_{n-1}|w_2)+p(nrel)P_n(t_{n-1}|w_2)}$$

...

$$\delta(w_{n-3},w_{n-2}) = \frac{p(rel)P_r(t_3|w_{n-2})}{p(rel)P_r(t_3|w_{n-2})+p(nrel)P_n(t_3|w_{n-2})}$$

$$\delta(w_{n-2},w_{n-1}) = \frac{p(rel)P_r(t_2|w_{n-1})}{p(rel)P_r(t_2|w_{n-1})+p(nrel)P_n(t_i|w_{n-1})}$$

$$\text{et } C_P(w_{n-1}) = \frac{p(rel)P_r(t_1)}{p(rel)P_r(t_1)+p(nrel)P_n(t_1)}$$

On a donc la valuation suivante de $V_{w_0}(P)$:

$$V_{w_0}(P)$$
$$= \Delta(\delta(w_0,w_1),$$
$$\Delta(\delta(w_1,w_2),$$
$$\dots$$
$$\Delta(\delta(w_{n-3},w_{n-2}),$$
$$\Delta(\delta(w_{n-2},w_{n-1}),C_P(w_{n-1})))\dots))$$
$$= \Delta(\delta(w_0,w_1),$$
$$\Delta(\delta(w_1,w_2),$$
$$\dots$$
$$\Delta(\delta(w_{n-3},w_{n-2}),$$
$$\frac{p(rel)P_r(t_2|t_1)P_r(t_1)}{p(rel)P_r(t_2|t_1)P_r(t_1)+p(nrel)P_n(t_2|t_1)P_r(t_1)})\dots))$$

$$= \dots$$
$$= \frac{p(rel)P_r(t_n|(t_1\dots t_{n-1}))\dots P_r(t_2|t_1)P_r(t_1)}{p(rel)\,P_r(t_n|(t_1\dots t_{n-1}))\dots P_r(t_2|t_1)\,P_r(t_1) + p(nrel)\,P_n(t_n|(t_1\dots t_{n-1}))\dots P_n(t_2|t_1)\,P_n(t_1)}$$

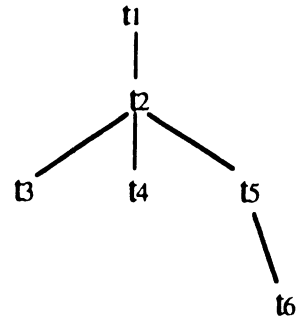
$$= \frac{p(rel)P_r(w_0)}{p(rel)P_r(w_0)+p(nrel)P_n(w_0)}$$

où $P_r(w_0) = P_r(t_n|(t_1\dots t_{n-1}))\,P_r(t_{n-1}|(t_1\dots t_{n-2}))\dots P_r(t_2|t_1)P_r(t_1)$

et $P_n(w_0) = P_n(t_n|(t_1\dots t_{n-1}))\,P_n(t_{n-1}|(t_1\dots t_{n-2}))\dots P_n(t_2|t_1)P_n(t_1)$

C'est exactement la même valuation des probabilités définie dans le modèle probabiliste dépendant. Ainsi, on peut conclure que le modèle probabiliste est une instance de notre modèle.

Nous reprenons l'exemple donné dans 2.3. Soit le document $D=w_0=(t_1,t_2,t_3,t_4,t_5,t_6)$ et l'arborescence de dépendance suivante:



En utilisant la valuation définie ci-dessus, on peut obtenir:

$$V_{w_0}(P) = \frac{p(rel)P_r(w_0)}{p(rel)P_r(w_0)+p(nrel)P_n(w_0)}$$

où $P_r(w_0) = P_r(t_1)\ P_r(t_2|t_1)\ P_r(t_3|t_2)\ P_r(t_4|t_2)\ P_r(t_5|t_2)\ P_r(t_6|t_5)$
$P_n(w_0) = P_n(t_1)\ P_n(t_2|t_1)\ P_n(t_3|t_2)\ P_n(t_4|t_2)\ P_n(t_5|t_2)\ P_n(t_6|t_5)$

Comme on peut observer, le calcul dans le modèle probabiliste dépendant est extrêmement compliqué, ce qui est un grand inconvénient. Dans la pratique, les probabilités ne peuvent qu'être estimées approximativement.

4.4. Les modèles sémantico-linguistiques

La distinction entre les trois phases conceptuelles (l'indexation des documents, l'interprétation de la requête, et l'évaluation) pour une interrogation est descriptible par notre modèle. L'indexation d'un document et l'interprétation d'une requête constituent respectivement certaines étapes dans les dérivations du document initial et dans les dérivations de la requête initiale. On peut illustrer cela par la figure suivante, où Q' et D' sont respectivement les représentations internes d'une requête et d'un document.

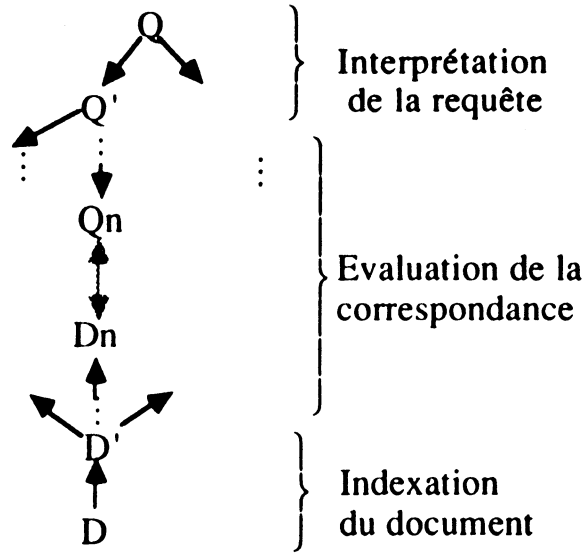


Fig.I.8. *Interrogation vue par rapport au modèle*

Théoriquement, la valuation de la correspondance entre la requête et le document doit s'effectuer entre le document initial et la requête initiale: R(D,Q). En réalité, la valuation est souvent effectuée entre D' et Q': R(D',Q'). Il est évident que D' et Q' ne sont pas toujours tout à fait équivalents à D et Q. Ce changement peut donc engendrer un certain écart.

Différentes stratégies de valuation de R(D,Q) sont souvent pré-établies dans le système. Ces valuations sont souvent des valuations classiques. Une évaluation de requête consiste donc à choisir la stratégie la mieux adaptée à une situation donnée (un utilisateur et un domaine déterminés). Par exemple, dans [Croft85], les valuations pré-établies sont celles du modèle booléen, du modèle probabiliste, du modèle booléen étendu, ...

Les valuations classiques de R(D,Q) étant déjà démontrées comme descriptibles par le modèle proposé, il est donc indirectement démontré que l'ensemble des valuations d'un tel système est inclus dans le modèle proposé.

En comparant la valuation fondée sur le modèle sémantico-linguistique, proposée dans certains SRI, on peut souvent observer leur incohérence, due à une définition souvent empirique.

Par exemple, la valuation définie dans IOTA n'est pas cohérente (cf.2.4). On ne peut décrire dans notre modèle, qu'une partie de la valuation: la valuation d'une proposition atomique P=t (terme d'indexation), et la valuation de $f_1 \wedge f_2$.

Cette comparaison montre l'avantage de l'approche commençant par la définition d'un modèle général, par rapport à une approche commençant par

la définition d'un modèle spécifique, fondé sur les besoins particuliers d'une application.

Dans IOTA, on remarque que les deux critères évoqués dans la formule générale de la correspondance sont tous pris en compte. Quand la requête est constituée d'un seul terme, par exemple, la correspondance est mesurée par la formule (pour Q=t):

$$R(D,Q) = \frac{F_1(Q)+F_2(Q)}{2} = \frac{REP(t,D)+REP(D,t)}{2}$$

dans laquelle REP(t,D) correspond à P'(t→D) (i.e. P'(Q→D)), REP(D,t) correspond à P(D→t) (i.e. P(D→Q)), et la fonction R est définie par:

$$R(D,Q) = \frac{P'(Q\rightarrow D)+P(D\rightarrow Q)}{2}$$

Un autre aspect intéressant dans IOTA est l'estimation de la typologie de l'utilisateur (spécialiste, moyen, débutant). La qualité d'une réponse est estimée relativement à la typologie de l'utilisateur, conduisant éventuellement à une reformulation automatique de la requête pour donner une autre réponse.

Cet aspect correspond bien à la formule générale proposée pour la valuation de la correspondance:

$$R(D,Q) = F\ [\ P_K(D\rightarrow Q),\ P'_K(Q\rightarrow D)\ ]$$

Etant donné que cet aspect n'est considéré dans IOTA que de façon non formelle, il est difficile d'effectuer une comparaison formelle avec le modèle proposé. Mais intuitivement, cet aspect peut être décrit par des définitions de P_K et de P'_K relatives à K.

4.5. Le modèle logique

La comparaison avec le modèle logique de Van Rijsbergen (cf.2.5) est directe, étant donné que le fondement de ce modèle est inclus dans celui du modèle proposé.

Par rapport au modèle logique, nous avons intégré un critère supplémentaire dans la valuation de la correspondance, et nous avons associé une signification encore plus étendue à l'implication document→requête. Cette implication est définie comme une implication conditionnelle dans le modèle logique. Dans notre modèle, elle est étendue pour tout type de *dérivation*. Ainsi, notre modèle peut recouvrir un nombre de cas beaucoup plus grand.

5. COMPARAISON AVEC DES MODELES ISSUS DES BASES DE DONNÉES

De nombreux outils informatiques sont utilisés dans les SRI, tels que ceux utilisés dans les bases de données, ceux de l'intelligence artificielle, etc.... Dans cette section, nous allons montrer que les bases de données déductives sont des outils adaptées pour les SRI.

5.1. Utilisation des bases de données classiques

Les systèmes de bases de données constituent des outils très efficaces pour gérer et manipuler de grands volumes de données. Ils présentent de grands avantages dans beaucoup de domaines dans lesquels le volume de données à traiter est important, tels que la CAO et les SRI.

L'utilisation des techniques de bases de données dans les SRI est pourtant très souvent restreinte à l'évaluation des attributs externes des requêtes (autres que le contenu), tels que l'attribut auteur, l'attribut date de publication, ... ([Blair88]). Cela est dû à la structuration similaire des attributs externes des documents par rapport à ceux des bases de données. On peut formaliser aisément les attributs externes d'un document sous forme relationnelle. Une requête d'un SRI portant sur les attributs externes peut être donc transformée en une requête de base de données.

Un SRI utilisant les bases de données de cette façon correspond au modèle booléen standard (ayant une valuation binaire). Le but est de trouver tout document satisfaisant l'implication suivante:

$$P(D \rightarrow Q) = 1$$

où l'implication $\rightarrow$ est restreinte en une implication logique et où D devient une représentation des attributs externes d'un document. Comme nous l'avons montré, le modèle booléen est un sous-modèle du modèle général proposé, cette méthode d'évaluation de base de données est aussi incluse dans notre modèle.

On donne un exemple simple pour illustrer cette utilisation. Considérons la relation suivante décrivant des documents:

DOCUMENT(NO, TYPE, TITRE, AUTEUR, DATE, ...)

Une instance de cette relation DOCUMENT(a,b,c,d,e,...) correspond à un document qui peut être représenté par l'expression booléenne suivante:

$D = (NO=a) \wedge (TYPE=b) \wedge (TITRE=c) \wedge (AUTEUR=d) \wedge (DATE=e) \wedge \ldots$

Etant donnée une requête portant sur des "livres" de "Kripke", elle peut être exprimée de façon équivalente par la requête de base de données suivante:

```
select    DOCUMENT
where     (TYPE = 'livre')
          and
          (AUTEUR = 'Kripke')
```

dont la condition est équivalente à la formule booléenne suivante:

$$Q=(TYPE='livre') \wedge (AUTEUR='Kripke')$$

L'évaluation de cette formule sélectionne donc tout document (relation) satisfaisant l'implication logique $D \rightarrow Q$.

Certaines utilisations ont porté sur l'attribut interne - l'attribut contenu ([Blair88, Desai86]). Mais dans ces utilisations, l'attribut contenu est banalisé comme un attribut externe, c'est-à-dire qu'il est représenté par un ensemble de mots-clés indépendants. Ces mots-clés ont les mêmes caractéristiques que les valeurs des attributs externes. Ainsi un mot-clé est considéré comme l'unique représentation d'une certaine sémantique. La présence ou l'absence d'un mot-clé implique la présence ou l'absence de la sémantique correspondante. Comme l'ont indiqué Deogun et Raghavan ([Deogun88]), les systèmes considérant l'attribut contenu de la même manière que les attributs externes ne sont pas vraiment des SRI.

A notre avis, la différence clé entre les attributs externes et internes est la suivante:

1). les valeurs des attributs externes sont indépendantes les unes des autres, tandis que les valeurs de l'attribut contenu (représentées par des descripteurs [Deogun88]) sont souvent dépendantes.

Pour un attribut externe, une valeur t se distingue de toute autre valeur $t' \neq t$. Une requête demandant t comme valeur de l'attribut ne peut pas être satisfaite par un document ayant t' comme valeur de l'attribut.

Par contre, si l'attribut contenu d'un document est représenté par un terme t (ex. "base de données"), ce document peut être en relation avec d'autres termes t' ($\neq t$) (ex. "banque de données). Une requête demandant t peut être satisfaite non seulement par les documents portant sur t, mais aussi

par les documents portant sur certains t'. Cela implique qu'il existe une dépendance entre les différentes représentations de l'attribut contenu.

Les attributs d'une BD conventionnelle sont fondés sur le même principe (l'unicité du nom) que les attributs externes d'un document: deux valeurs différentes sont considérées comme deux objets distincts. La dépendance de l'attribut contenu peut être donc difficilement prise en compte à travers une base de données conventionnelle.

En plus de la différence énoncée ci-dessus, sur la nature des données traitées dans les SRI et dans les bases de données conventionnelles, one peut mentionner également des différences sur les aspects suivants:

2).*Organisation des données*: Dans une base de données classique, les données sont identifiées par un nom d'attribut, et les attributs sont interconnectés par des relations pré-définies. Les données sont donc très structurées (par les schémas conceptuels). Dans le cas d'un SRI, il est souvent impossible de structurer des données aussi rigoureusement. Dans le cas d'un texte, par exemple, on peut être amené à représenter un contenu ayant une structure très complexe. L'organisation des bases de données classiques n'est pas suffisamment souple et riche pour les SRI.

3). *Critère d'évaluation*. L'évaluation dans les BD est fondée sur un seul critère: le critère d'exhaustivité $D \rightarrow Q$. L'autre critère des SRI (la spécificité $Q \rightarrow D$) peut difficilement être pris en compte par un système de BD.

4). *Valuation*. L'évaluation de la requête dans les bases de données est fondée sur des comparaisons exactes entre celle-ci et les données, conduisant à une valuation binaire. Dans le cas d'un SRI, on ne peut pas se limiter à une simple comparaison résultant en "vrai" ou "faux". La satisfaction est, dans la plupart des cas, valuée en terme de plausibilité. Le concept d'exactitude dans la comparaison n'est plus valable. C'est donc la valuation continue d'une fonction de vérité qui est nécessaire.

Ces différences montrent l'inadéquation des bases de données classiques pour les SRI. Les bases de données déductives montrent une grande similarité dans leur interrogation avec l'évaluation des SRI que nous avons décrite dans le modèle général. Elles permettent de réduire les différences énoncées précédemment entre BD et SRI.

5.2. Notion de base de données déductives

"Une base de données déductive est une base de données dans laquelle de nouveaux faits peuvent être dérivés de ceux qui sont explicitement introduits" ([Gallaire84]).

Depuis que cette notion de base est bien établie, de nombreux efforts sont consacrés à la réalisation d'un SGBD déductif efficace. Dans cette section, nous n'allons pas présenter la totalité des développements dans ce domaine, mais seulement l'aspect interrogation afin de le comparer avec le modèle général d'évaluation de SRI proposé.

Une base de données en général, peut être considérée de deux manières différentes via la logique: soit comme une *interprétation* de la logique du premier ordre, soit comme une *théorie* du premier ordre. Selon le premier point de vue, une requête est une formule devant être évaluée sur un ensemble de faits de la base lui donnant la valeur "vrai". Considérée sous l'angle de la théorie, une requête est un théorème devant être prouvé. Ces deux façons de considérer les bases de données formalisent respectivement le concept de base de données conventionnelle et de base de données déductive (BDD).

Soit K la théorie correspondant à une BDD (déterminée), K est formée d'un ensemble d'implications de la forme suivante:

$$P_1 \wedge P_2 \wedge \ldots \wedge P_n \Rightarrow C \qquad (n \geq 0)$$

où P_i et C sont des propositions atomiques de la logique du premier ordre.

On distingue trois groupes d'axiomes dans cette théorie: les axiomes avec n=0 sont appelés les *faits initiaux* et l'ensemble de ceux-ci constitue l'état initial de la base de données (noté BD), une partie des axiomes avec n>0 représente des *contraintes d'intégrité* (CI), et l'autre partie des axiomes ayant n>0 représente des *règles de déduction* ou des *lois de déduction* (LD).

Un fait quelconque (F) est dit *déductible* si, à partir des faits initiaux (BD), on peut prouver la vérité du fait avec les lois de déduction LD (sous les contraintes d'intégrité CI). On peut considérer cette déduction comme une déduction à partir de BD en utilisant les règles définies dans K:

$$BD \models_K F$$

Soit I une instanciation de toutes les variables existant dans la requête. Notons I(Q) l'instanciation correspondante de la condition de la requête Q. L'instanciation de la condition de la requête I(Q) constitue un fait ou une expression booléenne sur un ensemble de faits.

La réponse à la requête Q est constituée de toutes les instanciations possibles I des variables de la requête Q, telles que l'expression I(Q) soit vérifiée soit dans l'état initial de la base de données, soit dans un état déductible:

$$\{I : BD \models_K I(Q)\}$$

Par exemple, soient x une variable, b un individu de la base de données et Grandpère(a,x) une requête (qui recherche tous les petits-fils de a). La réponse à cette requête est constituée par toutes les instanciations possibles de Grandpère(x,b) vérifiées dans la base de données initiale ou déductible. Si les instances suivantes sont vérifiées:

Grandpère(a,b)
Grandpère(a,c)
Grandpère(a,d)

alors, la réponse à cette requête est donc {b,c,d}.

5.3. Comparaison de l'évaluation des BDD avec celle des SRI

Dans l'expression précédente, on peut définir un modèle $\mathcal{K}$par rapport à K. L'état initial de la base de données (BD) constitue le monde initial w_0.

Une proposition atomique P ($\in \mathcal{P}$) correspond au fait obtenu par l'instanciation d'une condition atomique de la requête, en remplaçant la variable dans la condition par un individu de la base. Par exemple Grandpère(a,b) est considéré comme une proposition atomique pour l'exemple précédent.

Une formule quelconque est définie de la façon suivante:

$$f ::= P \in \mathcal{P} \mid f_1 \wedge f_2 \mid \neg f \mid \Diamond f \mid \text{true}$$

Considérons l'ensemble des règles d'implication comme des relations de dérivation certaines. L'application d'une telle relation transforme donc un monde vers son monde possible.

$C_P(w)$ est une fonction binaire définie de la façon suivante:

$C_P(w) = 1$, si $P \in w$
$C_P(w) = 0$, sinon.

déductives est une instance de l'évaluation des SRI. En effet, on peut considérer un système de base de données déductive comme un SRI booléen étendu dans lequel on a intégré un ensemble de relations d'implication logiques.

Par rapport aux différences existant entre les BD classiques et les SRI (cf.5.1), bien que les BDD permettent de considérer une certaine dépendance entre les données, les différences existant entre BDD et SRI montrent que les BDD actuelles ne sont pas encore suffisantes:

1. L'évaluation dans les BDD est aussi fondée sur un seul des deux critères de la correspondance d'un document à une requête ($P(D \rightarrow Q)$ et $P'Q \rightarrow D)$).

2. La théorie des BDD est actuellement fondée sur la logique du premier ordre. Les relations prises en compte sont uniquement des implications logiques. Comme nous l'avons montré, la modélisation des relations existant dans le domaine des SRI en des relations d'implication logique n'est souvent pas réalisable.

   En effet, dans la description précédente des BDD, nous avons considéré les relations d'implication logique comme des relations de dérivation particulières. En réalité, l'application des relations d'implication logique ne conduit pas à un changement de monde. Les BDD décrites sont, en quelque sorte, des BDD étendues.

3. La valuation dans les BDD est toujours binaire. La notion d'incertitude nécessaire pour la plupart des SRI ne peut pas être établie dans les BD.

4. La possibilité de représenter un document quelconque par un fait d'une base de données reste encore très faible, étant donné qu'actuellement, la recherche sur les bases de données s'appuie plutôt sur une représentation relationnelle des données qui est loin d'être suffisamment riche pour les SRI.

La possiblité de représenter des documents sous forme de faits dans les BD est en train d'être augmentée avec les recherches sur les *objets complexes* (cf.la prochaine section). On espère que ces études permetteront le rapprochement progressif des BD avec les SRI dans un avenir proche.

Les trois premières raisons sont étroitement liées. Le fond est la restriction inhérente à la base théorique des BDD - la logique du premier ordre. Cette base théorique est beaucoup trop restrictive pour la plupart des systèmes modernes, tels que les systèmes experts, la CAO et les SRI, ... Pour que les BDD soient adaptées à ces systèmes, elles doivent être fondées sur une logique moins stricte, ou une logique d'un ordre plus haut. Ici, nous

$\mathcal{W}$ est l'univers du modèle, constitué d'un ensemble des faits. $\mathcal{W}$ est construit de la façon suivante:

1. L'état initial de la base BD constitue un monde dans $\mathcal{W}$.

2. Pour tout $w \in \mathcal{W}$, si $(P_1 \wedge P_2 \wedge \ldots \wedge P_n \Rightarrow C) \in K$ et $P_1 \in w$, $P_2 \in w$, ..., $P_n \in w$, alors $w'=w \cup \{C\} \in \mathcal{W}$.

On définit la fonction binaire $\delta(w,w')$ entre deux mondes quelconques comme suit:

$\delta(w,w')=1$, si $P_1 \in w$, $P_2 \in w$, ..., $P_n \in w$,
$w'=w \cup \{C\}$ et $(P_1 \wedge P_2 \wedge \ldots \wedge P_n \Rightarrow C) \in K$
$\delta(w,w')=0$, sinon

La valuation dans ce cas est binaire. La définition de l'opérateur Δ est impliquée par ses propriétés:

$\Delta(0,c)=0$ $\quad \Delta(c,0)=0$
$\Delta(1,c)=c$ $\quad \Delta(c,1)=c$

La fonction V: $\mathcal{F} \rightarrow \{0,1\}^{\mathcal{W}}$ est définie de la façon suivante:

- $V_w(P) = MAX[\ C_P(w),\ V_w(\Diamond P)]$, $P \in \mathcal{P}$
- $V_w(f_1 \wedge f_2) = MIN(V_w(f_1), V_w(f_2))$
- $V_w(\neg f) = 1 - V_w(f)$
- $V_w(\Diamond f) = \underset{w'}{MAX}(\Delta[\delta(w,w'), V_{w'}(f)])$, avec $w' \in \mathcal{W}$
- $V_w(true) = 1$

Etant donnée une requête Q à évaluer, chaque instanciation de Q est une expression booléenne des propositions atomiques. Elle est considérée comme une formule dans ce modèle: f=I(Q). Cette formule est valuée par la fonction V définie ci-dessus.

La réponse à la requête est constituée de toutes les instanciations de la requête valuées à 1 par rapport à la base de données initiale (w_0), c'est-à-dire celles qui sont soient directement vérifiées dans la base de données, soient déductibles à partir de celle-ci:

$$\text{Réponse} = \{I: V_{w_0}(I(Q)) = 1\}$$

A partir de cette description, on peut facilement conclure que l'évaluation des bases de données déductives est descriptible par le modèle proposé pour les SRI. Autrement dit, l'évaluation des bases de données

rejoignons l'opinion de Minski ([Minski88]) qui propose un élargissement du fondement des BDD de la logique du premier ordre vers une logique d'un ordre plus haut, pour s'adapter aux applications où les informations sont incomplètes ou imprécises, comme c'est le cas des SRI en général.

On a remarqué certains efforts dans cette direction dans les recherches menées en BDD. Par exemple, une interprétation de la négation est proposée en utilisant la logique modale ([Fitting85]). Dans cette interprétation, l'ensemble des valeurs de vérité est étendu de {vrai, faux} à {vrai, faux, indéfini}. Cette extension est très intéressante, car elle permet ensuite de développer la valeur "indéfini" avec une logique non classique, telle que la logique floue, la logique modale, etc.... Il est certain que cette extension sera poursuivie, aboutissant finalement à une logique plus souple, qui correspond à la perspective de Minski.

Pour notre objectif, on suggère ([Nie89b]) que les BDD soient fondées sur une logique modale non classique, telle qu'elle a été décrite dans notre modèle. Les BDD ainsi formées seront alors très adaptées pour l'implémentation des SRI. Dans [Abiteboul87b], une approche similaire est proposée. Celle-ci considère le traitement des informations incomplètes dans les bases de données comme des dérivations des mondes possibles de la logique modale.

5.4. Les objets complexes dans les bases de données

Le modèle de représentation relationnelle des données est souvent nommé "première forme normale" (1NF). Dans ce modèle, il est nécessaire que chaque attribut d'une relation soit atomique. Cela correspond à une structure "plate" pour les données.

Les structures proposées pour représenter les objets plus complexes sont des structures appelées "non-première forme normale" (N1NF) ([Makinouchi77]). Plusieurs modèles N1NF ont été simultanément proposés: le modèle NF^2 ([Jaeschke82]), le modèle V-relationnel ([Bançilhon82]), ainsi que les modèles plus généralisés comme le modèle de format ([Hull84]), le modèle de données logique ([Kuper85]), etc.... En résumé, les N1NF permettent plus ou moins une récurrence entre les définitions de relation et d'ensemble, en plus du modèle relationnel. En d'autres termes, une relation peut avoir des ensembles comme attribut, et un ensemble peut avoir des relations comme éléments.

Au niveau théorique, certains formalismes pour les objets complexes sont proposés ([Abiteboul87a]). Ils permettent encore plus de récurrence sur

la définition des ensembles et des relations. Ainsi, une relation peut être définie sur des ensembles, des relations et des atomes.

La comparaison des trois modèles est montrée dans le tableau suivant:

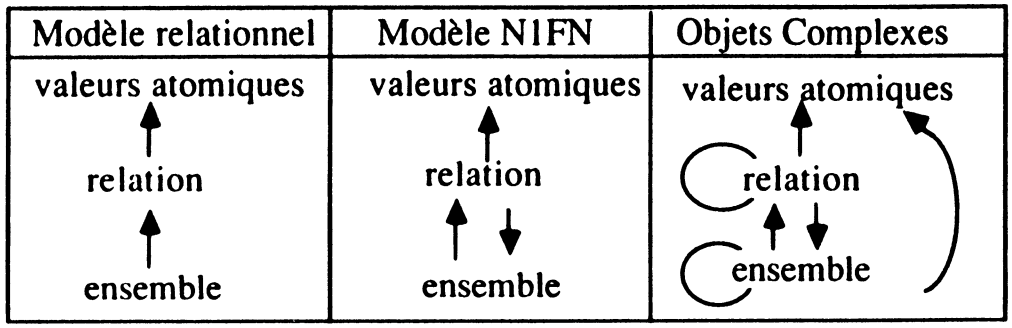


Fig.I.9. *Comparaison des modèles BD*

De multiples calculs sur des structures N1NF sont proposés pour compléter les calculs existant dans le modèle relationnel, tels que les fonctions nest/unnest ([Ozsoyoglu87], filter/pump ([Bancilhon87]), l'extension ([Abiteboul87a]), etc.... Ces fonctions ont des puissances comparables: les unes peuvent être décrites par les autres ([Abiteboul87a]). En bref, ces calculs permettent de transformer une structure bâtie sur un ensemble d'attributs en une autre structure. Nous montrons un exemple simple pour les fonctions nest/unnest ([Ozsoyoglu87]):

Soit une relation N1NF R (Fig.I.10), où l'attribut C est du type ensemble. La fonction NESTc(R) transforme R en R1. La fonction UNNESTc(R1) donne R2.

(R)	A	B	C
	a	b	{c1,c2}
	a	b	{c2,c3}

(R1)	A	B	C
	a	b	c1
	a	b	c2
	a	b	c3

(R2)	A	B	C
	a	b	{c1,c2,c3}

Fig.I.10

Théoriquement, avec le formalisme d'objets complexes tel qu'il est défini dans le tableau ci-dessus, on peut représenter tout type d'objet, tel que le contenu des documents. En réalité, les études sur les objets complexes se font toujours en imposant certaines restrictions pour en faciliter la manipulation et le contrôle ([Abiteboul87a]). Il semble aussi que les calculs proposés s'adaptent plutôt à une situation où une grande quantité des données ont des structures variant relativement peu, qu'à une situation où la variation de structure est très importante, car les manipulations sur la structure sont beaucoup plus coûteuses que celles sur les données. Il est donc difficile

d'adopter actuellement ce formalisme dans les SRI, sachant que les structures des documents d'un SRI sont très variées (notamment pour l'attribut contenu). Malgré cela, nous pensons que les études dans ce domaine sont nécessaires et complémentaires pour l'utilisation des bases de données déductives dans les SRI.

6. CONCLUSION

6.1. Le modèle général

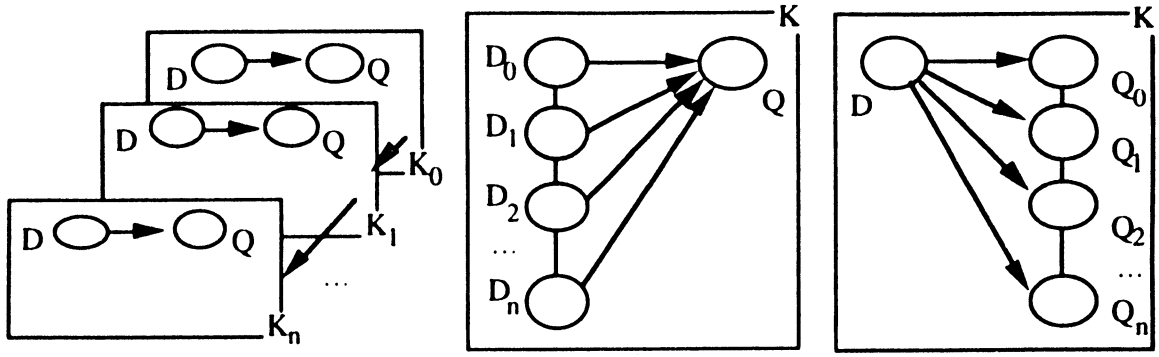
De nombreux modèles existent en recherche d'informations. Les modèles classiques sont très spécifiques. Les modèles plus récents possèdent une certaine généralité. En analysant ces modèles existants et leurs tendances, on voit facilement que les modèles sont de plus en plus généralisés. Cela est dû au fait que la notion de recherche d'informations est de plus en plus sophistiquée. Au départ, la notion de recherche d'informations était fortement liée aux techniques de recherche de données par rapport à une certaine organisation (telle que les fichiers inversés). A l'heure actuelle, cette notion implique beaucoup plus d'aspects, tels que la stratégie de recherche, l'organisation des connaissances, la correspondance des documents et des requêtes, etc.... Au niveau des applications, les SRI sont utilisés dans de plus en plus de domaines, du domaine très spécialisé au domaine bibliothécaire. Cette extension de la notion et des applications montre un besoin fortement marqué de modèles généraux moins dépendants des applications. Notre proposition d'un modèle général répond à ce besoin.

La généralité de notre approche a été démontrée par une représentation des modèles existants dans notre modèle, ce qui permet de conclure que notre modèle recouvre les modèles existants. La généralité de ce modèle est encore confortée par notre approche plus globale des principes d'incertitude.

En effet, avant notre formalisation des principes, la plupart des idées correspondantes étant déjà utilisées dans beaucoup de modèles. Le premier principe (de Van Rijsbergen) est surtout utilisé dans le domaine de la modélisation des connaissances ou de l'apprentissage: pour une requête donnée, l'utilisateur (ou le concepteur) indique les documents pertinents, le système révise ses connaissances pour que ces documents deviennent des réponses de la requête. La technique de révision des probabilités dans le modèle probabiliste peut être considérée comme une application de ce principe, ainsi que les classifications des documents en "clusters" orientés utilisateurs ([Deogun86, Yu85]).

Les idées des deux autres principes sont utilisées implicitement dans beaucoup de modèles fondés sur la logique multi-valuée ou la logique floue ([Bookstein80,Waller79]). En particulier, la combinaison des deux derniers principes permet la description des modèles linguistiques, qui procèdent à des transformations à la fois sur les documents et les requêtes.

Les idées des trois principes d'incertitude sont illustrées dans la figure suivante:



Principe d'incertitude 1 *Principe d'incertitude 2* *Principe d'incertitude 3*

Fig.I.11. *Illustration des principes d'incertitude*

La notion de déduction sémantique (conduisant à la dérivation de monde) joue un rôle essentiel dans notre modèle. Différentes définitions de la déduction sémantique peuvent conduire aux différents modèles existants. Dans cette partie, néanmoins, cette notion reste à un niveau général. En effet, comme on l'a déjà indiqué, la définition des déductions sémantiques dépend fortement de l'application. Nous ne pouvons pas inclure toutes les caractéristiques spécifiques dans ce modèle général. L'application présentée dans la partie II permet de concrétiser cette notion.

6.2. Structure possible des futurs SRI

La comparaison de notre modèle avec l'approche des bases de données laisse entrevoir au niveau théorique de bons moyens pour la réalisation des futurs SRI. Les bases de données déductives et le formalisme d'objet complexe visent à résoudre chacun un problème pour l'intégration des bases de données dans les SRI: les BDD pour leur nouveau mode de raisonnement lors de l'interrogation, le formalisme d'objet complexe pour sa capacité de représentation. On peut donc envisager la structure suivante pour les futurs SRI:

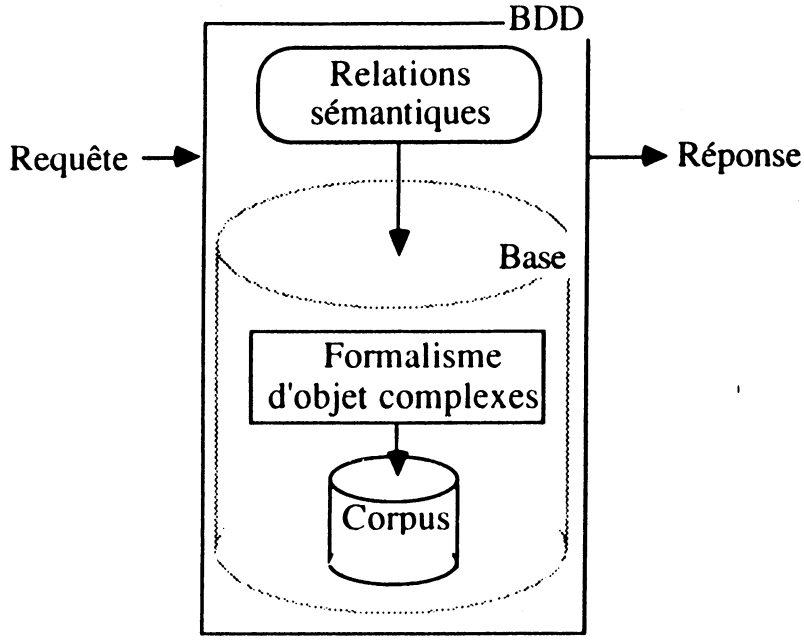


Fig.I.12. *Structure possible des futurs SRI*

PARTIE II

APPLICATION À UN SYSTEME INTELLIGENT DE RECHERCHE D'INFORMATIONS - LE PROTOTYPE RIME

---

1. INTRODUCTION

Dans cette partie, nous appliquons le modèle général que nous avons défini, à un système de recherche d'informations concret - RIME.

RIME est un prototype de Système de Recherche d'Informations appliqué à un contexte médical. Une interrogation est définie par l'ensemble de toutes les opérations nécessaires pour retrouver des documents en l'occurrence - des comptes rendus médicaux (CRM) radiologiques - à partir d'une requête formée par un médecin. Les comptes rendus médicaux et les requêtes sont tous à l'origine écrits en "langue naturelle" (un sous-ensemble particulier du français utilisé dans le milieu médical).

Etant donnés une requête et un ensemble de documents de ce type, l'exécution d'une opération d'interrogation nécessite une représentation intermédiaire (interne) de la requête et des documents. Ainsi, RIME doit être fondé sur une approche linguistique et trois processus sont donc nécessaires:

- l'indexation des documents,
- l'interprétation de la requête,
- l'évaluation de la requête.

Le domaine de RIME est un domaine médical restreint, où les connaissances sémantiques sont très bien formalisées et déterminées. Dans le langage médical, il existe beaucoup moins d'ambiguïtés que dans une langue naturelle prise dans sa totalité. Il est donc possible d'établir une représentation interne correspondant précisément à la *sémantique* d'une phrase dans ce langage pendant l'indexation des documents et l'interprétation de la requête. L'évaluation de la requête peut alors être fondée sur une comparaison sémantique précise.

En conclusion, l'opération d'interrogation de RIME comporte deux volets: un aspect "langue naturelle" et un aspect de modélisation sémantique.

Parmi les trois processus cités, le processus d'indexation des documents a fait l'objet d'une autre étude dans le groupe de recherche ([Berrut88]). Notre travail concerne les deux autres processus.

En ce qui concerne la relation entre RIME et le modèle général décrit dans la Partie I, l'interprétation d'une requête et l'indexation des documents peuvent être considérées respectivement comme certaines étapes dans les

transformations de la requête et des documents. En pratique, ces deux processus sont fortement marqués par des caractéristiques linguistiques.

Comme nous l'avons indiqué dans la partie I, lors de l'application du modèle linguistique, une certaine perte de précision est souvent engendrée au cours de l'interprétation des requêtes et de l'indexation des documents. Il existe donc une certaine différence entre la mesure de correspondance résultant de l'évaluation (entre la requête interprétée et les documents indexés) et la correspondance réelle entre la requête et les documents initiaux. Dans RIME, étant donné que les connaissances du domaine sont bien délimitées et que la langue naturelle traitée est peu ambiguë, la perte de précision peut être réduite au minimum.

Cette partie sera organisée comme suit:

Nous présentons d'abord le point de départ de notre travail, c'est-à-dire principalement le modèle de représentation interne des documents et le processus d'indexation étudié par C.Berrut ([Berrut88, Berrut89]). Après avoir défini le langage d'interrogation et le langage de représentation interne, nous présentons le processus d'interprétation des requêtes et le processus d'évaluation des requêtes proposés pour RIME. Ces deux processus sont réalisés dans un contexte sémantiquement restreint: les connaissances liées à la compréhension des comptes rendus médicaux.

2. MODELE SÉMANTIQUE DE REPRÉSENTATION INTERNE

Le modèle de représentation interne a été défini en coopération avec les spécialistes compétents dans le domaine de la radiologie ([Chiaramella87, Berrut88, Berrut89]). Il est issu d'un compromis entre les besoins ultérieurs des utilisateurs et les possibilités de traitement envisageables. La définition du modèle est inspirée de l'approche de Schank ([Schank80b, Schank81, Schank82], cf. Annexe 1).

2.1. Principe

• Notion de concept et de classe sémantique

Les documents dans RIME sont des comptes rendus médicaux, qui sont généralement constitués de trois parties, chacune concernant respectivement les *examens*, les *constats*, et éventuellement les *diagnostics* (cf.[Berrut88, Berrut89]). Ces parties sont constituées de descriptions d'éléments tels que des *signes*, des *lésions*, ..., eux-mêmes constitués d'autres éléments (*localisations*, ...). Une hiérarchie existe donc dans la définition du contenu sémantique d'un CRM:

- *examen* *constat* *diagnostic*
- *signe* *lésion* ...
- *localisation* ...
- ...

Un terme médical situé à un niveau quelconque dans cette hiérarchie est considéré comme un *concept* qui représente un certain phénomène médical. Par exemple, le terme "opacité" est un concept situé au niveau "signe", qui représente le phénomène "augmentation de densité".

Tous les concepts de même nature sont regroupés dans une même *classe sémantique* (par exemple, "opacité", "hypertrophie", ... sont regroupés dans la classe *signe* (SGN)). Les concepts d'une même classe sémantique possèdent donc les mêmes caractéristiques.

Théoriquement, un concept d'une classe sémantique peut toujours être considéré comme étant constitué par d'autres concepts plus fins reliés entre eux par une certaine relation sémantique ([Berrut88]). Par exemple, le concept "hypertrophie" peut être considéré comme une "augmentation" portant sur le "volume", les termes "augmentation" et "volume" étant des concepts plus fins et "portant sur" étant la relation entre eux. La décomposition de "augmentation" et de "volume" peut éventuellement se poursuivre ...

En pratique, il est donc nécessaire de définir une limite pour mettre fin à cette décomposition. Un concept plus fin ne doit être considéré que s'il est utile pour les traitements ultérieurs. Ainsi, cette limite doit être définie selon les besoins de l'application. Dans RIME, la limite est établie par la définition de certains concepts comme *concepts élémentaires* qui ne sont jamais décomposés dans cette application (ex: "augmenté", "volume").

A l'inverse de la notion de concept élémentaire, les concepts qui sont composés à partir des concepts élémentaires mis en relation sémantique, ou qui peuvent être décomposés en ceux-ci, sont appelés des *concepts complexes.* Par exemple, "opacité" est un concept complexe, car il peut être décomposé en "densité" et "augmenté" mis en relation par "a pour valeur": [a-pr-val](densité,augmenté).

• Contraintes sur la formation de concepts complexes

Liée à la notion de classe sémantique, la formation d'un concept complexe à partir d'autres concepts (complexes ou élémentaires) peut être réalisée de deux manières:

- un concept d'une certaine classe sémantique peut être décrit par un autre concept pour former un concept plus complexe de la même classe,

- deux concepts peuvent être mis en relation pour former un concept plus complexe d'une classe différente.

Dans le domaine médical, aucune de ces formations de concepts n'est faite d'une manière libre: elles doivent suivre certaines règles.

A l'aide du premier type de formation, par exemple, un signe (*opacité*) peut être décrit par une "localisation" ou par une "valeur qualitative", etc.... Si l'on représente la relation entre un signe et une localisation par un opérateur [p-sur] (porte sur), et la relation entre un signe et une valeur qualitative par l'opérateur [a-pr-val] (a pour valeur), on peut donc décrire les règles de formation d'un *signe* comme suit (<– désigne "formé par"):

signe <– [p-sur] (signe, localisation)
<– [a-pr-val] (signe, valeur qualitative)

Par exemple: [p-sur](opacité,poumon)
et [a-pr-val](opacité, tissulaire)

Grâce au second type de formation, un signe peut être formé par un "caractère physique" (ex: "densité") décrit par une "valeur qualitative impliquant un signe" (ex: "augmenté"), ce qui correspond à:

signe <– [a-pr-val] (caractère physique, valeur qualitative signe)

De cette manière, toute formation d'un concept complexe peut être décrite par deux sortes de règles comme ci-dessous:

<classe1> ::= [<opérateur>] (<classe1>, <classe2>)
<classe3> ::= [<opérateur>] (<classe1>, <classe2>)

ce qui signifie que deux concepts de <classe1> et de <classe2> peuvent être mis en relation, cette relation sémantique étant représentée par [<opérateur>] pour former un concept de <classe1> ou de <classe3>.

L'ensemble de ces règles constitue les *contraintes sémantiques* sur la formation de concepts complexes. Ces contraintes délimitent aussi toutes les représentations possibles des concepts dans RIME. Ainsi, la définition des contraintes est aussi une définition du *modèle de représentation* des concepts dans RIME (appelé ultérieurement *modèle sémantique*).

• **Notion de dépendance dans le modèle**

On peut remarquer dans les précédents exemples, que les opérateurs entre deux concepts impliquent une *dépendance* entre ceux-ci: étant donnée la formation [opérateur](concept1,concept2), le second concept décrit généralement (via l'opérateur) une localisation, une valeur qualitative, un contexte, ... attaché(e) au premier concept (cf. la définition dans 2.2). Le second concept (avec l'opérateur) constitue une *description* du premier concept. Ainsi, on peut considérer dans ces formations, que le premier concept est un *concept principal* et le second concept un concept *dépendant* de celui-ci.

Cette dépendance est particulièrement marquée dans le premier type de formation d'un concept, où la description d'un concept principal forme un concept de la même classe.

Par exemple, dans "[p-sur](opacité,poumon)", le signe "opacité" est décrit par la localisation "poumon", formant un signe plus détaillé "opacité pulmonaire".

Dans ce premier type de formation, on peut considérer que le concept principal ("opacité") représente une sémantique (un phénomène médical) moins détaillée que toute la description ("opacité pulmonaire"). Le concept principal est donc un *concept gouverneur*.

Cette dépendance est tout à fait comparable à celle du modèle de Schank (cf. Annexe 1). Dans l'approche de celui-ci, les propriétés des éléments sont tout d'abord décrites séparément par des *scripts*. Une phrase est interprétée comme une mise en *dépendance* des différents éléments grâce à des *primitives*. La notion de dépendance est impliquée par des primitives comparables aux opérateurs de RIME.

Dans l'approche de Schank, les contraintes sur la mise en dépendance des éléments sont représentées par une sorte de méta-connaissance spécifiant par exemple, que "l'agent d'une action ne peut être qu'un être animé". Les contraintes que l'on a définies dans RIME sur les formations de concepts jouent tout à fait le même rôle.

2.2. Définition

Les classes suivantes sont définies dans RIME (cf. Annexe 2 pour une description complète):

Classes sémantiques:

- caractère physique (CAR-PHY)
- constat (CONSTAT)
- diagnostic (DIAG)
- examen (EX)
- fonction (FCT)
- lésion (LESION)
- localisation (LOC)
  - constituant d'organisme (CONST-ORG)
    - élément de structure (ELT-STRUCT)
    - organe (ORG)
    - région (REGION)
    - détail (DETAIL)
- position (POS)
- signe (SGN)
- valeur qualitative (VAL-QUAL)
- valeur qualitative impliquant un signe (VAL-SGN)
- valeur quantitative (VAL-QUAN)
- ...

Remarquons qu'une classe sémantique peut en inclure une autre. Par exemple, la classe CONST-ORG inclut ELT-STRUCT, ORG, REGION et DETAIL. Un concept d'une de ces quatre dernières classes appartient donc aussi à la classe CONST-ORG.

Les opérateurs suivants ont été isolés comme primitives dans le modèle:

Opérateurs sémantiques:

Nature	Notation	Signification
booléen	$\vee, \wedge, \neg$	
opérateur de certitude	[prob]	(dont la certitude est [prob] [prob]$\in$ {[p],[pn]})
opérateurs sémantiques:		
causalité	[dû-à]	
déduction	[per-de-déd]	(permet de déduire)
démonstration	[montré-par]	
localisation	[p-sur]	(porte sur)
topologique	[en-rel-topo-avec]	(en relation topologique avec)
évaluation	[a-pr-val]	(a pour valeur)
position précise	[a-pr-val-loc]	(a pour valeur locative)

NB: [p] et [pn] denotent une description "probable" et "probablement-fausse", les deux valeurs de certitude possibles dans RIME à part la certitude et l'incertitude totales.

Etant donné que les opérateurs dans RIME sont binaires ou unaires, tout concept complexe correspond donc à une arborescence binaire (non stricte), dont les feuilles sont des concepts élémentaires et les noeuds des opérateurs. Par la suite, on présente, de façon descendante, la construction de quelques concepts.

Les concepts les plus complexes sont appellés *comptes rendus conceptuels* (CRC). Un CRC peut être construit de la manière suivante:

CRC ::= ∨(CRC, CRC) | ∧(CRC, CRC) | ¬CRC
| [prob] CRC
| CONSTAT
| DIAG
| [per-de-déd](CONSTAT,DIAG)

Les formations possibles d'un constat (CONSTAT) sont les suivantes:

CONSTAT ::= ∨(CONSTAT, CONSTAT)
| ∧(CONSTAT, CONSTAT)
| ¬ CONSTAT
| [prob] CONSTAT
| [dû-à] (CONSTAT, DIAG)
| [dû-à] (CONSTAT, CONSTAT)
| [en-rel-topo-avec] (CONSTAT, CONSTAT)
| [montré-par] (SGN, EX)
| SGN

Voici quelques exemples de CONSTAT:

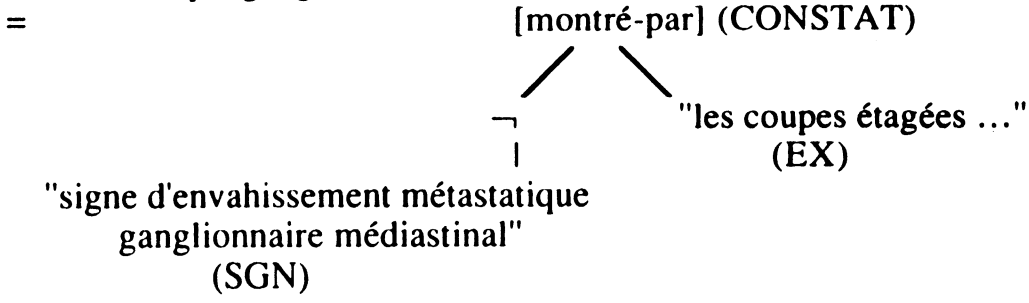
1. le bord antérieur de la vertèbre rongé par la tumeur
= [dû-à] (CONSTAT)

"rongement du bord antérieur de la vertèbre" (CONSTAT) — "la tumeur" (DIAG)

2. hypertrophie de densité tissulaire avec extension médiane dans l'orifice cervico-thoracique
= [en-rel-topo-avec] (CONSTAT)

"hypertrophie de densité tissulaire" (CONSTAT) — "extension médiane dans l'orifice cervico-thoracique" (CONSTAT)

3. les coupes étagées ... ne démontrent pas de signe d'envahissement métastatique ganglionnaire médiastinal



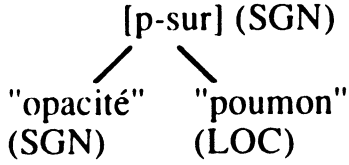
Un signe est défini de façon suivante:

SGN ::= ∧(SGN, SGN) | ∨(SGN, SGN) | ¬ SGN
| [prob] SGN
| [a-pr-val] (SGN, QUAL)
| [a-pr-val] (SGN, VAL-SGN)
| [p-sur] (SGN, LOC)
| [p-sur] (SGN, FCT)
| [en-rel-topo-avec] (SGN, LOC)
| SGN-ELE
| [a-pr-val] (CAR-PHY,VAL-SGN)
| [a-pr-val] (QUAL, VAL-SGN)

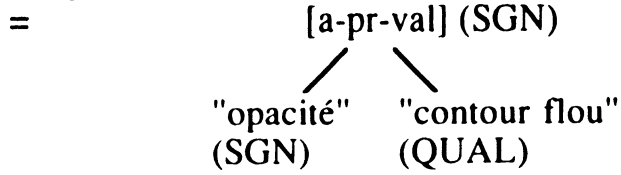
SGN-ELE ::= un concept élémentaire représentant un signe

Exemples:

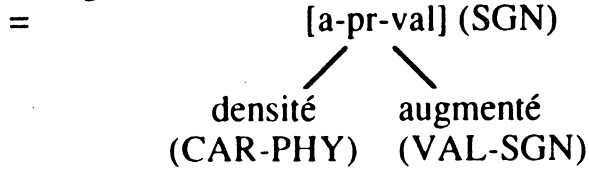
1. opacité pulmonaire



2. opacité à contour flou



3. augmentation de densité



Une description complète du modèle sémantique est donnée dans l'Annexe 3.

2.3. Caractéristiques

Les caractéristiques suivantes sont mentionnées dans [Berrut89]:

1. Les règles du modèle sont du type *hors contexte*: la partie gauche d'une règle contient un méta-symbole unique (CRC). Ainsi, la formation d'un concept n'est pas fonction de son contexte.

2. Les règles ne sont pas ambiguës: la structure d'un concept complexe ne peut pas correspondre à deux règles simultanément.

De plus, nous avons remarqué deux caractéristiques supplémentaires:

3. Comme on l'a montré précédemment, les opérateurs impliquent une dépendance entre les concepts connectés. En particulier, dans le cas où un premier concept est décrit (via un opérateur) par un second pour former un concept de la même classe, celui-ci est considéré comme *gouverneur*, et le second est considéré comme *dépendant*.

4. Le modèle sémantique a défini une hiérarchie entre les classes sémantiques.

Définissons la notion de *supériorité* comme suit:

> Si un premier concept est formé à partir d'un deuxième concept, la classe du premier est dite *supérieure* à celle du deuxième.

Par exemple, "*densité tissulaire*" est un cas où un caractère physique (CAR-PHY) - "densité" et une valeur qualitative (VAL-QUAL) - "*tissulaire*" forment une qualification (QUAL). Ainsi, nous considérons que la classe QUAL est supérieure à CAR-PHY et à VAL-QUAL.

On a donc:

CRC > CONSTAT
CONSTAT > SGN
DIAG > LESION
...

Si l'on trace toutes les formations décrites dans le modèle sémantique, on obtient le schéma de la figure II.1, où une flèche part d'une classe supérieure vers les classes inférieures.

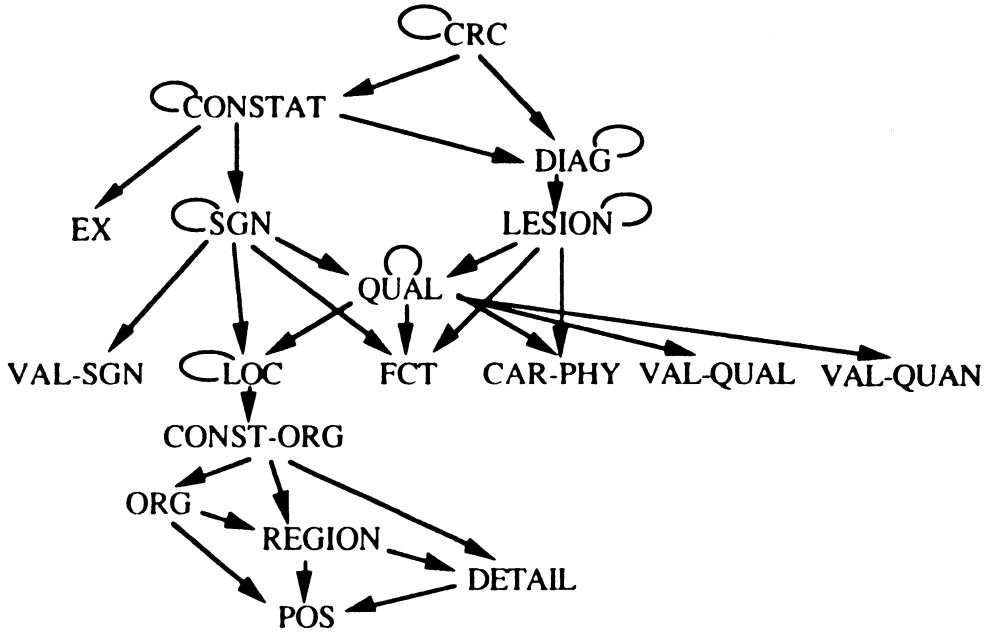


Fig.II.1. *Hiérarchie des classes sémantiques*

Remarquons qu'il n'existe pas de boucle entre deux classes différentes. Cela implique que, soit deux classes sémantiques sont indépendantes, soit l'une est supérieure à l'autre.

3. PRÉSENTATION DE L'INDEXATION DES DOCUMENTS

3.1. Documents

Les documents dans RIME (les comptes rendus radiologiques) sont rédigés par des spécialistes. Un compte rendu médical (CRM) est formé d'un ensemble de spécifications d'attributs, tels que le nom de patient, le nom de docteur, la date d'examen, ... et la description de l'observation médicale de l'examen. La description médicale est appelée *attribut interne* ou *attribut contenu*, les autres attributs étant des *attributs externes*. On donne un exemple de compte rendu dans Fig.II.2 (cf.Annexe 4 pour d'autres exemples).

Les CRM présentent les caractéristiques suivantes:

- Ils sont généralement très courts (une page)
- Ils comportent des termes très spécifiques au domaine
- Les phrases ont une syntaxe particulière mais très simple: elles sont souvent réduites à des groupes nominaux.

3.2. Processus d'indexation

L'objectif de l'indexation est de représenter l'attribut interne des CRM dans le modèle de représentation interne que l'on vient de définir. Schématiquement, l'indexation d'un CRM (cf. [Berrut88] pour les détails) est exécutée en trois phases, accomplissant les *tâches* suivantes:

Monsieur Durand le 8.12.1986

TOMEDENSITOMETRIE
(homme de 70 ans - découverte d'une opacité arrondie du lobe supérieur droit avec déformation du médiastin)

Les constatations sont les suivantes:
- opacité à contour bosselé, de densité tissulaire, apparemment homogène, en projection du segment apical du lobe supérieur droit.
- contact étroit avec le versant médiastinal.
- opacité ganglionnaire de plus de 15 mm de diamètre en projection de la loge de Baréty intéressant les derniers relais de la chaîne para-trachéale droite et allant jusqu'au niveau de la chaîne sus pulmonaire, c'est-à-dire dans la loge pré-carénale.

En conclusion:
Aspect TDM en faveur d'un cancer de siège périphérique avec extension ganglionnaire médiastinale et pédiculaire et probablement T3 pleural (plève médiastinale).

Docteur: Dupont

Fig.II.2. *Un exemple de CRM*

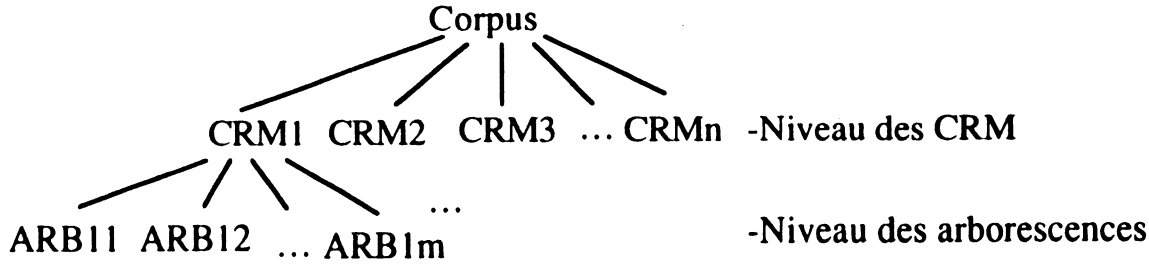
- *tâches infra-structurelles*: elles permettent l'identification morphologique des mots, et l'interprétation des mots isolés dans leur représentation interne.
- *tâches intra-structurelles*: elles permettent de regrouper les mots isolés d'une phrase selon le modèle de représentation interne pour former un concept complexe.
- *tâches inter-structurelles*: elles reconnaissent les liens implicites entre différentes phrases, et les représentent explicitement dans la représentation interne.

A la sortie de l'indexation, un ensemble d'arborescences binaires (ou une forêt) est établi pour chaque CRM: les feuilles sont des concepts élémentaires et les noeuds des opérateurs. La relation implicite entre les différentes arborescences est une conjonction ($\wedge$).

Le corpus est organisé à deux niveaux:

Au premier niveau, il est constitué d'un ensemble de CRM dont chacun est identifié par un numéro de CRM (NO-CRM) et l'ensemble des attributs externes.

Au second niveau, un CRM est constitué d'un ensemble d'arborescences dont chacune est identifiée par un numéro d'arborescence (NO-ARB).



Notons une arborescence issue de l'indexation par ARB_{ij} (avec i - le numéro du CRM, et j- le numéro de l'arborescence dans le CRM), l'attribut interne d'un CRM (D_i) peut donc être représenté après l'indexation par:

$$D_i = \{ARB_{i1}, ARB_{i2}, ARB_{i3},\ldots,ARB_{im}\}$$

ou exprimé sous forme de relation:

$$ARB_{i1} \in D_i$$
$$ARB_{i2} \in D_i$$
$$\ldots$$
$$ARB_{im} \in D_i$$

Le corpus peut donc être considéré comme C_{CRM} ou C_{ARB}:

$$C_{CRM} = \{D_1, D_2, D_3, \ldots, D_n\}$$
$$C_{ARB} = \{ARB_{11},ARB_{12},\ldots,ARB_{21},ARB_{22},\ldots,ARB_{n1},ARB_{n2},\ldots\}$$

4. INTERPRÉTATION DE LA REQUETE

4.1. Définition du langage d'interrogation

Compte tenu de la complexité d'une langue naturelle et de ses ambiguïtés, une interrogation en langue naturelle ne peut pas être définie sans simplification. La définition d'un langage "quasi-naturel" s'impose. Pour ne pas perdre en généralité, la définition du langage d'interrogation doit cependant tenir compte de la structure des requêtes que les utilisateurs peuvent

poser dans un contexte donné. Ainsi, nous analysons d'abord quelques aspects généraux des requêtes.

En général, un document peut être référencé par un ensemble de spécifications d'attributs, tels que dans RIME, le nom de patient, le nom de médecin (l'auteur de document), la date d'examen, ... la description médicale. Une requête pouvant être généralement considérée comme un ensemble de spécifications sur les attributs des documents que l'utilisateur souhaite obtenir, elle peut être décomposée de façon suivante:

REQUETE ::= spécification-d'attribut1 spécification-d'attribut2 ...
(où il existe une relation '$\wedge$' implicite entre les spécifications)

ou REQUETE ::= [spécification d'attribut]$^+$
où [...]$^+$ représente une répétition au moins une fois.

Voici un exemple typique de requête soumise à RIME:

Je voudrais les comptes rendus établis par Mr. Dupont depuis juillet 1985 sur le cancer du poumon.

Nous décrivons ci-dessous la syntaxe du langage d'interrogation, qui constitue un sous-ensemble du français permettant d'exprimer des requêtes de ce type.

On peut dégager de cette requête les spécifications d'attributs suivantes:

type de document - *compte rendu* (le seul type dans RIME)
auteur de document - *Dupont*
date - *depuis juillet 1985*
contenu - *le cancer du poumon*

Les trois premiers types d'attributs sont appelés *attributs externes* du document, par opposition au dernier qui est appelé *attribut interne*, ou *de contenu*. La première classe décrit des propriétés non liées au contenu sémantique. La seconde décrit un thème de recherche, faisant référence au contenu sémantique des documents.

A part les composants principaux énoncés ci-dessus, nous pouvons aussi remarquer d'autres composants présents dans la requête, tels que *Je voudrais, établis par, sur, ...* Ces composants servent à former une requête complète (*je voudrais*) ou à spécifier les attributs de sélection (*établis par, sur,...*). D'un autre point de vue, on peut aussi considérer ces composants comme *marqueurs* ou *introducteurs* des spécifications d'attribut. Par exemple, *je voudrais* introduit la requête complète, *établis par* introduit l'attribut auteur

de document, *sur* introduit l'attribut contenu. Ceci nous conduit à établir l'hypothèse suivante:

Hypothèse:
Une requête ou une spécification d'attribut de la requête est introduite par un *introducteur.*

En généralisant cette structure, on obtient la syntaxe suivante pour des requêtes:

REQUETE ::= INTRO-REQ REQ
REQ ::= ATT1 [INTRO-ATT ATT]*

où [...]* dénote une répétition éventuellement vide.

a). Distinction entre attributs externes et internes

La syntaxe précédente exprime qu'une requête est composée d'un introducteur de requête (INTRO-REQ), de l'attribut que l'utilisateur souhaite en réponse (ATT1) et d'un ensemble d'attributs (ATT) précédés d'un introducteur (INTRO-ATT). Si l'on sépare les attributs externes (EXT) de l'attribut interne (INT), on obtient:

REQ ::= EXT [INTRO-EXT EXT]* [INTRO-INT INT]

où [...] dénote une option.

b). Définition des attributs externes

Il est à remarquer que les introducteurs ne sont pas des éléments très variés. Par exemple, l'introducteur de l'attribut auteur appartient généralement à {*rédigé par, fait par, établi par, dont le docteur est, ...*}. Il est donc possible de limiter les valeurs possibles d'un introducteur dans un ensemble relativement restreint. On donne ci-dessous des ensembles possibles de valeurs définis pour certains introducteurs:

INTRO-REQ∈ {*existe-t-il, donnez moi, je voudrais, quel(s) est(sont), ...*}
INTRO-AUTEUR∈ {*rédigé par, fait par, établi par, dont le docteur est, ...*}
INTRO-INT∈ {*concernant, portant sur, à propos de, décrivant, qui montre ...*}

Les attributs externes ont une syntaxe relativement simple. Ils sont formés des valeurs atomiques de l'attribut, éventuellement connectées par des conjonctions (CONJ). Dans le cadre de notre étude, on n'accepte que deux conjonctions "et" et "ou" impliquant chacune un opérateur booléen:

EXT ::= EXT-ATOM [[,] CONJ EXT-ATOM]*

où EXT-ATOM est une valeur atomique de l'attribut externe correspondant.

Exemples: AUTEUR: Dupont, Laurent, ou Durand
DATE: après 1978 et avant 1987

c). Définition des attribut internes

En ce qui concerne l'attribut interne - le contenu, pour notre étude, sa syntaxe est définie de la façon suivante:

INT ::= GN [PRO-REL]

c'est-à-dire que l'attribut interne est formé d'un groupe nominal, suivi éventuellement d'une proposition relative. La syntaxe du groupe nominal et de la proposition relative, bien que simplifiée par rapport à la langue naturelle, présente une certaine complexité, permettant un niveau d'expression suffisant pour l'application:

GN ::= GNC [[ADV] EXP GNC]* [CONJ GN]
GNC ::= ART GNS [[ADV] PREP GNC]* [CONJ GNC]
GNS ::= NOM [[ADV] ADJ]* [[ADV] PREP GNS]* [CONJ GNS]
PRO-REL ::= PRO-REL1 [CONJ PRO-REL1]*
PRO-REL1 ::=*qui* VERBE [ADV] [GN]
| *que* GN VERBE [ADV] [PREP GN]
| PPASS [ADV] PREP GN
| PPRES [ADV] GN

où: GNC est un groupe nominal complet (un groupe nominal ayant un article au début, appelé groupe N3 en terme linguistique),
GNS est un groupe nominal simple (sans article au début),
EXPRE est une expression (locution) prépositionnelle (ex: au niveau de, en raison de), prédicative (dû à) ou adjective (secondaire à),
PREP une préposition,
CONJ une conjonction (et, ou),
VERBE un verbe conjugué,
PPRES est un participe présent de verbe (ex: entraînant),
PPASS est un participe passé de verbe (ex: entraîné),
ADV est un adverbe impliquant une certitude (ex. probablement).

Voici quelques exemples de tels groupes:

GNC: une augmentation de volume du foie

GNS: augmentation de volume du foie
PRO-REL: qui augmente le volume du foie

Les pronoms relatifs incorporés dans la définition sont *qui* et *que* pour l'instant. Nous pensons qu'une fois que l'on peut traiter ces deux pronoms relatifs, beaucoup d'autres (dont la portée est un groupe nominal) peuvent être traités de façon similaire. Par exemple, *dont* peut être considéré, à quelques détails près, comme *de qui*; *auquel* peut être considéré comme *à qui* (le groupe nominal référé devant être masculin singulier) ... Les deux pronoms relatifs considérés sont donc, en quelque sorte, assez représentatifs de l'ensemble des pronoms relatifs.

On a défini un langage d'interrogation qui permet un assez grand pouvoir d'expression. Ce langage peut être facilement étendu par la suite. Par exemple, on peut étendre les ensembles d'introducteurs, ou rajouter un attribut externe, ... Il est montré ([Nie87]) que la structure d'une requête ainsi définie est adaptée pour l'interrogation des SRI en général.

Pour terminer la définition du langage d'interrogation, nous donnons deux exemples complets typiques de requêtes que le langage défini permet d'exprimer:

1. *Je voudrais les comptes rendus spécifiant des kystes bronchogéniques qui siègent dans le médiastin postérieur et qui sont calcifiés.*

2. *Quel est le docteur qui a effectué l'examen de Mr. Dupont le 21 décembre 1987?*

Dans la première requête, l'attribut que l'utilisateur souhaite en réponse est un ensemble de "comptes rendus". La spécification porte sur l'attribut interne: "*des kystes bronchogéniques qui siègent dans le médiastin postérieur et qui sont calcifiés*". Cette requête est typique des SRI, qui demande une comparaison de la requête avec le contenu des documents (ce qui nécessite un mécanisme propre aux SRI), et dont la réponse est la totalité du document.

La seconde requête porte sur un ensemble d'attributs externes. La réponse n'est pas le document complet comme pour la première requête, mais seulement la valeur d'un attribut (docteur). Cette requête est comparable avec celles des bases de données. En effet, les attributs externes d'un CRM peuvent être organisés comme les attributs des relations dans les bases de données. Leur évaluation peut être donc fondée avec les opérations des ces dernières (select).

A ces deux types de requête, il s'ajoute naturellement un troisième type comportant à la fois des spécifications sur les attributs externes et sur

l'attribut interne. L'évaluation d'une telle requête comporte deux parties: l'une concerne les attributs externes et utilise des outils bases de données, l'autre porte sur l'attribut interne et utilise des outils propres aux SRI.

Cela montre que la frontière entre les SRI et les systèmes de bases de données n'est pas une séparation totale. L'utilisation des bases de données dans RIME sera décrite dans la partie d'évaluation des requêtes (cf.II.6).

4.2. Définition du langage de représentation interne des requêtes

Dans la section 2, nous avons défini une représentation arborescente binaire (le modèle sémantique) pour les attributs internes (contenus). Pour rester cohérent, nous définissons ici une représentation arborescente (non binaire) pour les requêtes.

Comme il a été indiqué précédemment, une requête peut être décomposée en deux ensembles d'attributs (ou éventuellement un seul des deux): les attributs externes et l'attribut interne. Ainsi, une requête est définie comme suit:

REQUETE ::= ∧(EXTERNES, INTERNE)
| EXTERNES
| INTERNE

EXTERNES est composé d'une suite (au moins un, et sans répétition d'un même attribut) d'attributs externes, tels que DATE, AUTEUR, PATIENT , ... Sa représentation interne est défini comme suit:

EXTERNES ::= ATTEXT
| ∧(ATTEXT [, ATTEXT]$^{+}$)

(NB: Ici, l'opérateur ∧ n'est plus nécessairement binaire)

ATTEXT ::= =(date, DATE)
| ⊇(auteur, AUTEUR)
| =(patient, PATIENT)

DATE ::= DATE1 | [op](DATE, DATE)

(où [op] est un opérateur booléen)

DATE1 ::= une valeur atomique de date de la forme: aammjj, où aa représente l'année, mm le mois, et jj le jour (ex:890712)
| [op-comp](aammjj)
(où [op-comp] ∈ {=, >, <, ≥, ≤})

AUTEUR ::= un nom d'auteur (ex: Dupont)
| [op](AUTEUR, AUTEUR)

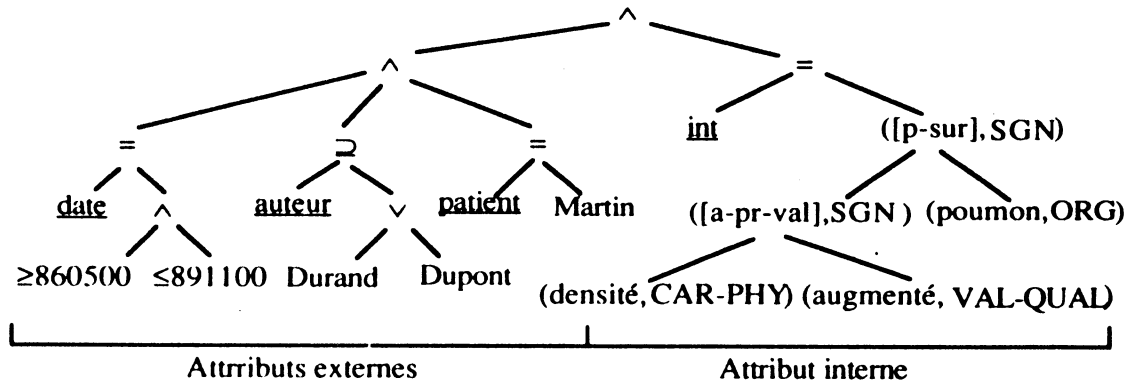
PATIENT ::= un nom de patient (ex: Martin)
| [op](PATIENT,PATIENT)

NB: Il existe une relation implicite entre les attributs externes. Elle est équivalente à un "et" logique n-aire.

L'attribut interne est représenté dans le modèle sémantique défini dans la section 2, c'est-à-dire en une arborescence binaire dont les noeuds sont des opérateurs et les feuilles des concepts élémentaires. A chaque composant de l'arborescence (concept élémentaire, sous-arborescence et arborescence elle-même) est associée sa classe sémantique.

INTERNE ::= arborescence défini par le modèle sémantique

En conclusion, une requête sous forme interne est une arborescence dont les noeuds sont des opérateurs de recherche et les feuilles des attributs de recherche:



4.3. Connaissances de base - Le dictionnaire

4.3.1. Généralités

Pour permettre l'analyse d'une requête, un *dictionnaire* est nécessaire pour fournir les informations de base concernant les mots ou les groupes de mots. Les informations nécessaires se situent aux niveaux syntaxique et sémantique. Pour un terme donné, les informations considérées peuvent être représentées de la façon suivante (les variables grammaticales sont représentées en italique):

- Pour les noms et les adjectifs:
  terme(*cat-gram, genre, nombre*, interprétation, classe-sém)

- Pour les prépositions et les groupes prépositionnels:
  terme(*cat-gram*, interprétation)

- Pour les verbes:
  terme(*cat-gram*, interprétation, classe-sém)

- Pour les articles:
  terme(*cat-gram, genre, nombre*)

- pour les participes:
  terme(*cat-gram, genre, nombre*, interprétation, classe-sém)

Au niveau syntaxique, le dictionnaire associe une catégorie syntaxique à chaque entrée, telle que "nom", "adjectif", "préposition", etc... et éventuellement d'autres variables grammaticales (genre, nombre).

Au niveau sémantique, les informations pour un terme donné sont constituées d'une interprétation et éventuellement de la classe sémantique de cette interprétation si elle est déterminée.

L'interprétation d'un terme est la représentation interne du terme. C'est une expression du modèle sémantique. La classe sémantique associée est un méta-symbole du modèle sémantique. Si un terme est interprété par une structure, cette structure doit vérifier les contraintes du modèle sémantique.

A l'intérieur d'une structure, aux composants interprétés en leur forme finale est associée leur classe sémantique.

Exemples:

entrée	**(cat-gram, genre, nombre, interprét., classe)**
bronche souche	(*nom, fém, sing*, bronche-souche, CONST-ORG)
bronchique	(*adj, mas-fém, sing*, bronche, CONST-ORG)
opacité	(*nom, fém, sing*, [a-pr-val]((densité,CAR-PHY), (augmenté,VAL-QUAL)), SGN)
	(cat-gram, interprétation)
de	(*prép*, [p-sur])
de	(*prép*, [a-pr-val])
de	(*prép*, [a-pr-val-loc])
	(cat-gram, interprétation)
dégager	(*verbe*, [per-de-déd])

entrée	**(cat-gram, interprétation,classe)**
dévier	(*verbe*, [dû-à]([p-sur]((déviation,SGN),LOC),LESION))
	(cat-gram, genre, nombre)
le	(*art, mas, sing*)
le	(*pp, mas, sing*)
	(cat-gram,genre,nombre,interprétation,classe)
rétracté	(*ppass, mas, sing*, rétracté, VAL-SGN)

4.3.2. Caractéristiques du dictionnaire

- **Interprétation non-terminale**

Une interprétation *non-terminale* désigne une représentation interne où tous les éléments ne sont pas substitués. Aucune classe sémantique n'est donc associée à cette interprétation. Par exemple, l'interprétation du verbe "dévier":

[dû-à]([p-sur]((déviation,SGN),LOC),LESION)

laisse deux éléments qui devront être substitués ultérieurement (représentés par les classes "LOC" et "LESION"). Cela signifie que le terme "dévier" a l'interprétation sémantique invoquée si ses opérandes (complément direct, sujet) appartiennent respectivement aux classes LOC et LESION. Dans le cas où la classe d'un concept non-substitué n'est pas déterminable, on représente le concept dans la structure par '0'. Par exemple, une interprétation de préposition "sans" est "[en-rel-topo-avec](0,¬(0))".

Une définition non terminale correspond donc à la spécification de contraintes sémantiques au niveau du contexte d'utilisation d'un mot dans le document. Elle doit être naturellement conforme au modèle sémantique (cohérence entre l'interprétation et sa classe sémantique notamment).

- **Mots vides**

Un mot vide est un mot qui n'apporte aucune sémantique intéressante pour la phrase. Son interprétation est vide. Par exemple, "image", "allure", "apparition", "très", ... sont des mots vides.

- **Interprétations multiples (polysémie)**

Certains termes possèdent plusieurs interprétations possibles. Pendant l'interprétation des requêtes, l'une d'entre elles doit être choisie pour aboutir

à une interprétation complète des requêtes. Dans le dictionnaire, on regroupe ces interprétations dans une liste, de sorte qu'un terme ne corresponde qu'à une seule entrée du dictionnaire.

Par exemple, la préposition "de" est représentée dans le dictionnaire comme suit:

de(*prép*, ([p-sur], [a-pr-val], [a-pr-val-loc]))

Un terme peut posséder dans certains cas, une signification implicite en plus de la signification explicite. Par exemple, l'adjectif "tissulaire" a une signification explicite qui est une valeur qualitative "tissulaire". Dans certains cas, cet adjectif peut aussi signifier "densité tissulaire". Ainsi, dans le dictionnaire, ces deux interprétations sont incluses dans le dictionnaire:

tissulaire(*adj, mas-fém, sing*, ((tissulaire, VAL-QUAL),
([a-pr-val]((densité,CAR-PHY),
(tissulaire,VAL-QUAL)),QUAL)))

4.4. Schéma de l'interprétation

L'objectif de l'interprétation d'une requête est de ramener la requête du langage d'interrogation (langage source, cf.4.1) dans le langage de représentation interne (langage cible, cf.4.2).

L'interprétation est effectuée de façon descendante: la requête est d'abord séparée en un ensemble de spécifications d'attributs. Les spécifications des attributs externes et la spécification de l'attribut interne sont ensuite interprétées respectivement par un processus approprié.

L'interprétation d'une requête peut donc être décomposée en trois phases:

- reconnaissance des spécifications des différents attributs dans la requête globale (séparation des attributs)
- interprétation des spécifications des attributs externes
- interprétation de la spécification de l'attribut interne

4.5. Interprétation globale

Une requête est composée d'un ensemble d'*introducteurs d'attribut* suivis chacun d'une *valeur d'attribut*. Sachant que les introducteurs ne sont pas très variés (même dans un contexte libre), il est donc possible de définir un ensemble fini de valeurs pour chaque introducteur. L'identification des attributs consiste alors à reconnaître ces introducteurs dans la requête. Cette

identification nécessite parfois une résolution des ambiguïtés, car un même introducteur peut parfois introduire plusieurs attributs différents.

Par exemple, "de" peut être introducteur de date (l'examen du 10 novembre 1987), d'auteur (les observations du Docteur Durand), ... et de l'attribut interne (l'examen du foie). Dans ce cas, on peut effectuer, par exemple, l'analyse suivante pour déterminer sa fonction:

1. Si "de" est suivi d'un nom de docteur (précédé par "docteur", "Dr." ...), c'est un introducteur de l'auteur.

2. Si "de" est suivi d'un nom de patient (précédé par "monsieur", "Mr.", ...), c'est un introducteur du nom de patient.

3. Si "de" est suivi d'une date, c'est un introducteur de la date d'examen.

4. Si "de" est suivi d'une description médicale, c'est un introducteur de l'attribut interne.

4.6. Interprétation des attributs externes

Les attributs externes des requêtes ont une syntaxe assez simple, qui est un ensemble de valeurs atomiques éventuellement connectées par des conjonctions (et, ou).

Quand la spécification d'un attribut externe ne correspond qu'à une valeur atomique de l'attribut (ex. DATE: en 1987), la transformation sous une forme interne peut être faite par un automate spécifique pour chaque attribut externe. Par exemple, soit l'attribut DATE: 15 janvier 1986, il sera transformé en: 860115.

Dans le cas où les attributs externes contiennent des conjonctions ("et", "ou"), ces conjonctions doivent être transformées en opérateurs de recherche (opérateurs booléens). Cette transformation implique deux phases: l'établissement d'une équivalence entre les conjonctions et les opérateurs de recherche, la détermination des portées ou des attachés (les éléments reliés par une conjonction) des conjonctions.

Selon l'attribut analysé, la signification des conjonctions peut être différente. Nous distinguons les deux cas suivants pour la conversion des conjonctions en opérateurs de recherche:

1er cas (dans l'attribut patient):

Dans cet attribut, la correspondance est la suivante:

et $\rightarrow \vee$
ou $\rightarrow \vee$

Vérifions cela dans les exemples suivants:

Je veux les CRM de Mr.X et de Mr.Y.
Je veux les CRM de Mr.X ou de Mr.Y.

Notons CRM(X) et CRM(Y) les CRM de Mr. X et ceux de Mr. Y. La réponse pour la première requête est constituée de CRM(X) et de CRM(Y): CRM(X) $\cup$ CRM(Y). La condition exprimée dans la requête est donc équivalente à "(patient=X)$\vee$(patient=Y)", ceci est "X$\vee$Y" dans notre représentation interne de la requête.

La réponse pour la seconde requête est la même, car quel que ce soit un document de Mr.X ou de Mr.Y, il doit être extrait.

2ième cas (dans les autres attributs):

Dans les autres attributs, la correspondance des conjonctions en opérateurs est définie de la façon suivante:

et $\rightarrow \wedge$
ou $\rightarrow \vee$

Cette correspondance peut être vérifiée dans les exemples suivants:

Je veux les observations de Dupont et Laurent sur les tumeurs de foie.
Je veux les observations de Dupont faites avant 1976 ou après 1986.

Pour déterminer les portées (attachés) d'une conjonction dans le cas où plus d'une conjonction apparaît dans le même attribut, les règles suivantes sont appliquées:

- La priorité d'attachement est donnée à celle qui n'est pas précédée d'une virgule par rapport à une autre précédée d'une virgule.

- Si cette priorité ne résout pas le problème, c'est-à-dire si plus d'une conjonction se trouve au même niveau (toutes avec ou sans une virgule avant), la priorité est imposée: et > ou.

Exemple: Dupont, et Durand ou Laurent et Legrand
=> Dupont $\wedge$ (Durand $\vee$ (Laurent $\wedge$ Legrand))

4.7. Interprétation de l'attribut interne

4.7.1. Principe

La définition syntaxique du langage d'interrogation et la définition du modèle sémantique de représentation interne impliquent respectivement la mise en oeuvre d'informations syntaxiques et sémantiques pour l'interprétation. On précisera ces aspects par la suite.

4.7.1.1. Informations syntaxiques

On remarque tout d'abord qu'une hiérarchie syntaxique est implicite dans la définition du langage d'interrogation: un GNC peut contenir des GNS, un GNS peut contenir des noms, ... Cette *hiérarchie syntaxique* reflète une hiérarchie de relations entre différents groupes de mots, permettant d'interpréter les groupes prépositionnels, par exemple. Ainsi, dans "le pot de fleur du bureau", la relation existant entre "pot" et "fleur" (GNS) est différenciée de celle existant entre "le pot de fleur" et "le bureau" (GNC).

On considère donc que les composantes d'une phrase correspondant aux plus bas niveaux de la hiérarchie syntaxique doivent être interprétées avant celles correspondant aux niveaux supérieurs. Sur le plan sémantique, on peut rapprocher cette hiérarchie de liens à une notion de "force" du lien sémantique induit entre ces constituants.

Dans [Graitson82], une approche similaire est proposée, qui segmente syntaxiquement une phrase à partir de *marqueurs*. Dans cet ouvrage, les marqueurs sont répartis en deux catégories: *marqueur faible* et *marqueur fort*. Les *marqueurs faibles* sont ceux qui séparent des mots susceptibles de former ensemble un groupe nominal identique à un "concept de base" du système. Ainsi, la préposition *de* et ses allomorphes *du*, *d'*, *des* sont définies comme des marqueurs faibles. Les *marqueurs forts* séparent des mots qui n'ont aucune chance de constituer ensemble un groupe nominal identique à un "concept de base". Ils correspondent, dans la plupart des cas, à une relation sémantique entre ces concepts. Ainsi, des expressions prépositionnelles comme "*au cours de*", "*au niveau de*", des expressions prédicatives comme "*dû à*", "*consistant en*", et certaines prépositions comme "*avec*", sont définies comme marqueurs forts. La phrase est segmentée d'abord par les marqueurs forts, ensuite par les marqueurs faibles.

Pour notre étude, on introduit la notion de *connecteur* plutôt que d'employer le terme *marqueur*, car les éléments servant à "marquer" la segmentation syntaxique impliquent aussi des relations. Plus un connecteur est fort, plus la relation qu'il implique est forte. Par rapport à [Graitson82], un

connecteur faible correspond à un marqueur fort, et un connecteur fort correspond à un marqueur faible.

Si l'on peut définir une telle hiérarchie pour les connecteurs, l'interprétation d'une requête peut alors être conduite de façon hiérarchique, de sorte à *attacher* d'abord les éléments connectés par un connecteur fort, ensuite, les éléments connectés par un connecteur faible (le terme "attacher" représente l'interprétation d'une connexion).

Dans [Graitson82], trois niveaux syntaxiques sont définis: au niveau le plus bas, les mots, au niveau intermédiaire, les mots séparés par les "marqueurs faibles", et au niveau le plus haut, les élément séparés par les "marqueurs forts". Pour notre interprétation, nous définissons 6 niveaux de connexion:

CONNEXION	**EXEMPLES**
0. mot isolé	-foie
1. NOM ADJ*	-bord antérieur, kyste ovarien droit
GNS ADJ*	-augmentation de volume importante
GNC ADJ*	-l'augmentation de volume du foie importante
2. GNS de GNS	-hypertrophie de densité tissulaire
GNS à GNS	-hypertrophie à densité tissulaire
3. GNC PREP GNC	-le lobe gauche du corps thyroïde
GNS PREP1 GNS	-hypertrophie avec extension médiane
4. GNC EXPRE GNC	-une opacité ganglionnaire au niveau de la chaîne para-trachéale droite
5. GNC PRO-PP	-une kyste bronchogénique siégeant dans le médiastin postérieur, -la tumeur circonscrite à la base pulmonaire
GNC PRO-REL	-la tumeur qui ronge le bord antérieur de la vertèbre

où PREP1 exclut les prépositions *de* et *à*.

Au niveau 0, il n'y a pas de connecteur. Au niveau 1, le connecteur est implicite - la connexion nom-adjectif ou adjectif-adjectif. Cette connexion est la plus forte. Au deuxième niveau, les connecteurs sont "de" et "à". La distinction entre ces prépositions et les autres est due au fait qu'elles représentent souvent une relation très forte entre deux noms ou groupes nominaux (cf.[Grevisse80]). La force de connexion diminue dans l'ordre croissant de la hiérarchie.

Le but de cette hiérarchisation syntaxique n'est évidemment pas de donner une détermination "correcte" des rattachements des éléments (car une détermination correcte implique aussi une vérification des conditions

sémantiques), mais plutôt de guider rapidement les attachements vers les solutions syntaxiquement les plus "probables".

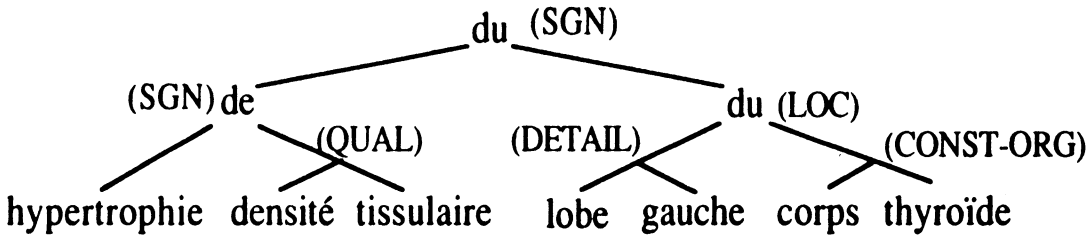
4.7.1.2. Informations sémantiques

Du côté sémantique, la définition du modèle sémantique implique des *contraintes sémantiques* pour la connexion des différents éléments: par exemple, deux éléments peuvent être connectés par un opérateur si l'ensemble correspond à la partie droite d'une règle du modèle sémantique. Cette connexion constitue un concept de la classe sémantique située à gauche de la règle.

On a aussi indiqué (cf.II.2.3) que le modèle sémantique établit également une hiérarchie parmi les différentes classes sémantiques. Par opposition à la hiérarchie syntaxique, on appelle cette hiérarchie la *hiérarchie sémantique*. Cette hiérarchie sémantique implique une hiérarchie dans les constructions des concepts médicaux. Cela nous conduit à faire l'hypothèse suivante: les concepts de bas niveaux sémantiques doivent être formés avant de les regrouper pour former les concepts de hauts niveaux. Autrement dit, les connexions sémantiques de bas niveaux sont plus fortes et doivent être interprétées avant les connexions de hauts niveaux. On peut montrer cette hiérarchie de connexion dans l'exemple suivant:

une hypertrophie de densité tissulaire du lobe gauche du corps thyroïde

On peut hiérarchiser la construction de cet concept de la façon suivante:



L'utilisation de la hiérarchie des classes peut être résumée comme ceci: si à un moment donné, plusieurs connexions se présentent, on choisit de former d'abord celle située au niveau sémantique le plus bas.

4.7.1.3. Méthode pour l'interprétation de l'attribut interne

Les informations syntaxiques et sémantiques décrites ci-dessus sont en réalité souvent compatibles: une connexion syntaxique forte est souvent une connexion sémantique forte et vice versa. Etant donné que le but de l'interprétation est d'aboutir à une représentation sémantique, les contraintes sémantiques sont donc des conditions à satisfaire absolument. Mais il s'avère qu'une analyse purement sémantique est souvent très coûteuse. Les informations syntaxiques peuvent aider à aboutir à une solution plus rapidement. Ainsi, notre interprétation est fondée sur une analyse syntaxico-sémantique.

Parmi les différents modèles d'analyse syntaxico-sémantiques, le modèle du GETA ([Yusoff87, Zajac86]) nous a particulièrement inspiré. Ce modèle construit simultanément plusieurs structures (syntaxique, sémantique, logique, ...). Selon la circonstance, on peut choisir le type d'information qui sera nécessaire pour résoudre le problème.

Nous proposons de suivre la hiérarchie syntaxique dans l'interprétation; à chaque moment, les conditions syntaxiques et sémantiques doivent être simultanément vérifiées. Plus précisément, on interprète d'abord les mots isolés, ensuite les connexions fortes et finalement les connexions faibles. S'il existe une suite de connexions du même niveau syntaxique, la connexion qui forme un concept de niveau sémantique le plus bas est interprétée en premier. Lorsqu'une connexion est interprétée, toutes les informations syntaxiques et sémantiques la concernant sont enregistrées pour les consultations ultérieures.

Nous donnons un exemple pour montrer le principe de l'interprétation. Soit l'attribut interne d'une requête:

rongement du bord antérieur de la vertèbre par la tumeur.

On y retrouve deux niveaux de connecteurs syntaxiques:

Connecteurs syntaxiques	Classe sémantique formée
1. *bord antérieur*	POS
3. *de la*	LOC
du	SGN
par	CONSTAT

Dans l'ordre, ces connecteurs ont une force de connexion décroissante et les classe sémantiques formées ont un niveau croissant. Au niveau 3, il y a trois connecteurs qui forment respectivement les concepts de classe LOC, SGN et CONSTAT. Etant donné que LOC, SGN et CONSTAT sont dans des niveaux sémantiques croissants, on interprète donc d'abord ceux qui forment

La condition sémantique à sayisfaire est la suivante: l'interprétation interne de l'adjectif et celle de l'attaché doivent pouvoir être connectées par une relation sémantique et cette connexion doit correspondre à l'une des contraintes sémantiques.

Remarquons que les variables grammaticales et la classe sémantique d'un groupe nominal sont souvent déterminées par le nom situé en tête. Par exemple, "*opacité du lobe moyen*" peut être considéré comme un nom féminin singulier de la classe SGN. Les variables grammaticales et la classe sémantique qui sont initialement attachées à "*opacité*" deviennent aussi celles de tout le groupe. Ainsi, les conditions pour attacher un adjectif à un groupe nominal sont souvent celles qui sont liées à l'attachement de l'adjectif au premier nom du groupe nominal. Cela permet d'attacher un adjectif à un groupe nominal sans que ce dernier soit interprété.

Parmi tous les candidats de l'attaché d'un adjectif, on choisit celui qui est le plus proche de l'adjectif. Par exemple, dans:

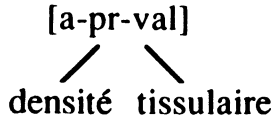
volume de poumon droit
opacité de poumon tissulaire

l'adjectif "droit" est attaché à "poumon", tandis que "tissulaire" est attaché à "opacité de poumon", car en tant que VAL-QUAL, il ne peut pas être attaché à "poumon" (CONST-ORG).

Quant à l'interprétation donnée pour l'adjectif et son attaché, 3 cas sont possibles:

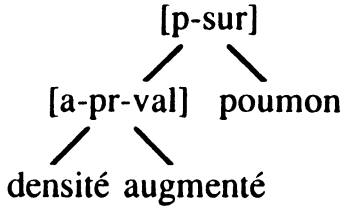
1. L'adjectif est une valeur qualitative (VAL-QUAL) (ex: *tissulaire*) pouvant décrire un SGN, une LESION, ou un CAR-PHY (la classe sémantique de l'attaché). L'interprétation interne de l'adjectif et celle de son attaché (si c'est un groupe nominal, il sera interprété ultérieurement) seront connectées par la relation sémantique [a-pr-val].

Exemple: *densité tissulaire* sera interprété comme:



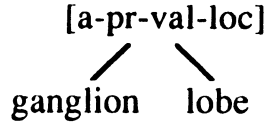
2. L'adjectif (ex: *pulmonaire*) est de la classe LOC (localisation) qui décrit un SGN, une LESION ou un CAR-PHY. L'interprétation interne de l'attaché et celle de l'adjectif seront connectées par la relation sémantique [p-sur].

Exemple: *opacité pulmonaire* sera interprété comme:



3. L'adjectif décrit la localité d'une sous-localité (par exemple, *lobe ganglionnaire*), l'interprétation interne de l'adjectif et celle de son attaché seront connectées par la relation sémantique [a-pr-val-loc].

Exemple: *lobe ganglionnaire* sera interprété comme:



Plusieurs adjectifs peuvent se succéder derrière un nom (ou un groupe nominal). Dans ce dernier cas, les derniers adjectifs peuvent être attachés soit au nom (ou groupe nominal) soit à l'adjectif précèdent. Dans le cas de deux adjectifs successifs, les situations suivantes sont possibles:

1. ((GN ADJ1) ADJ2)
2. (GN (ADJ1 ADJ2))
3. (GN PREP (GN ADJ1) ADJ2)

où GN est un GNC ou GNS (un groupe nominal complet ou simple).

Le premier cas exprime que les deux adjectifs portent sur le même nom (ou groupe nominal). Par exemple: *volume pulmonaire augmenté.* Dans le deuxième cas, le deuxième adjectif porte sur le premier adjectif, et le premier porte sur le nom. Par exemple: *kyste ovarien droit.* Dans la troisième, le deuxième adjectif porte sur un groupe nominal incluant la portée du premier adjectif. Par exemple: *volume du poumon droit augmenté.*

Quand un adjectif (précédé d'un autre adjectif) a simultanément plusieurs candidats satisfaisant les conditions syntaxiques et sémantiques, on en choisit un. Si ce choix conduit à un échec dans l'interprétation ultérieure, un retour arrière sera effectué pour choisir un autre candidat. Pour déterminer l'ordre du choix, on utilise la hiérarchie sémantique de la manière suivante:

> Une connexion de l'adjectif formant un concept d'une classe de bas niveau doit être choisie avant une autre connexion formant un concept d'une classe de haut niveau.

Schématiquement, l'attachement d'une suite d'adjectifs est effectué ainsi:

1. attacher le premier adjectif à un nom (ou un groupe nominal)
   le nom (le groupe nominal) et l'adjectif seront considérés comme un seul élément pendant l'attachement des adjectifs.
2. attacher l'adjectif suivant à l'un des éléments ci-dessous dont la connexion forme un concept d'une classe de plus bas niveau:
   - l'adjectif précédent
   - l'attaché de l'adjectif précédent
   - un groupe nominal incluant l'adjectif précédent et son attaché

   L'ensemble de l'adjectif et son attaché sera considéré comme un seul élément pendant l'attachement des autres adjectifs.
3. reprendre 2 jusqu'au dernier adjectif de la succession.
4. Si l'attachement ne peut pas achevé pour toute la succession des adjectifs, ou si un échec survient dans l'interprétation ultérieure, un retour arrière sera engendré pour choisir un autre alternatif de l'attaché.

Nous donnons deux exemples pour conclure cette section:

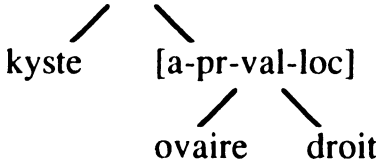
exemple1:

0. un kyste ovarien droit
1. *ovarien* (interprété comme *ovaire* de la classe de CONST-ORG) est attaché à *kyste* (LESION):

-> une [p-sur] (LESION) droit

2. *droit* (POS) ne peut être attaché qu'à *ovaire* (CONST-ORG):

-> une [p-sur] (LESION)

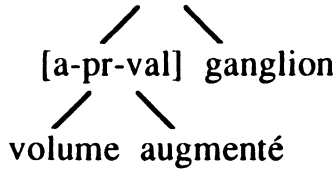


exemple2:

0. des hypertrophies ganglionnaires des loges latéro-trachéales droites carénaires

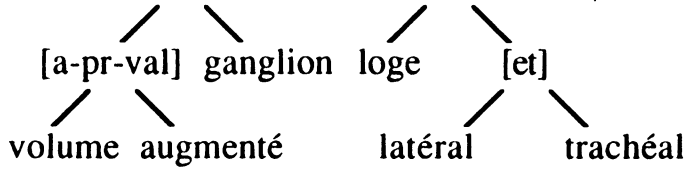
1. L'interprétation de *ganglionnaire* est *ganglion* de la classe CONST-ORG. Celle de *hypertrophie* est interprétée comme *[a-pr-val] (volume,augmenté)* de la classe SGN. *Hypertrophie* peut être connecté avec *ganglionnaire* par la relation [p-sur].

-> des [p-sur] (SGN) des loges latéro-trachéales droites carénaires



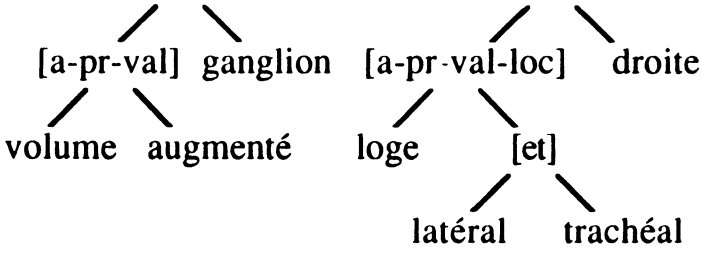
2. *latéro-trachéales* est interprété comme *[et](latéral,trachéal)* de la classe de POS. *Loges* est un DETAIL. Ils peuvent être connectés par la relation [a-pr-val-loc].

-> des [p-sur] (SGN) des [a-pr-val-loc] (DETAIL) droites carénaires



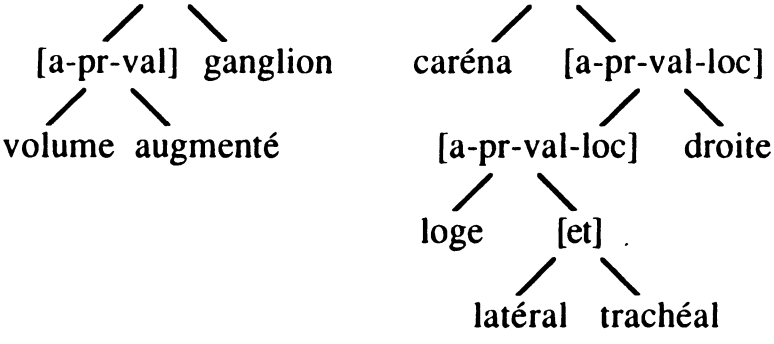
3. *droites* est une POS. Il peut être attaché à *loges latéro-trachéales* (DETAIL) (et pas à *latéro-trachéales*)

-> des [p-sur] (SGN) des [a-pr-val-loc] (DETAIL) carénaires



4. *carénaires* est interprété comme *caréna* de la classe de CONST-ORG. Il peut être connecté avec *loges latéro-trachéales droites* par la relation [a-pr-val-loc].

-> des [p-sur] (SGN) des [a-pr-val-loc] (ORG)



Si un adjectif est attaché à un mot creux, il doit prendre son interprétation par défaut. Par exemple, l'adjectif *convexe* possède une interprétation par défaut: [a-pr-val](forme,convexe). Ainsi, *image convexe* ("*image*" est un mot creux) sera interprétée en: [a-pr-val](forme,convexe).

Après l'interprétation d'une connexion, toutes les informations syntaxiques et sémantiques sont enregistrées. La connexion sera considérée comme un seul élément dans les étapes ultérieures de l'interprétation.

4.7.4. Connexions fortes prépositionnelles (niveau 2)

Ces connexions sont réalisées par les prépositions *de* et *à* dans un groupe nominal simple. Par exemple: *opacité nodulaire de 8 mm de diamètre, opacité à contour irrégulier.*

Analysons d'abord une seule connexion isolée. Elle a la forme syntaxique suivante:

NOM1 PREP NOM2

Les trois éléments ont chacun une interprétation distincte. PREP ("de" ou "à") peut être interprétée par plusieurs relations sémantiques ([p-sur], [a-pr-val], [a-pr-val-loc]). Le choix de l'opérateur sémantique dépend alors des classes sémantiques de NOM1 et NOM2. NOM1 et NOM2 ayant chacun une classe sémantique déterminée, un seul opérateur est possible.

Par exemple, pour *augmentation de densité*, l'interprétation de "de" en [a-pr-val] permet de connecter les interprétations internes de deux noms: *densité* (CAR-PHY), *augmenté* (VAL-QUAL).

Dans certains cas, les noms peuvent aussi avoir plusieurs interprétations possibles. Cela peut éventuellement conduire à plusieurs interprétations pour cette connexion. Cette ambiguïté peut généralement être résolue dans la suite du processus, lors de l'interprétation des connexions supérieures. Si cela aussi échoue, la requête sera considérée comme ambiguë et aboutira à plusieurs interprétations possibles. Ainsi, si une ambiguïté d'interprétation se produit, on associe toutes les interprétations possibles à ce groupe nominal.

Dans le cas d'une succession de connexions avec ces deux prépositions, un problème supplémentaire se pose pour déterminer les attachés de chaque préposition. Comme pour l'attachement d'adjectifs, un nom précédé d'une préposition peut avoir 3 sortes d'attachés possibles:

1. le nom (ou un groupe nominal) précédant la préposition,
2. l'attaché du nom (groupe nominal) précédent,
3. un groupe nominal incluant le nom (groupe nominal) précédent et son attaché.

Dans le cas de 3 prépositions, les situations suivantes sont possibles:

(A de (B de (C de D)))
((A de B) de (C de D))
(A de ((B de C) de D))
((A de (B de C)) de D)
(((A de B) de C) de D)

Pour déterminer le choix à faire, on utilise encore la méthode proposée dans 7.3.1: une connexion formant un concept de plus bas niveau sémantique est plus forte qu'une connexion formant un concept de plus haut niveau. Ainsi, pour les connexions situées à un même niveau syntaxique, on interprète d'abord celle qui forme un concept au niveau sémantique le plus bas possible. Ceci correspond à la stratégie suivante:

1. choisir trois éléments successifs correspondant au groupement NOM1 PREP NOM2 pour former un concept au niveau sémantique le plus bas possible.
2. Interpréter cette connexion: les trois éléments dans la connexion seront alors considérés comme un seul élément.
3. Recommencer 1, jusqu'à former un seul élément.

Exemple:

• hypertrophie de poumon de 8 mm de diamètre

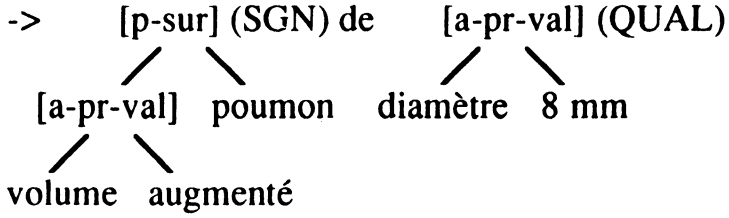
• Après l'interprétation des mots, on obtient:

-> [a-pr-val] (SGN) de poumon de 8 mm de diamètre
/ \ (CONST-ORG) (QUAN) (CAR-PHY)
volume augmenté

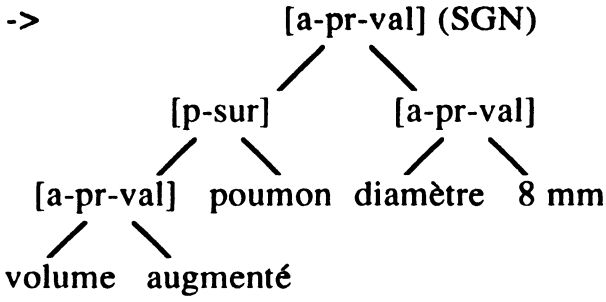
• "SGN de CONST-ORG" peut former un concept de classe SGN et "QUAN de CAR-PHY" peut former un concept de classe QUAL. Ce dernier est à un niveau sémantique plus bas que SGN. Ainsi, on interprète d'abord "8 mm de diamètre":

->[a-pr-val] (SGN) de poumon de [a-pr-val] (QUAL)
/ \ (CONST-ORG) / \
volume augmenté diamètre 8 mm

• dans cette nouvelle chaîne, seul "SGN de CONST-ORG" peut former un concept de classe SGN:



• Et finalement, "SGN de QUAL" forme un SGN:



4.7.5. Connexions faibles (niveaux 3 et 4)

Aux niveaux 3 et 4, les connexions concernent deux groupes nominaux (tous complets ou simples) reliés par une préposition ou une expression jouant un rôle similaire. L'interprétation des connexions de ces deux niveaux est fondée sur exactement le même principe que celui énoncé précédemment: le niveau 3 est interprété avant le niveau 4. Si plusieurs connexions de même niveau sont rencontrées, celle qui forme un concept au niveau sémantique le plus bas est interprétée en premier.

Exemple:

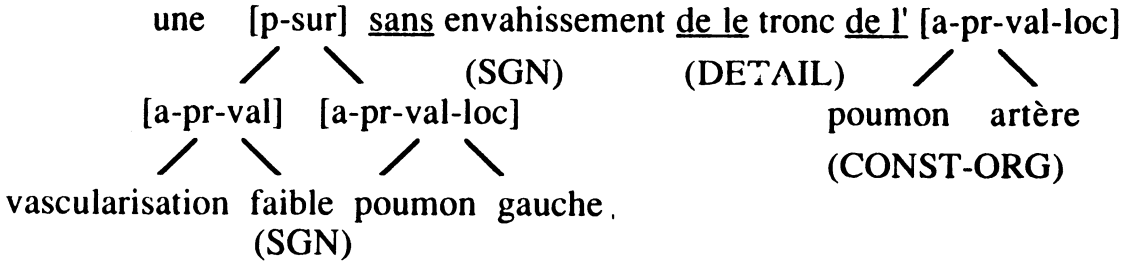
une olighémie pulmonaire gauche sans image apparente d'envahissement du tronc de l'artère pulmonaire

Dans le dictionnaire, les informations suivantes sont enregistrées:

une(*art,fém,sing*,vide)
olighémie(*nom, fém, sing*, [a-pr-val]((vascularisation,SGN),
(faible,VAL-QUAL)),SGN)
pulmonaire(*adj, mas-fém, sing*, poumon,CONST-ORG)
gauche(*adj, mas-fém, sing*,gauche,POS)
sans(*prép*,([a-pr-val](0,(0)), [en-rel-topo-avec](0,(0))))
image(*nom, fém, sing*,vide,-)
apparente(*adj, fém, sing*,vide,-)
envahissement(*nom, mas, sing*,envahissement,SGN)

tronc(*nom, mas, sing*,tronc,DETAIL)
artère(*nom, fém, sing*,artère,REG)

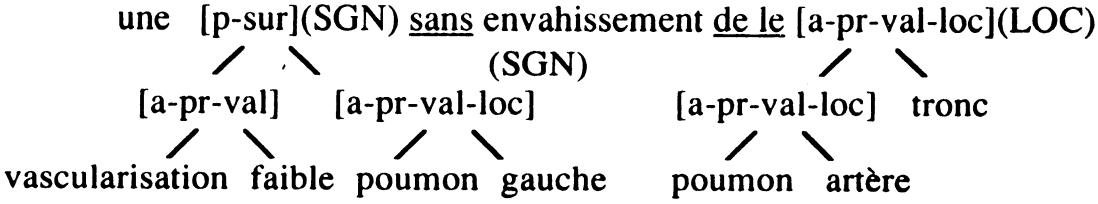
Après les interprétations du niveau 2, on obtient le résultat suivant:



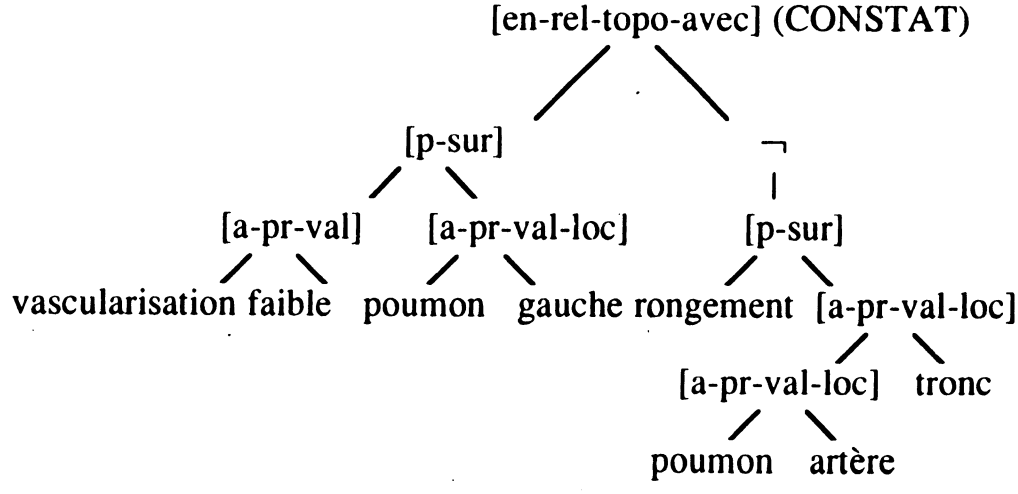
correspondant à la chaîne:

SGN sans SGN de le DETAIL de l' CONST-ORG

Dans cette chaîne, les trois connexions prépositionnelles sont toutes au niveau 3. Sémantiquement, la partie "DETAIL de l' CONST-ORG" forme un concept de plus bas niveau - CONST-ORG. Ainsi, cette connexion est interprétée la première. On aboutit à:



Dans cette nouvelle chaîne, "SGN de le LOC" forme un concept de la classe SGN, située à un niveau sémantique plus bas que "SGN sans SGN" - CONSTAT. Ainsi, la connexion "de le" est interprétée avant celle de "sans" et on obtient le résultat final de l'interprétation:



4.7.6. Interprétation des groupes verbaux

Les verbes peuvent exister dans une proposition relative décrivant un groupe nominal (l'antécédent). Les autres formes de verbes (les formes participe passé, participe présent) ont une fonction similaire.

Nous décrivons cette interprétation de la façon suivante:

- d'abord l'interprétation d'un verbe isolé,
- ensuite un syntagme verbal isolé,
- et finalement un syntagme verbal dans une requête.

• Interprétation d'un verbe isolé

Pour l'interprétation des verbes, nous adoptons une approche similaire à celle de [Schank80b, Schank81, Schank82]: un verbe représente un *trait* particulier dont éventuellement certains éléments doivent être complétés par les groupes de mots qui lui sont liés, tels que le sujet, les compléments.

Dans RIME, on peut distinguer 3 types de verbes:

- les verbes relationnels simples
  ex: *montrer, siéger, associer*
- les verbes relationnels complexes
  ex: *envahir, cerner, dépasser, augmenter*
- et les verbes outils
  ex: *découvrir, exister, apparaître, voir*

Les *verbes outils* sont ceux qui ont un trait sémantique vide. Les verbes relationnels simples expriment une relation sémantique entre le sujet et le complément. Les verbes relationnels complexes ont, dans le modèle, une interprétation réduite à un trait composé d'une arborescence décrivant un concept déterminé.

Exemples:

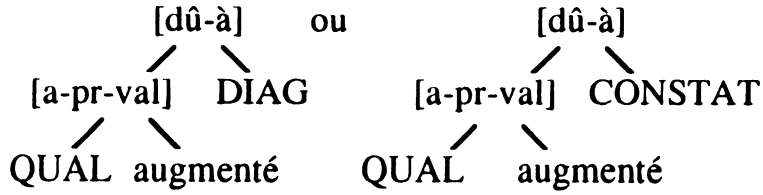
1. Verbes outils:
   Dans "*une atélectasie qu'on a découverte dans le lobe inférieur gauche*", le verbe "découvrir" est un verbe vide. Cette phrase est équivalente à "*une atélectasie dans le lobe inférieur gauche*".

2. Verbes relationnels simples (leur trait sémantique se réduit à un opérateur sémantique):

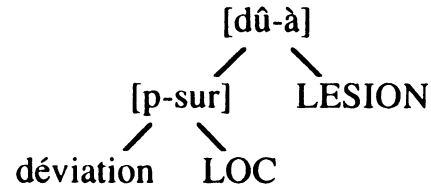
mettre en évidence, délimiter, dénombrer,	
dégager, faire apparaître, prendre la forme de ...	[montré-par]
traduire, signifier, être le signe de	
être la traduction de, laisser supposer ...	[per-de-déd]
être situé(e)(s) dans, se trouver dans ...	[p-sur]
entraîner, conduire, ...	[dû-à]
valoir, mesurer, être estimé(e)(s) à, ...	[a-pr-val]
...	

3.Verbes relationnels complexes (leur trait sémantique correspond à une structure de relation sémantique complexe):

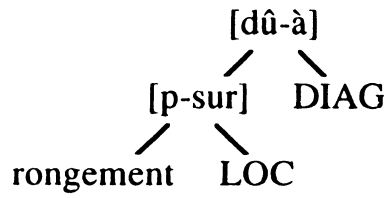
augmenter



dévier



ronger



• **Interprétation d'un syntagme verbal isolé**

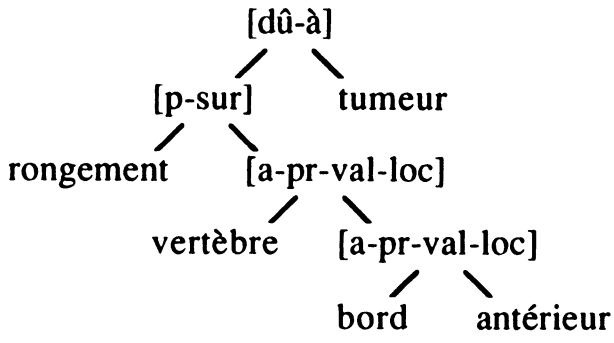
Ayant établi un trait pour chaque verbe, l'interprétation du syntagme verbal consiste à trouver parmi les éléments liés (le sujet, le complément) ceux qui doivent se substituer aux éléments non-terminaux du trait. Etant donné que ces éléments sont déjà interprétés auparavant et sont en général des concepts de classes sémantiques différentes, l'identification de ces éléments n'est donc pas très difficile à effectuer.

Par exemple, dans

le bord antérieur de la vertèbre qui est rongé par la tumeur ...,

l'antécédent du pronom relatif "qui" - *le bord antérieur de la vertèbre*, est une localité qui peut substituer LOC dans le trait de "ronger", et *la tumeur* est une

lésion et donc un diagnostic qui peut se substituer à DIAG dans le trait. Tout le syntagme verbal sera donc interprété en:



• **Interprétation d'un syntagme verbal dans une requête**

Dans la section précédente, l'antécédent du pronom relatif (ou du participe) est supposé déterminé. Dans une requête, l'antécédent n'est pas toujours la totalité du groupe nominal précédant le pronom relatif, mais souvent seulement une partie. Par exemple: dans

des opacités pulmonaires du lobe supérieur rongé par la tumeur

l'antécédent du participe passé "rongé" est seulement une partie du groupe nominal précédent - *le lobe supérieur.*

L'attribut interne d'une telle requête peut donc être représenté par:

GN1 PREP ANTECEDENT PRO-REL

où GN1 ne fait pas partie de l'antécédent.

Dans l'exemple ci-dessus,
GN1=*des opacités pulmonaires*,
PREP=*de*,
ANTECEDENT=*le lobe supérieur*,
et PRO-REL=*rongé par la tumeur*

Un problème supplémentaire est donc d'isoler le syntagme verbal, c'est-à-dire de déterminer l'antécédent du pronom relatif (ou du verbe pour le participe présent et le participe passé). Pour ce faire, il faut revenir à l'état précédant l'interprétation des connexions entre groupes nominaux complets, car l'antécédent doit être un groupe nominal complet, donc un élément dans ces connexions.

Les conditions syntaxiques et sémantiques pour déterminer un antécédent sont les suivantes:

- L'antécédent d'une proposition relative doit être un groupe nominal complet (ayant un article au début). Il doit être en accord en nombre et éventuellement en genre (pour le participe passé) avec le verbe (sauf pour le participe présent).
- Le syntagme verbal ainsi isolé doit substituer tous les éléments non-terminaux dans le trait du verbe, ou autrement dit, le trait sémantique du verbe doit pouvoir combiner tous les éléments dans le syntagme verbal.

Dans le cas où plusieurs candidats pour l'antécédent vérifient ces conditions, une ambiguïté subsiste. Elle ne peut pas être levée à ce niveau. On doit donc admettre toutes ces solutions et les proposer aux processus ultérieurs (le processus de recherche notamment).

Considérons l'exemple suivant:

une opacité nodulaire du lobe supérieur droit du poumon rongé par la tumeur.

Dans cette requête, deux groupes nominaux satisfont les conditions syntaxiques et sémantiques:

le poumon
le lobe supérieur droit du poumon

Cela aboutit aux deux solutions suivantes:

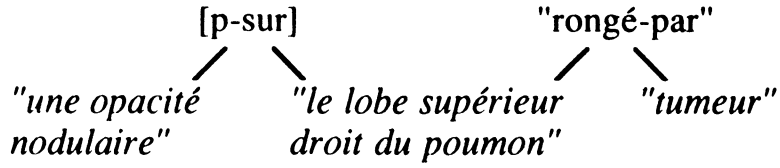
une opacité nodulaire du lobe supérieur droit de (le poumon) rongé par la tumeur
une opacité nodulaire de (le lobe supérieur droit du poumon) rongé par la tumeur

Ayant isolé un syntagme verbal, la question est suivante: comment interpréter toute la requête?

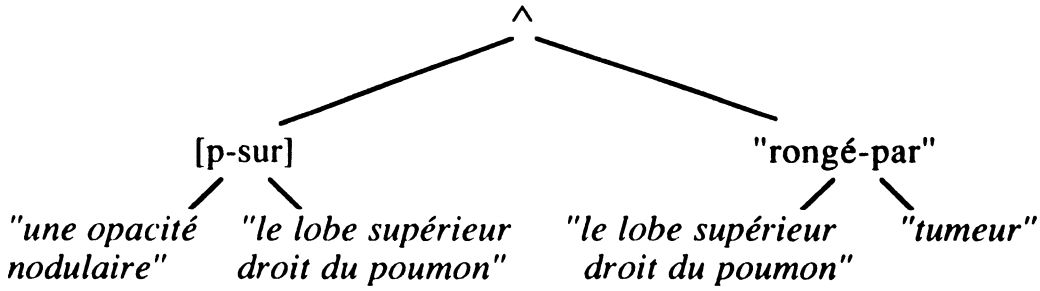
Remarquons que dans ces exemples (le second par exemple), l'antécédent (*le lobe supérieur droit du poumon*) est décrit deux fois:

- dans le groupe nominal devant la proposition relative:
  une opacité nodulaire de (le lobe supérieur droit du poumon)
- par la proposition relative:
  (le lobe supérieur droit du poumon) qui est rongé par la tumeur

Cela correspond aux descriptions sémantiques suivantes:



ce qui ne correspond pas à une arborescence. Pour les transformer en une arborescence, on sépare les deux descriptions et on les connecte par l'opérateur ∧, ce qui conduit à:



L'opérande gauche de l'opérateur ∧ a été déjà interprété au niveau 2. L'opérande droit est interprété comme un syntagme verbal isolé.

4.7.7. Interprétation des certitudes et de la négation

Les requêtes de l'utilisateur peuvent contenir une négation, telle que:
Quels sont les CRM contenant une opacité pulmonaire probable?
Quels sont les CRM décrivant une opacité pulmonaire de densité probablement non tissulaire?
Quels sont les CRM décrivant des processus expansifs avec calcifications, de densité tissulaire, qui ne soient pas des kystes bronchogéniques?
...

Dans RIME, on peut rencontrer trois types de certitude mis à part la certitude totale implicite: la certitude "probable", la certitude "probablement non", et l'incertitude totale. Nous les représentons dans les arborescences par les opérateurs suivants (unaires):

probable	→	[p]
probablement-non	→	[pn]
non	→	¬

Le problème qui se pose est de déterminer la portée de la valeur de certitude et de la négation, et de la représenter dans les arborescences.

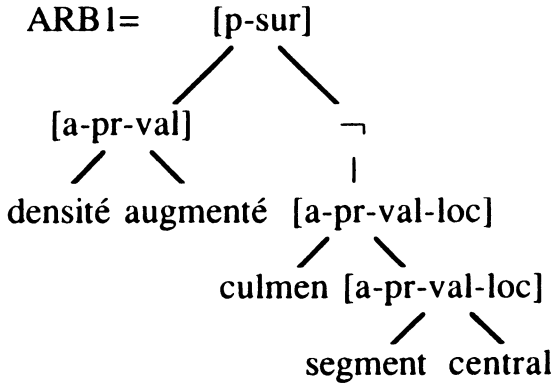
On distingue deux types de portée:

- sur un verbe,
- sur une partie d'un groupe nominal.

Par exemple:
un verbe - *l'opacité ne siège pas au niveau du segment ventral du culmen,*
une partie de groupe nominal - *une opacité pulmonaire de densité non tissulaire.*

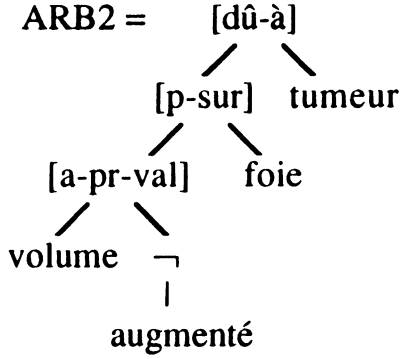
Quand la portée est un verbe relationnel simple, on considère que les certitudes et la négation portent sur l'opérateur correspondant.

Par exemple, dans le premier exemple, la portée de l'opérateur est la relation [p-sur] interprétée du verbe "siège". Remarquons que le sujet de la phrase n'est pas dans la négation, c'est-à-dire que "l'opacité" est vérifiée. On interprète donc cette phrase en:



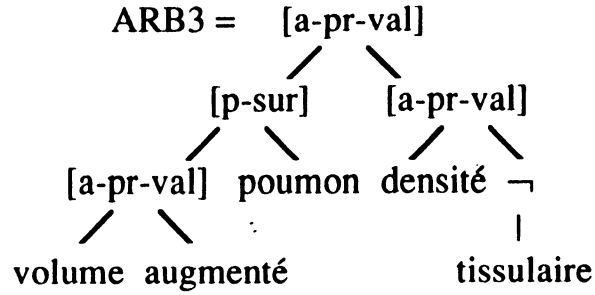
Quand la portée est un verbe relationnel complexe, les certitudes et la négation sont considérées comme portant sur le concept impliqué par le verbe.

Par exemple, dans "la tumeur n'augmente pas le volume du foie", le concept "augmentation" (interprété par "augmenté") est dans la négation. Cette phrase est donc interprétée en:



Quand les certitudes et la négation portent sur un groupe nominal ou un composant de celui-ci, c'est le concept correspondant qui est mis en cause.

Par exemple, dans *une hypertrophie pulmonaire de densité non tissulaire*, la négation porte seulement sur "tissulaire". On interprète ce groupe par l'arborescence suivante:



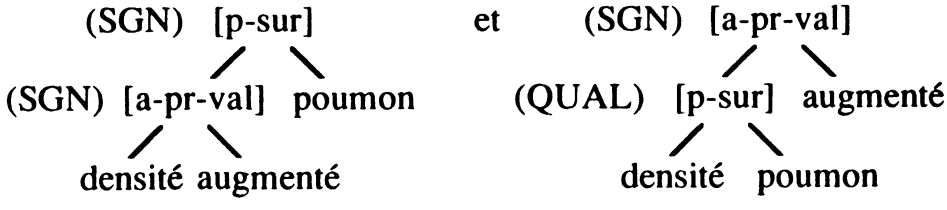
4.7.8. Standardisation des requêtes

1). Nous allons utiliser le modèle défini dans la partie I pour évaluer les requêtes. Nous choisissons le modèle de I.3.3.4.1, qui considère un document comme un monde et une requête comme une formule (cf.II.6). Avant l'évaluation, il est donc nécessaire de transformer les requêtes en formules, conformément à la définition donnée dans la partie I.

Une proposition atomique dans RIME correspond à une arborescence contenant uniquement des opérateurs sémantiques. Le premier aspect de la standardisation consiste à remonter les opérateurs de certitude et les opérateurs booléens par rapport aux opérateurs sémantiques. Ces phases sont appelées "structuration des opérateurs booléens" et "structuration des opérateurs de certitude".

2). Un autre aspect de la standardisation consiste à représenter les arborescences contenant uniquement des opérateurs sémantiques et sémantiquement équivalentes sous une même forme. Pour simplifier la discussion, nous appelons, par la suite, une arborescence contenant uniquement des opérateurs sémantiques une *arborescence sémantique.*

Nous remarquons une équivalence entre les deux arborescences suivantes correspondant respectivement "opacité pulmonaire" et "augmentation de la densité pulmonaire".



Au niveau sémantique, elles sont strictement équivalentes. Si ces arborescences ne sont pas représentées d'une façon standard, il est nécessaire, durant l'évaluation, de transformer une arborescence du document en toutes ses formes possibles pour pouvoir les comparer avec une arborescence de la requête, ou vise versa. Cela représente un coût considérable pour l'évaluation.

Pour diminuer le temps d'évaluation, on doit donc transformer les différentes arborescences sémantiques équivalentes en une forme standard, autrement dit, résoudre le problème de paraphrasage.

Le problème de représentation multiple d'une même sémantique peut être considéré à plusieurs niveaux:

- On peut considérer simplement l'équivalence au niveau structurel: une arborescence sémantique peut être transformée en une autre en modifiant seulement la structure d'une certaine manière.
- On peut considérer l'équivalence au niveau de la déduction: par exemple, si un ensemble de descriptions permettent de déduire une conclusion, on considère qu'une équivalence existe entre l'ensemble des descriptions et la conclusion déduite.

Le traitement du second cas nécessite des connaissances typiquement expertes. Cet aspect n'est pour l'instant pas intégré dans RIME. Nous nous restreindrons donc aux cas résultant du premier niveau. On appelle cette phase de standardisation "structuration des opérateurs sémantiques".

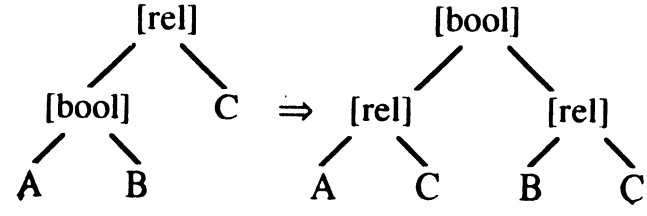
Pour préserver la cohérence, la même standardisation doit aussi être effectuée sur le résultat de l'indexation des documents.

4.7.8.1. Structuration des opérateurs booléens

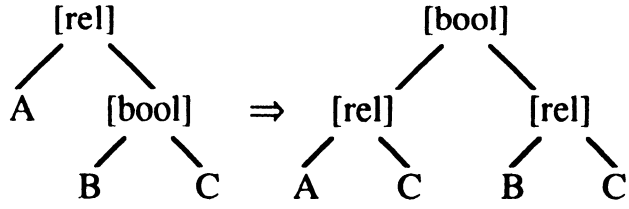
Soient [rel] un opérateur sémantique, [bool] un opérateur booléen binaire ($\vee$ ou $\wedge$), A, B, C des concepts quelconques.

La structuration consiste à appliquer successivement les transformations suivantes (transformant une arborescence de gauche en l'arborescence de droite):

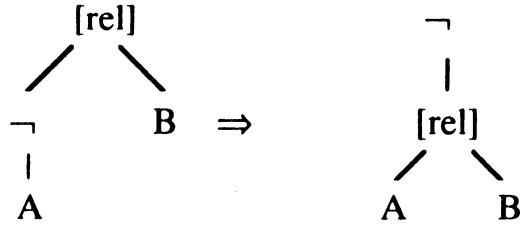
1.



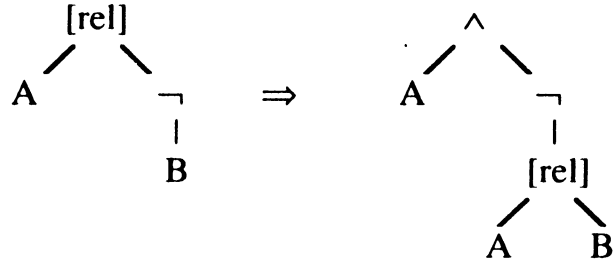
2.



3.

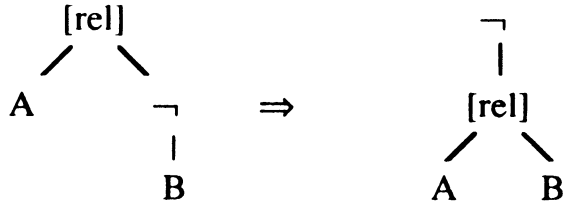


4.



(si A seul est significatif)

5.

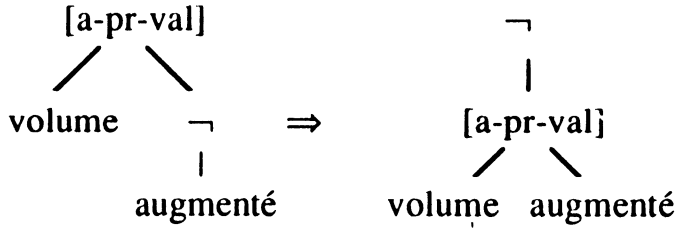


(si A seul n'est pas significatif)

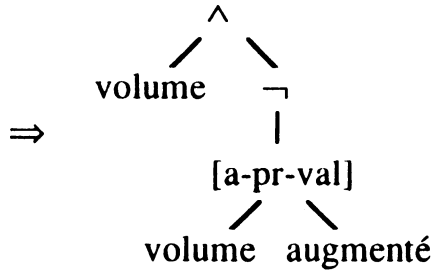
NB: - La transformation 3 est fondée sur le fait que le concept de gauche est un concept *principal* (cf.II.2). Une négation sur le concept principal porte donc aussi sur toute l'arborescence.
- Aux transformations 4 et 5 est associée une condition. Selon que le concept A seul a un sens ou non (par exemple, "volume" n'a pas de sens), on choisit l'une ou l'autre de ces transformations. Les concepts des classes sémantiques suivantes (pouvant apparaître sur la branche gauche d'un opérateur sémantique) sont considérés comme non significatifs pris isolément:

CAR-PHY	volume
LOC	poumon gauche
ORG	poumon
REGION	artère
DETAIL	lobe

Par exemple:



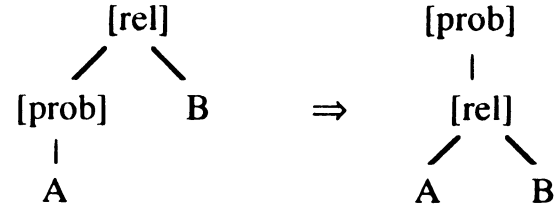
et non pas



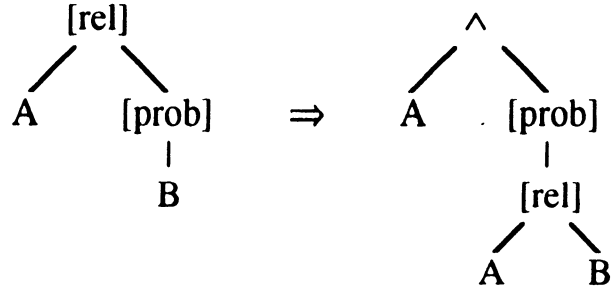
4.7.8.2. Structuration des opérateurs de certitude

Cette structuration est similaire à celle des opérateurs booléens. Elle correspond aux transformations suivantes (où [prob] représente [p] ou [pn]):

1.

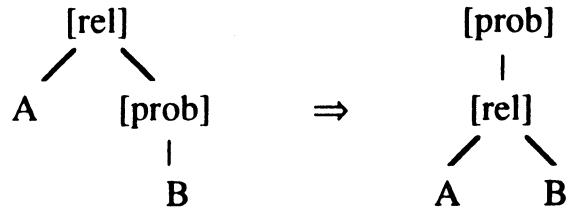


2.



(si A seul est significatif)

3.



(si A seul n'est pas significatif)

Après les structurations précédentes, il est possible que plusieurs négations et/ou opérateurs de certitude portent sur une même sous-arborescence. Par exemple, [p](¬([a-pr-val](densité,augmenté))).

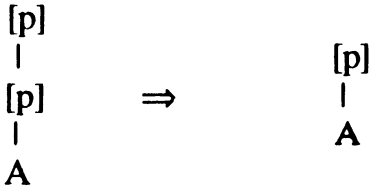
En effet, l'ensemble de ces négations et/ou opérateurs de certitude dénote une seule négation ou certitude sur la sous-arborescence. Dans l'exemple précédent, l'ensemble de [p] et de ¬ est équivalent à [pn] (probablement non). Ainsi, l'arborescence précédente peut être transformée en [pn]([a-pr-val](densité,augmenté)).

Pour toutes les combinaisons possibles des opérateurs de certitude et/ou des négations, le calcul existant dans le système S4 de la logique modale nous semble très adapté. Il propose les unifications d'opérateurs suivantes (où M dénote "probable"):

M·M→M	"probable" . "probable" → "probable"
M·M¬ → M¬	"probable" . "probablement-non" → "probablement-non"
M·¬ → M¬	"probable" . "non" → "probablement-non"
M¬·M → M¬	"probablement-non" . "probable" → "probablement-non"
M¬·M¬ → M	"probablement-non" . "probablement-non" → "probable"
M¬·¬ → M	"probablement-non" . "non" → "probable"
¬·M →¬	"non" . "probable" → "non"
¬·M¬ → 1	"non" . "probablement-non" → 1
¬·¬ → 1	"non" . "non" → 1

Dans RIME, l'opérateur M est représenté par [p] et l'ensemble de M et de ¬ par [pn]. Correspondant à ces unifications, on peut donc proposer les transformations suivantes:

1.



2.
```
[p]                 [pn]
 |                   |
[pn]      ⇒         A
 |
A
```

3.
```
[p]                 [pn]
 |                   |
¬         ⇒         A
 |
A
```

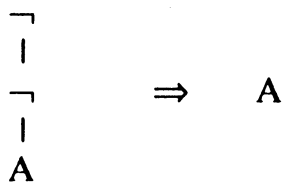
4.
```
[pn]                [pn]
 |                   |
[p]       ⇒         A
 |
A
```

5.
```
[pn]                [p]
 |                   |
[pn]      ⇒         A
 |
A
```

6.
```
[pn]                [p]
 |                   |
¬         ⇒         A
 |
A
```

7.
```
¬                   ¬
 |                   |
[p]       ⇒         A
 |
A
```

8.
```
¬
 |
[pn]      ⇒    A
 |
A
```

9.


Ces transformations permettent d'associer une seule négation ou certitude à une sous-arborescence.

4.7.8.3. Structuration des arborescences sémantiques

Rappelons les deux formes possibles d'arborescences (cf.II.2):

- Un concept d'une certaine classe peut être décrit par un autre concept pour former un concept de la même classe. Notons cette formation par:

CLASSE1 ::= [rel](CLASSE1, CLASSE2)

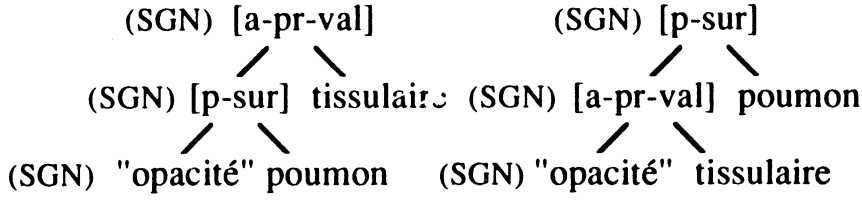
- Deux concepts sont connectés par un opérateur sémantique pour former un concept d'une classe différente:

CLASSE1 ::= [rel](CLASSE2, CLASSE3)

1). Pour la première forme, on a montré que le premier concept est un *gouverneur*, tandis que le second est un *dépendant*. Une telle arborescence peut donc être considérée comme une description du premier concept. On établit donc l'hypothèse suivante:

> Si un gouverneur est décrit par (a pour dépendants) plusieurs concepts sans changement de classe sémantique, l'ordre des descriptions (des dépendants) n'a pas d'importance.

Par exemple, les deux arborescences suivantes sont tout à fait équivalentes:



Ainsi, on peut imposer un ordre sur les descriptions des dépendants de façon que les descriptions d'un gouverneur par les mêmes dépendants soient représentées de la même façon.

En analysant les règles du modèle sémantique (cf.II.2), on peut s'apercevoir que les formations d'arborescences de ce type sont restreintes aux cas suivants:

CONSTAT ::= [en-rel-topo-avec](CONSTAT, CONSTAT)
CONSTAT ::= [dû-à](CONSTAT, CONSTAT)
CONSTAT ::= [dû-à](CONSTAT, DIAG)

DIAG ::= [en-rel-topo-avec](DIAG, DIAG)
DIAG ::= [dû-à](DIAG, DIAG)

SGN ::= [en-rel-topo-avec](SGN, LOC)
SGN ::= [p-sur](SGN, LOC)
SGN ::= [p-sur](SGN, FCT)
SGN ::= [a-pr-val](SGN, QUAL)
SGN ::= [a-pr-val](SGN, VAL-SGN)

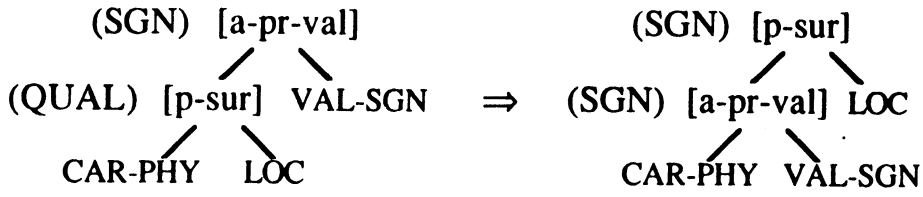
LESION ::= [en-rel-topo-avec](LESION, LOC)
LESION ::= [p-sur](LESION, LOC)
LESION ::= [p-sur](LESION, FCT)
LESION ::= [a-pr-val](LESION, QUAL)

LOC ::= [a-pr-val-loc](LOC, POS)
LOC ::= [en-rel-topo-avec](LOC, LOC)

ORG ::= [a-pr-val-loc](ORG, REGION)
ORG ::= [a-pr-val-loc](ORG, DETAIL)

Une arborescence (sous-arborescence) quelconque ainsi formée est transformée de façon à ce que les opérateurs soient dans l'ordre décrit ci-dessus, c'est-à-dire qu' une formation inférieure est prioritaire par rapport à une formation supérieure. Par exemple, une description via l'opérateur [a-pr-val] est prioritaire par rapport à une description via l'opérateur [p-sur]. Ainsi, la seconde arborescence donnée dans l'exemple précédent est considérée comme étant standard.

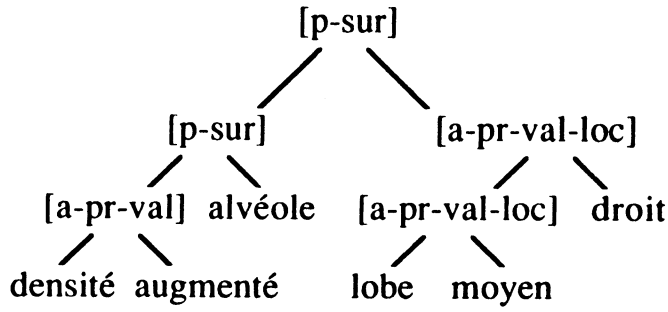
2). Sur le second type de formation d'arborescences (deux concepts forment un concept de classe différente), l'établissement de l'équivalence s'avère très délicat, car une modification de la structure peut modifier sa signification. Cela doit être effectué par des spécialistes du domaine médical. Dans le cadre de cette thèse, nous nous limitons au cas suivant qui correspond à l'exemple donné au début de la section:



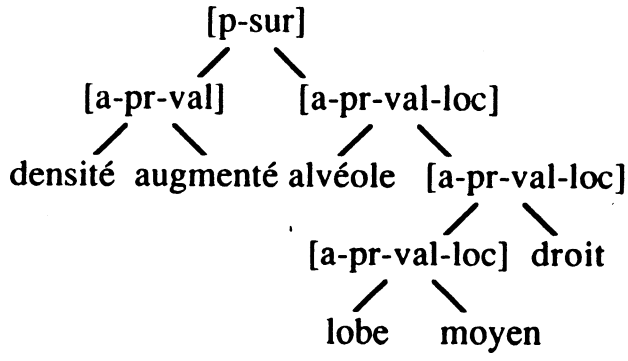
Après la transformation, la classe sémantique de toute l'arborescence n'est pas changée et les contraintes sémantiques sont respectées.

3). Analysons maintenant l'exemple suivant:

"opacité alvéolaire au niveau du lobe moyen droit" est interprété comme suit:



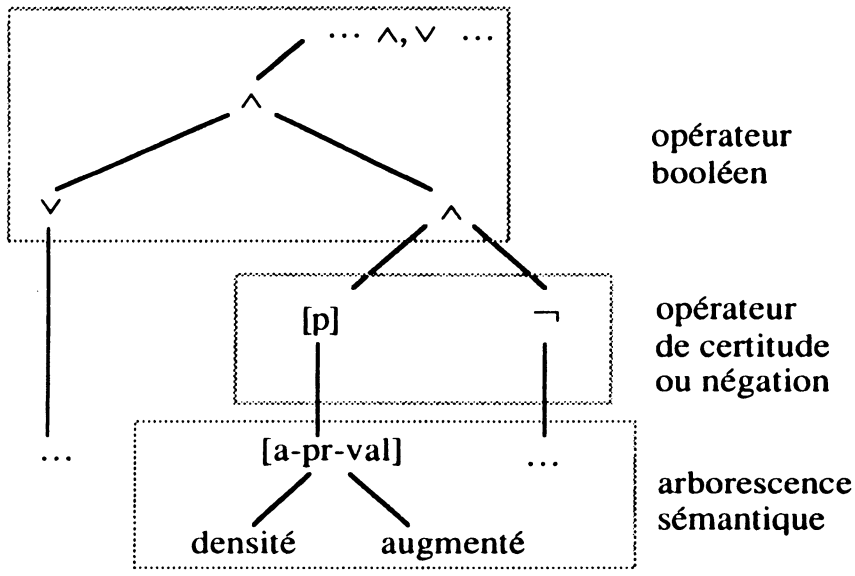
Dans cette arborescence, le signe "opacité" (interprété en [a-pr-val] (densité, augmenté)) est décrit par deux localisations: "alvéole" et "lobe moyen droit". Mais ces deux localisations ne sont pas indépendantes. En effet, elles appartiennent à deux classes sémantiques différentes (ORG et DETAIL), et sont complémentaires l'une à l'autre. Ainsi, on peut les regrouper ensemble pour former une localisation unique:



Il semble qu'il existe beaucoup d'autres cas d'équivalence d'arborescences dans le domaine de RIME. Mais une étude plus élaborée pour établir toutes les équivalences nécessiterait une bonne connaissance du domaine médical. Dans le cadre de cette thèse, nous nous limitons aux cas analysés précédemment.

4.7.8.4. Structure finale des requêtes

Après cette standardisation, nous obtenons la structure suivante pour les requêtes:



Remarquons qu'il n'existe au maximum qu'une seule négation ou certitude sur une arborescence sémantique, car plusieurs certitudes et/ou négations ont été unifiées pendant la structuration.

Etant donné que les documents doivent subir la même standardisation, ils ont la même structure finale. En ce qui concerne les documents, on remarque les particularités suivantes:

Si "A ou B" est décrit dans un document, on considère que l'auteur n'est sûr ni de A ni de B. Ainsi, A et B ne sont pas vérifiés dans le document avec une certitude totale (1). Mais par cette description, il est sous-entendu que A et B sont tous les deux "probables". On peut donc déduire que les assertions [p]A et [p]B sont vérifiées dans le document.

Ainsi, (A ou B) dans le document est interprété par [p]A∧ [p]B.

De cette façon, un document peut être donc considéré comme un ensemble d'assertions, chaque assertion étant une arborescence sémantique (correspondant à une proposition atomique), la négation de celle-ci, ou une arborescence sémantique avec un opérateur de certitude. Ainsi, on a la syntaxe suivante pour un document:

$D ::= \{A_1, A_2, \ldots\}$

$A_i ::= P \in \mathcal{P} \mid \neg P \mid [prob]\ P$

P ::= une arborescence sémantique
[prob] ∈ {[p],[pn]}

On peut également donner la syntaxe suivante pour une requête:

$$Q ::= P \in \mathcal{P} \mid Q_1 \wedge Q_2 \mid Q_1 \vee Q_2 \mid \neg P \mid [prob]\ P \mid true$$

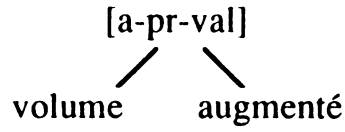
4.7.9. Quelques remarques sur l'interprétation des requêtes

• Résolution partielle du paraphrasage

Un mot représentant un concept complexe est décomposé, dans son interprétation interne, en une arborescence composée de concepts élémentaires et de relations sémantiques. Un tel mot est mis en équivalence avec une expression (un groupe de mots) qui sera interprétée par la même arborescence. On peut donc considérer que le problème du paraphrasage entre un tel mot et les expressions équivalentes est partiellement résolu.

Par exemple:
hypertrophie,
augmentation de volume,
volume augmenté
expansion du volume

seront tous interprétés par:



• Requêtes sur les concepts non-terminaux

Les utilisateurs peuvent aussi émettre des requêtes contenant des mots qui correspondent à des concepts non-terminaux, tels que le mot "signe". Par exemple:

Quels sont les CRM contenant des signes relatifs aux poumons?

Dans cet exemple, le mot *signes* ne correspond pas à un signe spécifique, mais à une classe sémantique. Ce mot ne doit pas être toujours considéré comme un mot vide comme dans l'indexation ([Berrut88]). Dans le cas de l'exemple, il doit être interprété relativement à la classe sémantique correspondant à la notion de signe (SGN):

[p-sur](SGN,poumon)

Si le mot "signe" est suivi d'un concept de la classe SGN, il peut être alors considéré comme un mot vide. Par exemple, *un signe d'opacité pulmonaire.*

5. EVALUATION DE LA REQUETE

L'essentiel du processus d'évaluation de RIME est la valuation de la *correspondance* entre un document (un CRM) et une requête. Examinons d'abord la signification de la *correspondance* dans le contexte de RIME.

Etant donnée une requête (Q) concernant des CRM satisfaisant certaines conditions sur des attributs externes et internes, un CRM (D) est considéré comme *correspondant* à Q quand il satisfait toutes les conditions sur les attributs. En termes de logique, le critère est la satisfaction de l'implication $D \rightarrow Q$ (le critère d'exhaustivité). L'autre critère - la spécificité - n'a pas d'importance dans cette application, car il n'est pas nécessaire que le CRM porte *exclusivement* sur le sujet de la requête pour qu'il réponde à la requête. En effet, RIME fait partie d'un type d'application où l'opération d'interrogation est fondée sur des besoins très précis. On peut comparer, en quelque sorte, ce type de SRI avec les systèmes question-réponse.

Ainsi, on peut définir la correspondance R entre un CRM et une requête de la façon suivante:

$$R(D,Q) = P(D \rightarrow Q)$$

Etant donnée une requête constituée de sous-requêtes externes et d'une sous-requête interne: $Q = Q_{EXT} \wedge Q_{INT}$, l'évaluation de la requête se fait en deux étapes en utilisant l'approche qui consiste à modifier le document. Considérons D comme le monde initial w_0, et la requête Q comme une formule f à valuer par rapport à ce monde w_0. Notons Q_{EXT} par f_{EXT}, et Q_{INT} par f_{INT}. Selon le modèle défini dans 3.3.3, on a donc:

$$\begin{aligned} P(D \rightarrow Q) &= V_{w_0}(f) = V_{w_0}(f_{EXT} \wedge f_{INT}) \\ &= MIN(V_{w_0}(f_{EXT}),\ V_{w_0}(f_{INT})) \end{aligned}$$

Sachant que la valuation de la vérification de la sous-requête externe est binaire ($\in \{0,1\}$), elle peut donc constituer une sorte de filtrage: seuls les documents satisfaisant la sous-requête externe vont être retenus pour la vérification de la sous-requête interne, car tout document ne satisfaisant pas la sous-requête externe donne pour valeur 0 à $V_{w_0}(f_{EXT})$. Il en est de même

pour $P(D \rightarrow Q)$. Les documents sélectionnés par la sous-requête interne vont alors prendre $V_{w_0}(f_{INT})$ comme valeur de la correspondance, car:

$$P(D \rightarrow Q) = MIN(V_{w_0}(f_{EXT}),\ V_{w_0}(f_{INT}))$$
$$= MIN(1,\ V_{w_0}(f_{INT})) = V_{w_0}(f_{INT}).$$

De ce fait, nous proposons le schéma d'évaluation suivant pour une requête:

- l'évaluation des attributs externes
- l'évaluation de l'attribut interne

Les attributs externes des CRM peuvent être organisés comme dans les bases de données (correspondance exacte), et évalués de manière classique (elle correspond à une évaluation du modèle booléen fondée sur le critère $D \rightarrow Q$). Comme nous l'avons montré dans la partie I, cette évaluation à l'aide des outils base de données est descriptible par le modèle général proposé.

Quant à l'évaluation de l'attribut interne, un mécanisme spécifique doit être conçu, car aucun outil existant n'est suffisant.

Pour évaluer l'attribut interne, deux approches sont envisageables selon le modèle général choisi. On retient l'évaluation considérant les documents comme des mondes initiaux pour la raison suivante:

Etant donnée une arborescence, il existe beaucoup plus d'arborescences spécifiques que les arborescences génériques.

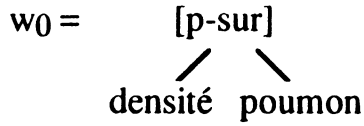
L'approche par modification de la requête pour l'évaluation de $P(D \rightarrow Q)$ consiste à transformer la requête initiale $w_0=Q$ en des requêtes w_i plus spécifiques qui sont à comparer avec le document D considéré comme une formule f à valuer. Pour une requête $w_0=Q$ donnée, il existe un nombre *infini* de w_i, et beaucoup d'entre elles correspondent à des descriptions qui ne peuvent pas exister dans la réalité. Ceci conduit à beaucoup de transformations inutiles.

Par exemple, soit la règle suivante (cf.II.6.2.2):

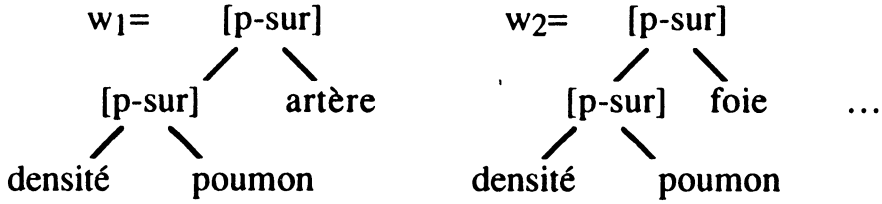
$$[\text{p-sur}](A,B) \Rightarrow A$$

Cette règle signifie qu'un concept A décrit par une localisation (B) est plus spécifique que le concept A sans la localisation.

Soit l'arborescence de la requête:



Les arborescences suivantes sont toutes plus spécifiques que la requête initiale et doivent être comparées avec le document:



Or la seconde arborescence (et beaucoup d'autres) ne peut pas exister dans la réalité. Elle ne peut donc correspondre à aucun document. La transformation de w_0 vers w_2 est donc superflue.

On peut éventuellement ajouter un contrôle supplémentaire pour interdire cette transformation. Mais cela représente un coût supplémentaire considérable pour le processus d'évaluation.

L'approche par modification du document consiste à transformer le document w_0=D en des documents w_n plus génériques. w_0 étant constitué des arborescences valides, on ne peut en déduire que des documents w_n valides. Ainsi, cette approche évite des transformations superflues. Elle est donc plus avantageuse par rapport à la précédente.

Dans cette optique, une évaluation consiste donc à effectuer des modifications sur les descriptions médicales du document en utilisant les connaissances du système.

5.1. Evaluation de la requête externe

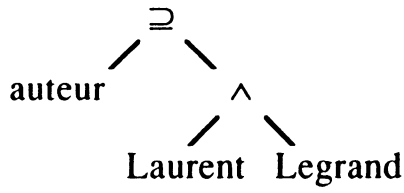
L'utilisation des bases de données dans le processus d'évaluation des SRI n'est pas récente. Certaines évaluations dans les SRI existants sont fondées sur les approches utilisées dans les systèmes de bases de données classiques ([Deogun86]). Dans ce cas, une certaine réorganisation est nécessaire, car tous les attributs des documents ne sont pas naturellement décrits en termes de valeur atomique. Par exemple, si un document est établi par plusieurs auteurs, l'attribut auteur doit correspondre à l'ensemble des auteurs, ce qui n'est pas à une valeur atomique. Dans une base de données relationnelle, cela conduirait à une duplication du tuple (une instance par auteur).

Les bases de données N1NF que l'on a présentées dans la partie I sont mieux adaptées à notre situation. Dans un modèle N1NF, il est possible de définir un attribut en tant qu'ensemble, ce qui nous permet de représenter naturellement les attributs tels que AUTEUR. Ainsi, on peut construire la relation CRM sur les attributs externes comme ci-dessous (où * marque un attribut du type ensemble):

CRM(	NUMERO,	PATIENT,	AUTEUR*,	DATE,	...)
ex:	0123	Dupont	{Laurent}	841207	
	1876	Durand	{Laurent,Legrand}	870811	

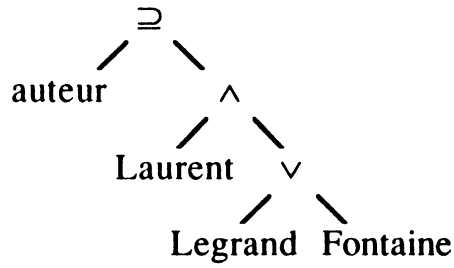
Les opérations proposées dans les modèles N1NF sont suffisantes pour évaluer les sous-requêtes externes sur l'ensemble des documents. On donne ci-dessous quelques exemples de requêtes transformées en des requêtes N1NF.

Exemples:
1. Donnez moi les documents établis par Laurent et Legrand.




```
=>    select  CRM
      where   AUTEUR* ⊇ {Laurent,Legrand}
```

2. Donnez moi les documents établis par Laurent, et Legrand ou Fontaine.




```
=>    select CRM
      where AUTEUR* ⊇ {Laurent, Legrand}
            or
            AUTEUR* ⊇ {Laurent, Fontaine}
```

Ces exemples montrent aussi qu'il existe une forte similitude de conception entre la représentation interne que nous avons choisie et la requête N1NF. Une représentation interne peut être donc très facilement convertie en une requête de bases de données par un automate approprié. Des travaux similaires ont été déjà effectués dans [Hendrix78, Waltz78,...].

L'évaluation des attributs externes dans le contexte d'une base de données classique s'avère donc plus difficile dans notre contexte: une transformation supplémentaire de la requête sous forme N1NF est nécessaire pour l'exprimer sous forme 1NF. Par exemple, si les CRM étaient organisés sous forme de relations classiques, la première requête correspondrait à la requête suivante des bases de données normalisées (1NF):

```
(select CRM
   where AUTEUR=Laurent
 intersect
   select CRM
   where  AUCTEUR=Legrand)
difference
(select CRM
  where     AUTEUR ≠ Laurent
          or AUTEUR ≠ Legrand)
```

ce qui est très complexe par rapport à la requête initiale.

L'évaluation des sous-requêtes externes sépare les CRM du corpus (C_{CRM}) en deux ensembles: ceux qui satisfont les conditions externes et ceux qui ne les satisfont pas. Notons l'ensemble des requêtes externes par Q_{EXT}. On représente les deux ensembles respectivement par: $S_{EXT}(Q_{EXT})$ et $N_{EXT}(Q_{EXT})$. La relation entre les deux ensembles est la suivante (où C_{CRM} est l'ensemble des CRM du corpus):

$$S_{EXT}(Q_{EXT}) \cup N_{EXT}(Q_{EXT}) = C_{CRM}$$

5.2. Evaluation de la requête interne

5.2.1. Modèle logique utilisé

Pour une raison d'efficacité, nous avons donc choisi l'approche par modification du document pour la valuation de $P(D \rightarrow Q)$, en utilisant le modèle de I.3.3.4.1 que l'on redéfinit de la façon suivante:

- $\mathcal{P}$ est l'ensemble de propositions atomiques. Une proposition atomique dans RIME est une arborescence sémantique.

Une formule bien formée (f) est définie comme suit:

$$f ::= P \in \mathcal{P} \mid f_1 \vee f_2 \mid f_1 \wedge f_2 \mid \neg P \mid \Diamond P \mid =_v P \mid \text{true}$$

$\mathcal{F}$ est constitué par toutes les formules bien définies: f$\in \mathcal{F}$.

En ce qui concerne une formule bien définie, nous pouvons faire les remarques suivantes:

1. Il n'existe pas de requête de la forme $\Diamond$f directement issue de l'interprétation d'une requête en langue naturelle. Cette requête n'est engendrée que par la valuation d'une proposition atomique (cf.la définition de $V_w(P)$).

2. La formule "true" correspond à une requête vide qui peut être satisfaite par un document quelconque.

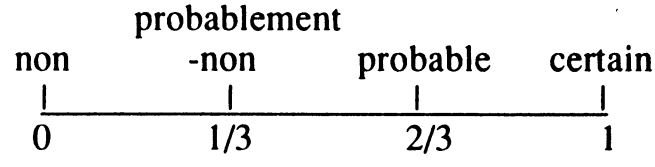
3. Dans l'interprétation précédente, nous avons isolé deux opérateurs de certitude: [p] et [pn]. Pour utiliser le modèle, il faut faire correspondre ces deux opérateurs avec l'opérateur $=_v$, c'est-à-dire qu'il faut définir une valeur v pour [p] et pour [pn].

Cette définition ne peut pas être établie de façon formelle, mais seulement de façon empirique. Néanmoins, elle doit permettre la vérification de certaines conditions:

- Pour préserver la cohérence avec le calcul du complément, si l'on choisit une valeur réelle p$\in$]0,1[pour représenter "probable", la certitude correspondant à "probablement non" doit être représentée par la valeur 1-p.

- Le choix de la valeur de p doit aussi remplir la condition (1-p) < p, car une description reconnue "probable" a plus de chance d'être vérifiée qu'une description reconnue "probablement fausse". Ainsi, 0,5 < p < 1.

Pour notre étude, nous donnons à p la valeur 2/3. Les certitudes que l'on peut retrouver dans les documents et les requêtes sont donc réparties de la manière suivante:



En résumé, nous définissons les correspondances suivantes:

[p] $\rightarrow$ $=_{2/3}$
[pn] $\rightarrow$ $=_{1/3}$

- $\mathcal{W}$ est l'univers du modèle, constitué d'un ensemble de mondes. La construction de $\mathcal{W}$ est présentée dans 6.2.3.

- $C_P(w) \in [0,1]$ est une fonction qui mesure la certitude de la vérification d'une proposition atomique ($P \in \mathcal{P}$) dans un monde (w). Cette fonction sera définie après avoir construit $\mathcal{W}$.

- δ: $\mathcal{W} \times \mathcal{W} \rightarrow [0, 1]$ est une valuation de la relation entre deux mondes. Elle vérifie les propriétés suivantes:

$$\delta(w,w)=0$$
$$\delta(w,w')>0 \wedge \delta(w',w'')>0 \Rightarrow \delta(w,w'')>0 \quad w,w',w'' \in \mathcal{W}$$

La définition de cette fonction dépend fortement de l'ensemble des connaissances définies dans RIME. Elle sera précisée dans 6.2.3, après la définition des connaissances utilisées.

- Δ: $[0,1]^2 \rightarrow [0,1]$ est une fonction qui combine la certitude d'une relation avec la certitude d'une vérification pour former une certitude d'une autre vérification. Cette fonction est définie comme la multiplication.

Relativement à cet opérateur, on peut définir un autre opérateur $\otimes$ pour que:

$$\Delta(\delta_1, \Delta(\delta_2, C)) = \Delta(\otimes(\delta_1, \delta_2), C)$$

L'opérateur $\otimes$ est également défini comme la multiplication.

- V: $\mathcal{F} \rightarrow [0,1]^{\mathcal{W}}$ est une fonction assignée à chaque monde ($w \in \mathcal{W}$), qui calcule la certitude de vérification d'une proposition ($f \in \mathcal{F}$) dans ce monde.

$V_w(f)$ est défini de la façon suivante:

- $V_w(P) = MAX[\ C_P(w),\ V_w(\Diamond P)]\ ,\ P \in \mathcal{P}$
- $V_w(f_1 \wedge f_2) = MIN(V_w(f_1), V_w(f_2))$
- $V_w(f_1 \vee f_2) = MAX(V_w(f_1), V_w(f_2))$
- $V_w(\neg P) = 1 - V_w(P)$
- $V_w(\Diamond P) = \underset{w'}{MAX}(\Delta[\delta(w,w'), V_{w'}(f)])$, avec $w' \in \mathcal{W}$
- $V_w(=_v P) = 1 - |V_w(P) - v|$
- $V_w(\text{true}) = 1$

Nous ne pouvons pas encore donner de définition précise à certains éléments du modèle, tels que $\mathcal{W}$, δ et $C_P(w)$. Leur définition dépend de l'application et donc des connaissances spécifiées dans l'application. Ainsi, par la suite, nous présentons d'abord les connaissances de RIME, avant de donner une définition précise des éléments du modèle.

5.2.2. Définition des connaissances dans RIME

Comme nous l'avons montré dans la partie I, les connaissances utilisées dans l'évaluation des requêtes sont des relations de dérivation. Ici, nous voulons établir un ensemble de relations de dérivation entre des assertions (descriptions médicales). Ces relations sont de la forme:

$$A_1 \Rightarrow_v A_2$$

ce qui signifie que l'assertion A_1 peut être dérivée vers A_2 avec la certitude de valeur v ($\in]0,1]$). Quand v=1, cette relation est représentée par $A_1 \Rightarrow A_2$.

1. Remarquons d'abord que les connaissances dans un domaine médical peuvent être très variées:

- celles-ci peuvent être des connaissances typiquement expertes, par exemple celles qui, à partir d'un signe, permettent de déduire un constat
- elles peuvent correspondre aux relations génériques entre concepts.

La définition du premier type de connaissances nécessite une compétence dans le domaine médical. Pour notre étude, nous avons adopté la seconde vision qui est une vision plus restreinte.

2. On remarque aussi que, dans le contexte de RIME, les relations de dérivation se divisent en deux types de relations différentes. Pour les introduire, nous distinguons d'abord deux types de description de concepts dans un CRM:

- la notion de *présence* ou de *description simple* d'un concept,
- et celle de *description sémantique.*

Une présence ou une description simple d'un concept signifie que le concept est vérifié dans le CRM, mais il n'est mis en relation sémantique avec aucun autre concept. C'est donc une description *isolée.*

Une description sémantique d'un concept signifie que le concept est lié par un opérateur sémantique avec un autre concept dans le CRM. C'est le cas

où le concept est un *concept principal* (situé sur la branche gauche d'un opérateur sémantique) d'une arborescence.

Par exemple, soit le concept "tumeur du poumon" (= [p-sur](tumeur, poumon)): il est décrit *isolément* dans un CRM et on constate simplement qu'une "tumeur du poumon" est *présente* (*décrit simplement*) dans le CRM. Il n'est mis en relation sémantique avec aucun autre concept. Par contre, le concept "tumeur" dans cette description est lié par une *relation sémantique* ([p-sur]) avec un autre concept ("poumon"). On dit alors que le concept "tumeur" est *décrit sémantiquement.*

Correspondant à ces deux notions, nous définissons deux relations "génériques" entre les concepts:

1. Si la présence d'un premier concept sous-entend celle d'un second concept, le premier *peut être dérivé* (noté par $\Rightarrow$) vers le second. C'est la même *relation de dérivation* que celle décrite dans la partie I.

2. Si dans une description sémantique quelconque,

   a). un premier concept peut être remplacé par un second sans violer les contraintes sémantiques,

   b). l'arborescence obtenue après le remplacement représente une sémantique plus générique que celle de l'arborescence initiale,

   on dit que le premier peut être *substitué* (noté par $\equiv>$) au second. On appelle la relation entre les deux concepts une *relation de substitution.* Cette relation est plus stricte que la relation de dérivation. Cette relation est comparable avec la relation de *congruence* dans les études des mathématiques discrètes (cf.[Stanat77, Hindley86]). Nous les comparerons dans la section 5.2.2.5.

L'utilité de cette seconde relation est de pouvoir établir la relation de dérivation dans le contexte suivant:

Soit une arborescence dans laquelle un composant peut être substitué à un autre. L'arborescence obtenue après la substitution est une arborescence pouvant être dérivée de l'arborescence initiale.

Par exemple:

$$[rel](A,B) \Rightarrow [rel](A',B) \qquad \text{si } A \equiv> A'$$

ce qui peut conduire à la relation de dérivation suivante:

[p-sur]([a-pr-val]("opacité",tissulaire),poumon)
⇒ [p-sur]("opacité",poumon)

car [a-pr-val]("opacité",tissuliare) ≡> "opacité".

Ayant défini ces deux relations, nous réalisons l'analyse par rapport à chaque type d'arborescence.

5.2.2.1. Relations entre concepts élémentaires

Il existe dans RIME certaines relations génériques entre les concepts élémentaires. Ces relations constituent un ensemble particulier de relations de dérivation et de substitution.

Pour notre étude, nous avons repris les relations génériques telles qu'elles sont étudiées dans REMEDE ([Dachelet82]) (cf. Annexe 5). Nous donnons quelques exemples ci-dessous:

tronc
- *thorax*
  - *médiastin*
  - *côte*
  - ...
- *dos*
  - *rachis*
    - ...

Une relation générique dans ce cas est évidemment une relation de dérivation, car la présence d'un concept élémentaire plus spécifique sous-entend celle du concept élémentaire plus générique. Ainsi, on a:

médiastin ⇒ thorax
côte ⇒ thorax
thorax ⇒ tronc
...

Ces relations génériques satisfont aussi les deux conditions suffisantes pour la relation de substitution:

- un concept élémentaire plus générique et un concept élémentaire plus spécifique appartiennent à la même classe sémantique

- la sémantique correspondant à un concept élémentaire plus spécifique implique celle d'un concept plus générique.

Ainsi, ces relations génériques sont aussi des relations de substitution:

médiastin ≡> thorax
côte ≡> thorax
thorax ≡> tronc
...

5.2.2.2. Dérivation entre arborescences sémantiques

Etant donnée une arborescence [rel](A,B) décrite dans un CRM, la description de cette relation sémantique implique l'existence de A et de B. On peut donc concrétiser la relation de dérivation comme suit (pour un opérateur sémantique [rel] quelconque):

$$[rel](A,B) \Rightarrow A \qquad [rel](A,B) \Rightarrow B$$

Un exemple de la relation de dérivation est le suivant:

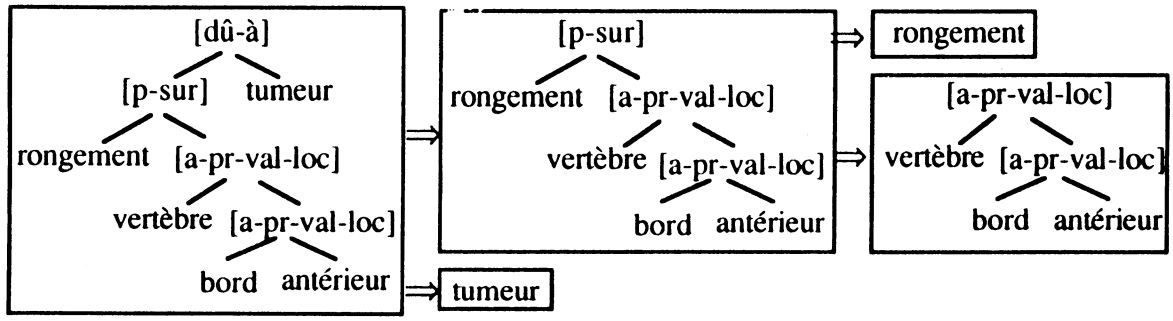


Fig.II.3. *Relation de dérivation*

ce qui signifie que la présence de "rongement du bord antérieur de la vertèbre par la tumeur" sous-entend la présence de "rongement du bord antérieur de la vertèbre", et la présence de celui-ci sous-entend la présence de "rongement" et du "bord antérieur de la vertèbre".

5.2.2.3. Substitution entre arborescences sémantiques

Comparons maintenant les conditions portant sur une relation de substitution et les propriétés attachées à la description d'un concept *gouverneur*.

Soit une arborescence sémantique [rel](A,B) dans laquelle A est un gouverneur:

1. Le concept gouverneur appartient à la même classe sémantique que toute la sous-arborescence. C'est une condition suffisante pour que le remplacement de l'arborescence par le gouverneur dans une arborescence quelconque, n'engendre pas de violation des contraintes sémantiques (condition 1 de la relation de substitution).

2. Le concept gouverneur A peut constituer seul une description moins spécifique que toute la proposition [rel](A,B), car la description du gouverneur A par son dépendant (B) ne fait que lui ajouter une spécification supplémentaire. La sémantique de [rel](A,B) est plus spécifique que celle de A. Cette condition garantit que, si la proposition [rel](A,B) apparaît comme un composant d'une autre assertion, le remplacement de [rel](A,B) par son gouverneur A dans l'assertion aboutira à une nouvelle assertion *dérivée* de l'assertion initiale.

Ainsi, nous constatons la relation de substitution suivante:

Si A est un gouverneur dans [rel](A,B), alors [rel](A,B)≡>A

ou classe(A)=classe([rel](A,B)) ⇒ [rel](A,B)≡>A.

Un exemple de relation de substitution est le suivant:

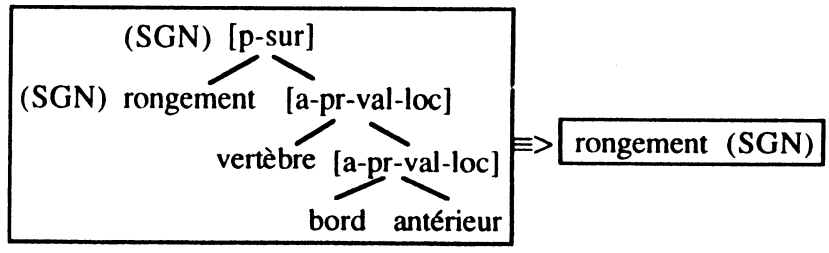


Fig.II.4. *Relation de substitution*

Cette relation signifie que, si la description "rongement du bord antérieur de la vertèbre par une tumeur" apparaît dans un document, alors on peut dériver l'assertion "rongement par une tumeur".

Dans la section 5, on a donné une liste complète de toutes les descriptions contenant un concept gouverneur. On peut également établir une liste correspondante pour les relations de substitution:

[en-rel-topo-avec](CONSTAT, CONSTAT') ≡> CONSTAT
[dû-à](CONSTAT, CONSTAT') ≡> CONSTAT
[dû-à](CONSTAT, DIAG) ≡> CONSTAT

[en-rel-topo-avec](DIAG, DIAG') ≡> DIAG
[dû-à](DIAG, DIAG') ≡> DIAG

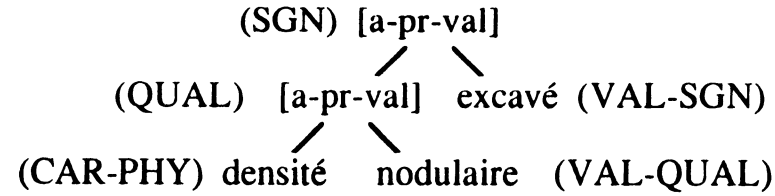
[en-rel-topo-avec](SGN, LOC) ≡> SGN
[p-sur](SGN, LOC) ≡> SGN
[p-sur](SGN, FCT) ≡> SGN
[a-pr-val](SGN, QUAL) ≡> SGN
[a-pr-val](SGN, VAL-SGN) ≡> SGN

[en-rel-topo-avec](LESION, LOC) ≡> LESION
[p-sur](LESION, LOC) ≡> LESION
[p-sur](LESION, FCT) ≡> LESION
[a-pr-val](LESION, QUAL) ≡> LESION

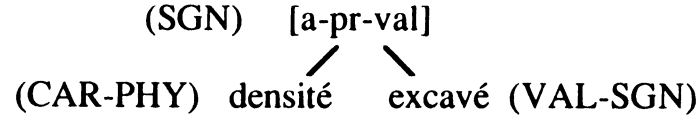
[a-pr-val-loc](LOC, POS) ≡> LOC
[en-rel-topo-avec](LOC, LOC) ≡> LOC

[a-pr-val-loc](ORG, REGION) ≡> ORG
[a-pr-val-loc](ORG, DETAIL) ≡> ORG

Les relations entre une arborescence et son gouverneur constituent les conditions suffisantes mais non nécessaires pour la relation de substitution. En effet, les deux conditions relatives aux relations de substitution peuvent être satisfaites par d'autres relations moins strictes que des descriptions de gouverneur. Relativement à la première condition, le remplacement d'un concept par un autre de classe différente ne conduit pas nécessairement à une violation des contraintes sémantiques. Par exemple, si dans l'arborescence suivante,



on remplace "[a-pr-val](densité,nodulaire)" par "densité", l'arborescence suivante est obtenue:



On constate que l'arborescence obtenue a préservé la vérification des contraintes sémantiques, et qu'elle représente une sémantique plus générique que celle de l'arborescence initiale. Dans ce contexte particulier, il existe donc une relation de substitution entre [a-pr-val](CAR-PHY,VAL-QUAL) et CAR-PHY.

Une extension de la relation de substitution dans ce sens nécessite non seulement une étude beaucoup plus élaborée, mais aussi une compétence dans le domaine médical. Dans le cadre de notre étude, nous nous restreignons aux cas suivants:

[a-pr-val](CAR-PHY,VAL-QUAL) ≡> CAR-PHY
[a-pr-val](CAR-PHY,VAL-QUAN) ≡> CAR-PHY
[a-pr-val](CAR-PHY,VAL-SGN) ≡> CAR-PHY
[a-pr-val](QUAL,VAL-SGN) ≡> QUAL
[p-sur](CAR-PHY,LOC) ≡> CAR-PHY
[p-sur](CAR-PHY,FCT) ≡> CAR-PHY
/*les règles ci-dessus sont valides sous la condition suivante: après la substitution, l'arborescence peut constituer un concept sans violer les contraintes sémantiques.*/
[a-pr-val-loc](ORG,REGION) ≡> REGION
[a-pr-val-loc](ORG,DETAIL) ≡> DETAIL
[a-pr-val-loc](REGION,DETAIL) ≡> DETAIL

On peut remarquer que la condition pour que les deux substitutions soient possibles est une sorte de "*look forward*", qui prend en compte les niveaux supérieurs.

5.2.2.4. Relations liées aux opérateurs de certitude

Soit l'assertion ($=_v$ A) dans un document. Cette assertion signifie que A est vérifié dans le document avec la certitude v. Si l'on veut transformer cette assertion en une assertion certaine A, la transformation apporte donc la certitude v. Autrement dit, on a la relation suivante:

$$(=_v A) \Rightarrow_v A$$

Remarquons aussi que ($=_v$A) et A ont la même classe sémantique. Ainsi, cette relation de dérivation est également une relation de substitution:

$$(=_v A) \equiv>_v A$$

5.2.2.5. Propriétés

• La relation de substitution est une relation de dérivation particulière.

$$\text{si } A \equiv>_v B \text{ alors } A \Rightarrow_v B$$

En effet, une relation de substitution remplit des conditions plus strictes qu'une relation de dérivation. Une des conditions pour la relation de substitution entre A et B est que A peut être dérivé en B.

• Les relations de dérivation et de substitution sont non-réflexives.

$$A \not\equiv> A \qquad\qquad A \not\Rightarrow A$$

La définition de cette propriété sur les relations est destinée à éviter des transformations inutiles: les transformations d'une arborescence en elle-même. Elle peut donc augmenter l'efficacité du processus de la vérification détaillée. Cette définition correspond bien à la propriété de non-réflexivité définie sur le modèle présenté dans la partie I ($\delta(w,w)=0$).

• Les relations de dérivation et de substitution ne sont pas symétriques.

$$\text{si } A \equiv> B \text{ alors } B \not\equiv> A$$
$$\text{si } A \Rightarrow B \text{ alors } B \not\Rightarrow A$$

Cette propriété, avec la précédente, permet d'éviter les boucles dans les transformations: des transformations d'un monde vers lui-même.

• Les deux relations sont dépendantes.

Soit une description dans document contenant A comme composant. Si $A \equiv> B$, on peut alors remplacer A par B dans cette description pour obtenir une autre description dérivée de la description initiale. Ainsi, nous pouvons avoir, en général, la déduction suivante:

Soit $\lambda_C(A/B)$ le résultat du remplacement de A par B dans C.

$$\text{Si } A \equiv>_v B \text{ et A est un composant de C, alors } C \Rightarrow_v \lambda_C(A/B)$$

La définition de cette liaison entre les deux relations permet de ramener la relation de substitution sous la forme d'une relation de dérivation entre propositions atomiques.

Remarquons que si C et $\lambda_C(A/B)$ ont la même classe sémantique, la relation de dérivation précédente est aussi une relation de substitution. On a donc:

Si $A \equiv>_v B$, A est un composant de C et $\lambda_C(A/B)$ a la même classe que C, alors $C \equiv>_v \lambda_C(A/B)$

Cela implique, dans une certaine mesure, la propriété de transitivité de la relation ≡>, c'est-à-dire qu'une succession de substitutions dans une arborescence conduit à une arborescence pouvant substituer l'arborescence initiale. Dans ce cas, la relation de substituion est une relation de congruence (cf.Stanat77, Hindley86]), qui possède la propriété suivante:

Soit A≡>B une relation quelconque. Si pour un contexte C[A] quelconque qui contient A, C[A]≡>C[B] est aussi vérifiée, alors ≡> est une relation de congruence (sur le contexte C).

5.2.3. Définition du modèle dans RIME

Il nous reste à préciser la définition des éléments suivants pour que le modèle soit complètement défini:

- l'univers du modèle $\mathcal{W}$
- la fonction $C_P(w)$
- la relation $\delta(w,w')$ entre deux mondes.

5.2.3.1. Construction de l'univers $\mathcal{W}$

Ayant défini l'ensemble des connaissances de RIME (noté par K), nous pouvons construire l'univers du modèle $\mathcal{W}$ de la façon suivante:

1. Un document D existant dans le système constitue un monde initial. Ce monde appartient à $\mathcal{W}$;

2. Si $w \in \mathcal{W}$, $A_1 \in w$ et $(A_1 \Rightarrow_v A_2) \in K$, alors $w'=\lambda_w(A_1/A_2) \in \mathcal{W}$.

L'univers ainsi construit est donc constitué de l'ensemble des mondes initiaux correspondant aux documents existants, et de l'ensemble des mondes possibles dérivables à partir des mondes initiaux en utilisant les connaissances définies.

Correspondant à la structure des documents obtenue après la restructuration, on peut définir la structure suivante pour un monde:

$w ::= \{A_1, A_2, \ldots, \}$
$A ::= A' \mid \neg A' \mid (=_v A')$
A' ::= une arborescence sémantique.
$v \in]0,1[$

5.2.3.2. Définition de la fonction Cp(w)

Cette fonction mesure la validité d'une proposition atomique dans un monde. C'est une valuation directe sans utiliser les relations de dérivation.

Etant donnée une proposition atomique P à valuer, la valuation de cette fonction est déterminée selon l'existence ou non d'une assertion dans w pouvant correspondre à P ou ($=_v$P):

$$Cp(w) = \begin{cases} 1, & \text{si } P \in w \\ v, & \text{si } (=_v P) \in w \\ 0, & \text{autrement.} \end{cases}$$

5.2.3.3. Définition de l'opérateur δ

Soient deux mondes w et w' dans $\mathcal{W}$.

Si $A_1 \in w$, $(A_1 \Rightarrow_v A_2) \in K$ et $w' = \lambda_w(A_1/A_2)$,
alors $\delta(w,w') = v$.

Si l'on remplace une assertion dans un monde par une assertion dérivée, on obtient un monde possible. La certitude de la relation entre les deux mondes est déterminée par la certitude de la relation de dérivation utilisée pour la dérivation du monde possible.

Si l'on considère les relations de substitution séparément des relations de dérivation, on peut calculer la relation entre mondes directement à partir de celles-ci:

Si $(A_1 \equiv >_v A_2) \in K$, A_1 est un composant de $A \in w$,
$A' = \lambda_A(P_1/P_2)$, et $w' = \lambda_w(A/A')$
alors $\delta(w,w') = v$.

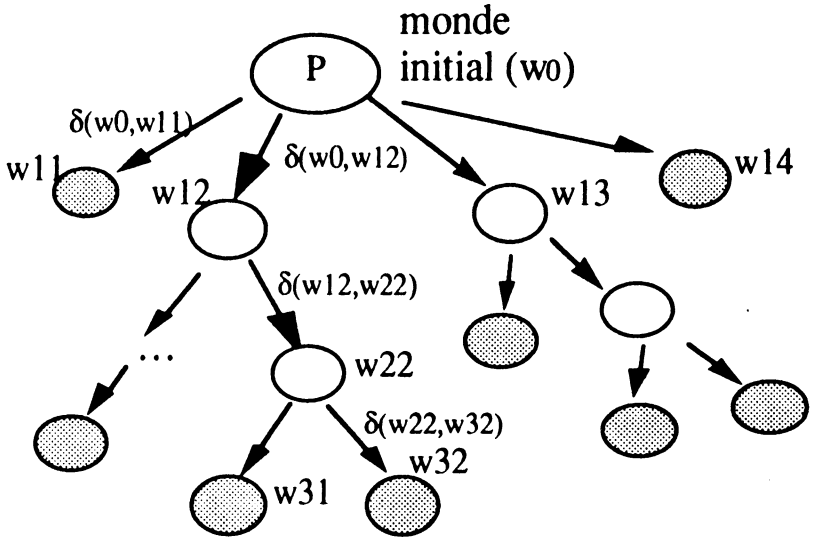
Si l'on remplace un composant A_1 d'une assertion A de w par A_2, et que A_1 peut être substituée à A_2, alors le monde w' obtenu est un monde possible de w, la certitude de la relation entre deux mondes étant déterminée par la certitude de la relation de substitution entre A_1 et A_2.

5.2.4. Méthode d'évaluation proposée par le modèle

Etant donnés un document considéré comme un monde initial et une requête considérée comme une formule à valuer par rapport à ce monde

initial, le modèle propose de décomposer la valuation de la formule en celles des propositions atomiques constituant la formule.

Pour la valuation d'une proposition atomique P par rapport à un monde initial w_0: $V_{w_0}(P)$, w_0 doit être transformé en tous ses mondes possibles, le processus se répétant pour chaque monde obtenu, en utilisant les connaissances définies. Comme nous l'avons montré dans la partie I, cette valuation correspond à l'établissement d'une arborescence de transformation, dans laquelle les noeuds correspondent aux mondes possibles et les liens aux transformations:



Par rapport à chaque noeud w_i de l'arborescence, on obtient une valeur de certitude pour $V_{w_0}(P)$ via le chemin de transformation depuis w_0 jusqu'à w_i (cf.I.3.3.3):

$$V_{w_0}^{w_i}(P) = \Delta[\delta^i(w_0,w_i), C_P(w_i)]$$
$$= \prod_{j=0}^{i-1} \delta(w_j, w_{j+1}) \cdot C_P(w_i)$$

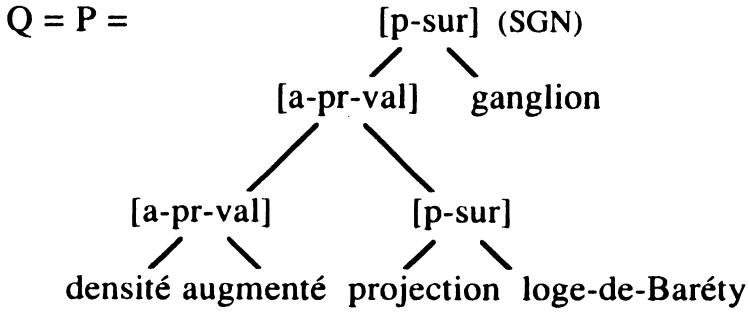
Parmi toutes ces valeurs de certitude obtenues par rapport à chaque noeud w_i de l'arborescence de transformation, la valeur maximale est assignée à $V_{w_0}(P)$ (cf.formules (5) et (5') de la partie I):

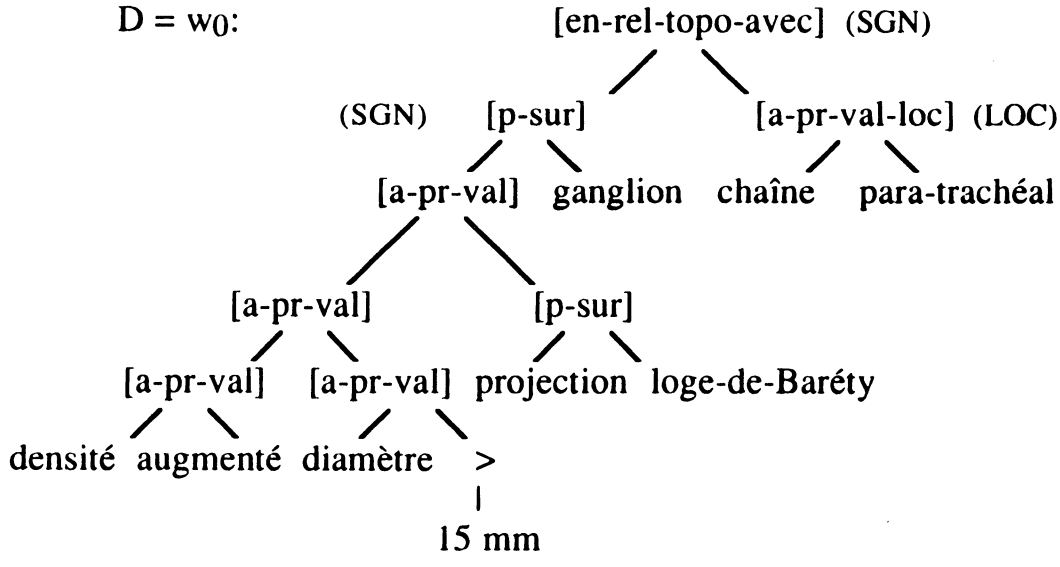
$$V_{w_0}(P) = \max_{w_i \in \text{arborescence}} (V_{w_0}^{w_i}(P))$$
$$= \max_{w_i \in \text{arborescence}} (\prod_{j=0}^{i-1} \delta(w_j, w_{j+1}) \cdot C_P(w_i))$$

On montre ci-dessous quelques exemples de comparaison entre une arborescence de document et une arborescence de requête.

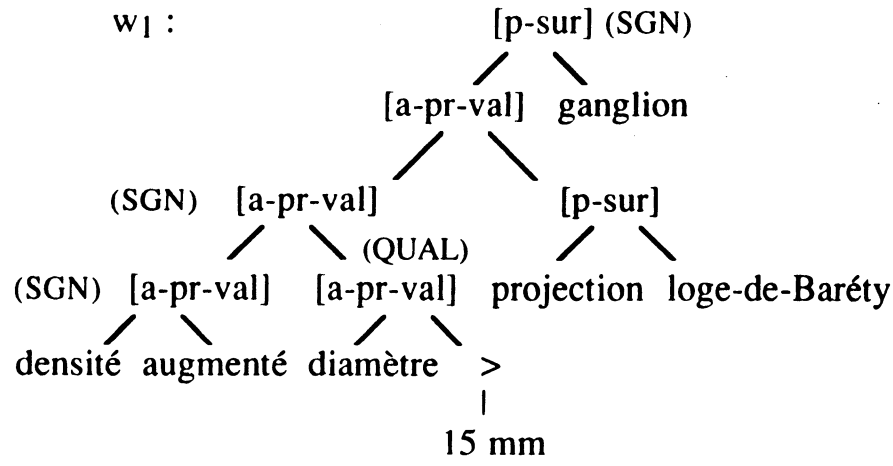
Soit la requête sur "opacité ganglionnaire en projection de la loge de Baréty" et un document décrivant une "opacité ganglionnaire de plus de 15 mm de diamètre en projection de la loge de Baréty intéressant la chaîne para-trachéale".

La requête et le document sont respectivement interprétés par les arborescences suivantes:

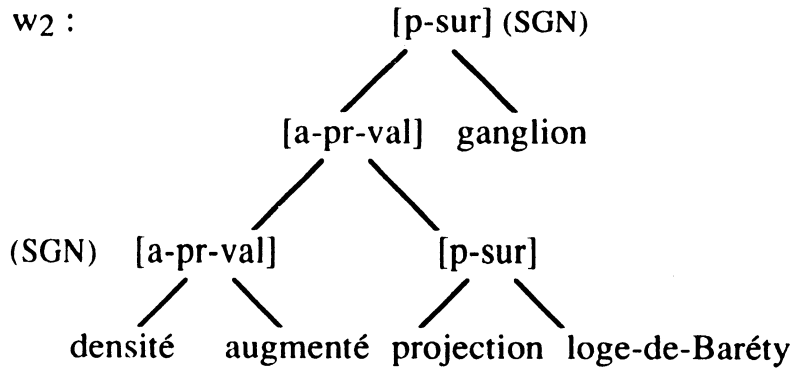




La description A constitue un monde initial. Ce monde peut être transformé en plusieurs mondes possibles, parmi lesquels, le suivant, obtenu en utilisant la relation de dérivation [en-rel-topo-avec](SGN,LOC)⇒SGN:



Dans la partie gauche de cette arborescence, on peut appliquer la relation de substitution [a-pr-val](SGN,QUAL) ≡> SGN, cette substitution conduisant à l'arborescence suivante dérivable (avec une certitude égale à 1) de l'arborescence précédente:



Dans ce monde, $C_P(w_2)$ est valuée à 1, puisque cette arborescence est identique à celle de P. Ainsi, par ce chemin de transformation, on donne la valeur suivante à $V^{w_2}_{w_0}(P)$:

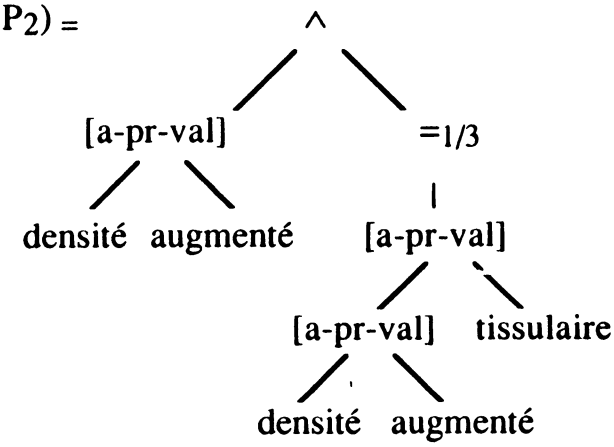
$$V^{w_2}_{w_0}(P) = \prod_{j=0}^{2} \delta(w_j, w_{j+1}) \cdot C_P(w_2) = 1$$

Ce chemin de transformation est un chemin optimal qui donne la valeur maximale (1) à $V_{w_0}(P)$.

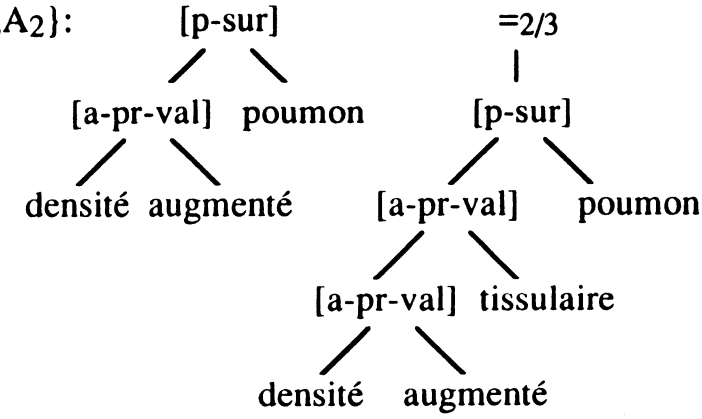
Nous montrons un autre exemple concernant l'opérateur de certitude.

Soient la requête portant sur "opacité probablement non tissulaire" et le document décrivant "une opacité pulmonaire probablement tissulaire". Ils sont représentés par les arborescences ci-dessous:

$Q = f = P_1 \wedge (=_{1/3} P_2) =$



$D=w_0=\{A_1,A_2\}$:

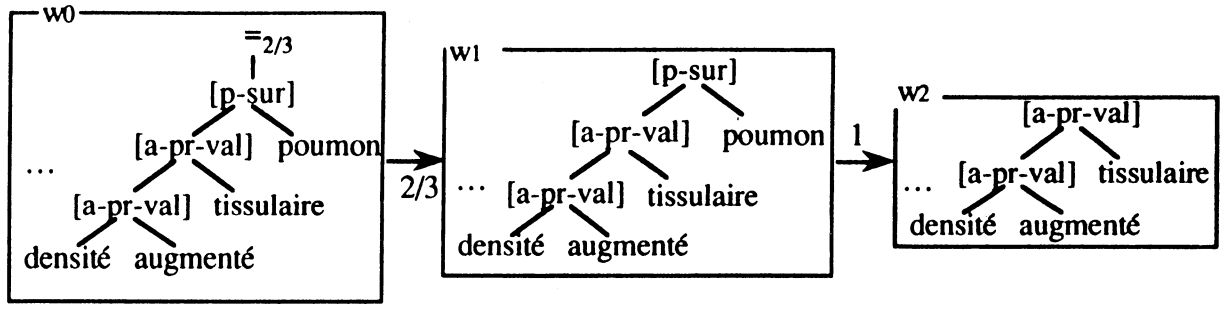


La valuation de la formule f se décompose en la valuation des deux propositions atomiques P_1 et P_2. La valuation de la première proposition atomique P_1 est tout à fait similaire à l'exemple précédent. Elle est vérifiée par la première assertion du document. On a donc $V_{w_0}(P_1)=1$.

On détaille par la suite la valuation de la seconde proposition atomique P_2.

Les transformations sur la première assertion A_1 de D ne peuvent évidemment pas aboutir à la satisfaction de P_2.

Considérons seulement la seconde assertion A_2. Cette assertion peut être progressivement transformée de la façon suivante (on ne montre que les transformations utiles pour la vérification de P_2):



Les chemins de transformation donnent les valeurs suivantes à $V^{w_i}_{w_0}(P_2)$:

via $w_0 \rightarrow w_1$: $\quad V^{w_1}_{w_0}(P_2) = \delta(w_0,w_1)\cdot C_P(w_1) = \frac{2}{3}\cdot 0 = 0$

via $w_0 \rightarrow w_1 \rightarrow w_2$: $\quad V^{w_2}_{w_0}(P_2) = \delta(w_0,w_1)\cdot\delta(w_1,w_2)\cdot C_P(w_2)$

$$= \frac{2}{3}\cdot 1 \cdot 1 = \frac{2}{3}$$

Parmi toutes ces valuations, la valeur maximale est assignée à $V_{w_0}(P_2)$. Ainsi:

$$V_{w_0}(P_2) = \underset{A_i \in \text{arborescence}}{MAX}(V^{w_i}_{w_0}(P_2)) = \frac{2}{3}.$$

La valuation de $(=_{1/3}P)$ est:

$$V_{w_0}(=_{1/3}P_2) = 1 - |V_{w_0}(P_2) - \frac{1}{3}| = 1 - |\frac{2}{3} - \frac{1}{3}| = \frac{2}{3}.$$

Ainsi, la valuation de f est la suivante:

$$V_{w_0}(f) = MIN(V_{w_0}(P_1), V_{w_0}(=_{1/3}P_2)) = \frac{2}{3}$$

5.2.5. Application pratique de la méthode dans RIME

Par rapport à la méthode proposée par le modèle, l'application dans RIME est plus restreinte. Cette restriction nous permet de développer un algorithme plus efficace dans un cas particulier. Les restrictions portent notamment sur les faits suivants:

1. Soient une proposition atomique P à valuer et un document contenant un ensemble d'assertions. Pour que P soit vérifiée dans D, il faut qu'une de ses assertions vérifie la proposition. Ainsi, nous pouvons effectuer une vérification sur chaque assertion. Autrement dit, la première opération consiste à transformer le monde initial constitué d'un document complet en

des mondes possibles correspondant à chacune des assertions contenues dans le document. La certitude de la relation entre le monde initial et ces mondes possibles est égale à 1.

Pour permettre cette opération, il faut élargir la définition de l'univers du modèle $\mathcal{W}$ de façon à ce que chaque assertion d'un document puisse constituer à elle seule un monde possible.

Pour des raisons de simplicité, nous notons par la suite, $\{A_i\}$ le monde constitué de l'assertion A_i.

On peut donc décomposer de la façon suivante la valuation d'une proposition $V_{w_0}(P)$ par rapport à un monde initial correspondant à un document:

$$V_{w_0}(P) = \underset{A_i \in w_0}{MAX}\ V_{\{A_i\}}(P), \text{ avec } \{A_i\}=w_i \in \mathcal{W} \text{ et } \delta(w_0,w_i)=1.$$

2. Nous faisons l'hypothèse que les phénomènes non décrits dans un document sont faux par rapport au document. Cela implique que tout phénomène non décrit est considéré comme si sa négation était décrite dans le document.

Cette hypothèse est fondée sur l'idée que les CRM et les requêtes ne concernent que des phénomènes anormaux. Ainsi, la négation d'une description ou une description non spécifiée dans le document correspond à un phénomène "normal".

Avec cette hypothèse, une proposition atomique (décrivant un phénomène anormal) ne peut pas être satisfaite par la négation d'une description ou une description non spécifiée dans le document (correspondant à un phénomène normal). Ainsi, la valuation de toute proposition atomique par rapport à une négation donne 0.

L'hypothèse faite est quelque peu simpliste, quoi qu'elle corresponde bien aux cas que l'on a rencontrés. Plus généralement, un médecin peut très bien s'intéresser aux phénomènes "normaux". Les phénomènes non décrits devraient être considérés comme étant normaux, c'est-à-dire correspondant à la "norme".

Par exemple, si "opacité pulmonaire" n'est pas décrite (explicitement ou implicitement) dans un document, on devrait considérer que la "densité du poumon" correspond à la valeur normale. Si l'on pose une requête demandant les documents decrivant "un poumon de densité normale", ce document devrait figurer dans la réponse.

Or, cette référence à la notion de normalité a besoin d'un ensemble de connaissances supplémentaires pour décrire la "norme". Ces connaissances n'étant pas encore disponibles actuellement dans RIME, nous avons donc fait cette hypothèse simpliste.

3. Dans la pratique, l'opérateur MAX dans la valuation de $V_w(P)=MAX(C_P(w),V_w(\Diamond P))$ joue également un rôle de contrôle des transformations:

si la proposition est directement vérifiée avec la valeur de certitude 1 dans un monde (i.e. $C_P(w)=1$), il n'est plus nécessaire de continuer les transformations sur ce monde en ses mondes possibles, car la valeur finale de $V_{w_0}(P)$ sera 1.

En effet, si $C_P(w)\neq 0$ (w ne contient qu'une assertion), les transformations sur l'assertion de w ne sont plus nécessaires, car on a déjà ramené l'assertion à une forme directement comparable avec P. Les transformations supplémentaires n'aboutiront pas à une valeur de $V_w(P)$ supérieure à $C_P(w)$.

Vérifions cela dans le cas où $0<C_P(w)<1$ avec $w=\{A\}$.

Si $C_P(\{A\}) = v$, cela implique que, $A= (=_v P)$. Dans ce cas, on peut continuer à transformer $A= (=_v P)$ en P. Cette transformation a une certitude v. La valeur assignée à $V_{\{A\}}(P)$ est donc $\Delta[\delta(\{A\},\{P\}),C_P(\{P\})] = v$. Elle est identique à $C_P(\{A\})$.

4. Dans le cas où l'assertion A et la proposition atomique P correspondent toutes les deux à une arborescence sémantique, la valuation de $V_{\{A\}}(P)$ est binaire, car les transformations sur une arborescence sémantique ont toujours une valeur de certitude binaire. Il en est de même pour la validité $C_P(w)$ d'une proposition par rapport à une telle assertion.

5. Si l'assertion $A = [rel](A_1,A_2)$ et la proposition $P = [rel'](P_1,P_2)$, alors P peut être vérifiée (avec certitude 1) dans le monde constitué par A dans les trois cas suivants:

- si [rel]=[rel'], et A_1 peut être successivement substituée à P_1 (ou si A_1 est identique à P_1), et A_2 peut être successivement substituée à P_2 (ou si A_2 est identique à P_2), alors P peut être vérifiée dans {A},

- étant donné que $[rel](A_1,A_2) \Rightarrow A_1$, si P peut être vérifiée dans $\{A_1\}$ alors P peut aussi être vérifiée dans {A},

- étant donné que [rel](A_1,A_2) $\Rightarrow$ A_2, si P peut être vérifiée dans {A_2} alors P peut aussi être vérifiée dans {A}.

En effet, dans les deux derniers cas, nous avons utilisé les relations de dérivation [rel](A_1,A_2) $\Rightarrow$ A_1 et [rel](A_1,A_2) $\Rightarrow$ A_2.

Dans le premier cas, si A_1 peut être successivement substituée à P_1, alors [rel](P_1,A_2) constitue un monde possible ayant une relation certaine avec le monde {A}. Si A_2 peut être aussi substituée à P_2, alors [rel](P_1,P_2) constitue un monde possiblement possible ayant une relation certaine avec {A}. Si [rel]=[rel'], alors P sera vérifiée dans ce monde.

Après cette analyse, nous pouvons proposer un algorithme spécifique à RIME, qui correspond à l'idée suivante:

Soient l'assertion du document A et la proposition atomique P. Elles peuvent prendre les formes suivantes:

A ::= A' | $\neg$A' | $=_v$A' (où A' est une arborescence sémantique)
A'::= concept élémentaire | [rel](A'$_1$,A'$_2$)
et P ::= concept élémentaire | [rel](P_1,P_2)

0. Si $C_P(\{A\}) \neq 0$ alors $V_{\{A\}}(P) = C_P(\{A\})$.

1. Sinon, si A= ($=_v$A'), alors $V_{\{A\}}(P) = v \cdot V_{\{A'\}}(P)$

2. Sinon, si A = $\neg$A', A ne peut satisfaire aucune proposition. Ainsi, $V_{\{A\}}(P)=0$.

Sinon, A est une arborescence sémantique (A').

3. Si P et A sont des concepts élémentaires, alors si A=P ou A$\Rightarrow$P, $V_{\{A\}}(P)=1$, sinon $V_{\{A\}}(P)=0$.

4. Si A est un concept élémentaire et P ne l'est pas, alors $V_{\{A\}}(P)=0$.

5. Si P et A sont des arborescence sémantiques, soient A=[rel](A'$_1$,A'$_2$) et P=[rel'](P$_1$,P$_2$),

$V_{\{A\}}(P) = MAX([SUB(A'_1,P_1) \cdot SUB(A'_2,P_2)], V_{\{A'1\}}(P), V_{\{A'2\}}(P))$

où SUB(A_1,P_1) est une fonction qui mesure la possibilité de substituer successivement A_1 à P_1. Cette fonction est définie par la suite.

6. Si P est un concept élémentaire et A =[rel](A'$_1$,A'$_2$), alors $V_{\{A\}}(P)$ = MAX($V_{\{A'1\}}(P)$, $V_{\{A'2\}}(P)$).

La fonction SUB(A,P) est binaire, car cette fonction n'est appelée que pour vérifier si une sous-arborescence sémantique peut être substituée par une autre. Les relations de substitution définies entre arborescences sémantiques sont toutes binaires. Cette fonction est valuée de la façon suivante:

1. si A = P alors SUB(A,P) = 1.

2. sinon, si (A ≡> P)∈ K (c'est-à-dire que A et P correspondent aux deux éléments d'une relation de substitution), alors SUB(A,P)=1.

   Le cas où A et P sont des concepts élémentaires, est aussi inclus dans celui-ci.

3. sinon, si A est constituée d'un concept élémentaire, et P ne l'est pas, alors SUB(A,P)=0.

4. sinon, A et P sont des arborescences sémantiques. Soit A=[rel](A'$_1$,A'$_2$) et P=[rel'](P$_1$,P$_2$), on a deux possibilités pour substituer A à P:

   a). si A'$_1$ peut être substituée à P$_1$, A'$_2$ peut être substituée à P$_2$ et [rel]=[rel'], alors A peut être substituée à P. Dans ce cas:

   $$SUB_1 = SUB(A'_1,P_1) \cdot SUB(A'_2,P_2).$$

   b). dans le cas où [rel](A'$_1$,A'$_2$) ≡> A$_1$, si A'$_1$ peut être substituée à P, alors A peut aussi être substituée à P. Ainsi, dans ce cas:

   $$SUB_2 = SUB(A'_1,P).$$

   Parmi ces deux possibilités, on choisit la valeur maximale pour l'assigner à SUB(A,P):

   $$SUB(A,P) = MAX(SUB_1, SUB_2).$$

Ces analyses correspondent à une méthode de valuation "top-down": pour vérifier une proposition atomique constituée d'une arborescence, on cherche d'abord à vérifier l'opérateur, ensuite les deux opérandes. Si l'opérateur est vérifié, il est nécessaire que les deux opérandes de l'assertion puissent être substituées respectivement aux deux opérandes de la proposition, pour que la proposition soit vérifiée dans son intégralité par l'assertion.

Correspondant à ces analyses, on peut écrire les algorithmes suivants:

```
fonction: V{A}(P): ∈ [0,1]
début
 c ← CP(A)
 si c > 0 alors V ← c
 sinon si A = (=vA') alors        V← v'·V{A'}(P)
 sinon si A= ¬A' alors V← 0
 sinon /* A est une proposition atomique */
       si A et P sont des concepts élémentaires ∧ (A⇒P∨A=P) alors V←1
       sinon si A est un concept élémentaire et P ne l'est pas alors V ← 0
       sinon soit A= [rel](A'1,A'2)
         début
           si P = [rel'](P1,P2) alors
            début
              si ([rel]=[rel'])
              alors V1 ← SUB(A'1,P1) · SUB(A'2,P2)
              sinon V1←0
              si V1=1 alors V←1
               sinon
                 début
                   V2←V{A'1}(P)
                   si V2=1 alors V←1
                     sinon
                       début
                         V3=V{A'2}(P)
                         V←MAX(V1,V2, V3)
                       fin
                 fin
            fin
           sinon /* P est un concept élémentaire */
            début
              V1← V{A'1}(P)
              si V1=1 alors V←1
              sinon
                    début
                      V2←V{A'2}(P)
                      V←MAX(V1,V2)
                    fin
            fin
         fin
fin.
```

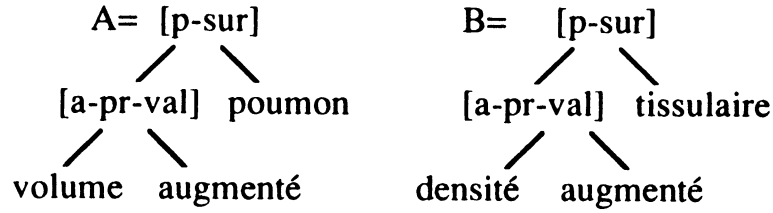
Fig.II.5. *Algorithme pour la mesure de la vérification*

fonction: SUB(A,P) ∈ [0,1]
début
si A=P alors SUB←1
sinon si (A≡>P)∈ K alors SUB←1
sinon si A est un concept élémentaire et P ne l'est pas alors SUB←0
sinon /* A et P sont des arborescences sémantiques */
début
soient A=[rel](A'$_1$,A'$_2$) et (P=[rel'](P$_1$,P$_2$))
si [rel]=[rel'] alors SUB$_1$←SUB(A'$_1$,P$_1$)·SUB(A'$_2$,P$_2$)
si SUB$_1$=1 alors SUB ← 1
sinon si [rel](A'$_1$,A'$_2$) ≡> A$_1$
alors SUB←SUB(A'$_1$,P)
sinon SUB←0
fin
fin.

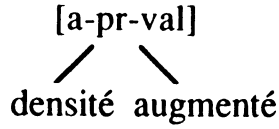
Fig.II.6. *Algorithme pour mesurer la possibilité d'une substitution*

5.2.6. Discussion

Comme on peut le remarquer, une telle évaluation est extrêmement coûteuse, car non seulement toutes les transformations possibles sur une assertion d'un document doivent être effectuées, mais il en est de même sur toutes les assertions, y compris celles qui n'ont aucun rapport avec la proposition à valuer. Par exemple, soient deux assertions d'un document:



Pour valuer la proposition atomique suivante dans le document



il est clair que les transformations sur l'assertion A ne peuvent pas aboutir à une vérification de la proposition. Ces transformations sont donc superflues.

Pour éviter ces transformations inutiles, nous pouvons effectuer une sorte de *présélection* pour choisir les assertions du CRM étant *susceptibles* de

correspondre à la proposition à valuer. Cette présélection peut largement réduire les transformations à effectuer.

Pour ne pas introduire de *silence* (c'est-à-dire le fait que des documents pertinents ne soient pas présents dans la réponse), il est nécessaire que la condition pour la *présélection* par rapport à une proposition atomique soit impliquée par la proposition atomique.

Notons $P_{pré}$ la condition de la présélection pour la proposition atomique P. On doit avoir $P \Rightarrow P_{pré}$ ($P_{pré}$ doit être une condition moins sélective que P), car ainsi, tout document A satisfaisant Q satisfait aussi $Q_{pré}$.

Pour définir plus précisément la nature de cette présélection, remarquons le fait suivant:

Une proposition atomique est constituée de deux sortes d'éléments: les concepts élémentaires et les opérateurs sémantiques entre ceux-ci. Pour qu'une sous-requête soit totalement vérifiée dans une assertion (A), il est nécessaire que tous les concepts élémentaires de P soient présents dans A, et que ces concepts soient connectés de la même façon dans P que dans A.

La présélection est définie justement sur la première partie de cette condition: tous les concepts élémentaires d'une proposition atomique P doivent apparaître dans une assertion A ou être dérivables à partir des concepts élémentaires de A pour que A puisse dériver P. On appelle ces concepts élémentaires les *concepts nécessaires.*

L'ajout de la présélection divise la vérification en deux étapes:

1). la vérification de $P_{pré}$ sur toutes les assertions du document ($A \in D$) pour sélectionner celles qui sont susceptibles de pouvoir dériver P,

2). la vérification de P sur les assertions présélectionnées.

Remarquons que cette présélection est analogue à celle opérée via les attributs externes, lorsqu'ils existent: si $Q = \wedge(Q_{EXT}, Q_{INT})$, on a bien $Q_{pré} = Q_{EXT}$, et $Q \Rightarrow Q_{pré}$.

Un autre facteur pouvant accélérer le processus de la vérification est le modèle sémantique.

Comme nous l'avons défini, à chaque concept (élémentaire ou complexe) est associée une classe sémantique. Une proposition atomique constituée d'un concept d'une classe donnée ne peut être vérifiée que par une assertion constituée d'un concept d'une certaine classe (la même classe ou une

classe supérieure, cf.II.2.3). La comparaison des classes peut aussi éliminer beaucoup de transformations inutiles: les assertions comprenant des concepts de classes non *compatibles* ne doivent pas être transformées.

5.2.7. Optimisation

5.2.7.1. Présélection

Les concepts nécessaires sont définis comme les concepts élémentaires qui doivent apparaître dans une assertion ou qui sont dérivables à partir des concepts élémentaires existant dans l'assertion. Nous concrétisons cette notion par rapport à toute forme de proposition atomique. Nous notons les concepts nécessaires pour une proposition atomique P par NEC(P).

Deux cas de figures sont possibles pour une proposition atomique:

1er cas: La proposition atomique P est un concept élémentaire: P = *a*

Si une telle proposition atomique est satisfaite par A, le concept élémentaire doit être soit *contenu* dans D, soit *dérivable* à partir d'un concept de A. Ainsi, la fonction NEC(P) est définie par le concept élémentaire:

$$NEC(P)=\{a\}.$$

2ième cas: La proposition atomique P est de la forme [rel](P_1,P_2).

Pour qu'une telle proposition atomique soit satisfaite par une assertion (A), il est donc nécessaire que les sous-propositions P_1 et P_2 soient satisfaites séparément dans l'assertion A. Pour cela, il est nécessaire que les concepts nécessaires de P_1 et de P_2 soient tous présents dans A. Ainsi, les concepts nécessaires pour [rel](P_1,P_2) sont les suivants:

$$NEC(P) = NEC(P_1) \cup NEC(P_2).$$

Par exemple, pour que la proposition atomique [a-pr-val](densité, tissulaire) soit satisfaite par une assertion, il faut que les concepts *densité* et *tissulaire* soient tous les deux présents dans l'assertion. Ainsi, NEC(Q) = {densité, tissulaire}.

Notons CE(A) l'ensemble des concepts élémentaires présents dans l'assertion A. L'assertion (A) est sélectionnée si chaque concept nécessaire de NEC(P) est présent dans CE(A) ou dérivable d'un concept de CE(A). Le résultat de la présélection est le suivant:

$$\{A \mid \forall t \in NEC(P)\ \exists t' \in CE(A)((t'=t) \vee (t' \Rightarrow t))\}$$

Cette opération ressemble beaucoup à une opération de sélection dans les bases de données classiques. Néanmoins, il existe une différence dans le critère de sélection: dans les bases de données classiques, les sélections sont fondées uniquement sur le critère d'*égalité* (t'=t), tandis que dans RIME, la sélection s'effectue aussi sur le critère de dérivation ($t' \Rightarrow t$). Il est donc nécessaire d'ajouter une transcription pour interpréter la relation de dérivation. On propose la transcription suivante:

Chaque concept élémentaire dans l'ensemble des concepts nécessaires (NEC(P)) doit être remplacé par la conjonction de ce concept avec les concepts pouvant dériver vers celui-ci:

$\forall t \in NEC(P)$, remplacer t par $t \vee t_1 \vee \ldots \vee t_n$, avec $t_1 \Rightarrow t$, $\ldots$ $t_n \Rightarrow t$

L'ensemble des concepts élémentaires nécessaires d'une proposition atomique $\{t_1, t_2, t_3, \ldots\}$ correspond à une conjonction des termes. La transcription consiste à remplacer chaque concept t_i de cette conjonction par $t_{i0} \vee t_{i1} \vee t_{i2} \vee \ldots \vee t_{in}$, où t_{ij} est un concept pouvant dériver vers t_i et $t_{i0} = t_i$. Finalement, on obtient une requête de présélection ayant la forme suivante:

$$(t_{10} \vee t_{11} \vee t_{12} \vee \ldots t_{1m}) \wedge (t_{20} \vee t_{21} \vee t_{22} \vee \ldots t_{in}) \wedge \ldots$$

Une opération "select" des bases de données correspondant à cette requête, sélectionne aussi les arborescences dont les concepts élémentaires peuvent dériver vers les concepts nécessaires de la proposition atomique initiale. Cette opération correspond à l'évaluation de la requête de bases de données suivante:

```
select  A    /* assertion */
from    D    /* document */
where ((t10 ∈ CE(A)) or ( t11 ∈ CE(A)) or ...)
          and
          ((t20 ∈ CE(A)) or ( t21 ∈ CE(A)) or ...)
          and
          ...
```

Par exemple, la requête "*Je voudrais les CRM décrivant des tumeurs du poumon*" est interprétée en

[p-sur](tumeur, poumon)

donc NEC(P) est défini par: {tumeur, poumon}

Etant donné que le concept *tumeur* est dérivable vers les concepts élémentaires suivants:

xanthome, hygroma, kyste, polykystose, pseudokyste, polype, polypose, papillome, adénome, neurinome, cancer, dégénérescence, épithélioma, sarcome, lymphosarcome, blastome, tératome

la présélection correspond à la sélection suivante des bases de données:

```
select  A
from    D
where ((tumeur∈ CE(A)) or (xanthome∈ CE(A))
        or (hygroma∈ CE(A)) or (kyste∈ CE(A))
        or ...)
      and
      (poumon∈ CE(A)))
```

Dans notre application, la relation de dérivation entre les concepts élémentaires est réduite à un nombre de niveaux très limité (5 niveaux au maximum) (voir Annexe 5). Le nombre de concepts inclus dans un autre est aussi relativement limité. La complexité de la requête de présélection des bases de données finalement obtenue sera donc limitée. Grâce à la rapidité de l'opération "select" des bases de données, cette présélection peut être très rapide. Ainsi, cette présélection constitue un "filtrage" efficace dont le rôle est de sélectionner les assertions "susceptibles" d'être dérivées vers une proposition atomique à valuer.

5.2.7.2. Contrôle de la classe sémantique

Nous analysons ici les circonstances dans lesquelles il peut exister une relation de dérivation et une relation de substitution. Cela permettra d'effectuer uniquement les transformations *utiles* pour la vérification de la proposition.

- **Sur les dérivations**

Pour qu'une assertion A puisse dériver vers une proposition atomique P, il est nécessaire que A ait la même classe sémantique que P ou qu'il existe dans A un composant A' de même classe sémantique que P. Autrement dit, la vérification d'une proposition atomique ne s'effectue que sur A et les composants de A ayant la même classe sémantique que la proposition atomique P.

Comme nous l'avons indiqué dans II.2.3, il existe une hiérarchie parmi les classes sémantiques. Une proposition atomique d'une classe sémantique

donnée ne peut exister que dans les assertions de même classe sémantique ou d'une classe sémantique supérieure.

Ainsi, étant données une assertion et une proposition atomique à vérifier, on peut directement sélectionner les sous-assertions dont la classe sémantique est identique à celle de la proposition atomique pour pouvoir la comparer avec la proposition. Si l'assertion est d'une classe qui n'est ni identique à celle de la proposition ni supérieure à celle-ci, la possibilité de dériver la proposition atomique à partir de cette assertion est directement valuée à 0.

Après avoir repéré les composants de la même classe sémantique que la proposition à valuer, il faut dériver toute l'assertion vers ces composants, de façon à ce que la proposition soit valuée par rapport à chacun des composants.

• **Sur la possibilité de substitution**

Etant données une proposition atomique et une assertion constituée d'une proposition atomique, pour que l'assertion puisse être substituée à la proposition, il est nécessaire qu'elles soient de classes sémantiques *compatibles*, c'est-à-dire qu'il y ait au moins une relation de substitution entre les deux concepts correspondants.

1. En général, une proposition ne peut être substituée qu'à une proposition de même classe sémantique.

Cependant, sous certaines conditions, une proposition d'une certaine classe sémantique supérieure peut être substituée à une proposition de classe inférieure.

2. Comme nous l'avons défini dans le modèle sémantique, une proposition d'une classe donnée peut constituer à elle seule une proposition d'une classe supérieure. Par exemple, une proposition de la classe LESION constitue une proposition de la classe DIAG. Ainsi, SUB(A,P) peut être assigné à 1 dans le cas où A est de classe DIAG et P de classe LESION.

3. Nous avons défini certains cas particuliers pour la relation de substitution entre deux arborescences de classes différentes (cf.6.2.2.2). Par exemple, [a-pr-val](CAR-PHY,VAL-QUAL)≡>CAR-PHY, dans le cas où la substitution ne viole pas de contraintes sémantiques aux niveaux plus hauts.

En résumé, nous pouvons définir toutes les possibilités de substitution par une relation "substituable" (où substituable(A,B) signifie que la substitution d'un concept de la classe A par un concept de la classe B est possible):

substituable(classe-de-substitué, classe-de-substituant)

x	x
CONSTAT	SGN
SGN	QUAL
DIAG	LESION
QUAL	CAR-PHY
QUAL	VAL-QUAL
QUAL	VAL-QUAN
LOC	ORG
ORG	REGION
REGION	DETAIL

(où x représente une classe quelconque)

Nous définissons ensuite la propriété de transitivité sur cette relation:

$$\text{substituable}(x,y) \wedge \text{substituable}(y,z) \Rightarrow \text{substituable}(x,z).$$

Comme pour la vérification de la dérivation, la vérification de la possibilité de substitution ne s'effectue qu'entre deux concepts pour lesquels il existe une relation "substituable" entre les classes sémantiques.

5.2.8. Algorithmes proposés

On dérive des algorithmes précédents, ceux donnés ci-dessous, en incorporant les contrôles des classes sémantiques. Ces algorithmes valuent la certitude de la vérification de la proposition par rapport à une assertion du document filtré par la présélection.

Dans ce nouveau contexte, la fonction $V'_{\{A\}}(P)$ est définie comme suit:

$$V'_{\{A\}}(P) = \begin{cases} 0, & \text{si } (\text{classe}(A) \neq \text{classe}(P)) \wedge \\ & \quad \text{non}(\text{supérieur}(\text{classe}(A),\text{classe}(P)) \\ V_{\{A\}}(P), & \text{sinon} \end{cases}$$

où $V_{\{A\}}(P)$ est la fonction définie précédemment.

La fonction SUB'(A,P) est définie comme suit:

$$\text{SUB}'(A,P) = \begin{cases} 0, & \text{si non}(\text{substituable}(\text{classe}(A),\text{classe}(P))) \\ \text{SUB}(A,P), & \text{sinon} \end{cases}$$

où SUB(A,P) est la fonction définie précédemment.

```
fonction: V'{A}(P): ∈ [0,1]
début
 c ← CP(A)
 si c > 0 alors V' ← c
 sinon si classe(A)≠classe(P) ∧ non(supérieure(classe(A),classe(P)))
       alors V' ← 0
/* la suite est la même que pour V{A}(P) */
 sinon si A = (=vA') alors      V'← v·V'{A'}(P)
 sinon si A= ¬A' alors V'← 0
 sinon /* A est une proposition atomique */
    si A et P sont des concepts élémentaires ∧ (A⇒P ∨ A=P) alors V' ← 1
    sinon si A est un concept élémentaire et P ne l'est pas alors V' ← 0
    sinon soit A= [rel](A'1,A'2)
      si classe(A)≠classe(P) alors V←MAX(V'{A'1}(P),V'{A'2}(P))
      sinon
        début
          si P = [rel'](P1,P2) alors
            début
              si ([rel]=[rel'])
              alors V1 ← SUB'(A'1,P1) · SUB'(A'2,P2)
              sinon V1←0
              si V1=1 alors V'←1
               sinon
                 début
                   V2←V'{A'1}(P)
                   si V2=1 alors V'←1
                     sinon
                       début
                         V3=V'{A'2}(P)
                         V'←MAX(V1,V2, V3)
                       fin
                 fin
            fin
          sinon /* P est un concept élémentaire */
            début
              V1← V'{A'1}(P)
              si V1=1 alors V'←1
              sinon
                    début
                      V2←V'{A'2}(P)
                      V'←MAX(V1,V2)
                    fin
            fin
        fin
fin.
```

fonction: SUB'(A,P) ∈ [0,1]
début
si A=P alors SUB'←1
sinon si non(substituable(classe(A),classe(P))) alors SUB'←0
/* la suite est identique que celle du SUB(A,P) */
sinon si (A≡>P)∈ K alors SUB'←1
sinon si A est un concept élémentaire et P ne l'est pas alors SUB'←0
sinon /* A et P sont des arborescences sémantiques */
début
soit A=[rel](A'$_1$,A'$_2$) et (P=[rel'](P$_1$,P$_2$))
si [rel]=[rel'] alors SUB$_1$←SUB'(A'$_1$,P$_1$)·SUB'(A'$_2$,P$_2$)
si SUB$_1$=1 alors SUB' ← 1
sinon si [rel](A'$_1$,A'$_2$) ≡> A$_1$
alors SUB'←SUB'(A'$_1$,P)
sinon SUB'←0
fin
fin.

5.3. Organisation globale de l'évaluation

Nous avons présenté le processus suivant pour la valuation d'une formule (requête) par rapport à un monde initial (document):

La valuation de la formule est décomposée en des valuations de propositions atomiques.

Pour valuer la vérification d'une proposition atomique dans un monde initial, les assertions du monde initial doivent subir tout d'abord la présélection pour choisir les assertions susceptibles de satisfaire la proposition.

On considère ensuite que chaque assertion choisie constitue à elle seule un monde possible du monde initial. La proposition est vérifiée dans ce monde possible par les algorithmes proposés.

Ce processus doit être effectué sur tous les documents du corpus, c'est-à-dire qu'il se répète autant de fois qu'il y a de documents dans le corpus.

Nous pouvons également proposer une approche ensembliste équivalente:

1. La vérification d'une formule par rapport aux documents du corpus est également décomposée en des vérifications de propositions atomiques.

2. Pour la vérification d'une proposition atomique (P), l'ensemble des assertions du corpus est soumis à la présélection pour choisir un ensemble A(P) d'assertions susceptibles de satisfaire P.

3. Toutes les assertions présélectionnees dans A(P) sont ensuite comparées individuellement avec la proposition atomique pour valuer la fonction $V_{\{A\}}(P)$. A chaque assertion est associée une valeur de certitude: $V_{\{A\}}(P)$. On associe aux assertions non sélectionnées dans A(P) la valeur de certitude 0. On obtient donc un ensemble

$$\{\langle A, V_{\{A\}}(P)\rangle : A \in \text{ l'ensemble d'assertions du corpus}\}.$$

4. Pour un document D donné, la certitude de la vérification de P par rapport à ce document prend comme valeur la maximale des valeurs assignées aux vérifications de P par rapport à chaque assertion $A_i \in D$. Ainsi, on obtient un autre ensemble dans lequel à chaque document est assigné une valeur de certitude:

$$\{\langle D, V_D(P)\rangle : D \in \text{ l'ensemble de documents du corpus}\}$$

où $$V_D(P) = \underset{A_i \in D}{MAX}(V_{\{A_i\}}(P)$$

5. On procède finalement au regroupement des ensembles de documents selon la structure de la formule.

- Si $\{\langle D,V_D(f_1)\rangle : D \in \text{corpus}\}$ et $\{\langle D,V_D(f_1)\rangle : D \in \text{corpus}\}$ sont connus, alors
  $\{\langle D,V_D(f_1 \wedge f_2)\rangle : D \in \text{corpus}\} = \{\langle D, MIN[V_D(f_1),V_D(f_2)]\rangle : D \in \text{corpus}\}$

- Si $\{\langle D,V_D(f_1)\rangle : D \in \text{corpus}\}$ et $\{\langle D,V_D(f_1)\rangle : D \in \text{corpus}\}$ sont connus, alors
  $\{\langle D,V_D(f_1 \vee f_2)\rangle : D \in \text{corpus}\} = \{\langle D,MAX[V_D(f_1),V_D(f_2)]\rangle : D \in \text{corpus}\}$

- Si $\{\langle D,V_D(f_1)\rangle : D \in \text{corpus}\}$ est connu, alors
  $\{\langle D, V_D(\neg f_1)\rangle : D \in \text{corpus}\} = \{\langle D, (1-V_D(f_1))\rangle : D \in \text{corpus}\}$

- Si $\{\langle D, V_D(f_1)\rangle : D \in \text{corpus}\}$ est connu, alors
  $\{\langle D, V_D(=_v f_1)\rangle : D \in \text{corpus}\} = \{\langle D, (1-|V_D(f_1)-v|)\rangle : D \in \text{corpus}\}$

5.4. Réponse à une requête

La réponse à une requête est d'abord constituée de l'ensemble des documents ayant une correspondance de valeur 1 avec la requête. Si l'utilisateur n'est pas satisfait de cette réponse, les documents fournis doivent être complétés avec les documents ayant une valeur de correspondance $\in]0,1[$, dans l'ordre décroissant de leur valeur de correspondance, jusqu'à ce que l'utilisateur soit satisfait.

5.5. Discussion sur l'évaluation

L'utilisation des outils bases de données est devenue maintenant très courante pour l'évaluation des requêtes d'un SRI à cause de leur efficacité. Comme dans les utilisations classiques, nous avons intégré les outils des bases de données pour l'évaluation des attributs externes d'une requête. Nous avons aussi voulu intégrer les outils des bases de données pour *aider* le processus d'évaluation de l'attribut interne (dans la présélection), sans banaliser ce dernier comme un attribut externe (comme cela est réalisé dans certains systèmes). Pour ce faire, une certaine "déduction" est effectuée pour que les requêtes de bases de données engendrées puissent refléter la caractéristique de multi-représentation d'une sémantique de l'attribut interne.

L'objectif d'une évaluation de l'attribut interne est de vérifier si un document peut être transformé en une forme équivalente à la requête (avec une certitude). Etant donné le coût élevé de la transformation, il est nécessaire d'intégrer certains mécanismes pour accélérer le processus.

Dans notre étude, nous avons proposé une présélection afin de filtrer certains documents qui ne peuvent pas satisfaire la requête, ce qui peut permettre d'accélérer de façon sensible l'opération d'évaluation. Mais la présélection proposée n'est pas le seul moyen de réaliser cette accélération. Beaucoup d'autres techniques sont étudiées et utilisées dans d'autres SRI, par exemple les techniques statistiques ([Bookstein77, Salton83a, Yu79]) qui permettent de donner directement les documents pouvant satisfaire à une requête donnée. Pour l'évaluation de RIME, ces techniques peuvent être utilisées en concurrence avec la présélection, ce qui permettra une évaluation encore plus efficace.

Dans la partie "interprétation des requêtes", les requêtes sont standardisées sous une forme qui permet une séparation entre l'évaluation des opérateurs sémantiques, celle des opérateurs booléens et celle des opérateurs de certitude (les opérateurs booléens et les opérateurs de certitude sont situés à des niveaux supérieurs par rapport aux opérateurs sémantiques). Mais cela engendre aussi une certaine redondance dans la représentation finale des

requêtes, donc une redondance pour le processus d'évaluation. Par exemple, la standardisation de [rel](A,¬B) en ∧(A,¬([rel](A,B))) (cf.5) conduit à une double présence de A dans la représentation standardisée, et donc une double évaluation de A. Cette redondance se traduit par une perte d'efficacité. Il faudrait donc une sorte de contrôle permettant d'éviter ou réduire la redondance dans l'évaluation. Par exemple, pour l'exemple précédent, on pourrait comparer [rel](A,B) uniquement avec les arborescence qui ont vérifié A auparavant, car les arborescences ne vérifiant pas A ne peuvent pas vérifier [rel](A,B).

Actuellement, nous ne disposons que d'un ensemble de connaissances très limité.

1. Nous n'avons pas une "norme" pour le domaine traité. Une hypothèse simpliste a donc été faite concernant les descriptions absentes d'un document.
2. Le domaine médical est connu pour ses connaissances déductives, par exemple, la déduction d'un constat à partir d'un ensemble de signes.

Ces aspects seront étudiés dans le futur développement de RIME. Une fois que RIME aura intégré ces aspects, ce système, de part son utilisation très dense des connaissances, prendra encore plus d'intérêt. Il conduira également à un rapprochement entre les SRI et certains systèmes développés dans le domaine de l'intelligence artificielle, notamment les systèmes experts.

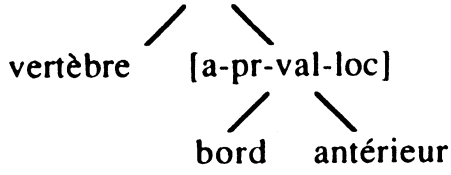
LOC ("de la"), puis ceux qui forment SGN ("du") et finalement ceux qui forment CONSTAT ("par").

Le résultat des différentes étapes est montré ci-dessous:

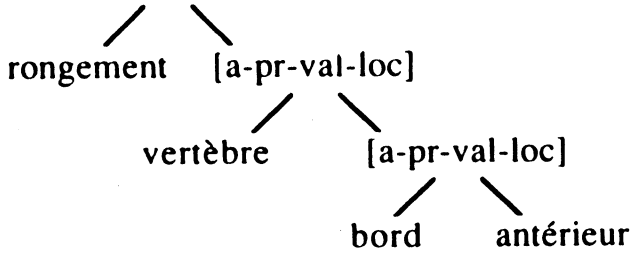
0. le rongement du bord antérieur de la vertèbre par la tumeur

1. le rongement du [a-pr-val-loc] (DETAIL) de la vertèbre par la tumeur

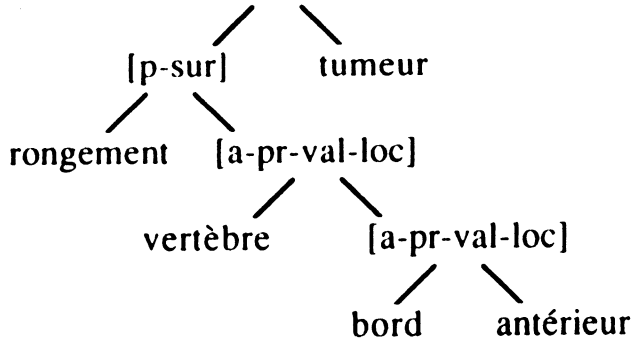
2. le rongement du [a-pr-val-loc] (ORG et donc LOC) par la tumeur



3. [p-sur] (SGN) par la tumeur



4. [dû-à] (CONSTAT)



En comparant le processus d'interprétation avec celui de l'indexation (cf.II.3.2), on trouve une bonne cohérence. Cela garantit qu'une requête subit des transformations similaires à celles des documents. On espère donc que la correspondance mesurée entre une requête interprétée et un document indexé sera très proche de la correspondance réelle entre la requête initiale et le document initial.

4.7.2. Interprétation des mots (Niveau 0)

Cette interprétation a pour objectif d'associer les informations syntaxiques et sémantiques à chaque composant de base (mots et expressions correspondant à une entrée du dictionnaire) de la requête.

Voici deux exemples montrant les résultats obtenus après la consultation du dictionnaire:

1. le lobe antérieur du (=de le) foie

le((*art, mas, sing*),(*pp,mas,sing*))
lobe(*nom, mas, sing*, lobe,DETAIL)
antérieur(*adj, mas, sing*, antérieur, POS)
de(*prép*, ([p-sur], [a-pr-val], [a-pr-val-loc]))
foie(*nom, mas, sing*, foie, CONST-ORG)

2. une opacité du lobe moyen rétractée avec bronchogramme aérique pathologique

une(*art, fém, sing*)
opacité(*nom, fém, sing*, [a-pr-val]((densité,CAR-PHY),
(augmenté,VAL-QUAL)),SGN)
de(*prép*,([p-sur],[a-pr-val],[a-pr-val-loc]))
le(*art, mas, sing*)
lobe(*nom, mas, sing*, lobe,DETAIL)
moyen(*adj, mas, sing*, moyen,POS)
rétractée(*ppass, fém, sing*, rétracté, VAL-QUAL)
avec(*prép*, ([en-rel-topo-avec],[a-pr-val],[et]))
bronchogramme(*nom, mas, sing*, bronchogramme, SGN)
aérique(*adj, mas-fém, sing*, [a-pr-val]((densité,CAR-PHY),
(aérien,VAL-QUAL)), QUAL)
pathologique(*adj, mas-fém, sing*, pathologique, VAL-QUAL)

4.7.3. Attachement d'adjectif (Niveau 1)

Un adjectif peut décrire un nom, un groupe nominal ou un autre adjectif. L'attachement d'un adjectif nécessite de déterminer d'abord l'attaché de l'adjectif.

Prenons d'abord le cas d'un seul adjectif (pas de succession d'adjectifs comme: *opacité du lobe moyen rétractée*). La condition syntaxique pour être l'attaché d'un adjectif est la suivante: l'attaché doit être un nom ou un groupe nominal situé juste avant l'adjectif, satisfaisant les accords en nombre et en genre avec l'adjectif.

PARTIE III

RÉALISATION ET EXPÉRIMENTATION

1. RÉALISATION

1.1. Généralités

Les fonctions d'analyse et d'évaluation des requêtes de RIME proposées dans la partie II ont été réalisées, dans le cadre d'un prototype restreint, en C-Prolog sur la machine Gould UTX/32 du laboratoire. La réalisation a pour but de démontrer la validité de l'approche proposée dans la partie II. Ainsi, certaines simplifications ont été faites au niveau du prototype, notamment sur les points suivants:

- Sur le vocabulaire

Le vocabulaire que l'on a considéré est un sous-ensemble du vocabulaire utilisé dans le domaine de RIME. Nous ne nous sommes pas occupé de l'aspect gestion du vocabulaire. Un système de dictionnaire comportant cette fonction (cf. [Palmer90]) est étudié dans le groupe de recherche.

- Sur les attributs externes

Nous n'avons pris en compte que trois attributs externes: le *nom du patient*, les *médecins* et la *date d'examen*. Ces trois attributs ont tous une syntaxe relativement simple qui est la suivante:

• Pour les attributs patient et docteur: un nom propre ou une connexion de noms propres par les conjonctions (et, ou).

ATTRIBUT ::= NOM [CONJ NOM]*

• Pour l'attribut date d'examen: une expression de date ou une connexion de dates par les conjonctions. Une expression de date peut être une simple date (ex. *juillet 89*) ou une date précédée d'une expression telle que "avant", "après", "jusqu'à" ....

DATE ::= EXP-DATE [CONJ EXP-DATE]*
EXP-DATE ::= [INTER] DATE-SIMPLE
INTER ::= avant | après | entre...et | jusqu'à | à partir de
DATE-SIMPLE ::= une date

- Sur l'attribut interne:

Nous avons considéré tous les niveaux syntaxiques du langage d'interrogation défini dans II.4. La simplification n'est faite que sur certains détails. Par exemple, nous avons pas traité les verbes vides. Ceux-ci n'apportent pas une précision significative sur les besoins des utilisateurs, mais sont simplement une commodité.

La réalisation peut être globalement illustrée par le schéma suivant:

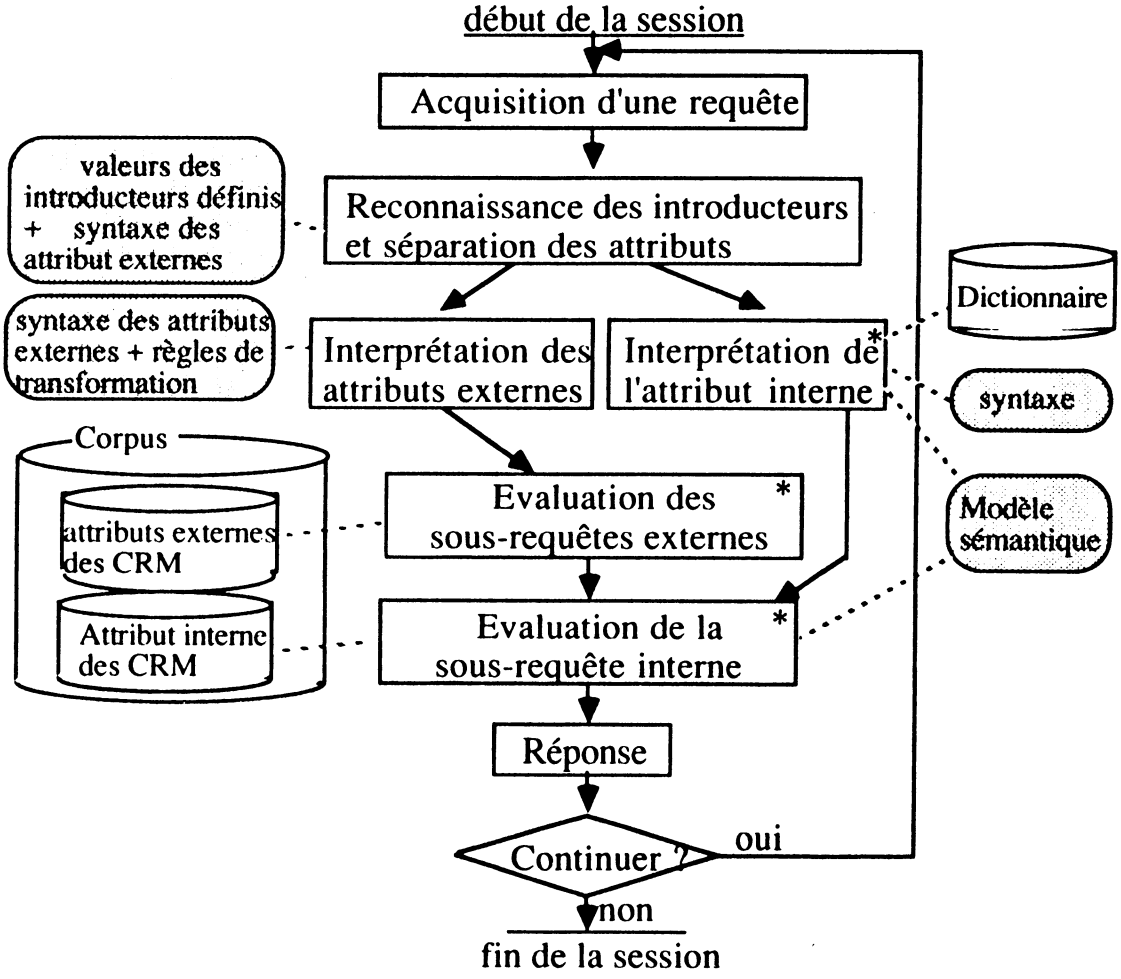


Fig.III.1. *Schéma général du prototype réalisé*

(NB: les lignes en pointillés représentent l'utilisation d'un outil)

• Acquisition d'une requête: ce module accepte une chaîne de caractères de l'utilisateur et la transforme en une chaîne de mots.

• Reconnaissance des introducteurs et séparation des attributs: ce module analyse la requête jusqu'à rencontrer un mot ou un groupe de mot qui

correspond à un introducteur. Une analyse contextuelle simple est ensuite engagée pour déterminer (ou confirmer dans le cas non ambigü) la nature de l'introducteur.

Ce module repose sur la définition des introducteurs et une définition de la syntaxe des attributs externes.

• Interprétation des attributs externes: chaque attribut externe, reconnu et isolé par le module précédent, est transformé en une représentation interne par une automate simple. Celle-ci correspond à la définition donnée dans la partie II.

• Les autres modules (marqués d'une étoile dans le schéma) sont plus complexes. Ils sont présentés plus en détail par la suite.

• Une session d'interrogation peut comporter plusieurs requêtes. Dans ce cas, on considère qu'une requête précédée d'une autre requête porte sur le résultat de la requête précédente.

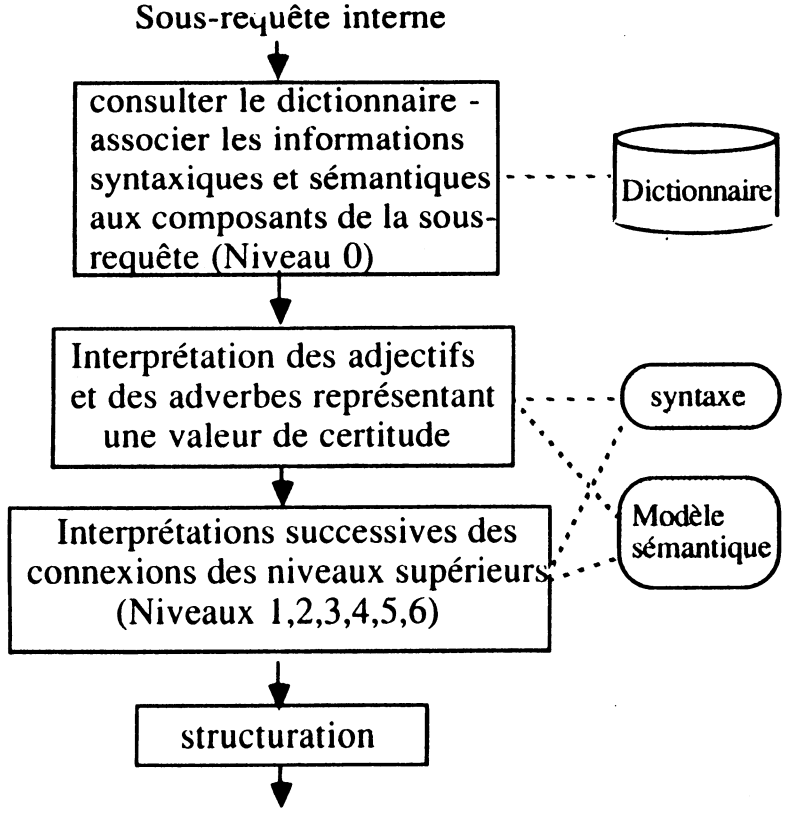


Fig.III.2. *Schéma de l'interprétation de l'attribut interne*

1.2. Interprétation de l'attribut interne

L'interprétation de l'attribut interne suit la méthode proposée dans la partie II, c'est-à-dire que l'interprétation commence par le traitement des connexions les plus fortes (syntaxiquement et sémantiquement), et se poursuit ensuite par le traitement des connexions plus faibles. Ce module correspond au schéma de Fig.III.2.

Avant de présenter le mécanisme des différentes étapes, nous décrivons d'abord son environnement.

1.2.1. L'environnement de travail

1.2.1.1. Le dictionnaire

Les informations syntaxiques (les variables grammaticales) et sémantiques (la représentation interne et sa classe sémantique) sont associées à chaque entrée du dictionnaire. Elles sont organisées selon la structure suivante en Prolog:

mot(_entrée,_cat_gram,_genre,_nombre,_sémantique).

Le dictionnaire est constitué par un ensemble de clauses ayant cette structure. Pour des raisons de simplicité, nous avons associé à toutes les entrées du dictionnaire la même structure (dans la partie II, la structure varie selon la nature de l'entrée). L'absence d'une information dans la structure est représentée par '0'. Par exemple, les prépositions n'ont ni genre ni nombre. Leur genre et nombre sont représentés par '0'.

L'élément "_sémantique" dans cette structure peut avoir plusieurs formes possibles:

- Un mot vide a une sémantique 'vide'.

- Un mot représentant un opérateur (par exemple: une préposition, un verbe relationnel simple) est interprété par l'ensemble des opérateurs qu'il peut représenter.

- Un mot représentant un trait sémantique ou une structure complexe (par exemple: verbe relationnel complexe) est interprété par une structure arborescente dont les éléments à substituer ultérieurement sont représentés soit par '0', soit par la classe sémantique de l'élément si elle est déterminée.

- Un mot représentant un concept déterminé a une représentation sémantique composée de sa représentation interne et de la classe sémantique de celle-ci. L'interprétation dans ce cas est donc un doublet, représenté comme suit:

[_rep_interne, _classe_sémantique]

Nous donnons quelques exemples ci-dessous.

- article:

mot(un, *ar*, *m*, *s*, VIDE).

- adjectif:

mot(tissulaire, *a*, *mf*, *s*, [tissulaire, VAL-QUAL]).

- adverbe:

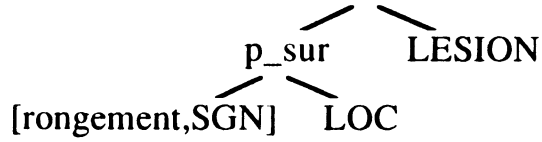
mot(probablement, *ad*, *0,0*, [p]).

- nom:

mot(foie, *n*, *m*, *s* [foie, CONST-ORG]).

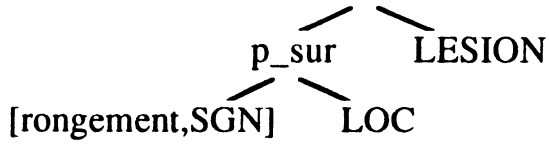
- verbe:

mot(montre, *v*, *0*, *s*, [montre_par]).
mot(ronge, *v*, *0*, *s*, [du_a]).



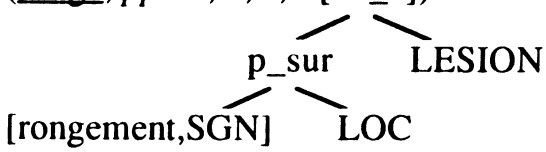
- participe présent:

mot(rongeant,*ppres*,*0*,*0*, [du_a]).



- participe passé:

mot(rongé, *ppass*,*m*, *s*, [du_a]).



- préposition:
  mot(de,*p,0,0*,[p_sur,a_pr_val,a_pr_val_loc]).

- expression:
  exp([au,niveau,de],*exp,0,0,* [p_sur,a_pr_val_loc]).

Nous n'avons enregistré dans le dictionnaire que les mots masculins singuliers. Les autres formes sont engendrées par un analyseur morphologique simple. Cette analyse morphologique est une simulation d'une analyse plus complète qui est étudiée dans [Palmer90].

Dans la version actuelle du prototype, le dictionnaire contient près de 1.000 entrées.

1.2.1.2. La représentation des comptes rendus médicaux indexés

L'indexation des CRM est constituée de deux parties, concernant respectivement les attributs externes et l'attribut interne.

Les attributs externes d'un CRM sont organisés dans une relation, représentée par la clause Prolog suivante:

crm(_n_crm,patient(_pat),docteur(_liste_doct),date(_dat)).

où "_n_crm" est le numéro du CRM,
"_pat" est le nom du patient,
"_liste_doct" est la liste de noms des docteurs,
"_dat" la date d'examen.

Par exemple:

crm(12,patient('René'),docteur(['Coulomb','Rose-Pittet']),date(860312)).

L'attribut interne est organisé comme un ensemble d'arborescences dont chacune est représentée par une relation "cont" ayant la forme suivante:

cont(_n_crm, _n_arb, _texte, [_rep_int, _classe], _CE).

où "_n_crm" est le numéro du CRM,
"_n_arb" est le numéro de l'arborescence
"_texte" est le texte original correspondant dans le CRM
"_rep_int" est la représentation interne arborescente du texte
"_classe" est la classe sémantique de "_rep_int",
"_CE" est la liste des concepts élémentaires de "_rep_int"

L'ensemble de relations "cont" est établi à partir de l'ensemble de relations "int" saisies manuellement:

int(_n_crm,_n_arb,_chaine_car).

où "chaine_car" est "_texte" sont des chaînes de caractères.

L'élément "[_rep_int,_classe]" est issu d'une interprétation automatique similaire à celle de la requête. "_CE" est généré ensuite à partir de "_rep_int".

Par exemple:

int(12,124,"une hypertrophie ganglionnaire").

⇓

cont(12,124,'une hypertrophie ganglionnaire',
[p_sur([a_pr_val([volume,CAR_PHY],[augmenté,val_qual]),SGN],
[ganglion,CONST_ORG]),SGN],
[augmenté,ganglion,volume]).

1.2.1.3. Le modèle sémantique

Le modèle sémantique de la partie II est implémenté par un ensemble de règles, exprimées sous forme de clauses ayant la forme suivante:

Expression de la partie II		Expression dans la réalisation
CONSTAT ::= [dû-à](CONSTAT,DIAG)	→	du_a(constat,diag,constat).
CONSTAT ::= SGN	→	herite(sgn,constat).

La vérification de l'accord sémantique est effectuée par l'opération

accord_classe(_opérateur,_classe1,_classe2,_classe3),

qui vérifie si deux concepts de *_classe1* et de *_classe2* ou de leurs classes héritées respectives (par la relation herite) peuvent être connectés par *_opérateur* pour former un concept de *_classe3*.

Par exemple, dans l'interprétation du groupe nominal "l'opacité du poumon" où "opacité" et "poumon" sont respectivement reconnus comme un "sgn" et un "const_org", et la préposition "de" peut être interprétée par un des opérateurs: [p_sur, a_pr_val, a_pr_val_loc], la vérification suivante est d'abord engagée:

accord_classe(**p_sur**, **sgn**, **const_org**, _classe)

Etant donné que la règle: **p_sur(sgn, const_org**, _classe3) n'existe pas dans le modèle sémantique qui permet de mettre directement "sgn" et "const_org" en accord sémantique via "p_sur", la classe "const_org" est généralisée en classe "loc" en utlilisant la règle: herite(const_org, loc). Les trois composants de ce groupe nominal peuvent donc correspondre à la règle:

p_sur(sgn, loc, sgn)

la vérification est donc réussie, unifiant "_classe" avec "sgn".

Si les deux concepts ne sont pas en accord sémantique via le premier opérateur, on tente de les accorder en utilisant les autres opérateurs.

1.2.2. Le processus d'interprétation des requêtes

L'interprétation des connexions de différents niveaux est analogue à celle présentée dans la partie II. D'un point de vue générale, les processus se déroulent de la façon suivante:

1. La requête en entrée pour chaque niveau d'interprétation est une chaîne de groupes de mots. Un groupe de mots peut être composé d'un seul mot. Chaque groupe de mots de la chaîne correspond à une expression de la requête déjà reconnue et interprétée. Le processus d'interprétation cherche donc à regrouper successivement les différents groupes de mots de la chaîne.

2. Le processus d'interprétation d'un niveau donné cherche dans cette chaîne une connexion ayant la syntaxe correspondante. Si cette connexion peut être interprétée par une arborescence vérifiant les règles du modèle sémantique, les éléments de cette connexion sont regroupés pour former un seul groupe. Les informations syntaxiques et sémantiques concernant ce nouveau groupe sont également enregistrées afin de permettre à une consultation à tout moment lors d'une étape ultérieure.

3. Si à un niveau donné, il existe une suite de connexions (ex. A de B de C), la connexion produisant un concept au niveau sémantique le plus bas est d'abord interprétée.

4. L'existence de plusieurs solutions possibles à une étape donnée implique qu'une ambiguïté ne peut pas être résolue à ce niveau. On retient alors toutes les solutions et on les soumet aux étapes ultérieures. Si à la fin de l'interprétation, plusieurs solutions existent, la requête est considérée comme ambiguë, et toutes les solutions sont connectées par l'opérateur 'ou'.

5. Si à la fin de l'interprétation, la requête est transformée en une chaîne composée d'un seul groupe de mots, l'interprétation réussit. Sinon, elle

échoue, car la connexion entre les groupes restants n'est pas reconnue par le système.

1.3. Structuration

La structuration des documents et des requêtes est celle décrite dans la partie II. La seule différence est sur l'organisation des documents:

Dans la partie II, un document est constitué d'un ensemble d'*assertions* ne contenant pas d'opérateurs booléens ($\vee$, $\wedge$).

Tandis que dans la réalisation, nous considérons qu'un document est constitué d'un ensemble d'*arborescences*. Chaque arborescence est interprétée d'une phrase du CRM. Cette arborescence peut éventuellement contenir l'opérateur $\wedge$ et $\neg$. Une *arborescence* peut donc correspondre à un sous-ensemble d'assertions. Mais ce changement d'organisation des documents n'affecte pas le processus d'évaluation.

Ainsi, par la suite, les opérations portent sur les arborescences plutôt que sur les assertions.

1.4. Evaluation des sous-requêtes externes

Dans la partie II, nous avons proposé d'évaluer les sous-requêtes externes par des opérations d'accès similaires à celles des bases de données. Actuellement, notre prototype n'a pas encore intégré ces fonctionnalités. Nous avons donc simulé ces opérations en Prolog.

Cette évaluation correspond au schéma de la figure III.3.

Le filtrage sur l'attribut "patient" est fondé sur une comparaison entre le nom de patient de la requête et celui du CRM, ceux-ci ayant des valeurs atomiques.

Le filtrage sur l'attribut "date" repose sur une comparaison de la date du CRM avec l'expression de la sous-requête sur la date.

L'attribut "docteur" est du type ensemble. La comparaison entre la sous-requête et les docteurs d'un CRM est donc fondée sur l'appartenance.

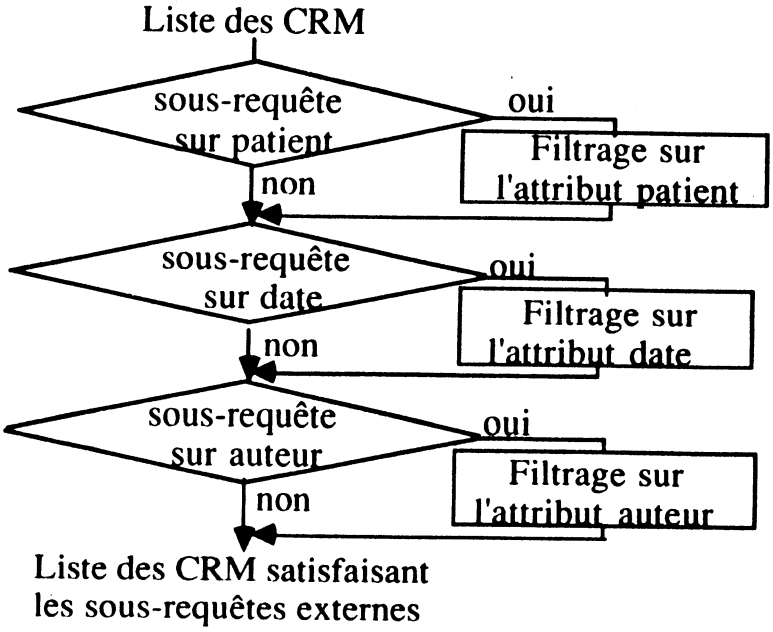


Fig.III.3. *Schéma de l'évaluation des sous-requêtes externes*

1.5. Evaluation de la sous-requête interne

L'évaluation d'une sous-requête interne se décompose en celles des propositions atomiques contenues dans la sous-requête.

L'évaluation d'une proposition atomique est résumée dans le schéma de la figure III.4.

Comme l'évaluation des sous-requêtes externes, la présélection doit être théoriquement effectuée par des opérations classiques en bases de données. Pour la même raison que précédemment, nous avons simulé cette opération en Prolog. Nous avons créé une structure de fichier inverse: les arborescences des CRM sont indexées par les concepts élémentaires. La présélection est donc effectuée via des opérations de conjonction d'arborescence.

Le module "vérification détaillée" vérifie si une arborescence présélectionnée peut satisfaire la proposition atomique. Cette vérification correspond aux algorithmes décrits dans la partie II, à la différence qu'il faut ajouter dans l'algorithme de $V_{\{A\}}(P)$ la comparaison de la proposition P avec une arborescence de la forme $A=\wedge(A_1,A_2)$, pour que $V_{\{A\}}(P)$ soit valué à:

$$MAX(V_{\{A_1\}}(P), V_{\{A_2\}}(P)).$$

Cet ajout est dû à la différence sur l'organisation des documents.

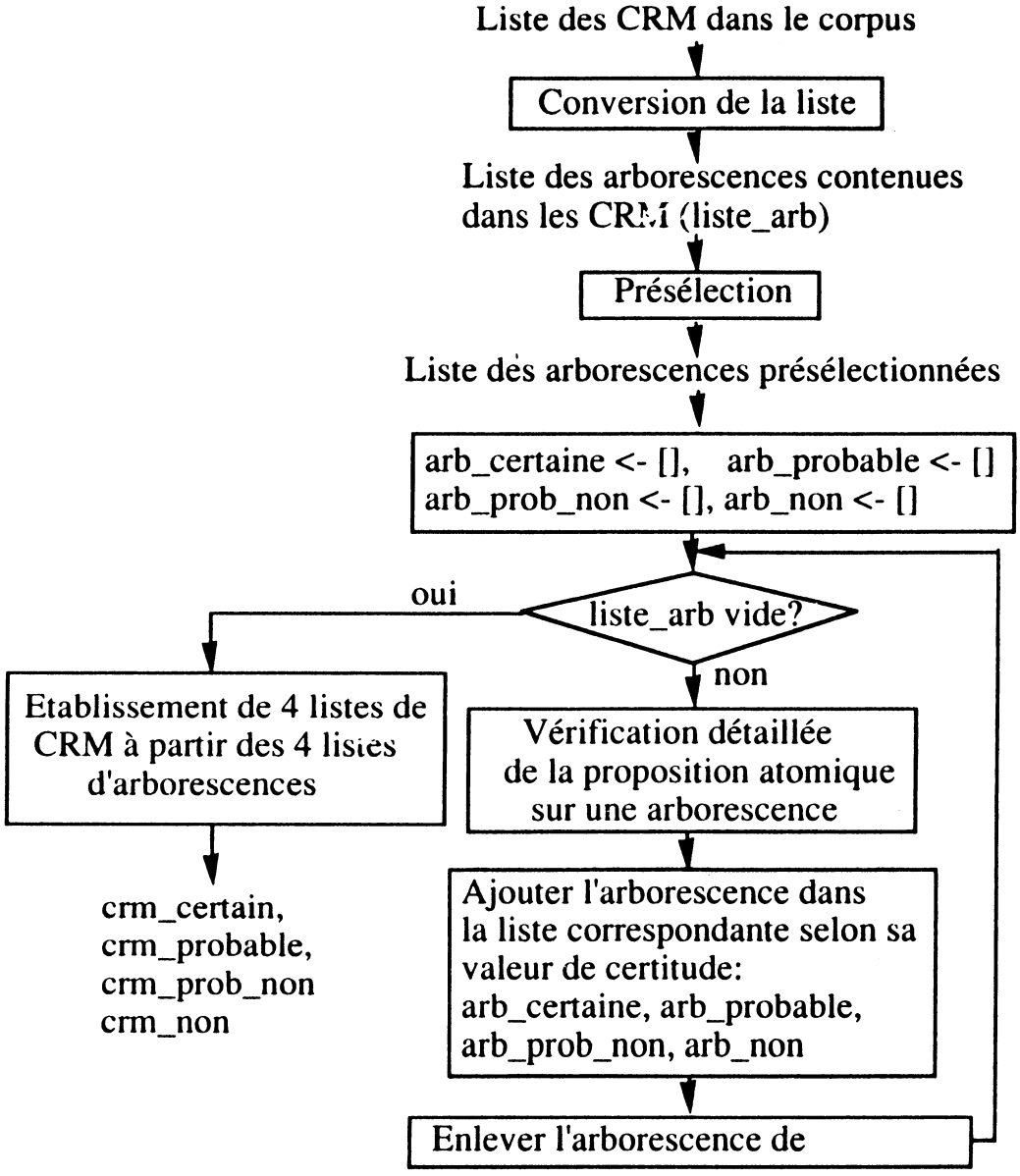


Fig.III.4. *Schéma de l'évaluation d'une proposition atomique*

Une autre particularité dans la réalisation est que nous n'avons que 4 valeurs de certitude dans les documents et dans les requêtes: certaine (1), probable (2/3), probablement-non (1/3) et non (0). Dans chaque arborescence du document, une sous-arborescence peut avoir au maximum un opérateur de certitude, ainsi que dans les requêtes. Avec la valuation $V_w(f)$ définie dans le modèle d'évaluation, toutes les vérifications d'une proposition atomique par une arborescence est valuée par une de ces 4 valeurs, ainsi que la correspondance entre le document et toute la requête. Nous pouvons donc décomposer les ensembles:

{⟨ARB, $V_{\{ARB\}}(P)$⟩: ARB∈ l'ensemble d'arborescences du corpus}
{⟨D, $V_D(P)$⟩: D∈ l'ensemble de documents du corpus}

en quatre ensembles chacun:

arb_certaine, arb_probable, arb_prob_non, arb_non
crm_certain, crm_probable, crm_prob_non, crm_non

1.6. Un exemple de session

Nous donnons ici un exemple complet de session. Les mots en italique gras (ex. ***prolog***) sont des entrées sur le clavier. Les phrases entre parenthèses (ex. /***index est le fichier inverse***/) sont des commentaires. Le reste est le résultat affiché.

Script started on Wed Jan 10 18:13:04 1990
{1}***prolog***
C-Prolog version 1.5
| ?- ***[prog].***
prog consulted 102076 bytes 3.6 sec.
yes
| ?- ***charger.***
dico consulted 87500 bytes 3.2 sec.
crm consulted 6860 bytes 0.216672 sec.
corpus consulted 146084 bytes 4.43333 sec.
imp consulted 2608 bytes 0.150003 sec.
index consulted 28992 bytes 1.01666 sec.
/* le fichier "imp" est celui sur l'implication entre concepts élémentaires */
/* index est le fichier inversé */
yes
| ?- ***q.***
Entrez votre requete, S.V.P.
|: ***"je veux des crm a partir de 1987 concernant des opacit{s ganglionnaires m{diastinales".***

/* A). phase d'interprétation de la requête, avec affichage des */
/* résultats intermédiaires des différentes sous-phases */

{{{{{ Apres la separation des attributs }}}}}
[nb(des),type(CRM),date([a,partir,de,870000]),int([des,opacit{s,ganglion naires,m{diastinales])]

{{{{{ Apres la transformation des attributs externes }}}}}
[nb(T),type(crm),date(sup_eg(870000))]

{{{{{ Apres la reconnaissance des elements }}}}}
info([des],ar,mf,p,vide)
info([opacit{s],s,f,p,[a_pr_val([densit{,car_phy],[augment{,val_qual]),sgn
])
info([ganglionnaires],a,mf,p,[ganglion,const_org])
info([m{diastinales],a,f,p,[m{diastin,const_org])

{{{{{ Apres la transformation des adjectifs }}}}}
info([des],ar,mf,p,vide)
info([opacit{s,ganglionnaires,m{diastinales],s,f,p,[p_sur([p_sur([a_pr_val
([densit{,car_phy],[augment{,val_qual]),sgn],[ganglion,const_org]),sgn],[
m{diastin,const_org]),sgn])

{{{{{ Apres la transformation de GNS }}}}}
info([des],ar,mf,p,vide)
info([opacit{s,ganglionnaires,m{diastinales],s,f,p,[p_sur([p_sur([a_pr_val
([densit{,car_phy],[augment{,val_qual]),sgn],[ganglion,const_org]),sgn],[
m{diastin,const_org]),sgn])

{{{{{ Apres la transformation de GNC }}}}}
info([des,opacit{s,ganglionnaires,m{diastinales],s,f,p,[p_sur([p_sur([a_pr
_val([densit{,car_phy],[augment{,val_qual]),sgn],[ganglion,const_org]),sg
n],[m{diastin,const_org]),sgn])

{{{{{ Apres la transformation de verbe }}}}}
info([des,opacit{s,ganglionnaires,m{diastinales],s,f,p,[p_sur([p_sur([a_pr
_val([densit{,car_phy],[augment{,val_qual]),sgn],[ganglion,const_org]),sg
n],[m{diastin,const_org]),sgn])

/* résultat de l'interprétation */

{{{{{ Arborescence de requete }}}}}

densit{ (car_phy)
a_pr_val (sgn)
augment{ (val_qual)
p_sur (sgn)
ganglion (const_org)
a_pr_val_loc (const_org)
m{diastin (const_org)

@@@@@ Analyse a utilise 0.383339 secondes @@@@@

==> Taper un retour chariot pour continuer ->

/* B). phase d'évaluation */

/*1ère sous-phase: évaluation des attributs externes */

===== Apres evaluation des att-ext =====
[10,12,13,14,23,24,26,27,28,29,30,32,33,34,35,36] **/*liste des CRM vérifiant les attributs externes */**
==> Continuer (c.) ou un listing detaille (l.) ->*l.*

/* affichage des CRM vérifiant les attributs externes */
crm(10,patient(Antoine),docteur([Coulomb]),date(870316),[101,102,103,104,105,106,107,108,109])
crm(12,patient(Annie),docteur([Rose-Pittet]),date(870130),[124,125,126,127,128,129])
crm(13,patient(Anne),docteur([Coulomb]),date(871028),[131,132,133,134, 135,136,137])
crm(14,patient(Bertrand),docteur([Rose-Pittet]),date(870303),[141,142,143, 144,145,146,147,148,149])
crm(23,patient(Louis),docteur([Rose-Pittet]),date(870203),[232,233,234,235,236,238,239])
crm(24,patient(Francois),docteur([Rose-Pittet]),date(870201),[240,241,242, 243,244,245,246,247,248,249,250])
crm(26,patient(Serge),docteur([Paramelle,Rose-Pittet]),date(870210),[261, 262,263,264,265,266,267])
crm(27,patient(Ren{),docteur([Coulomb]),date(870211),[271,272,273,274,275,276,277,278,279])
crm(28,patient(Nathalie),docteur([Coulomb]),date(870316),[280,281,282,283,284,285,286,287,288,289,290])
crm(29,patient(Denis),docteur([Coulomb]),date(870310),[291,292,293,294,295,296,297,298,299])
crm(30,patient(Patrick),docteur([Coulomb]),date(870310),[301,302,303,304,305,306,307,308,309,310,311,312])
crm(32,patient(Philippe),docteur([Coulomb]),date(870217),[321,322,323,324,325,326,327,328])
crm(33,patient(Patrice),docteur([Rose-Pittet]),date(870323),[331,332,333, 334, 335])
crm(34,patient(Domonique),docteur([Villars,Rose-Pittet]),date(870323),[341,342,343,344,345,346,347,348,349])
crm(35,patient(Florence),docteur([Rose-Pittet]),date(870106),[351,352,353,354,355])

crm(36,patient(Pierre),docteur([Rose-Pittet]),date(870323),[361,362,363,
364,365,366,367])

==> Taper un retour chariot pour continuer ->

@@@@@ Evaluation externe a utilise 0.116661 @@@@@

/* 2ème sous-phase: évaluation de l'attribut interne */
/* 1). présélection */

@@@@@@ La preselection a utilise 1.58334 secondes @@@@@@

===== Les arborescences suivantes sont preselectionnees =====
[244,279,309,328,333,355] **/*liste des arborescences contenues dans les CRM sélectionnés par les attributs externes */**

==> Continuer (c.) ou un Listing detaille (l.) ->***l.***

244->l'hypertrophie ganglionnaire {voqu{e par des opacit{s de densit{
tissulaire de 6 mm de diam}tre dans la loge lat{ro-trach{ale gauche et
dans la loge m{diastinale ant{rieure gauche
279->des opacit{s en projection du segment ventral du lobe ant{rieur
gauche associ{es a une hypertrophie ganglionnaire des loges
m{diastinales ant{rieures gauches et loges lat{ro-trach{ales gauches et
loges sous-car{naires
309->la pr{sence d'opacit{s ganglionnaires sup{rieures a 10 mm de
diam}tre sur la loge de Bar{ty et sur la chaine m{diastinale ant{rieure
gauche et para-trach{ale gauche
328->les opacit{s nombreuses ganglionnaires m{diastinales de moins de 8
mm de diam}tre
333->pas d'opacit{ ganglionnaire sup{rieure a 10 mm de diam}tre au
niveau des loges m{diastinales
355->l'opacit{ hilaire gauche sans hypertrophie ganglionnaire
m{diastinale

==> Taper un retour chariot pour continuer ->

/* 2). vérification détaillée */

===== Apres la verification detaillee =====
Pour la proposition atomique suivante:
p_sur([a_pr_val([densit{,car_phy],[augment{,val_qual]),sgn],[a_pr_val_l
oc([ganglion,const_org],[m{diastin,const_org]),const_org])

1. ARB Certaines:
[309,328]

2. ARB Probables:
[]

3. ARB Probabalement Non:
[]

4. ARB Non:
[244,279,333,355]

==> Continuer (c.) ou un listing detaille sur (o./p./pn./n.) ->***c.***

======= Apres l'evaluation globale de requete ======
1. Les CRM certains =>
[30,32]
2. Les CRM probables =>
[]

==> Continuer (c.), listing sur Certains (o.) ou sur Probables (p.) ->***o.***

/* affichage de la réponse: 2 CRM considérés certains */

--------- CRM No: 30 ----------
crm(30,patient(Patrick),docteur([Coulomb]),date(870310),[301,302,
303,304,305,306,307,308,309,310,311,312])

=> l'art{riopathie des membres inf{rieurs.
=> un carcinome {pidermoide du lobe sup{rieur gauche en projection du segment apico-dorsal.
=> les coupes jonctives.
=> le processus expansif de densit{ tissulaire h{t{rog}ne en projection du segment apico-dorsal du culmen.
=> le bombement scissural expliqu{ par un large adossement scissural.
=> l'{paississement de la scissure gauche.
=> l'opacit{ nodulaire de 8 mm de diam}tre en projection du segment apico-dorsal du culmen.
=> l'{paississement de la pl}vre sur le versant pari{tal sugg{rant un T3 pleural et non pari{tal.
=> la pr{sence d'opacit{s ganglionnaires sup{rieures a 10 mm de diam}tre sur la loge de Bar{ty et sur la chaine m{diastinale ant{rieure gauche et para-trach{ale gauche.
=> la tomodensitom{trie d{montrant l'extension trans-scissurale.
=> un T3 probable pleural.

=> une opacit{ nodulaire isol{e en projection du segment apical du lobe sup{rieur droit.

==> Taper un retour chariot pour continuer ->

--------- CRM No: 32 ---------
crm(32,patient(Philippe),docteur([Coulomb]),date(870217), [321,322, 323, 324,325,326,327,328])

=> la calcification en projection du culmen.
=> l'opacit{ excav{e de la moiti{ sup{rieure de l'h{mithorax gauche.
=> la st{nose maligne bourgeonnante de la bronche culminale et de la bronche segmentaire apicale du lobe inf{rieur gauche.
=> l'opacit{ mixte en projection du segment apical du lobe inf{rieur gauche.
=> une opacit{ de densit{ solide a tendance alv{olaire en projection du sous-segment interne accompagn{e de l'alv{ogramme a{rique et bronchogramme a{rique en projection du sous-segment externe.
=> une opacit{ par comblement alv{olaire du segment apical de la lingula.
=> des opacit{s calciques correspondant a la calcification.
=> les opacit{s nombreuses ganglionnaires m{diastinales de moins de 8 mm de diam}tre.

==> Taper un retour chariot pour continuer ->

=== Les CRM certains ===
[30,32]
=== Les CRM probables ===
[]
==> Continuer (c.), listing sur Certains (o.) ou sur Probables (p.) ->***c.***

@@@@@ Evaluation interne a utilise 2.49993 secondes @@@@@

@@@@@ Evaluation totale a utilise 2.73326 secondes @@@@@

========== Interrogation terminee ==========
==> Continuer sur la meme requete (c.), une autre requete (a.),
ou fin de session (f.) ->***f.***

/* arrêt de la session */

[Prolog execution halted]

15.6u 1.0s 2:00 13% 0+40k 24+0io 147pf+0w
{2}exit
{3}
script done on Wed Jan 10 18:15:16 1990

2. EXPÉRIMENTATION

Nous avons expérimenté le prototype sur un ensemble de 36 CRM, soit environ 300 arborescences. A partir de cette expérimentation, nous pouvons faire le bilan suivant relativement aux caractéristiques du prototype:

• Sur le processus d'interprétation

Bien que le langage d'interrogation quasi-naturel possède une syntaxe assez simple, la plupart des phrases originales des CRM peuvent être décrites. Les performances du processus d'interprétation ont été largement démontrées dans l'expérimentation: les interprétations des phrases des CRM correspondent tout à fait aux interprétations manuelles. A partir de cette expérience, nous pouvons voir que la qualité d'une interface en langue naturelle n'est pas fonction de sa complexité, mais de son efficacité quant aux problèmes à traiter.

• Sur le processus d'évaluation

1. Le prototype donne une très bonne précision: les CRM extraits pour une requête sont presque tous pertinents pour la requête.

2. Il donne un très bon rappel: tout CRM décrivant explicitement le même fait que la requête est retrouvé. Cette estimation est uniquement fondée sur le critère du "sens explicite". Comme il est indiqué dans la partie II, nous n'avons pas pris en compte les connaissances "déductives" qui permettraient, par exemple, de déduire "opacité pathologique" à partir de "opacité pulmonaire de plus de 10 mm de diamètre". Cet aspect déductif reste à être exploité dans la suite de ce travail.

Pour analyser les performances du prototype, nous avons soumis trois types de requêtes à RIME: des requêtes portant uniquement sur l'attribut interne - type A (environ 200), les mêmes requêtes portant sur l'attribut "date" dont la valeur est "à partir de 1987" - type B, et les requêtes portant uniquement sur des attributs externes - type C (13 requêtes).

L'attribut interne dans les deux premiers types de requêtes est choisi de façon aléatoire, depuis des requêtes simples comme:

Je veux des crm montrant des opacités.

jusqu'à des requêtes complexes telles que:

Je veux des crm montrant des opacités pulmonaires supérieures à 10 mm de diamètre au niveau du lobe supérieur droit.

Nous résumons les résultats de ces interrogations dans les figures suivantes où la complexité est mesurée par le nombre de concepts élémentaires différents existant dans la représentation interne de la requête:

A). requêtes portant sur l'attribut interne:

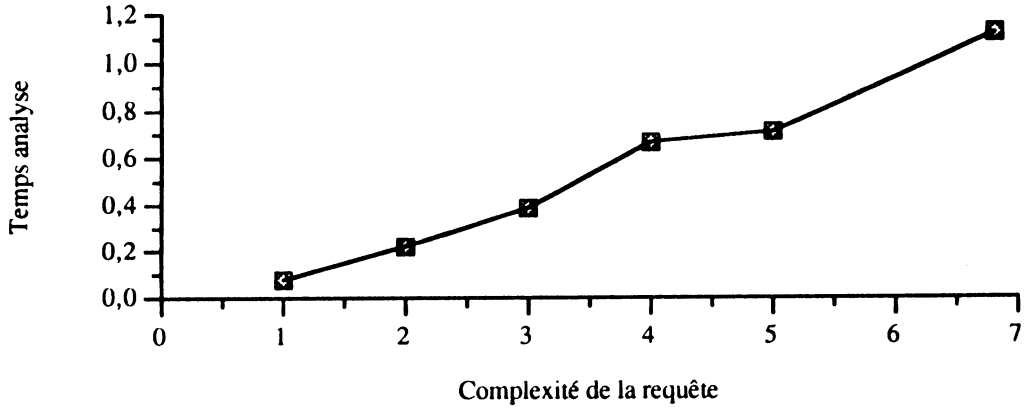


Fig.III.5.*Temps de l'analyse*

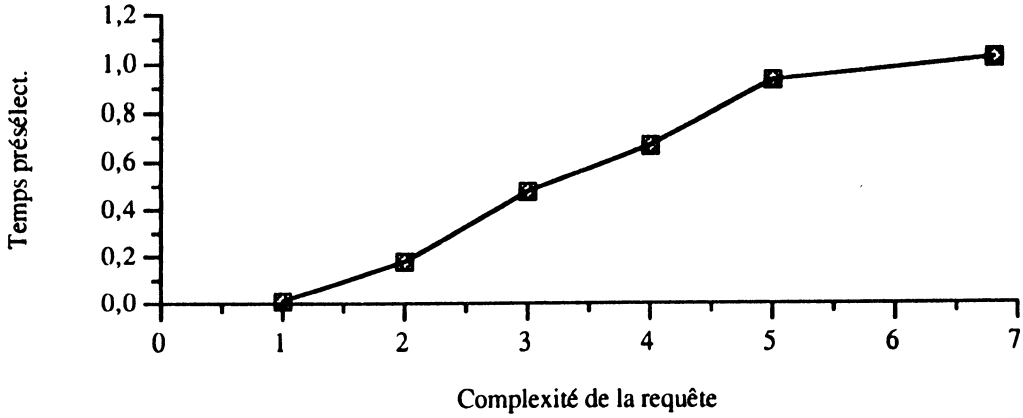


Fig.III.6. *Temps de la présélection*

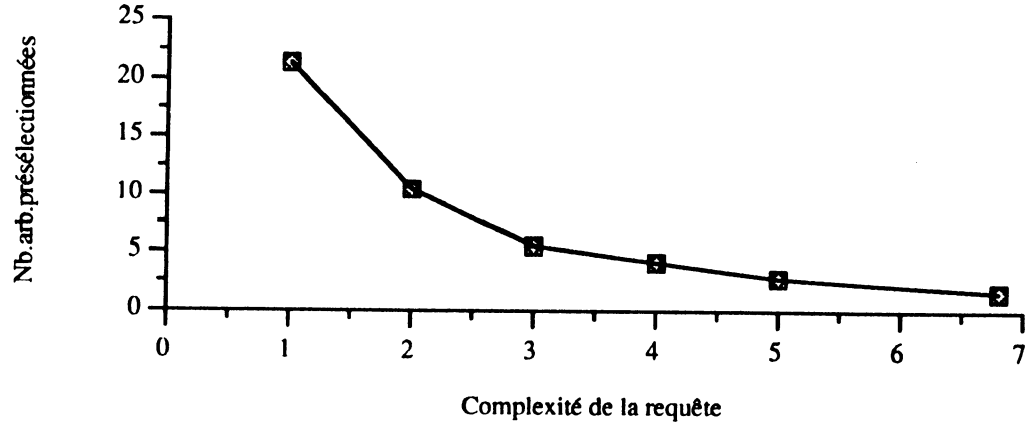


Fig.III.7. *Nombre d'arborescences présélectionnées*

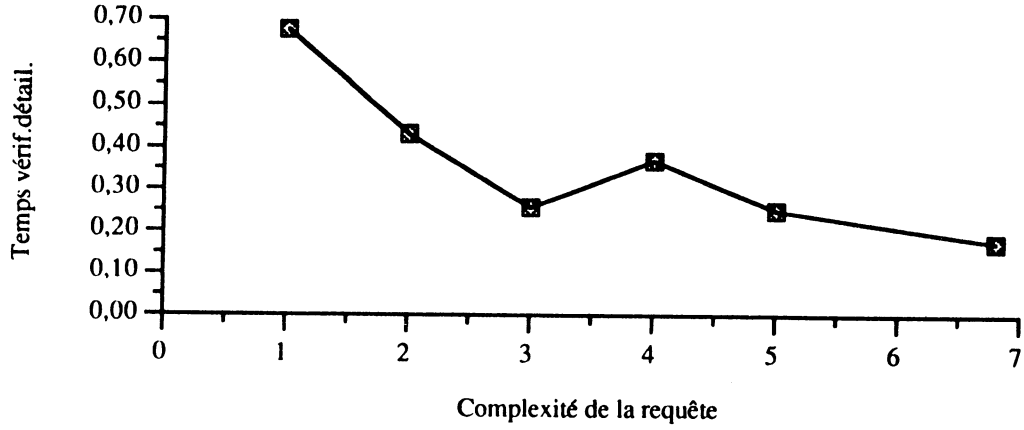


Fig.III.8. *Temps de la vérification détaillée*

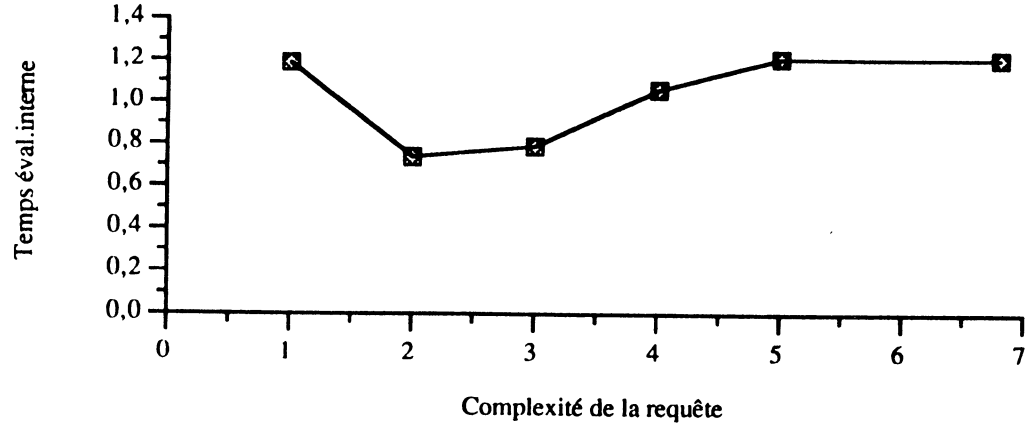


Fig.III.9. *Temps de l'évaluation totale (de l'attribut interne)*

B). requêtes portant sur l'attribut interne et l'attribut "date":

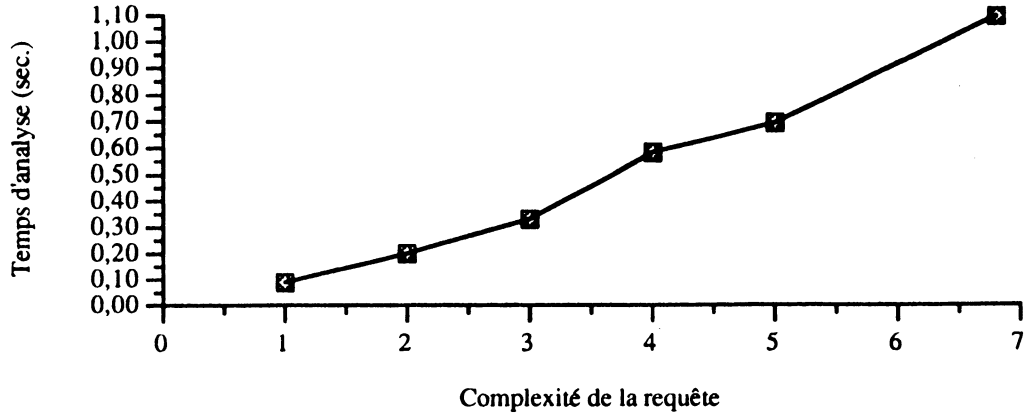


Fig.III.10. *Temps de l'analyse*

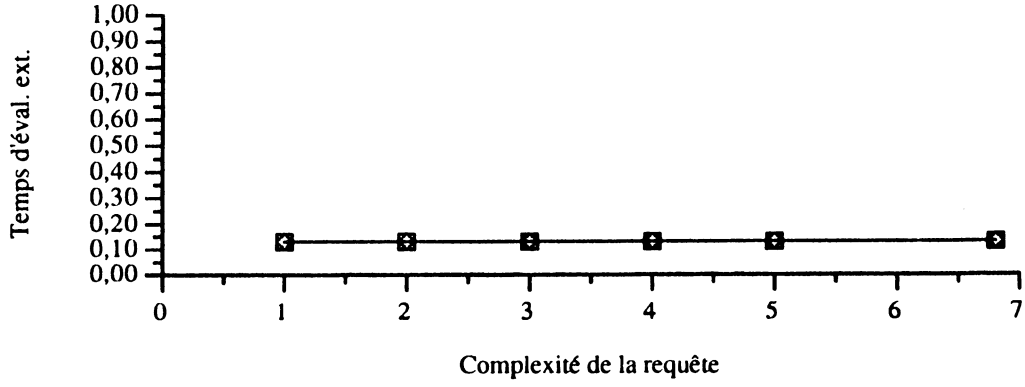


Fig.III.11. *Temps de l'évaluation des attributs externes*

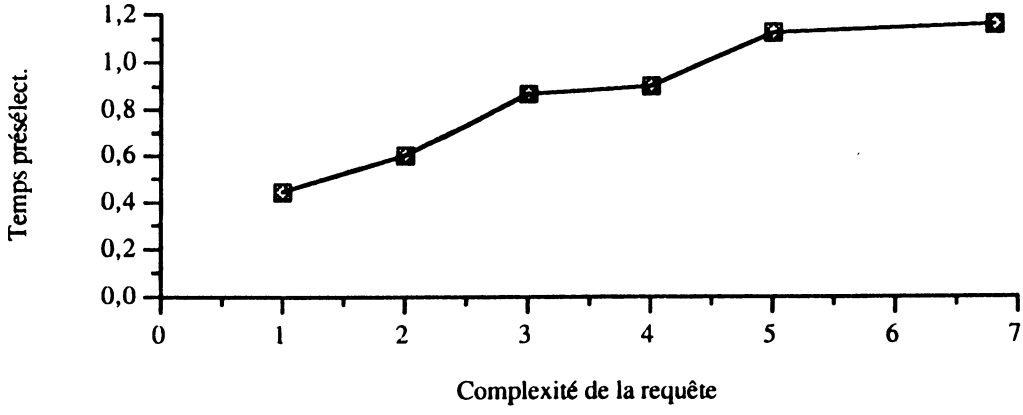


Fig.III.12. *Temps de la présélection*

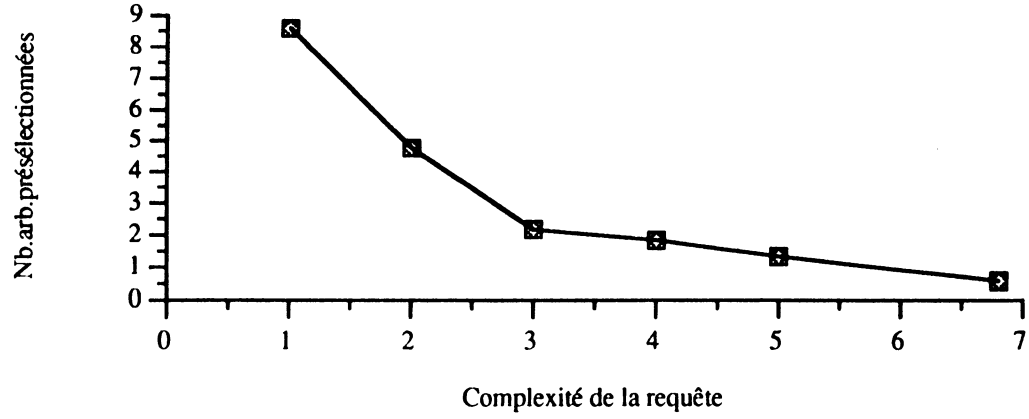


Fig.III.13. *Nombre d'arborescences présélectionnées*

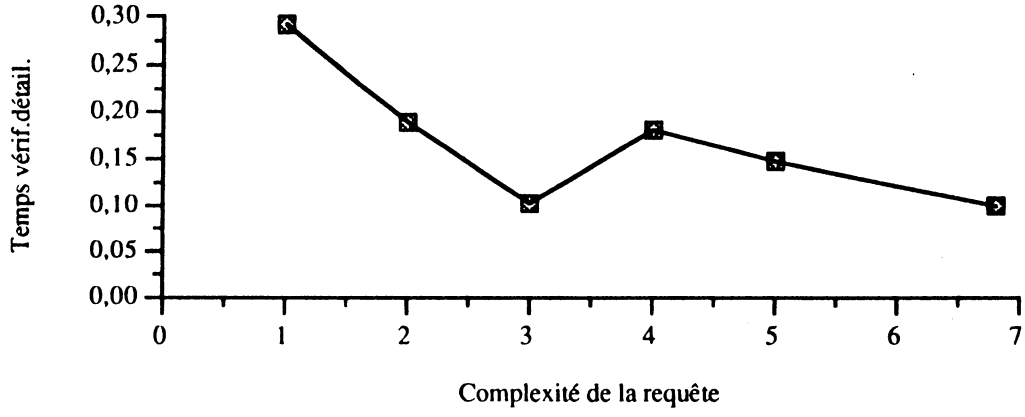


Fig.III.14. *Temps de la vérification détaillée*

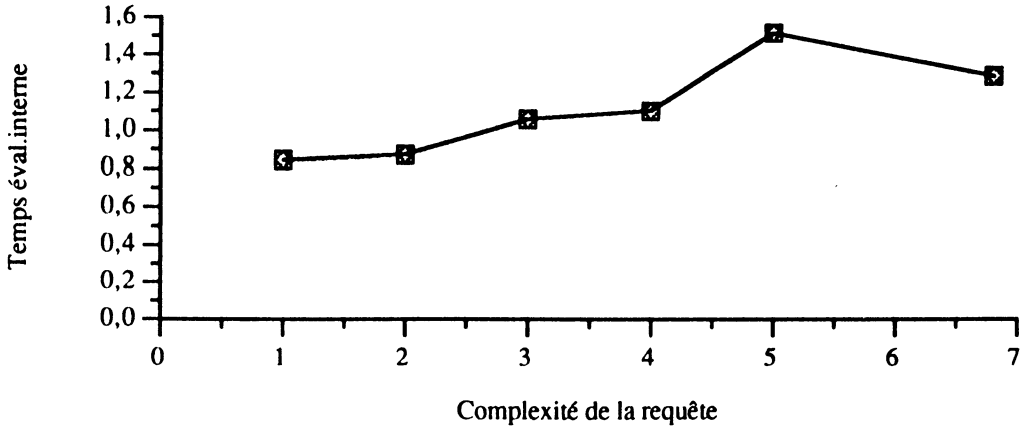


Fig.III.14. *Temps de l'évaluation de l'attribut interne*

C). requêtes portant sur des attributs externes:

Temps d'analyse	Temps d'évaluation	Nb.crm séléct.
0	0,116661	23
0	0,0499954	2
0	0,116669	16
0	0,0833359	16
0	0,183334	36
0	0,0666656	7
0,0166626	0,0666656	0
0	0,0500031	1
0,0166626	0,0833359	12
0	0,0333328	0
0,0166702	0,0333328	1
0,0166702	0,216667	21
0	0,0833282	3
moyenne= 0,0051281	0,0910251	10

A partir de ces résultats, nous pouvons donner les conclusions suivantes:

1. Le temps d'interprétation d'une requête est proportionnel à la complexité de la requête (Fig.III.5 et III.10). Mais ce temps reste acceptable.

2. Le temps d'évaluation des attributs externes (cf. tableau ci-dessus) reste faible mais la sélection est assez forte (1/4). La sélection sur les attributs externes est donc efficace et peu coûteuse.

3. Le temps de présélection est aussi proportionnel à la complexité de la requête (Fig.III.6 et III.12), car plus une requête est complexe, plus la requête de présélection est complexe et nécessite donc de temps.

4. Le nombre d'arborescence présélectionnées diminue quand la complexité de la requête augmente (Fig. III.7 et III.13).

5. Le temps de la vérification détaillée n'est pas proportionnel à la complexité de la requête (Fig.III.8 et III.14).

En effet, la vérification détaillée est influencée par deux facteurs: le nombre d'arborescences soumises à la vérification détaillée et la complexité de la requête. D'un côté, plus la requête est complexe, plus la présélection est forte, et moins on obtient d'arborescences à soumettre à la vérification détaillée. La complexité est un facteur qui diminue le temps de la vérification détaillée. D'un autre côté, plus la requête est complexe, plus la vérification sur

une arborescence de CRM est coûteuse. La complexité de la requête est donc aussi un facteur qui augmente le temps de la vérification détaillée. Le temps total est donc un compromis entre ces deux facteurs. Voilà pourquoi le temps de la vérification détaillée est irrégulier par rapport à la complexité de la requête.

Le temps pour l'évaluation de l'attribut interne est voisin de la somme des temps de la présélection et de la vérification détaillée. On peut observer une légère différence entre la somme de ces deux temps et celui de l'évaluation de l'attribut interne. Rappelons que l'évaluation d'une sous-requête interne est composée de deux parties: une vérification détaillée sur les sous-arborescences composées d'opérateurs sémantiques, et une évaluation de la sous-requête globale. Cette différence correspond au temps d'évaluation de la requête globale.

6. En comparant les résultats des interrogations de type A et de type B, on peut voir l'influence de l'attribut "date" (dont la valeur est "à partir de 1987"):

a). Le temps de l'analyse de la requête reste analogue (Fig.III.5 et III.10).

b). Le nombre d'arborescences présélectionnées est largement diminué (un peu plus d'une moitié) (Fig.III.7 et III.13).

c). Le temps de la vérification détaillée est aussi diminué de plus de la moitié (Fig.4 et 10). On montre aussi que le temps de la vérification détaillée est proportionnel au nombre d'arborescences soumises.

d). Le temps de la présélection augmente (Fig.III.6 et III.12). Ainsi, les avantages précédemment cités ne sont pas clairement présents dans le temps d'évaluation de l'attribut interne.

La cause de cette augmentation est la simulation des opérations de bases de données par celles de Prolog dans la présélection. Dans la réalisation actuelle, la présélection est fondée sur une organisation du type fichier-inverse. Une présélection est donc une série d'intersections d'ensembles (de numéros d'arborescence). L'existence des attributs externes nécessite donc une opération d'intersection supplémentaire avec les CRM sélectionnés par les attributs externes. Cette intersection supplémentaire dure environ 0,4 seconde. En comparant les temps pour les requêtes de complexité 1, on peut voir clairement cette différence. En conclusion, l'augmentation du temps de la présélection est liée à l'implémentation. Quand un SGBD sera utilisé pour la présélection, cette augmentation disparaîtra.

Pour donner une comparaison plus significative, nous avons reconstitué la figure suivante de l'évaluation de l'attribut interne pour les requêtes du type B, en prenant les mêmes temps de présélection que pour les requêtes du type A. En comparant cette figure avec la Fig.III.9, on voit que le temps d'évaluation de l'attribut interne diminuera en fonction des attributs externes, surtout pour les requêtes simples (car dans ce cas, le temps d'évaluation de l'attribut interne est fortement influencé par celui de la vérification détaillée, celui-ci étant largement diminué à cause de la diminution du nombre d'arborescences présélectionnées). Cette comparaison montre l'influence des attributs externes sur toute l'évaluation fondée sur une base de données.

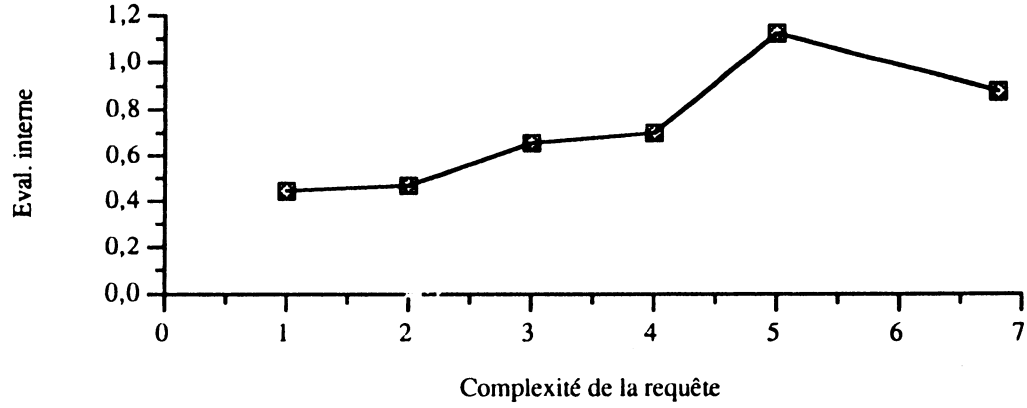


Fig.III.15. *Temps estimé pour l'évaluation de l'attribut interne des requêtes de type B*

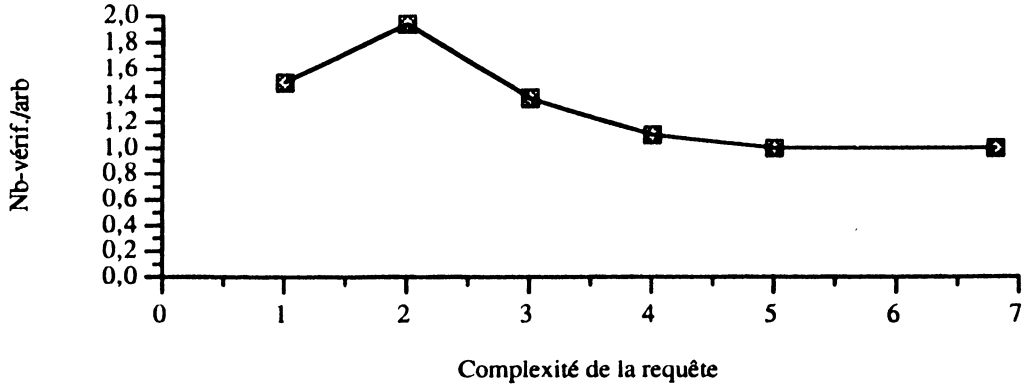


Fig.III.16. *Nombre d'essais pour vérifier une sous-arborescence*

Une vérification détaillée commence par un contrôle sur les classes sémantiques (étape1). La vérification se poursuit (étape2) quand le contrôle est réussi. Pour analyser les performances de ce processus, nous avons fait une statistique sur le nombre de poursuites (étape2) par vérification d'implication

entre une arborescence de CRM présélectionnée et une sous-arborescence (ne contenant que des opérateurs sémantiques) de la requête. Le résultat est montré dans Fig.III.16.

On observe que le nombre d'essais diminue quand la complexité de requête augmente, sauf pour les requêtes de complexité "1" dont la vérification détaillée d'un concept élémentaire est un cas particulier. Quand la requête est complexe, une vérification ne nécessite qu'un essai.

On observe qu'il faut environ 1,4 essais en moyenne pour compléter une vérification détaillée pour une sous-arborescence de la requête, ce qui est très peu. Ce résultat signifie que la présélection sur les concepts élémentaires suivie du contrôle des classes sémantiques représente un critère de sélection presque suffisant. Les moyens supplémentaires à mettre en oeuvre pour contrôler le processus de vérification détaillée n'apportent pas nécessairement les avantages escomptés si l'on compare l'amélioration que cela apporte avec le surcoût du contrôle.

3. CONCLUSION

• Sur la phase interprétation de la requête

Le processus d'interprétation a été utilisé pour indexer les CRM du corpus. La plupart des phrases dans ces CRM sont interprétées correctement.

Le processus d'interpréttaion a été montré performent et efficace. D'une part, l'utilisation simultanée des informations syntaxiques et sémantiques a permis de résoudre les problèmes d'ambiguïtés. D'autre part, le processus se déroulant par rapport à la structure syntaxique des phrases, le résultat est très rapidement obtenu. Cela montre que la méthode d'analyse des phrases proposée dans la partie II est envisageable dans un système réel.

Malgré sa simplicité, le langage d'interrogation est presque suffisant - la plupart des phrases originales dans les CRM peuvent être décrites dans ce langage. Mais nous avons aussi rencontré des phrases dont l'interprétation nécessite une légère modification de la structure de la phrase.

Par exemple, pour pouvoir interpréter la phrase "mise en évidence, au niveau du lobe supérieur droit, de multiples images nodulaires de petites dimensions", il a fallu la modifier en "de multiples images nodulaires de petites dimensions mises en évidence au niveau du lobe supérieur droit", car la structure de la phrase originale n'est pas décrite dans la syntaxe du langage d'interrogation.

Dans le système RIME réel, ce langage d'interrogation devrait donc être légèrement élargi.

Quant à l'aspect vocabulaire, le système de dictionnaire étudié dans le groupe de recherche ([Palmer90]) apportera beaucoup d'intérêt. Cet aspect n'a pas été considéré dans le prototype.

La définition actuelle du modèle sémantique n'est pas encore complète. Certaines classes sémantiques ne sont pas incluses, par exemple, la classe "opération chirurgicale". Or il existe dans les CRM des concepts de cette classe.

Nous avons trouvé une incohérence dans une précédente version du modèle sémantique. La formation suivante n'est pas décrite dans cette version:

SGN ::= [a-pr-val](CAR-PHY, VAL-QUAL)

Or, le terme "opacité" est considéré comme un SGN, mais a une représentation interne correspondant à cette formation dans la même version:

([a-pr-val]((densité,CAR-PHY),(augmenté,VAL-QUAL)),SGN)

Pour compléter la formation correspondant à celle de "opacité" (et beaucoup d'autres), nous avons isolé une partie des anciennes valeurs qualitatives (VAL-QUAL) en VAL-SGN (ex. augmenté, rétracté, ...). Nous ajoutons la formation suivante:

SGN ::= [a-pr-val](CAR-PHY, VAL-SGN)

Cette création d'une nouvelle classe résoud le problème d'incohérence dans ce cas particulier. Mais le problème de cohérence doit être vérifié, de façon plus systématique, sur toute la définition du modèle sémantique de façon.

En conclusion, une étude plus approndie, en collaboration avec des médecins, devrait porter sur la définition du modèle sémantique.

- **Sur la phase évaluation de la requête**

Les résultats obtenus sont très encourageants. Malgré la diminution de performance due à la simulation des opérations de base de données avec Prolog, le temps reste acceptable. Le coût de l'évaluation d'une requête est d'environ 1 seconde en moyenne. Si l'on admet l'hypothèse que le temps d'évaluation augmente proportionnellement à la taille du corpus, on peut

prévoir que le prototype actuel peut effectuer une évaluation de requête (sans les attributs externes) sur un corpus de 10,000 CRM en moins de 5 minutes.

L'expérimentation nous permet donc de conclure que l'approche proposée dans la partie II est envisageable.

Relativement aux résultats obtenus, on peut faire les remarques suivantes concernant des points à développer dans le futur:

1. L'évaluation des sous-requêtes externes par des techniques de bases de données peuvent largement améliorer les performances.

2. Le parallélisme permettrait d'exécuter simultanément la vérification détaillée sur plusieurs sous-arborescences composées d'opérateurs sémantiques. Le temps de la vérification détaillée pourrait être sensiblement diminué.

3. Il existe des redondances entre les vérifications détaillées, car plusieurs sous-arborescences peuvent avoir des composants en commun (surtout après la standardisation). La vérification détaillée séparée de ces sous-arborescences engendre une redondance. L'optimisation des vérifications détaillées permettrait sans doute de réduire la redondance.

PARTIE IV

CONCLUSION

1. SUR LE MODELE

La définition du modèle est un problème clé des SRI. La qualité du modèle influence directement les performances du système.

Dans la première partie, nous avons proposé un modèle général fondé sur la logique modale floue. Par rapport aux modèles existants, le modèle proposé a montré sa généralité et sa richesse quant à l'aspect sémantique: non seulement tous les modèles existants peuvent être obtenus à partir du modèle proposé, mais aussi et surtout le modèle proposé a intégré l'aspect sémantique dans sa définition. Or dans les modèles existants (intelligents), la sémantique est toujours prise en compte dans un composant non intégré au modèle, avec des approches particulières telles que celle de "feedback".

Ces nouvelles caractéristiques du modèle proposé permettent:

- d'une part, de mieux comprendre le comportement d'un SRI et d'unifier les modèles existants sur une base commune et solide, permettant la comparaison de différents modèles existants sur leurs définitions plutôt que sur leurs résultats seulement;

- d'autre part, de bien situer l'aspect sémantique dans les SRI, ce qui permet d'apprécier la tendance actuelle de la recherche et offre un critère aux futurs développements.

La validité du modèle a été doublement vérifiée: par la comparaison avec les modèles existants, et par son application dans un cas concret (RIME). La première vérification montre que le modèle est valide sur le plan théorique, tandis que l'expérimentation du RIME nous a permis de conclure que le modèle est aussi valide dans un cas pratique.

La définition du modèle n'est qu'une première configuration. Certains aspects restent à développer:

- Les connaissances du système sont actuellement représentées de multiples façons. Ces connaissances ne sont pas toujours directement utilisables dans l'évaluation des requêtes. Dans les recherches actuelles, beaucoup d'études se concentrent sur les relations spécifiques, les relations génériques, les relations de voisinage, etc. Or la relation utilisée dans le modèle est celle d'implication. Il est donc nécessaire d'unifier les relations définies dans la recherche avec celle du modèle, ce qui nécessite des études approfondies.

- Nous avons dégagé deux critères de jugement sur la correspondance entre document et requête. Les deux critères sont coordonnés par un paramètre dans le modèle. Chaque système particulier possède une définition particulière

du paramètre. Il a été montré que le jugement varie selon le domaine d'application et selon l'utilisateur. Le paramètre défini dans le modèle doit varier de la même manière. Ainsi, dans une application particulière, il serait intéressant de donner une définition du paramètre variable selon l'utilisateur, voire de le faire définir plus ou moins par l'utilisateur lui-même.

- Comme nous l'avons montré dans la dernière remarque, l'évaluation d'un modèle doit varier selon l'utilisateur et selon le domaine d'application. L'intérêt de prendre en compte une classification des utilisateurs pour améliorer les performances du système est maintenant bien acceptée. Mais il n'a pas été encore question, bien que cela ne paraîtrait pas moins intéressant, de classifier les domaines d'applications. Cette classification pourraient donner une vision globale sur toutes les applications, qui aiderait à affiner la modélisation des SRI. Une étude plus approfondie et plus systématique à ce propos pourrait beaucoup faciliter la réalisation des SRI.

- Il a été montré que l'approche bases de données déductives constitue une voie très intéressante pour la réalisation d'un SRI intelligent. Bien que le développement actuel de celles-ci ne nous permette pas encore de les utiliser directement pour implémenter des SRI, cete approche constituera sans nul doute un moyen de réalisation privilégié dans un proche futur.

2. SUR LE SYSTEME RIME

Les études sur RIME ont un double objectif:

- montrer un exemple d'utilisation du modèle proposé dans un cas particulier, et ainsi vérifier sa validité.

Le processus d'évaluation des requêtes du prototype RIME a été fondé sur modèle proposé. L'expérimentation montre que la méthode d'évaluation du modèle donne de très bons résultats, ce qui constitue une démonstration de la validité du modèle proposé.

- montrer l'intérêt des SRI dans les domaines où les données sont représentées de façon précise, et l'intérêt d'une interface en langue naturelle.

La plupart des SRI existants sont utilisés dans les applications où les documents et les connaissances sont représentés de manière approximative. RIME se distingue donc des autres SRI par sa représentation très précise des connaissances et des données, ce qui permet au processus d'évaluation d'être fondé sur une comparaison fine de données. Cette expérimentation montre que les SRI sont aussi applicables dans des domaines où la représentation des connaissances doit être précise. Ces deux différents types de domaines d'application reflètent parfaitement la potentialité des SRI.

Le prototype RIME a démontré l'intérêt d'une interface en langue naturelle: dans l'expérimentation, les requêtes soumises sont analogues à des phrases figurant dans les CRM sous leur forme originale, ce qui signifie que les médecins ayant rédigé ces documents peuvent utiliser directement le système sans aucune formation préalable.

Les études sur RIME se sont déroulées dans un cadre limité. Comme nous l'avons indiqué dans la partie III, certains aspects n'ont pas été suffisamment étudiés:

- Le langage d'interrogation doit être légèrement élérgi pour décrire tous les phrases possibles dans des CRM.

- Il faudrait intégrer un système de dictionnaire dans RIME (le système étudié dans le groupe de recherche [Palmer90]) pour prendre en compte l'aspect gestion du vocabulaire. Celui-ci n'a pas été considéré dans le prototype.

- Le modèle sémantique utilisé dans le prototype est un modèle restreint. D'une part, certaines formations admises des concepts médicaux ne sont pas décrites. D'autre part, le problème d'ambiguïté n'est pas totalement résolu. Ce modèle sémantique doit donc être complété et raffiné.

- La sélection sur les attributs externes et la présélection sur les concepts élémentaires doivent être effectuées par des opérations de bases de données. Dans le prototype actuel, ces opérations sont simulées en Prolog, ce qui affecte beaucoup les performances. Dans le futur développement de RIME, l'intégration d'un SGBD (Oracle, par exemple) est prioritaire.

- Les connaissances du système que l'on a prises en compte sont très limitées: elles sont restreintes à une partie des relations existant entre ces concepts élémentaires. Les connaissances du type déductif n'ont pas été considérées. Mais elles sont beaucoup utilisées dans le domaine médical et elles sont relativement bien formalisées. Le développement ultérieur de RIME devrait donc porter aussi sur l'élargissement de ce type de connaissances du système.

ANNEXES

ANNEXE 1. TRAITEMENTS DE LANGUE NATURELLE

Les recherches sur les traitements des langues naturelles en informatique remontent à l'apparition de l'ordinateur. De multiples approches ont été depuis développées, parmi lesquelles on distingue souvent trois catégories selon les informations utilisées et l'objectif visé: approche syntaxique, approche sémantique, et approche syntaxico-sémantique.

1.1. Approches Syntaxiques

Les approches syntaxiques visent à la reconnaissance de la structure syntaxique des phrases. Elles disposent des informations lexicales spécifiant la catégorie syntaxique (ou grammaticale) de chaque composant élémentaire (mot) dans la phrase, et des informations syntaxiques (représentées par une grammaire) concernant la compositios des phrases (sujet, objet, groupe nominal, etc). L'objectif des approches syntaxiques est alors de trouver une structure admise par la grammaire qui correspond à la phrase à analyser.

Parmi ces approches, on peut mentionner les grammaires de chaîne de Harris ([Harris68]), les grammaires transformationnelles de Chomski ([Chomski57]), les grammaires de cas de Fillmore ([Fillmore68]), etc. A titre d'exemple, nous allons considérer les grammaires transformationnelles de Chomski un peu plus en détail.

Sous forme standard, la théorie de Chomski suppose que la formulation d'une idée se déroule en deux temps:

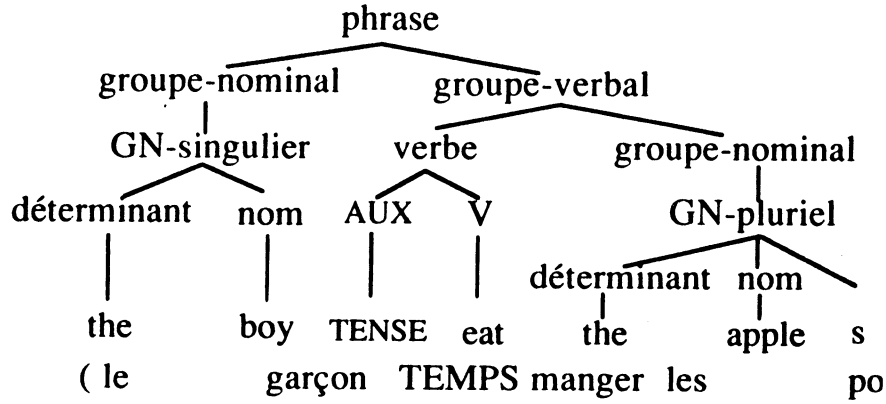
- d'abord la création d'une *structure profonde*,
- ensuite la génération de la *structure de surface* à partir de la structure profonde.

Une structure profonde détermine "l'interprétation sémantique" d'une phrase et une structure de surface détermine son "interprétation phonétique" ([Chomsky65]). Dans cette approche, la création de la structure profonde est exprimée à travers une *grammaire*, dit "générative", c'est-à-dire qu'elle doit avoir une définition de longueur finie, capable de représenter un nombre infini de phrases possibles dans cette langue et de donner à chacune une description de structure qui caractérise la sémantique de la phrase. A cause de la grande variété d'expression existant dans les langues naturelles, de multiples structures de surface peuvent correspondre à une structure profonde. Implicitement, il existe certaines *règles de transformation* permettant d'obtenir d'autres structures de surface à partir d'une structure de surface donnée. L'approche de

Chomski est essentiellement fondée sur une grammaire et un ensemble de règles de transformation.

Par la suite, nous donnons un exemple concret pour illustrer le fonctionnement de cette grammaire. Supposons que la grammaire suivante de structure soit utilisée pour obtenir l'arborescence de dérivation ci-dessous:

P ::= GN GV
GN ::= GNS | GNP
GNS ::= DET NOM
GNP ::= DET NOM *'s'*
GV ::= VERBE GN
VERB ::= AUX V
DET ::= *'the'*
NOM ::= *'boy'* | *'apple'*
V ::= *'eat'*
...



Pour générer la phrase "*The boy ate the apples*" (Le garçon mangea les pommes), on doit appliquer la transformation qui change "TENSE + *eat*" en "*eat* + PAST"; une autre règle de transformation morpho-phonétique transforme ensuite "*eat* + PAST" en "*ate*". Une telle transformation est une *transformation obligatoire*: elle opère sur la morphologie des mots en prenant compte les accords en temps et en nombre. Une autre sorte de transformation, appelée *transformation optionnelle*, opère sur la structure syntaxique de la phrase et l'ordre des mots. Par exemple, pour changer une phrase de la forme active en forme passive, la règle suivante est appliquée:

GROUPE-NOMINAL1 + AUX + V + GROUPE-NOMINAL2
⇒ GROUPE-NOMINAL2 + (AUX + *be*) + (en V) + *by* + GROUPE-NOMINAL1

où (en V) dénote la forme participe passé du verbe.

Les approches syntaxiques sont souvent des moyens efficaces pour analyser la structure des phrases étant donné qu'il existe une grammaire peu volumineuse et, relativement à la sémantique, peu variée dans une langue donnée. Mais dans beaucoup d'applications, ce que visent ces approches (la

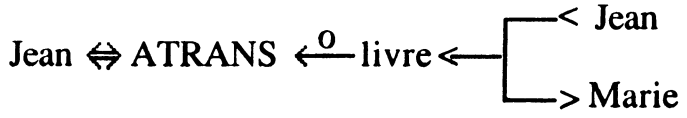
structure syntaxique) ne doit pas être considéré comme la finalité du traitement de la langue naturelle. Dans les SRI, par exemple, une opération de recherche ne se ramène pas à une recherche de documents contenant des phrases ayant la même structure syntaxique que la requête. L'objectif des SRI est de retrouver les documents représentant la même sémantique que la requête. Ainsi, les approches syntaxiques sont considérées uniquement comme des outils servant à une analyse sémantique.

1.2. Approches Sémantiques

Une partie des études sur ces approches ont d'abord eu pour but la compréhension de textes de petite taille ([Shanck80b]), ce qui explique l'objectif de ces approches: la compréhension de la sémantique d'un texte. Beaucoup de modèles ont été développés dont le but est de simuler la compréhension humaine. Une de ces approches, la théorie de dépendance conceptuelle ([Shanck80b]), est destinée à donner une représentation totale du sens du texte. Cette théorie est fondée sur l'hypothèse qu'il existe un "interlangage" dans l'être humain, c'est-à-dire un langage indépendant de tous les langages réels, capable de donner une représentation du le sens. Une des caractéristiques qui attire l'attention dans cette théorie est sa représentation explicite des connaissances implicites.

Dans la théorie de Shanck, la compréhension d'un texte en langue naturelle consiste d'abord à lui donner une "représentation conceptuelle", puis à inférer des informations implicites en utilisant les connaissances acquises dans les "scripts" (sénarios).

La représentation conceptuelle d'une phrase est composée de "primitives conceptuelles" reliées par des "dépendances conceptuelles". Par exemple, la phrase "Jean donne un livre à Marie" est représentée par



où "Jean", "livre", "Marie" et "ATRANS" sont des primitives conceptuelless, et ⇔, ←, $[^{<}_{>}$ dénotent des dépendances conceptuelles. Ce diagramme signifie: Jean exécute une action ATRANS (transfert de possession) sur l'objet "livre" dans le sens de Jean à Marie.

Dans la théorie de Shanck, la représentation des évènements est considérée comme le coeur de la représentation du sens ([Schank81]). Toutes les informations attachées à un évènement, sont décrites dans un *script*. La notion de script est illustrée par l'exemple suivant::

chaque évènement a
un ACTEUR
une ACTION effectuée par l'acteur
un OBJET sur lequel l'action est effectuée
une DIRECTION dans laquelle est effectuée l'action.

A partir de la représentation conceptuelle obtenue précédemment, le système essaie d'instancier le script, soit par des informations données dans la phrase, soit par des inférences du système. Par exemple, le verbe "donner" correspond au script de gauche. La phrase "Jean donne un livre à Marie" instancie ce script pour fournir celui de droite:

ACTEUR :	ACTEUR : Jean
ACTION : ATRANS	ACTION : ATRANS
OBJET :	OBJET : livre
DIRECTION: DE	DIRECTION: DE Jean
VERS	VERS Marie

La description des évènements en scripts permet, dans une certaine mesure, l'enchaînement d'histoires et la déduction de certaines informations manquantes.

Par rapport aux approches purement syntaxiques, les approches sémantiques permettent d'effectuer un traitement plus fin et plus profond sur la langue naturelle. Elles touchent le sens véhiculé d'une langue, ce qui est l'objectif de beaucoup d'applications de traitement de la langue naturelle. Par contre, elles ne s'intéressent pas, ou très peu au véhiculant, c'est-à-dire la forme de la langue. Le principal argument dans ce type d'approches est l'importance de la sémantique dans les traitements de langue naturelle, au mépris souvent de la syntaxe. Ainsi on voit apparaître un inconvénient relativement commun à ces approches: l'inefficacité parfois de l'analyse. Dans beaucoup de cas, une simple analyse syntaxique permet de mieux (ou plus rapidement) connaître le rôle des composants de phrase et les relations existant entre eux. Elle donne, en quelque sorte, un racourci pour l'analyse sémantique. Mais elle est souvent réfusée dans les approches purement sémantiques. Un autre inconvénient de ces approches est la complexité de leur analyse. Etant données que les informations à considérer sont souvent volumineuses, il n'est pas possible de faire une telle analyse, ce qui explique pourquoi les applications réussies des approches purement sémantiques sont toutes sur les domaines très restreints.

1.3. Approches syntaxco-sémantiques

L'interêt s'est naturellement tourné vers une collaboration entre la syntaxe et la sémantique afin de bénéficier des avantages de chacune. Beaucoup

de travaux ont été effectués dans cette optique. Par la suite, nous présentons une de ces approches - celle du GETA.

Le système ARIANE du GETA, conçu pour la traduction automatique, établit pour une phrase à analyser, une structure multi-niveaux qui contient plusieurs types d'informations sur la phrase. Celles-ci sont disponibles à chaque instant durant la traduction ([Vauqois75, Yusoff87, Zajac86]). Une structure multi-niveaux peut être vue sous plusieurs angles: au niveau des classes syntaxiques et syntagmatiques, au niveau des fonctions syntaxiques, au niveau des relations logiques et au niveau des relations sémantiques.

- Le niveau des classes syntaxiques et syntagmatiques décrit un parenthésage de la chaîne d'entrée. Il est représenté en une arborescence décorée: les feuilles correspondant aux mots (verbes, substantifs, ...) décrivent les catégories syntaxiques, et les noeuds décrivent les catégories des groupes syntagmatiques (groupes nominaux, groupes verbaux,...).

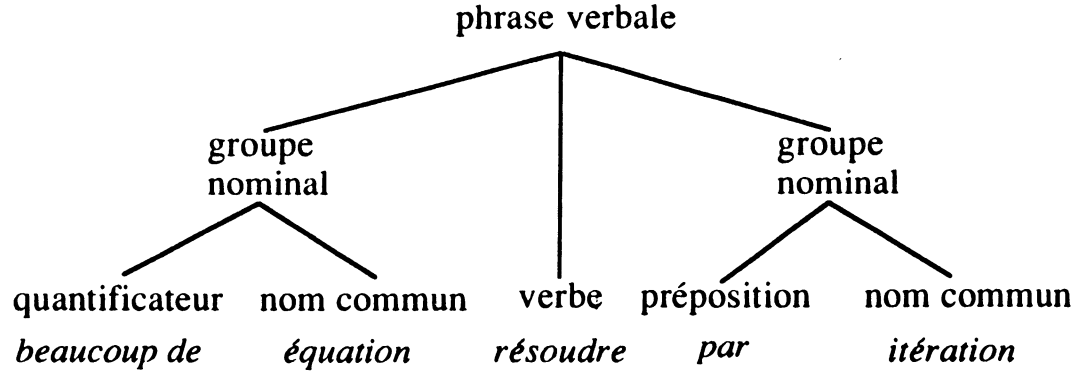
- Le niveau des fonctions syntaxiques représente une fonction syntaxique (sujet, objet, attribut, ...) par une relation entre deux noeuds relatifs à cette fonction.

- Le niveau des relations logiques décrit une phrase en terme de prédicat et d'arguments.

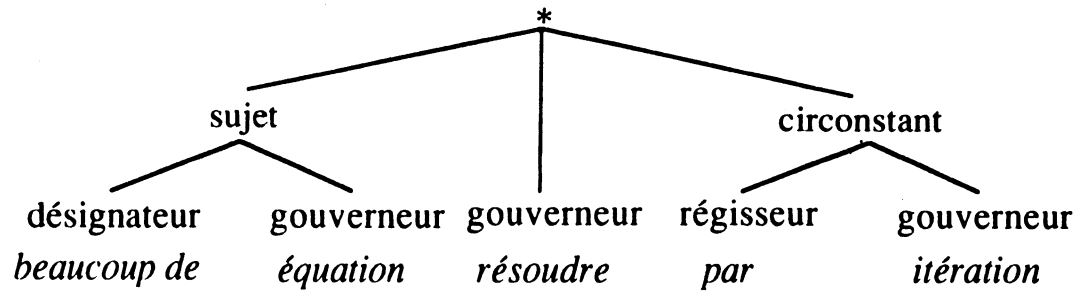
- Le niveau des relations sémantiques décrit les relations sémantiques (cause, condition, ...), qui particularisent les relations logiques, ou expriment des relations entre le prédicat et un groupe n'étant pas d'argument (groupes circonstanciels, par exemple).

Un exemple est donné ci-dessous pour illustrer les différents niveaux de la structure de représentation pour la phrase "beaucoup d'équations sont résolues par itération".

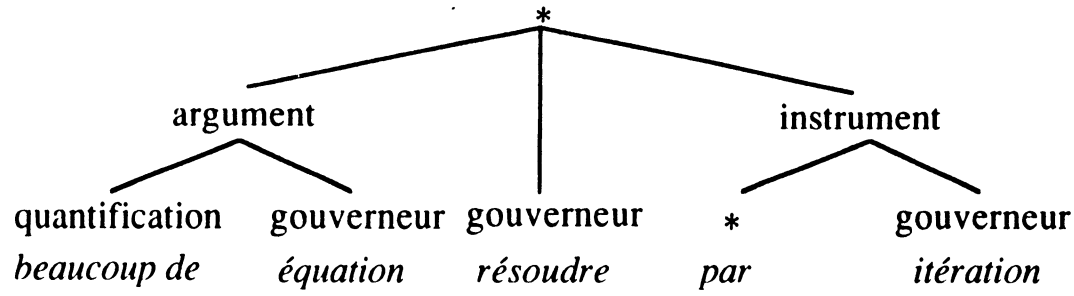
- Niveau des classes syntaxiques et syntagmatiques:



- Niveau des fonctions syntaxiques:



- Niveau des relations logico-sémantiques:



L'établissement simultané de plusieurs types d'informations offre une large possiblité à l'analyse. On peut choisir à tout moment d'utiliser le type d'informations le plus utile pour la situation. L'efficacité de cette approche dépend d'une stratégie de coopération pour choisir les informations à utiliser selon les circonstances.

ANNEXE 2. CLASSES SÉMANTIQUES

caractère physique (CAR-PHY)
examen (EX)
    examen général (EX-GEN)
    étape d'examen (ETAPE-EX)
    technique spécifique à un examen général (TECH-EX)
fonction (FCT)
    fonction attachée à un organe (FCT-ORG)
    fonction élémentaire (FCT-ELEM)
    fonction complexe (FCT-COMP)
lésion (LESION)
    lésion élémentaire spécifique non-active (LES-SPEC-NON-ACT)
    lésion élémentaire non-spécifique active
        (LES-NON-SPEC-ACT)
    lésion élémentaire non-spécifique active
        (LES-ELEM-NON-SPEC-ACT)
    lésion complexe (LES-COMP)
localité (LOC)
    constituant d'organisme (CONST-ORG)
        élément de structure (ELT-STRUCT)
        organe (ORG)
        région (REGION)
        détail (DETAIL)
position (POS)
signe (SGN)
    signe spécifique (SGN-SPEC)
    signe non-specifique (SGN-NON-SPEC)
valeur qualitative (VAL-QUAL)
    attribut physique (ATT-PHY)
    valeur sémantique qualitative (VAL-SEM-QUAL)
    pertinence médiacale (VAL-MED) ([{normal, amormal})
valeur qualitative (VAL-QUAL)

ANNEXE 3. MODELE SÉMANTIQUE

CRC ::= [op] (CRC, CRC)
[prob] CRC
CONSTAT
DIAG
[perm-de-déd] (CONSTAT, DIAG)

CONSTAT ::= [op] (CONSTAT, CONSTAT)
[prob] CONSTAT
[dû-à] (CONSTAT, DIAG)
[dû-à] (CONSTAT, CONSTAT)
[en-rel-topo-avec] (CONSTAT, CONSTAT)
[montré-par] (SGN, EX)
SGN

DIAG ::= [op] (DIAG, DIAG)
[prob] DIAG
[dû-à] (DIAG, DIAG)
[en-rel-topo-avec] (DIAG, DIAG)
LESION

SGN ::= [op] (SGN, SGN)
[prob] SGN
[a-pr-val] (SGN, QUAL)
[p-sur] (SGN, LOC)
[p-sur] (SGN, FCT)
[en-rel-topo-avec] (SGN, LOC)
[a-pr-val] (CAR-PHY, QUAL-SGN)
SGN-ELE

SGN-ELE ::= signe élémentaire
QUAL-SGN ::= valeur qualitative impliquant un signe

LESION ::= [op] (LESION, LESION)
[prob] LESION
[a-pr-val] (LESION, QUAL)
[p-sur] (LESION, LOC)
[p-sur] (LESION, FCT)
[en-rel-topo-avec] (LESION, LOC)
LESION-ELE

LESION-ELE ::= lésion élémentaire

LOC ::= [op] (LOC, LOC)
[en-rel-topo-avec] (LOC, LOC)
CONST-ORG

CONST-ORG ::= ORG
REGION
DETAIL

ORG ::= [a-pr-val-loc] (ORG, POS)
[a-pr-val-loc] (ORG, DETAIL)
[a-pr-val-loc] (ORG, REGION)
ORG-ELE

ORG-ELE ::= nom d'organe

REGION ::= [a-pr-val-loc] (REGION, POS)
[a-pr-val-loc] (REGION, DETAIL)
REGION-ELE

REGION-ELE ::= région de constituant d'organisme

DETAIL ::= [a-pr-val-loc] (DETAIL, POS)
DETAIL-ELE

DETAIL-ELE ::= détail de constituant d'organisme
POS ::= position
FCT ::= fonction

QUAL ::= [op] (QUAL, QUAL)
[prob] QUAL
[a-pr-val] (CAR-PHY, VAL-QUAL1)
[a-pr-val] (CAR-PHY, VAL-QUAN)
VAL-QUAL1

VAL-QUAL1 ::= [a-pr-val] (VAL-QUAL, VAL-QUAN)
VAL-QUAL

CAR-PHY ::= [p-sur] (CAR-PHY, LOC)
[p-sur] (CAR-PHY, FCT)
CAR-PHY-ELE

CAR-PHY-ELE ::= caractère physique élémentaire
VAL-QUAL ::= valeur qualitative
VAL-QUAN ::= valeur quantitative

ANNEXE 4. EXEMPLES DES CRM

Exemple 1:

Monsieur XXX
le 8.12.1986

TOMEDENSITOMETRIE
(homme de 70 ans - découvert d'une opacité arrondie du lobe supérieur droit avec déformation du médiastin)

Les constataions sont les suivantes:

- opacité à contour bosselé, de densité tissulaire, apparemment homogène, en projection du segment apical du lobe supérieur droit.

- contact étroit avec le versant médiastinal correspondant.

- opacité ganglionnaire de plus de 15mm de diamètre en projection de la loge de Barety intéressant donc les derniers relais de la chaîne para-tranchéales droite et allant jusqu'au niveau de la chaîne sus pulmonaire, c'est-à-dire dans la loge pré-carénale.

En conclusion:

Aspect TDM en faveur d'un cancer de siège périphérique avec extension ganglionnaire médiastinale et pédiculaire et probablement T3 pleural (plève médiastinale).

Exemple 2:

Professeur M. ...

Monsieur ... Albert
le 14.11.83

TOMODENSITOMETRIE (Bilan d'extension d'un carcinome ganglionnaire de la bronche lobaire supéreiur droite - Opacité nodulaire controlatérale au niveau de la pyramide basale gauche)

Les constatations sont les suivantes:

- condensation pulmonaire en projection du lobe supérieur droit avec bombement du versant scissural; après injection intra-veineuse de produit de contraste, hypodensités nombreuses traduisant probablement des phénomène de nécrase.

- Extension médiastinale prouvée par plusieurs vignes:

. comblement de la loge précarénaire parsun nodule interposé entre la face postérieure de la veine cave supérieure de l'aorte et la face antérieure de la bifurcation tranchéale.

. épaississement de la bande bronchique droite postérieure.

. opacité anormale située au niveau de la partie haute de la loge de Baréty, en arrière du tronc artériel brachio-céphalique peu avant sa division (coupe No. 3).

Pr. ...

Exemple 3:

Monsieur ...

THOMODENSITOMATRIE THORACIQUE

(homme de 61 ans - antécédent de caverne tuberculeuse bi-apicale - apparition récente d'une opacité expansive du sommet droit - cytologie : cellules malignes napithiennes - fibroscopie : sténose de l'origine de la bronche lobaire supérieure droite ainsi que segment apico-dorsale)

Les constatations sont les suivantes:

- processus expansif en projection du segment apical du lobe supérieur droit. Densité tissulaire. Diamètre axiale transverse maximum de 65 mm.
- contact étroit avec le versant droit du médiastin, notamment au contact du tronc veineux brachio-céphalique gauche et de l'origine de la veine cave supérieure.
- hypertrophie ganglionnaire supérieure à 10 mm, au niveau de la loge de Baréty (groupe 4R).
- opacité de siège pédiculaire supérieur, à disposition péri-bronchique correspondant bien aux données de la fibroscopie et pouvant traduire l'adénopathie et/ou l'extension tumorale pédiculaire.
- on ne constate pas d'extension extra-pleurale au niveau du sommet.

EN CONCLUSION

Volumineux cancer de siège périphérique, en projection du segment apical du lobe supérieur droit, avec extension ganglionnaire pédiculaire et médiastinale omo-latérale. Le contact étroit avec le versant droit du médiastin fait craindre un envahissement direct à ce niveau, notamment en regard du tronc veineux brachio-céphalique droite.

Dr. ...

ANNEXE 5. THÉSAURUS DE GÉNÉTIQUES DE REMEDE (une partie)

...

développement
- mal dévelopement
  - asymétrie
  - dysmorphie
  - dystrophie
  - anarchie
  - dysplasie
- sous développement
  - hypoplasie
  - hypotrophie
  - atrésie
- non développement
  - aplasie=agénésie
  - défaut
- destruction
  - séquestration=necrose
  - atrophie
  - lacune
- sur développement
  - hyperplasie
  - hypertrophie
  - mémi-hypertrophie

caractère physique
- longueur
- largeur
- hauteur
- épaisseur
- profondeur
- densité

...

tumeur
- tumeur bénigne
  - xanthome
  - hygroma
  - tumeur embaryonnaire
  - kyste
    - kyste branchial
    - polykystose
  - pseudokyste
  - polype
    - polypose

papillome
adenome
gliome
neurinome
tumeur intermédiaire
tumeur maligne (= cancer)
dégénérescence
épithéliome
sarcome
lymphosarcome
blestome
tératome
...
cou
nuque
membre
membre supérieur
creux axillaire
épaule
omoplate
clavicule
bras
humérus
coude
avant bras
cubitus
radius
poignet
carpe
main
paume
métacarpien
doigt
pouce
index
médius
annulaire
autriculaire
membre inférieur
hanche
cuisse
femur
creux poplite
génou
rotule
jambe

- tibia
- perone
- cheville
- tarse
- pied
- plante
- métatarsien
- orteil
- phalange

tronc

- thorax
- médiastin
- côte
- sternum
- grand pectoral
- mamelon
- glande mammaire
- dos
- rachis
- rachis cervical
- rachis dorsal
- rachis lombaire
- sacrum
- sacrococcyx
- coccyx
- bassin
- région pubienne
- pubis
- ilon=aile ilique

parois abodominale

- grand droit

diaphagme

appareil cardio vasc

- coeur
- cornaire
- parois cardiaque
- péricarde
- myocarde
- endocarde
- cavite cardiaque
- oreillette
- septum interauriculaire
- ventriculaire
- septum interventriculaire
- valvule cardiaque=orifice cardiaque
- valvule tricuspide=orifice tricuspide

- valvule mitrale=orifice mitral
- valvule aortique=orifice aortique
- orifice pulmon

gros vaisseux
- artère
  - aorte
- veine
  - veine cave
- vaisseux pulmonaire
  - artère pulmonaire
  - veine pulmonaire

appareil respitatoire
- larynx
- trachée
- bronche
- poumon
  - lobe pulmonaire
  - scissure
- plevre

appareil digestif
- pharyx
- desophage
- abdomen
  - estomac
    - pylore
      - prépylore
  - duodenum
  - intestin
    - intestin grele
    - colon
    - appendice
    - prectum
    - anus
    - anorectum
  - mésentère
  - péritoine

foie
- voie biliaire
  - voie biliaire interahépatique
  - voie biliaire extra-hépatique
    - choledoque
    - vesicule biliaire
    - sphincter d'oddi

...

BIBLIOGRAPHIE

[Abiteboul87a] S.Abiteboul, S.Grumbach, Bases de Données et Objets Structurés. T.S.I. Vol. 6, No.5, 1987

[Abiteboul87b] S.Abiteboul, P.Kanellakis and G.Grahne, On the representation and querying of sets of possible worlds, Proceedings of SIGMOD'87, may 1987, ed. U.Dayal and I.Traiger, SIGMOD record, Vol.16, No.3, Dec 1987

[Aczél75] J.Aczél and Z.DarÓczy,On Measures of Information and Their Characterizations. Vol.115, dans la série: Mathématics in science and engineering, ed. R.Bellman, Academic Press, NewYork, San Francisco - London, 1975

[Bancilhon87] F.Bancilhon, T.Briggs, S.Khoshafian, et P.Valduriez, FAD - a powerful and simple database language. Proc. of Very Large Data base Conference, 1987

[Barr82] A.Barr, Feigenbaum E., The Handbook of Artificial Intelligence, Pitman, Vol.1, 1982

[Berrut87] C.Berrut, P.Cinquin, J.Nie, G.Monoz, Y.Chiaramella, J.Demongeot et M.Coulomb, Modélisation sémantique de comptes rendus radiologiques, Actes d'Intelligence artificielle et santé, Toulouse, 1987

[Berrut88] C.Berrut, Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés - Le prototype RIME et son application à un corpus médical, thèse de l'Université Joseph Fourier - Grenoble I, décembre 1988

[Berrut89] C.Berrut, Y.Chiaramella, Indexing madical report in a multimedia environment: the RIME experimental approach, ACM SIGIR 89, Boston, June 1989

[Bookstein77] A.Bookstein and D.Kraft, Operations research applied to document indexing and retrieval decisions, Journal of the ACM, vol.24, No.3, pp.418-427, July 1977

[Bookstein80] A.Bookstein, Fuzzy Requests: An approach to Weighted Boolean Searches. Journal of the American Society for Information Science, July 1980

[Bruandet85] M.F.Bruandet, Modèle partiel de connaissances pour un système de recherche d'informations, Actes des conférences int. RIAO85, Grenoble, Mars 1985

[Chang78] C.L.Chang, Deduce 2: Further Investigations of Deduction in Relational Data Bases. in Logic and Databases, ed. H. Gallaire, J.Minker, Plenum Press, 1978

[Chang86] C.L.Chang and A.Walker, PROSQL: A prolog programming interface with SQL/DS. 1st int. Workshop on Expert Database Systems, Benjamin/Cummings Pub., L.Kerschberg ed. 1986

[Chen76] P.P.Chen, The Entity-Relationship Model - Toward a Unified View of Data. ACM Transition on Database Systems, 1(1), pp.9-36, March 1976

[Chiaramella86] Y.Chiaramella, B.Defude, M.F.Bruandet, et D.Kerkouba, IOTA: A full text information retrieval system. ACM Conference on Research and Development in Information Retrieval, ed. F.Rabitti, Pisa, September 1986

[Chomsky65] N.Chomsky, Aspects of the Theory of Syntax, M.I.T., Cambridge, The M.I.T. Press, 1965, (traduction française: Aspects de la théorie syntaxique, Paris, Edition du Seuil, 1971)

[Codd79] E.F.Codd, Extending the Database Relational Model to Capture More Meaning. ACM Transition on Database Systems, 4(4), pp397-434, December 1979

[Croft83] W.B.Croft, Wolf R. and Thompson R., A network organization used for information retrieval. Proc. 6th ACM-SIGIR Conference on Research and Development in Information Retrieval, Bethesda, 1983

[Croft85] W.B.Croft, An expert assistant for a document retrieval system. Proceedings RIAO85, Grenoble, Mars 1985

[Croft88] W.B.Croft, T.J.Lucia and P.R.Cohen, Retrieving documents by plausible inference: A preliminary study, 11th international conference on research and development in information retrieval, ed.Y.Chiaramella, ACM-SIGIR88, Grenoble, June 1988

[Dabrowski75] M.Dabrowski, A General Model of Distribution of Objects in Information Retrieval Systems. Information Systems, Vol.1, Pergamon Press, 1975

[Davis77] R.Davis, J.King, An overview of production systems, Machine Intelligence 8, pp.300-332, 1977

[Defude86] B.Defude, Etude et Réalisation d'un Système Intelligent de Recherche d'Informations: Le Prototype IOTA, Thèse INPG, 1986

[Delobel82] C.Delobel, M.Adiba, Bases de Données et Systèmes Relationnels, Dunod, Paris, 1982

[Deogun86] J.S.Deogun and V.V.Raghavan, User-oriented Document Cluster: A Framework for learning in Information Retrieval. Proc. ACM-SIGIR 86, Pisa, Italy, September 1986

[Deogun88] J.S.Deogun and V.V.Raghavan, Integration of information retrieval and database management systems. Information processing & Management, Vol.24, No.3, 1988

[Desai87] B.C.Desai, P.Goyal and F.Sadri, Non-first Normal Form Universal Relations: An Application to Information Retrieval Systems, Information Systems, Vol.12, No.1, pp.49-55, 1987

[Dubois80] D.Dubois & H.Prade, Fuzzy sets and systems: Theory and applications. Academic Press, 1980

[Duda77] R.O.Duda, P.E.Hart, N.J.Nilsonn, G.L.Sutherland, Semantic network representation for rule based inference system, Stanford Research Institute, Menlo Park, California, 1977

[Esprit86] CEE Esprit, Integration of Logic Programming and Data Bases. Proc. Venice Int. Workshop on the intergration of logic and databases, 1986

[Fitting85] M.Fitting, A Krepke-Kleene semantics for logique programs, Journal of Logic Programming, Vol.4, pp.295-312, 1986

[Gallaire84] H.Gallaire J.Minker J.M.Nicolas, Logic and Database: A Deductive Approach. Computer Surveys, Vol.16, No.2, June 1984

[Gardarin87] G.Gardarin E.Simon, Les Systèmes de Gestion de Bases de Données Déductives. Vol.16, No.5, 1987

[Giger88] H.P.Giger, Concept Based Retrieval in Classical IR Systems. 11th international conference on research and development in information retrieval, ed.Y.Chiaramella, ACM-SIGIR88, Grenoble, June 1988

[Graitson82] M.Graitson, Aspects du traitement computationnel du langage médical, Liège, thèse, 1982

[Grevisse80] M.Grevisse, Le Bon Usage, Hatier, Paris, dernière éd. 1980

[Hindley86] J.R.Hindley, J.P.Seldin, Introduction to combinators and λ-calculus, London Mathematical Society Students Texts 1, Cambridge University Press, 1986

[Hendrix78] G.G.Hendix, E.D.Sacerdoti, D.Sagalowicz and J.Slocum, Developing a natural language interface to complex data, ACM Transactions on Database Systems, Vol.3, No.2, pp.105-147, June 1978

[Hughes68] G.Hughes, Gresswill M., An Introduction to Modal Logic. Methuen, 1968

[Jacobs82] B.E.Jacobs, On Database Logic. Journal of the A.C.M. Vol.29, No.2, pp.310-332, April 1982

[Joloboff78] V.Joloboff, Unification d'arborescences - Evaluation sémantique d'énoncés en langue naturelle, thèse de U.S.M.Grenoble et I.N.P.G., septembre 1978

[Kerkouba84] D.Kerkouba, Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles - Application à un corpus technique, thèse INPG, Nov.1984

[Kerkouba85] D.Kerkouba, Indexation automatique et aspects structurels du texte, RIAO85, Grenoble, mars 1985

[Kok88] A.J.Kok and A.M.Botman, Retrieval based on User Behaviour. 11th international conference on research and development in information retrieval, ed.Y.Chiaramella, ACM-SIGIR88, Grenoble, June 1988

[Kuipers75] B.Kuipers, A frame for frames: representing knowledge for recognition. in Representation and Understanding, ed. Bobrow and Collins, Academic Press, N.Y., 1975

[Laurière81] J.L.Laurière, Représentation et utilisation des connaissances, TSI, Vol.1, No.1 et 2, 1981

[Lopez79] M.Lopez, Communication en langue naturelle avec un système d'aide à la conception d'assemblages physiques: un essai d'utilisation des réseaux sémantiques partitionnés. thèse docteur ingénieur, INPG, 1979

[Minski88] J.Minski, Perspectives in deductive databases. Journal of Logic Programming, 1988:5, p33-60

[Munoz87] G.Munoz-Baca, Stockage et exploitation de dossiers médicaux multimédia au moyen d'une base de données généralisée - Projet Tigre, thèse USTMG, Grennoble, Juillet 1986

[Nie87] J.Nie, A user interface in quasi natural language for an information retrieval system, Proc. of 11th int. Online Information Meeting, London, December 1987

[Nie88] J.Y.Nie, An outline of a general model for information retrieval systems, 11th international conference on research and development in information retrieval, ed.Y.Chiaramella, ACM-SIGIR88, Grenoble, June 1988

[Nie89a] J.Y.Nie, A general information retrieval model based on modal logic, Information Processing & Management, Vol.25, No.5, pp.477-491, 1989

[Nie89b] J.Y.Nie, Information retrieval systems and their integration of deductive data bases, Beijing International Symposium for Young Computer Professionals, Beijing, August 1989

[Nicolas78] J.M.Nicolas and H.Gallaire, Database: Theory vs. interpretation. Logic and Databases, ed. H.Gallaire, J.Minski, Plenum, New York, pp.33-54, 1978

[Novak76] G.S.Novak, Computer understanding of physics problems stated in natural language, Technical report NL-30, Computer Science Dept, University of Texas at Austin, 1976

[Ozsoyoglu87] Z.M. Ozsoyoglu and L-Y.Yuan, A new normal form for nested relations, ACM Transactions on Database Systems, 12(1), March 1987

[Palmer90] P.Palmer, Outils de traitement linguistique adapté à l'indexation automatique de textes libres, Thèse de l'Université Josoph Fourier, à paraître

[Reiter84] R.Reiter, Towards a Logical Reconstruction of Relational Database Theory. On conceptual Modeling, eds. M.Brodie,

J.Mylopoulos, J.W.Schmidt, Spring-Verlag, Berlin and New York, 1984

[Salton71] G.Salton(editor), The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1971

[Salton83a] G.Salton, M.J.McGill, Introduction to Modern Information Retrieval. International Student Edition, 1983

[Salton83b] G.Salton, E.A.Fox and H.Wu, Extended Boolean Information Retrieval. Communications of the ACM, Vol.26, No.12, 1983

[Schank80a] R.C.Schank, M.Lebowitz, L.Birnbaum, An Integrated Understander, American Journal of Computational Linguistics, Vol.6, No.1, January-March 1980

[Schank80b] R.C.Schank, Language and Memory, Cognitive Science, 4(3), pp243-284, 1980

[Schank81] R.C.Schank and C.K.Riesbeck(ed.) Inside Computer Understanding: Five Programs Miniatures, Hillsdale, New Jersey, 1981

[Schank82] R.C.Schank, Rewind and memory organisation: an introduction to MOPs, dans Strategies for natural language processing, ed.Lehnert W.G., Ringle M.H. Lawrence Evlbaum Associates, Londre, 1982.

[Schotch75] P.K.Schotch, Fuzzy modal logic. Proceedings of the 1975 international Symposium on Multiple-valued logic, Indiana university, Bloomington

[Stanat77] D.F.Stanat, D.F.McAllister, Discete mathematics in computer science, Prentice-Hall, Inc., 1977

[Stonebreaker86] M.Stonebreaker and A.L.Rowe, The Postgres Papers. UC Berkeley, ERL M86/85, Berkeley, November 1986

[Tsichritzis88] D.Tsichritzis (ed.), Active Object Environments. Centre Universitaire d'Informatique, Université de Genève, 1988

[Tuner84] R.Tuner, Logics for Artificial Intelligence. Elis Horwood limited, 1984

[Valduriez87] P.Valduriez, Objets Complexes dans les Systèmes de Bases de Données Relationnelles. T.S.I. Vol.6, No.5, 1987

[vanRijsbergen79] van Rijsbergen C.J., Information Retrieval. 2nd edition, London, Butterworths, 1979

[vanRijsbergen86] C.J.van Rijsbergen, A Non-classical Logic for Information Retrieval. Computer Journal, Vol.29, No.6, 1986

[Vauqois85] B.Vauquois, C.Boitet, Automated Translation at Grenoble University, Computational Linguistics,11/1, pp28-36, January - March 1985.

[Waller79] W.G.Waller and D.H.Kraft, A Mathematical Model of a Weighted Boolean Retrieval System. Information Processing & Managament, Vol.15, Pergamon Press Ltd. 1979

[Waltz78] D.L.Waltz, An English language question answering system for a large relational database, Communication of the ACM, Vol.21, No.7, july 1978

[Wilks75] Y.Wilks, An Intelligent Analyser of English, CACM 18/5, pp265-274, May 1975.

[Woods70] W.A.Woods, Transition Networks Grammar for Natural Language Analysis, Comm. of the ACM, Vol.13/Nov. 10/Oct. 1970

[Yu79] C.T.Yu, W.S.Luk and M.K.Siu, On models of information retrieval processes, Information systems, Vol.4, pp205-218, 1979

[Yu85] C.T.Yu, Adaptive document clustering. Proceedings of the 8th Annual international ACMSIGIR Conference on Research and Development in Information Retrieval, Montréal, Canada, 1985

[Yusoff87] Z.Yusoff, The Linguistic Approch at GETA: a Synopsis, Technologos, No.4, pp93-110, Printemps 1987

[Zajac86] R.Zajac, Etudes des possibilités d'intégration homme-machine dans un processus de traduction automatique, thèse doctorale, INPG, Juillet 1986.

[Zeman75] J.J.Zeman, Modal logic. Oxford, 1975

8501602

AUTORISATION DE SOUTENANCE

DOCTORAT 3ème CYCLE, DOCTORAT INGENIEUR,
DOCTORAT DE L'UNIVERSITE JOSEPH FOURIER - GRENOBLE 1

Vu les dispositions de l'Arrêté du **16 avril 1974**,

Vu les dispositions de l'Arrêté du **5 juillet 1984**,

Vu les rapports de M Joseph Sifakis

M Keith Van Rijsbergen

M Jianyun NIE est autorisé(e)
à présenter une thèse en vue de l'obtention du Doctorat de
l'Université Joseph Fourier - Grenoble I

Grenoble, le 09 JUIL. 1999

Le Président de l'Université
Joseph Fourier - Grenoble 1

A. NEMOZ

UNIVERSITE JOSEPH FOURIER

Résumé:

La définition d'un modèle d'évaluation est le problème clé d'un Système de Recherche d'Informations. De nombreux modèles existent, qui sont généralement spécifiques à un type d'application particulier et avec lesquels la prise en compte de la sémantique est difficile. Dans la première partie de cette thèse, nous dégageons d'abord deux critères pour la valuation de la correspondance entre un document et une requête: l'exhaustivité et la spécificité du document pour la requête. Nous définissons ensuite un modèle général fondé sur la logique modale floue pour la valuation des deux critères. Ce modèle est comparé avec quelques modèles existants pour démontrer sa généralité.

Dans la seconde partie de la thèse, le modèle proposé est appliqué au processus d'interrogation du prototype RIME pour la recherche d'informations médicales. Ce prototype possède une interface en langue quasi-naturelle (un sous-ensemble du français). Un processus d'interrogation se décompose en deux parties: l'interprétation des requêtes en langue quasi-naturelle et l'évaluation des requêtes en utilisant le modèle général précédemment défini. Ces deux parties sont étudiées en détail. Une réalisation est finalement présentée, ainsi que son expérimentation sur un corpus médical.

Mots-clefs:

système de recherche d'informations
modélisation logique
logique modale
informatique médicale
représentation sémantique
interrogation en langue naturelle
intelligence artificielle
base de données déductives