Ontologies et services aux patients : Application à la reformulation des requêtes

Radja Messai

Université Joseph Fourier – Grenoble I **École Doctorale EDISCE**

Ontologies et services aux patients : Application à la reformulation des requêtes

THÈSE

présentée et soutenue publiquement le 09 novembre 2009

pour obtenir le grade de

Docteur de l'Université Joseph Fourier – Grenoble I

(Spécialité : Ingénierie de la cognition, de l'interaction, de l'apprentissage et de la création)

par

Radja Messai

Composition du jury

Président :	Pr. Christophe Roche	Professeur des universités, Polytech'Savoie
Rapporteurs :	Dr. Sylvie Calabretto	Maître de Conférences (HDR) à l'INSA de Lyon
	Pr. Geneviève Lallich-Boidin	Professeur à l'Université Claude Bernard – Lyon 1
Examinateurs :	Pr. Mireille Mousseau	PU-PH à l'Université Joseph Fourier – Grenoble 1
	Dr. Ismaïl Timimi	Maître de Conférences à l'Université Lille III
	Dr. Michel Simonet	Chargé de recherche CNRS (Directeur de thèse)

Laboratoire des Techniques de l'Ingénierie Médicale et de la Complexité
TIMC-IMAG — UMR 5525

Remerciements

Quel bonheur d'en arriver là et quel plaisir d'écrire ces quelques lignes pour remercier les gens qui ont contribué à l'aboutissement de ce travail.

En premier lieu, je tiens à remercier Michel Simonet, Directeur de l'équipe Osiris du laboratoite TIMC-IMAG, de m'avoir accueilli au sein de l'équipe et encadré durant ma thèse. Son aide, sa disponibilité et surtout sa gentillesse ont été un vrai moteur pour l'aboutissement de cette thèse. Je le remercie également d'avoir cru en moi et de m'avoir appris beaucoup de choses sur les activités de recherche.

Je remercie sincèrement Madame Geneviève Lallich-Boidin, Professeur à l'Université Claude Bernard - Lyon 1 et Madame Sylvie Calabretto, Maître de Conférences (HDR) à l'INSA de Lyon, d'avoir accepté d'être les rapporteuses de ma thèse.

Je remercie également Monsieur Christophe Roche, Professeur à l'Université de Savoie, Madame Mireille Mousseau, PU-PH à l'Université Joseph Fourier - Grenoble 1, Monsieur Ismaïl Timimi, Maître de Conférences à l'Université Lille III et Mme Celia Boyer Directrice exécutive de la fondation Health On the Net, d'avoir accepté de participer au jury de ma thèse.

Je remercie également Professeur Régis Beuscart Directeur du laboratoire CERIM à Lille de m'avoir accueili dans son équipe. Je remercie sincèrement mes collègues du CERIM pour leur bonne humeur : Nathalie, Jean-Marie, Jean-Marc, Alain, Patrick, Renaud, Arnaud et Julien.

Mes remerciements vont également à La Ligue Nationale Contre le Cancer et AGARO Association Grenobloise de la Recherche en Oncologie pour avoir financé ce travail de recherche.

Je remercie mes collègues et amis de l'équipe Osiris pour leur soutien et leurs encouragements : Houda, Shokoh, Lamis, Samer, Nora, Badia, Delphine, Gayo, Séverine, Hai, Sylvain et Rémi. Je remercie mes amis de l'équipe TIMC : Nourredine, Cheikh et Redha.

Une pensée à Mme Gindre et à ses enfants Carole et Maurice. Merci de m'avoir accueilli dans votre famille.

Enfin, je tiens à remercier très sincèrement toutes les personnes de mon entourage. Mes formidables amis : Camille, Lina, Sara, Nazim, Fethi, Soraya et Souhila. Mon cher mari Lyes pour sa patience et son soutien sans faille. Ma fille Cyrine pour le réconfort et le bonheur qu'elle m'apporte. Mon père pour ses encouragements et sa présence. Enfin le plus grand merci pour ma mère, pour sa douceur, son amour, ses conseils, ses encouragements et sa présence. Merci maman, je t'aime de tout mon cœur.

à maman, la lumière de ma vie
à papa, le meilleur père du monde
à mes chers frère et sœurs
à mon amour, Lyes
à ma raison de vivre, Cyrine

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

ARDA	Advanced Research and Development Activity
CHV	Consumer Health Vocabulary
CISMeF	Catalogue et Index des Sites MEdicaux Francophones
CUI	Concept Unique Identifier
DAML	DARPA Agent Markup Language
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
EVS	Entreprise Vocabulary Service
FMA	Foundational Model of Anatomy
GO	Gene Ontology
HON	Health On the Net
ICD	International Classification of Diseases
IHTSDO	Health Terminology Standards Development Organisation
KIF	Knowledge Interchange Format
LD	Logiques de description
LSI	Latent Semantic Indexing
MeSH	Medical Subject Headings
NCBO	National Center for Biomedical Ontology
NCI	National Cancer Institute
NCIT	National Cancer Institute's Thesaurus
NIST	National Institute of Standards and Technology
NLM	National Library of Medicine
OBO	Open Biomedical Ontologies
OIL	Ontology Interchange Language
OWL	Web Ontology Language
PA	Propagation d'Activation
RDF	Resource Description Framework
RI	Recherche d'Information
SKOS	Simple Knowledge Organisation System
SNOMED CT	Systematized NOmenclature of MEDicine - Clinical Terms
SNOMED RT	Systematized NOmenclature of MEDicine - Reference terminology
SRI	Système de Recherche d'Information
SUI	String Unique Identifier
TAL	Traitement Automatique de la Langue
TIA	Terminologie et Intelligence Artificielle
TREC	Text REtrieval Conference
UMLS	Unified Medical Language System
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language

Introduction

L'information médicale à destination du grand public peut avoir un impact considérable sur les plans personnel, social et économique. Cependant, les stratégies de communication de cette information souffrent encore de plusieurs lacunes, et le grand public a peu bénéficié de l'explosion de la recherche médicale, de la littérature scientifique et même des connaissances médicales de base. Deux facteurs principaux peuvent limiter la communication ou le transfert de cette information : l'accessibilité physique et l'accessibilité conceptuelle à l'information. L'accessibilité physique à l'information a été améliorée grâce aux campagnes d'information de proximité (hôpitaux, médecins, pharmaciens, numéros verts d'information), l'utilisation des médias (télévision, radio, journaux et magazines) et les technologies d'information comme Internet. Cependant, l'accessibilité conceptuelle à l'information, qui correspond à la capacité de trouver, comprendre et interpréter l'information médicale, n'a pas beaucoup progressé.

De plus en plus de personnes veulent jouer un rôle actif dans leur parcours de santé. Ils cherchent de l'information médicale sur leur condition de santé, sur les derniers essais cliniques ou sur des personnes qui ont eu la même expérience. Ce groupe est appelé "les usagers de santé" parce que leur comportement et leurs besoins sont similaires à ceux de n'importe quel autre usager qui a besoin d'un produit ou d'un service *traditionnels*. Les organisations de santé et les professionnels du domaine essayent de satisfaire ces besoins grâce à la diffusion massive d'information à travers les publications, les différents médias et récemment le Web. Cependant, et malgré la quantité d'information croissante, la méconnaissance du domaine médical limite le bénéfice que peuvent retirer les usagers de santé de cette information [Stavri, 2001].

Contexte et motivation du travail

Le sujet de cette thèse concerne principalement les aspects terminologiques de l'accessibilité des usagers de santé à l'information médicale. Comment les usagers de santé formulent leurs problèmes de santé ? Comment les termes médicaux utilisés par le grand public diffèrent de ceux utilisés par les professionnels de santé ? Quels sont les concepts et les termes partagés par les deux publics ? Est-il justifié de séparer les terminologies utilisées par les professionnels de santé et celles utilisées par les usagers de santé ? Comment peut-on bénéficier d'une ressource onto-terminologique destinée aux usagers de santé dans la reformulation des requêtes ?

Le mouvement des usagers de santé, qui se focalise sur la transmission de l'information médicale au grand public, est un mouvement relativement récent. *"Même si rien ne peut se substituer à l'expertise de votre propre médecin, aucune prescription n'a plus de valeur que la connaissance"* [Pifalo et al., 1997, p. 21]. Mais même trouver une information pertinente, précise et actuelle parmi les volumes grandissants de la documentation imprimée et plus de 17 000 sites Web de santé [1] reste un défi considérable pour la plupart des usagers de santé [Fox & Rainie, 2000].

1. Ces données concernent les sites Web en anglais. Nous n'avons pas pu trouver ce type d'information pour

Comme il y a de plus en plus d'usagers de santé qui recherchent de l'information médicale, il est important d'explorer comment les non-professionnels expriment les concepts médicaux par le langage. Une meilleure compréhension permettra des améliorations dans la conception et l'implémentation des ressources destinées à ce public, en termes des mécanismes et des contenus de la recherche d'information. Peu de travaux ont été consacrés à cette question. Notre recherche a utilisé les théories et les approches dérivées des domaines de la terminologie et de la linguistique cognitive pour comprendre les termes et les concepts que les usagers utilisent pour exprimer des *idées* médicales. Les résultats ne vont pas seulement fournir une vision sur les processus cognitifs impliqués, mais ont également des retombées pratiques sur la recherche d'information, le traitement automatique du langage et d'autres problèmes terminologiques.

Le problème terminologique

La flexibilité expressive est une propriété fondamentale de la langue. Les concepts peuvent être exprimés de différentes manières, utilisant des combinaisons différentes de mots et de phrases. Cette propriété est utile pour produire de nouvelles expressions pour décrire de nouvelles idées ou de nouveaux procédés. Cependant, une telle flexibilité a un coût : elle peut également conduire à *"une mauvaise communication"* quand les personnes utilisent des termes différents pour désigner la même chose, problème connu sous le nom de *"problème terminologique"*. Ainsi, une étude a montré qu'il y a une probabilité de moins de 20% pour que deux individus utilisent les mêmes mots pour décrire le même objet [Furnas et al., 1987]. Une telle variabilité diminue les chances de bonne correspondance entre les termes des requêtes et les termes d'indexation.

Traditionnellement, l'indexation manuelle utilisant les thésaurus et les vocabulaires contrôlés a aidé à diminuer l'éventail des termes nécessaires pour des recherches fructueuses. Cependant, la faible correspondance entre les indexations réalisées par des indexeurs différents a soulevé plusieurs questions quant à l'efficacité d'une telle méthode [Salton & McGill, 1983]. Plus récemment, l'indexation automatique a été utilisée en se basant sur des méthodes qui pour la plupart sont de nature statistique. Ce type de méthodes se base sur des mots qui ne sont ni très fréquents ni très rares dans un ensemble de documents, pour les représenter. En vérité, l'indexation automatique déplace le problème terminologique des mots sélectionnés par les indexeurs à ceux utilisés par les auteurs ; les utilisateurs doivent encore *deviner* quels termes ont été utilisés pour pouvoir formuler des requêtes efficaces.

Vocabulaire médical

Les professionnels de santé souffrent également de problèmes de communication dans le domaine médical. Même si la médecine moderne a des bases scientifiques solides qui peuvent faciliter la standardisation terminologique, les différentes spécialités, les sous-domaines, les zones d'incertitude et le manque de connaissances, excluent la possibilité d'une terminologie médicale unifiée. En pratique, les vocabulaires sont créés et adaptés pour servir différents besoins et tâches nécessitant des niveaux de granularité et des perspectives différents, allant des statistiques de mortalité et de morbidité aux systèmes de facturation, aux dossiers patients électroniques : *"... la terminologie n'est pas une fin en elle-même, mais s'intéresse à des besoins sociaux et essaye d'optimiser la communication entre les spécialistes et les professionnels ..."* [Cabré, 1999, p. 10]. Même si UMLS regroupe les vocabulaires médicaux utilisés par les professionnels de santé, le problème terminologique continue à toucher les professionnels de santé et leurs systèmes d'information.

les sites francophones.

Les terminologues font la distinction entre le langage général LG (language for general purposes) et le langage de spécialité LS (language for specific purposes). Le LG est utilisé et compris dans le discours de tous les jours et *"contient un vocabulaire pour parler de concepts et d'actions largement partagés"* [Haas & Hert, 2000, p. 54]. Par contre, le LS est *"un ensemble complet de phénomènes linguistiques qui se produisent dans une sphère de communication définie et limitée par des sujets, des intentions et des conditions spécifiques"* [Hoffmann, 1979, p. 16]. Haas et Hert définissent le LS comme *"les termes pour communiquer sur ... une expertise commune"* [Haas & Hert, 2000, p. 54] dans une communauté limitée (e.g., les professionnels de santé). Alors que le LG est acquis essentiellement à travers les interactions sociales et l'éducation scolaire, le LS est acquis au cours des interactions de groupes de discours de spécialistes, des formations académiques spécialisées ou par l'expérience. Par exemple, les patients avec des maladies chroniques peuvent arriver à bien connaître le LS à cause d'une exposition prolongée à cette terminologie, même sans formation académique. Des exemples de paires de termes LG/LS dans le domaine médical sont : *œil rose/conjonctivite*, *infection de l'oreille/otite* et *attaque cardiaque/infarctus du myocarde*. Alors que tout le monde (y compris les spécialistes) utilise le LG, seuls les membres de sous-populations limitées utilisent différents LS. Les frontières entre le LG et le LS sont floues –*"La différence entre le langage général ... et le langage de spécialité est difficile à établir"* [Cabré, 1999, p. 71]– mais peuvent être mieux définies pour certains domaines. Ainsi, une étude a trouvé que les sciences exactes utilisent plus de termes spécifiques que les sciences sociales et humaines, qui elles partagent beaucoup de termes avec le LG [Haas, 1997]. Des termes peuvent être également "empruntés", mais acquièrent parfois des sens différents (e.g., *migraine* dans le LG désigne parfois une douleur à la tête, mais possède une définition particulière en LS [1]). Par conséquent, une comparaison entre des terminologies doit se faire à la fois au niveau des concepts et au niveau des termes. Deux termes similaires ne correspondent pas à deux concepts similaires et vice-versa.

Groupes de discours

Dans notre recherche, nous avons fait la distinction entre trois populations : les usagers de santé, les médiateurs de santé et les professionnels de santé.

Les usagers de santé

Les usagers de santé ont des besoins d'information médicale avec l'objectif de gérer leur santé au quotidien. Chaque personne est susceptible d'avoir le profil d'usager de santé, car le besoin de services et de produits médicaux est devenu vital et s'informer sur sa santé devient indispensable. Dans cette étude, la catégorie des usagers exclut les médiateurs et les professionnels. Les usagers de santé sont caractérisés dans la section 1.1.

Les médiateurs de santé

Bien que les membres de ce groupe viennent de diverses professions (santé, journalisme, éducation, marketing et communication) et de différents types d'organisations (gouvernementales, privées à but caritatif et commerciales), ils partagent le même objectif : communiquer l'information médicale au grand public. Le type d'information communiquée peut être informationnel, persuasif ou commercial (tableau 1).

1. Mal de tête particulier touchant généralement la moitié droite ou gauche de la tête, associé à une crainte de la lumière et du bruit, et à des troubles digestifs (Vidal de la famille).

Type d'information	Description
Informationnel	Améliorer la compréhension d'un sujet de santé
Persuasif	Influencer les idées ou les comportements liés à un sujet de santé
Commercial	Commercialiser des services ou des produits de santé

TABLE 1 – Types d'information fournis par les médiateurs de santé

Les professionnels de santé

Les professionnels de santé sont définis comme des individus qui ont reçu une formation spéciale pour intégrer une profession liée à la santé. Les membres de ce groupe ont été en contact avec le LS dans le cadre de leur instruction et de leur profession.

Objectifs

La construction de terminologies destinées aux usagers de santé n'est pas une tâche facile. Contrairement à la terminologie médicale, qui contient un vocabulaire plus stable et mieux identifiable, le langage des usagers est dynamique et influencé par l'histoire de chaque individu. De ce fait, l'identification du vocabulaire utilisé par les usagers de santé revient d'abord à l'identification de groupes d'usagers qui utilisent plus ou moins les mêmes termes pour parler des mêmes concepts. Les recherches actuelles suggèrent qu'une grande quantité d'expressions utilisées par le grand public est suffisamment stable pour pouvoir constituer un vocabulaire standard [Zeng et al., 2002, Tse & Soergel, 2003]. Dans un contexte plus précis (les utilisateurs d'Internet dans notre cas), et pour une tâche précise (la recherche d'information médicale), il existe un certain consensus, condition nécessaire pour la construction d'une terminologie. Dans [Zeng & Tse, 2004], un vocabulaire pour les usagers de santé est défini par : *"un ensemble de termes qui représentent les concepts médicaux (et leurs relations) utilisés dans une communication médicale pour une tâche particulière (e.g., la recherche d'information) par un pourcentage significatif de personnes d'un groupe de discours particulier (e.g., les utilisateurs d'un système de recherche d'information destinée au grand public)"*.

Notre travail a commencé dans le contexte du projet *"Les mots au service des patients atteints du cancer du sein"* financé par *La Ligue Nationale Contre le Cancer* et *La Fédération Hospitalière de France* dans le cadre de leur appel à projets annuel. D'une manière générale, le projet visait à identifier les termes et les expressions utilisés par les patients atteints du cancer du sein et leur entourage pour parler de leur maladie.

Au sein de ce projet, notre travail répondait à deux objectifs principaux :

- La construction d'une ontologie du cancer du sein destinée aux usagers de santé : La méthodologie adoptée est fondée sur la sémantique différentielle [Roche, 2003, Bachimont et al., 2002]. Les étapes suivies fournissent un apport méthodologique pour la construction de ressources ontologiques destinées à un public profane dans des domaines de spécialité.
- L'utilisation de cette ontologie dans la reformulation de requêtes des usagers de santé. Les concepts de l'ontologie et les relations qu'ils entretiennent entre eux sont utilisés par la méthode de propagation d'activation pour induire les concepts les plus représentatifs d'une requête donnée. Cette méthode est particulièrement utile dans un domaine où les usagers ne savent pas forcément comment formuler leurs requêtes.

Au cours de notre travail, nous nous sommes également attachés à répondre aux questions suivantes :

1. Quelles sont les caractéristiques des termes et des concepts utilisés par les médiateurs de santé et de ceux utilisés par les usagers de santé ?
2. Quelles sont les différences entre ces familles de termes et concepts ?
3. Quelle est la meilleure manière d'utiliser les relations sémantiques au sein d'une ontologie pour la reformulation des requêtes ?

Pour atteindre ces objectifs, nous avons opéré certains choix méthodologiques, que nous explicitons dans la suite de cette introduction.

Pourquoi Ontologie ?

Au début de notre thèse, nous avons centré notre travail sur la construction d'une ontologie dans le domaine du cancer du sein. Cependant, après l'étude de l'état de l'art et des différentes définitions attribuées au terme "ontologie", nous pensons que nos travaux s'inscrivent plus dans le champ des ontoterminologies. Cette notion a été introduite par *Christophe Roche* [Roche, 2007]. l'ontoterminologie est un néologisme créé à partir des termes ontologie et terminologie : il désigne cette approche qui place l'ontologie au centre de la terminologie. Une approche où l'ontologie joue un rôle fondamental à double titre : pour la construction du système notionnel et pour l'opérationnalisation de la terminologie. Une ontoterminologie est une terminologie dont le système notionnel est une ontologie du domaine [1]. Néanmoins, nous avons gardé dans ce manuscrit le terme ontologie pour caractériser la structure ontoterminologique que nous avons construite, pour deux raisons. D'une part le terme "ontologie" est celui qui désigne habituellement ce type de construction, et d'autre part nous n'avons pas suffisamment mis l'accent sur le travail terminologique au sens strict pour justifier l'appellation d'ontoterminologie, qui serait pourtant la plus appropriée.

Méthodologie et matériel

Le cadre général de notre approche peut être caractérisé par les points suivants : travail sur corpus, construction de l'ontologie, application de l'ontologie dans la reformulation des requêtes.

Travail sur corpus

L'avènement de la linguistique de corpus (Corpus Linguistics), rendu possible grâce aux progrès techniques qui ont augmenté à la fois les capacités de stockage et de traitement de l'information, s'est accompagné d'une renaissance des méthodes statistiques d'analyse des données textuelles [Church & Mercer, 1993]. On peut définir un corpus comme un grand ensemble de textes électroniques sélectionnés et variés. Toutefois, il subsiste trois grandes questions autour de cette notion [Péry-Woodley, 1995] :

- Taille : Quelle est la bonne taille pour un corpus et comment la mesurer (nombre de textes, nombre d'occurrences de formes) ?
- Texte : textes entiers ou échantillons de textes de taille constante ?
- Choix : comment sélectionner les textes entrant dans la composition du corpus pour qu'il soit représentatif de la langue (ou de la variété de langue) à étudier ? Cette dernière question est centrale dans le domaine de la linguistique de corpus.

1. Source Wikipédia

Nous avons utilisé deux corpus de textes acquis à partir d'Internet, manuellement pour le premier et automatiquement pour le second. Le premier corpus est le corpus des médiateurs de santé, issu de sites Web d'information sur le cancer du sein. Le deuxième corpus est celui des usagers de santé, issu de deux forums de patients atteints du cancer du sein. En effet, Internet constitue la plus grande source de textes électroniques non structurés ou faiblement structurés. La méthode de construction de corpus est détaillée dans le Chapitre 4.

Construction de l'ontologie

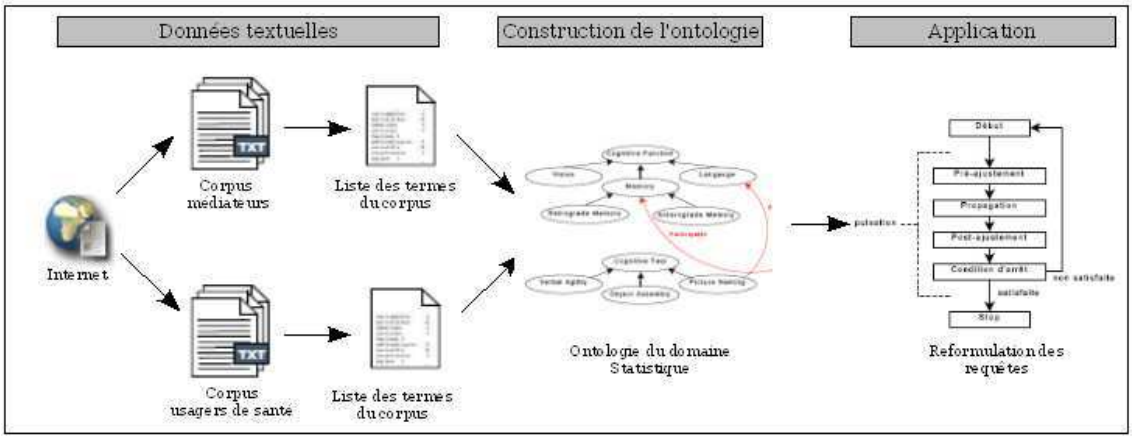
La construction de l'ontologie s'appuie sur l'analyse des corpus textuels construits. L'ensemble des n-grammes (séquence de n mots consécutifs) est extrait et analysé à l'aide d'un concordancier, un outil qui permet à l'utilisateur de formuler des requêtes sous forme d'expressions régulières dans le corpus et d'afficher toutes les parties du corpus contenant cette expression, pour enrichir l'ontologie au fur et à mesure de leur examen. Nous avons également comparé les termes et les concepts issus de chaque corpus en utilisant plusieurs critères quantitatifs ou qualitatifs.

Application de l'ontologie

Plusieurs recherches actuelles s'intéressent aux relations sémantiques pour explorer le contexte qui entoure la requête d'un utilisateur pour l'aider dans la formulation de sa requête [Slaughter, 2002, Soergel et al., 2004]. Dans la même optique, l'ontologie construite a été le pivot de notre application informatique de reformulation de requêtes. Elle a été utilisée comme un réseau sémantique de concepts et de relations entre eux. Des poids différents ont été assignés aux relations entre les concepts selon leur collocation dans le corpus des usagers de santé. Ces poids numériques ont été ensuite utilisés dans l'algorithme de propagation d'activation pour calculer les concepts les plus représentatifs d'une requête donnée. Les étapes de réalisation de cette application sont détaillées dans le chapitre 7.

Organisation du mémoire

Le schéma suivant présente les différentes étapes de notre travail :



La première partie du mémoire décrit le contexte général de la recherche. Dans le chapitre 1, nous commençons par définir la notion d'usager de santé. Nous dressons ensuite un panorama d'études autour des usagers de santé et de leurs besoins en information médicale.

Le chapitre 2 est consacré à la représentation des connaissances, principalement aux solutions terminologiques et ontologiques. Nous présentons un bilan des ressources terminologiques et ontologiques existantes en médecine. Nous faisons également un état de l'art des formalismes de représentation et des méthodes de construction d'ontologies. Finalement, nous faisons l'inventaire des différents types d'outils d'ingénierie ontologique et de traitement automatique du langage pouvant aider à la construction d'ontologies.

Le troisième chapitre présente les notions et concepts de base de la recherche d'information. Nous présentons les principaux modèles utilisés pour implémenter les systèmes de recherche d'information (SRI) et nous abordons la notion de recherche d'information à base d'ontologies.

La deuxième partie du mémoire est consacrée à la construction de l'ontologie du cancer du sein destinée aux usagers de santé. Dans le chapitre 4, nous détaillons les étapes de construction de corpus utilisés comme source de connaissances pour cette ontologie. Dans le chapitre 5, nous présentons les différentes étapes de construction de cette ontologie, les résultats obtenus ainsi que l'étude quantitative et qualitative des différents éléments de cette ontologie.

La dernière partie présente l'application de cette ontologie à la reformulation des requêtes des usagers de santé. Le chapitre 6 présente un état de l'art des techniques de propagation d'activation dans la recherche d'information. Le chapitre 7 détaille les étapes de réalisation de notre application utilisant l'ontologie pour la reformulation des requêtes. Nous terminons par la présentation d'un ensemble de perspectives possibles à ce travail de recherche.

Première partie

Contexte de la recherche

Chapitre 1

L'information médicale et les usagers de santé

1.1 Les usagers de santé

Selon [Tse & Soergel, 2003], un usager de santé est une personne qui cherche des informations sur des sujets ou des services et produits de santé, dans le but de prendre une décision à propos d'un problème médical personnel. Le terme anglais utilisé pour désigner les usagers de santé est *health consumer*, littéralement *consommateur de santé*. cette dénomination n'est pas anodine ; et le terme "consommateur" peut être justifié par les recherches actuelles qui visent à établir des modèles de comportement d'usagers en essayant de prévoir le comportement des consommateurs face à des choix multiples et qui sont influencés par des sources d'information externes (ex. publicité, statistiques) et des facteurs internes (ex. attractivité - *perceived desirability*). Une recherche d'information active est considérée comme partie intégrante du processus de décision : le consommateur, face à plusieurs choix possibles, évalue son attirance pour chaque option, les bénéfices et les risques possibles de chacune, avant de décider d'une action [Lenz, 1984]. Par exemple, Neelamegham et Jain (1999) ont développé un modèle économique de consommation pour les services (*experience goods*), tels que le choix d'un film dans une vidéothèque :

> Les usagers peuvent évaluer la qualité de ces biens seulement après avoir fait leur choix. Donc, ils vont probablement s'appuyer largement sur des influences psychologiques, telles que leurs préférences, et des informations extérieures telles que des incitations verbales ou le bouche-à-oreille [Neelamegham & Jain, 1999, p. 373].

Les services et les produits de santé semblent entrer dans cette catégorie. Les usagers, dans ce type de modèle économique, sont considérés comme des chercheurs d'information qui font leurs choix en se basant sur leurs préférences personnelles et leurs interactions sociales. Le modèle du consommateur pour les services est donc approprié pour l'étude des décisions personnelles sur les questions de la santé.

Les préférences personnelles et les relations sociales jouent un rôle très important dans le choix et l'adoption d'un traitement. Les conséquences de cette action ne peuvent être évaluées avant qu'elle n'ait eu lieu. Rudd et Glanz ont appliqué la théorie du CIP *"Consumer Information Processing"* pour étudier comment les usagers de santé utilisent l'information médicale. Dans le modèle CIP, les variables prises en compte concernent la capacité à traiter l'information, la disponibilité des ressources, le niveau de motivation, la capacité de trouver et d'interpréter l'information, les compétences dans la recherche d'information, et le choix des ressources.

> Rudd et Glanz ont conclu que les usagers de santé utilisent un critère de satisfaction plutôt qu'un critère d'optimisation pour décider d'arrêter un processus de recherche d'information. Ils arrêtent leur recherche dès qu'ils trouvent une alternative satisfaisante, plutôt que continuer jusqu'à l'obtention de la meilleure. Les usagers veulent traiter la moindre quantité d'information possible afin de prendre rapidement des décisions [Rudd & Glanz, 1990, p. 123].

Cependant, si nous voulons traiter le problème de l'accès des usagers de santé à l'information médicale, ce modèle demeure insuffisant parce qu'il pose plusieurs hypothèses qui ne sont pas toujours valides :

- l'information médicale recherchée est disponible,
- l'information médicale est accessible,
- l'information médicale est compréhensible.

En effet, il arrive souvent que les usagers de santé ne sachent pas si l'information médicale est disponible, où la trouver, ou même quelle information ils doivent chercher. Quoi qu'il en soit, ce modèle demeure un point de départ intéressant pour représenter les usagers de santé.

Les usagers s'intéressent de plus en plus à la santé, surtout après les récentes (r)évolutions de la médecine. Premièrement, les avancées dans la recherche biomédicale ont révolutionné notre compréhension de la santé et des maladies. Une conséquence de ces avancées est l'explosion de la littérature biomédicale. La base de données de citations bibliographiques MEDLINE de la NLM (National Library of Medicine) contient, à ce jour, plus de 11 millions d'enregistrements [1]. Par ailleurs, le temps nécessaire pour transférer une recherche du laboratoire à la vie de tous les jours est en moyenne de 17 ans [Balas & Boren, 2000] ; ce qui confirme le fossé entre le monde des chercheurs et des professionnels de santé et celui du grand public. Deuxièmement, les coûts de santé continuent d'augmenter. Par exemple, en France, le coût des prescriptions de médicaments est passé de 23.6 milliards d'euros en 2000 à 31.9 milliards d'euros en 2006 (INSEE, 01 mai 2008). Selon une étude réalisée par la Société Française de Médecine Générale en 2002, la durée moyenne de consultation relevée par le médecin généraliste en France est de 16.18 minutes [Kandel et al., 2004]. Cependant, dans un article paru dans *l'Humanité* le 27 décembre 2005, le Collectif Interassociatif Sur la Santé (CISS) estime que la durée moyenne d'une consultation médicale est de 7 minutes. Ces études montrent que le système de santé en France est en train de changer et qu'il va probablement ressembler, dans quelques années, à celui des États-Unis. Vu le coût des soins et le temps de consultation très court, les usagers de santé en France vont probablement adopter le même comportement que ceux des États-Unis, à savoir un taux élevé d'auto-médication et une part importante donnée à la gestion personnelle de leur propre santé. Dans une enquête récente réalisée auprès d'un panel représentatif de médecins généralistes bretons composant le Baromètre des Pratiques en Médecine Générale, 84% des médecins estimaient avoir à effectuer des tâches qui ne relevaient pas de leur qualification. Il s'agissait notamment de tâches liées à l'information et à l'éducation du patient (64.3%), à la coordination des soins (58.3%) ou encore à la gestion du dossier du patient (56%) [Bataillon et al., 2006]. Cette enquête montre qu'une grande partie des médecins ne considèrent pas que l'information des patients fait partie de leur travail. Les patients se trouvent ainsi *obligés* de trouver eux-mêmes plus d'informations sur la santé. Finalement, le développement et la démocratisation d'Internet et des technologies d'information ont augmenté le taux d'accès aux ressources médicales. Le taux d'accès à Internet n'a cessé d'augmenter au cours des 15 dernières années dans l'ensemble des pays industrialisés, pour atteindre 54 % en France et 70 % aux États-Unis en 2007 [2]. En France, la part d'internautes

1. `http://medline.cos.com`
2. `http://www.internetworldstats.com/stats4.htm`

ayant déjà fait des recherches d'information concernant la santé sur Internet a été estimée à 30% par une enquête de l'INSEE en 2005 [Frydel, 2006] et à 50% en Ile-de-France par une précédente enquête de l'Inserm [Renahy, 2007]. L'effet de tous ces facteurs est une demande accrue des usagers à l'information médicale. Ce phénomène est relativement récent et peu de recherches ont été faites pour caractériser les besoins et le comportement des usagers. Traditionnellement, les professionnels de santé ont été la principale source de ce type d'information [Spadaro, 2003]. Cependant, avec les récents changements dans le modèle économique et social du système de santé, tels que des consultations médicales de plus en plus courtes, des remboursements de soins moins importants et des changements culturels significatifs dans les attitudes envers la santé, de plus en plus de personnes recherchent activement de l'information médicale pour prendre des décisions plus salutogènes.

La place d'Internet dans la société et dans la vie courante est devenue incontournable. Pour améliorer la recherche d'information destinée aux patients sur le Web, il faut développer des stratégies et des outils qui prennent en compte les besoins des usagers et qui en premier lieu s'adaptent à leur langage et à leur compréhension des choses.

Le problème de recherche d'information existe aussi du côté des professionnels de santé. Face à la quantité énorme de la littérature scientifique qui ne cesse d'augmenter, il faut développer des stratégies très efficaces pour trouver l'information pertinente à une requête donnée. On a compté qu'il faudrait cinq années de lecture (à raison de deux articles par jour) à un scientifique pour qu'il se mette à jour dans son domaine. Pour ce faire, on a supposé que seulement 1% de la totalité de la littérature qui apparaît chaque année lui est utile [Dwivedi et al., 2003]. Le professeur Alex Markham, président du NCRI (Britain's National Cancer Research Institute), a affirmé que plus de 80% de l'information disponible sur les recherches du cancer ne trouve pas son chemin vers la communauté concernée. Le NCRI et un groupe de partenaires qui inclut le gouvernement, le CRUK (Cancer Research UK), le conseil des recherches médicales et The Wellcome Trust, envisagent de développer une plate-forme informatique internationale qui permettra l'accès et l'analyse des données médicales de différentes disciplines. Ils ont lancé en Mars 2004 une initiative pour améliorer l'échange d'information entre les groupes de recherche sur le cancer en encourageant ces derniers à rendre leurs données accessibles aux autres [Pincock, 2004]. Ce projet est soutenu par les organisations internationales de recherche sur le cancer et US NCI (United-States National Cancer Institute).

Nous constatons que la question posée sur l'accès à l'information médicale a une portée mondiale. Un travail similaire en France serait d'un grand bénéfice pour les professionnels de santé ainsi que pour les patients, surtout lorsque l'on sait qu'il y a actuellement un grand nombre de sources francophones de l'information de santé sur Internet.

Dans ce qui suit, nous essayons de montrer à travers plusieurs études les besoins des patients, particulièrement ceux atteints du cancer du sein, en information médicale. Ces données nous serviront à soulever des questions ainsi que des perspectives pour répondre à de tels besoins.

1.2 Les besoins des usagers de santé en information médicale

La recherche d'information médicale est l'une des tâches les plus courantes des utilisateurs d'Internet. Ceux-ci ont tendance à rechercher l'information médicale sur des moteurs de recherche populaires plutôt qu'au travers des sites spécialisés. Une étude menée par le NCI (National Cancer Institute) aux États-Unis [Bader & Theofanos, 2003] a tenté de cerner les besoins spécifiques des usagers quant aux informations recherchées sur le cancer, la manière dont ils expriment leurs requêtes et le niveau de détail qu'ils souhaitent avoir. L'étude a été réalisée sur le moteur de

recherche Ask, très utilisé aux États-Unis (35 millions requêtes/mois), en analysant un mois de requêtes liées au cancer. Cette étude a montré que les usagers utilisent le langage naturel pour exprimer leurs requêtes. Ils posent des questions en utilisant des phrases complètes ou des mots-clés. Les catégories les plus recherchées sont :

Cancer : 78.37%
Recherche générale : 10.26%
Traitement : 5.04%
Diagnostic et test : 4.36%
Cause/Risque/Lien : 1.64%
Soutien : 0.33%

Les types de cancer sur lesquels on recherche le plus d'information sont classés de la manière suivante :

Digestif/Gastro-intestinal/Intestin : 15.0%
Sein : 11.7%
Peau : 11.3%
Génital : 10.5%
Hématologique/Sang : 9.2%
Gynécologique : 9.0%
Poumon : 7.8%
Tissu mou/Muscle : 6.6%
Lymphome : 5.6%

Nous remarquons que le cancer du sein est le cancer sur lequel on pose le plus de requêtes si on sépare les trois cancers de la première catégorie.

Une autre étude, menée aux États-Unis sur 188 femmes atteintes d'un cancer du sein, a examiné ce qui caractérise les femmes utilisant Internet [Fogel et al., 2002]. Dans cet échantillon, 41.5% des patientes utilisent Internet principalement pour rechercher l'information médicale. Celles qui utilisent Internet diffèrent des non-usagers par leur revenu, leur niveau d'étude et leur race/ethnie. Les usagers semblent avoir un plus haut revenu ainsi qu'un niveau d'études plus élevé et semblent appartenir à la race blanche. L'âge, le temps écoulé depuis le diagnostic et le stade de la maladie ne semblent pas avoir d'effet notable. Plusieurs patientes souhaitent fortement être bien informées et participer à leur choix de traitement. Une première étude montre que 79 à 96% des patientes préfèrent avoir plus d'informations que celles données par leurs médecins [Cassileth et al., 1980]. Une seconde étude a montré que seulement 19% des 232 patientes étaient satisfaites des informations fournies par leurs médecins [Wiggers et al., 1990].

Selon l'INSERM, en France, plus de 3/4 des internautes consultent des sites de santé. L'internaute en quête d'informations médicales est plutôt une femme (67.9%), d'âge moyen (la moitié à entre 29 et 53 ans), au niveau d'étude élevé, avec une bonne expérience d'Internet et confronté(e) à un problème de santé, personnel ou d'un proche [Renahy et al., 2007].

Une autre étude [Satterlund et al., 2003], également menée aux États-Unis, a montré que 36 à 55% des utilisateurs d'Internet de la population américaine y accèdent pour rechercher de l'information médicale. Le cancer est l'une des maladies sur laquelle on recherche le plus d'information sur Internet. Deux cent vingt quatre femmes récemment diagnostiquées avec un cancer du sein de stade I, II ou III ont participé à cette étude. Elles ont été interrogées 8 et 16 mois après le diagnostic. Huit mois après le diagnostic, les trois sources d'information les plus utilisées par les femmes étaient les livres (64%), Internet (49%) et les vidéos (41%). Cependant, après seize mois, la source d'information la plus citée était Internet (40%) suivie par les livres (33%) et enfin La Société Américaine du Cancer. Les femmes continuaient à utiliser Internet pour rechercher des informations, même après avoir terminé leur traitement. Ainsi, Internet continue

à jouer un rôle important pour les survivants du cancer après la fin de leur traitement.

Ces chiffres reflètent l'importance de l'information dans la vie des patients. Dans nos recherches, nous n'avons pas pu trouver d'étude comparable pour la France. Néanmoins, on peut penser qu'un tel travail mettrait en lumière la même réalité et montrerait la nécessité d'envisager des services destinés aux patients pour améliorer la recherche d'information.

Dans une analyse de 24 sondages publiés [Eysenbach, 2003], l'auteur estime que 39% des personnes atteintes du cancer dans les pays développés utilisent Internet directement, et 15 à 20% l'utilisent indirectement à travers la famille et les proches. L'auteur distingue aussi quatre types d'utilisation d'Internet : communication (courrier électronique, ...), communauté (groupes virtuels de support, forums, ...), contenu (information médicale sur le Web) et e-commerce. L'auteur a également proposé un schéma qui relie les trois premiers services fournis par Internet et les résultats possibles sur la communauté atteinte du cancer (voir Fig 1.1).

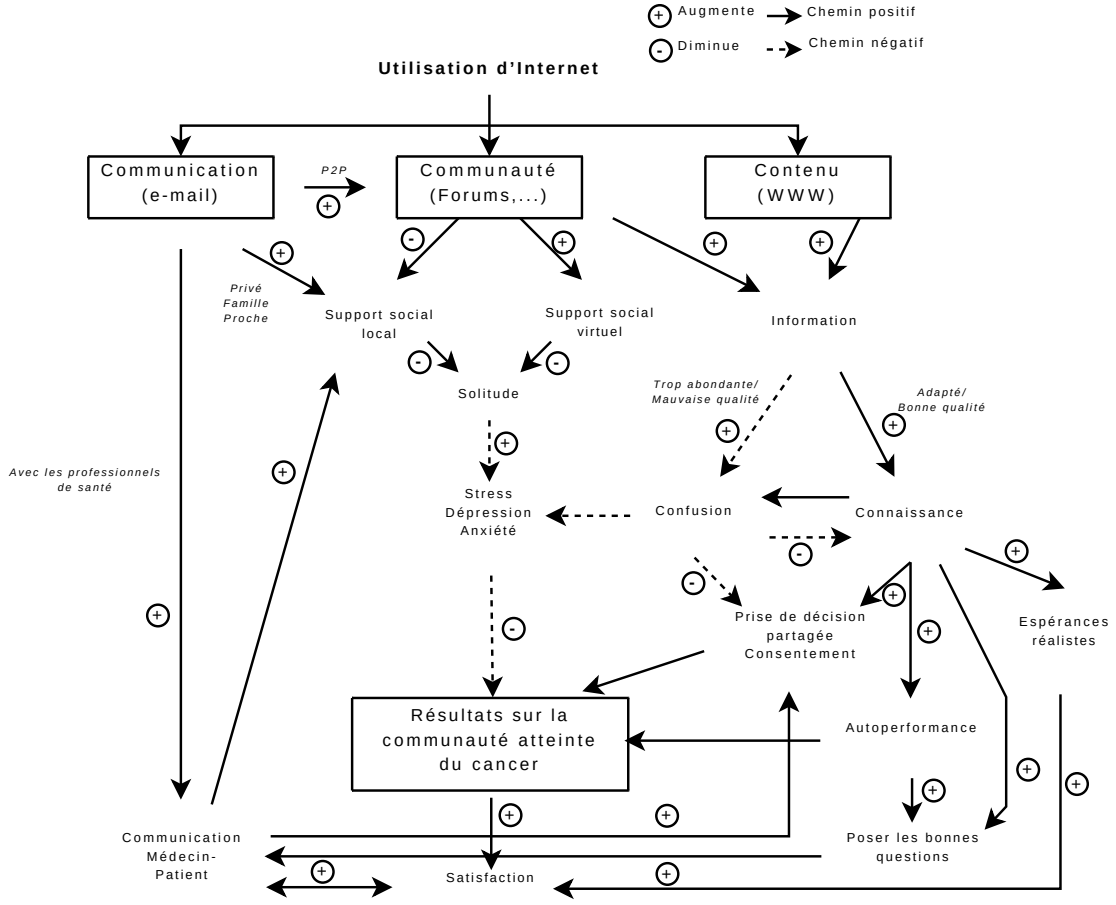


FIGURE 1.1 – Internet et les résultats possibles sur la communauté atteinte du cancer

En 2005, selon le 9ème sondage réalisé par HON (Health On The Net) [1] 44.5% des personnes qui utilisent Internet à des fins médicales le font depuis 7 ans ou plus, tandis que l'on observe seulement 5% de nouveaux accès dans les 6 derniers mois. L'analyse du sondage a montré que 54% des participants ont déjà discuté le résultat de leur recherche sur Internet avec leur médecin. La plupart des patients participants (88.2%) ont rapporté que rechercher l'information médicale

1. `http://www.hon.ch/survey/survey2005/res.html`

sur Internet améliore la qualité des consultations avec leur médecin. Plus de la moitié (53%) utilisent Internet pour rechercher une seconde opinion sur un diagnostic médical.

Une étude récente réalisée en 2005 par Harris Interactive a rapporté que le nombre des citoyens américains qui ont déjà utilisé Internet pour rechercher l'information médicale a atteint 117 millions, contre 111 millions l'année précédente (based on July 2004 U.S. Census estimate released January 2005).

1.3 La qualité de l'information médicale sur le Web

Dans ce qui suit, nous nous intéressons à la qualité de l'information médicale présente sur le Web et au rapport éventuel avec le type des sites. En effet, dans l'extrême majorité des cas et contrairement à l'édition papier validée par les comités de lecture de revues scientifiques, l'information disponible sur ce réseau n'a pas été évaluée. La facilité de création d'un site Web a fait que les sources de l'information médicale sont très hétérogènes, variant d'un site commercial, à un site professionnel, gouvernemental, voire personnel, d'où la nécessité d'une critique "sérieuse" de cette information.

La Fondation "Health On the Net" (La Santé sur Internet), une fondation sans but lucratif dont le siège est à Genève, vise à promouvoir le développement et les applications de nouvelles technologies d'information notamment dans les domaines de la médecine et de la santé. Elle a édité un code de conduite destiné aux sites Web de santé, le HONcode. Celui-ci est une charte dont le but est de certifier certains sites Web médicaux et de santé. En adhérant aux principes et en affichant le sceau actif HONcode, le propriétaire d'un site Web s'engage à respecter les huit principes de bonne conduite élaborés par la fondation. Cette charte permet de s'assurer qu'un site Web respecte certains standards, notamment en termes de vérifiabilité des informations. Néanmoins, elle ne garantit pas la véracité des informations présentes sur un site, le contrôle du contenu ne faisant pas partie des critères de sélection. Les demandes d'accréditation sont volontaires et les sites accrédités sont revisités annuellement par HON. L'objectif est de former un espace de confiance dans lequel les citoyens peuvent consulter les informations qui les intéressent avec un certain degré de confiance.

Une étude menée à Rouen [1] a proposé un ensemble de critères pouvant être utilisés pour évaluer la qualité de l'information de santé sur Internet. Le travail a été réalisé par un groupe pluridisciplinaire comportant des médecins, des ingénieurs, des bibliothécaires et des juristes. Ces critères peuvent être utilisés de deux façons : 1) par les cyber-citoyens pour améliorer leur esprit critique ; 2) par les maîtres-toile (webmasters) des sites de santé francophones pour en augmenter la qualité. Le groupe a défini 49 critères, répartis en huit catégories : crédibilité, contenu, hyperliens, design, interactivité, aspects quantitatifs, déontologie et accessibilité. Chaque critère peut être considéré comme soit un critère essentiel, soit un critère important soit un critère mineur. Il en est déduit un score global du site. Il ressort de leur étude que la question posée sur l'évaluation de la qualité d'un site Web n'est pas facile et que les scores peuvent être différents d'un expert à un autre.

Une seconde étude a été menée aux États-Unis pour déterminer les caractéristiques des sites populaires relatifs au cancer du sein et si les sites les plus populaires sont d'une qualité d'information supérieure [Meric et al., 2002]. Le moteur de recherche Google a été utilisé pour générer la liste des sites sur le cancer du sein. Google range les résultats de recherche en fonction des mesures de liens de popularité (le nombre des liens à un site à partir d'autres sites). Les sites

1. `http://www.chu-rouen.fr/netscoring`

les plus populaires selon le rangement de Google semblent contenir plus d'informations que ceux moins populaires :

- Essais cliniques en cours (27% vs 21%).
- Résultats de recherches (12% vs 3%).
- Opportunités d'un soutien psychosocial (48% vs 23%).

Les mesures de qualité telles que l'affichage de la profession de l'auteur, l'attribution ou les références, l'actualité de l'information, ne diffèrent pas entre les deux groupes de sites. L'étude a montré que la popularité des sites est associée à leur type plutôt qu'à la qualité de leur contenu. De ce qui précède, on conclut que nous sommes aussi confrontés à un problème de qualité. La quantité d'information sur le Web devient énorme, mais on ne dispose pas de critères établis pour contrôler sa qualité.

Cette question est illustrée par une étude qui a examiné si l'information sur l'imagerie par mammographie présente sur les sites Web est équilibrée et indépendante de la source d'information, et reflète des conclusions récentes [Jorgensen & Gotzsche, 2004]. Au cours de cette étude, 27 sites Web ont été analysés. 13 sites de groupes partenaires et 11 sites d'institutions gouvernementales recommandent tous la mammographie. En revanche, les 3 sites des organisations de consommateurs mettent en question une telle pratique. Ainsi, il semble que les groupes partenaires et les institutions gouvernementales favorisent l'information qui soutient la mammographie. Les dangers majeurs d'une telle pratique, qui sont le sur-traitement et le sur-diagnostic, sont mentionnés seulement par quatre sites de ces groupes, mais le sont par tous les sites des organisations de consommateurs. De plus, l'information sélectionnée était parfois erronée ou trompeuse et ne reflétait pas les nouveaux résultats d'essais, à l'exception des sites des organisations de consommateurs, où l'information était plus équilibrée et plus compréhensible. L'importance d'une information équilibrée est soulignée par une étude qui a trouvé que 61% des femmes décident elles-mêmes d'avoir une mammographie et 26% prennent la décision avec leur médecin [Jorgensen & Gotzsche, 2004].

Ces chiffres montrent l'importance de disposer d'une information objective qui soit à la fois d'actualité et de qualité. L'information médicale peut influencer le choix des patients et joue un rôle important dans leur vie. Il y a donc un besoin réel de traiter cette question.

1.4 Les problèmes de la communication médicale

La disponibilité de l'information n'implique pas son accessibilité conceptuelle. Comme la plupart des gens ne possèdent pas une formation médicale, leur compréhension de la maladie dépendra largement de leurs acquis culturels. Rothschild [Rothschild, 1998, p. 299] a défini la culture comme un mélange d'influences incluant : *"la race, l'ethnie, l'origine, l'origine de la famille..., la religion, la langue, les croyances, l'histoire personnelle et familiale, et le statut économique"*.

La recherche sur la terminologie des usagers de santé est un domaine relativement jeune [McCray et al., 1999, Patrick et al., 2001, Zeng et al., 2001]. Une session à la conférence MEDINFO en 2001 a réuni plusieurs chercheurs dans ce domaine, qui ont conclu que le développement d'un vocabulaire des usagers de santé doit être basé sur des recherches qui tiennent compte de leurs besoins et de la manière dont ils sont exprimés.

Le développement des vocabulaires médicaux destinés au grand public est essentiel pour faciliter la recherche d'information, la compréhension des documents médicaux et la réalisation d'autres tâches nécessitant une recherche dans des documents. Bien qu'il soit depuis longtemps reconnu que les patients et les professionnels de santé expriment et comprennent différemment

les concepts médicaux, cette différence de langage continue à affecter la communication médicale entre les médecins et leurs patients, la recherche d'information et finalement la prise de décision médicale. Les recherches actuelles se sont attachées à identifier et caractériser les termes utilisés par les patients et à observer les effets de la disparité entre les deux langages, spécialisé et non spécialisé (ou profane), mais peu d'efforts ont été faits pour développer des terminologies destinées aux patients. Ce problème de différence de langage devient d'autant plus urgent à traiter que de plus en plus de personnes sont à la recherche d'informations médicales sur leur situation et qu'elles prennent leur santé plus en charge.

Dans la littérature, les recherches sur la compréhension par les usagers de santé des termes et concepts médicaux se sont essentiellement focalisées sur des listes discrètes de termes dans différentes spécialités, par exemple, 17 termes en oncologie [Chapman et al., 2003], 6 paires de mots (spécialisé et profane) parmi les patients d'un service d'urgence [Lerner, 2000] et 50 termes médicaux parmi les patients d'une unité de chirurgie et les membres de leurs familles [Spees, 1991]. Ces études confirment l'hypothèse que les usagers de santé ont des difficultés à comprendre le langage médical. Cependant, elles ne fournissent aucune méthodologie pour identifier les termes alternatifs que les patients ont le plus de chance d'utiliser, d'identifier et de comprendre.

Une étude a essayé d'évaluer la compréhension par les patients des termes utilisés par les médecins durant les consultations médicales en oncologie [Chapman et al., 2003]. Les termes et les phrases ont été sélectionnés à partir de 50 enregistrements vidéo de consultation, et ont été utilisés dans un sondage de 105 personnes choisies au hasard. Le questionnaire se composait de scénarios contenant des termes de diagnostic/pronostic potentiellement ambigus, des questions de compréhension avec choix multiples et des images sur lesquelles les personnes devaient localiser des organes censés être affectés par un cancer. Les participants devaient aussi donner un taux reflétant leur niveau de confiance dans leurs réponses. Près de la moitié des personnes de l'échantillon ont compris les euphémismes qui désignent une métastase du cancer ex. "seedlings", "spots in the liver" (44% et 55% respectivement). Soixante-six pour cent ont été conscients que le terme "métastase" veut dire que le cancer est en train de se disséminer, mais 52% seulement ont compris que "la tumeur est en train de progresser" était une mauvaise nouvelle ; pourtant les participants étaient suffisamment confiants dans leurs réponses. La connaissance concernant la localisation des organes était variable ; 94% ont identifié correctement les poumons mais seulement 46% ont pu localiser le foie. Les résultats de cette étude laissent penser qu'une partie significative du grand public ne comprend pas les phrases utilisées dans les consultations du cancer et qu'on ne peut pas lui attribuer une connaissance anatomique, même élémentaire. Par contre, le niveau de confiance élevé indique que poser la question aux patients s'ils ont compris peut donner une réponse qui est susceptible d'être surestimée, comme l'ont déjà signalé les auteurs de [Chan & Woodruff, 1997].

Un rapport de l'Institut de Médecine aux États-Unis, intitulé *"la culture sanitaire : Une ordonnance pour mettre fin à la confusion"*, a montré que près de la moitié de la population américaine, soit 90 millions de personnes, montre des difficultés à comprendre et à utiliser l'information médicale, et qu'il y a un taux plus élevé d'hospitalisation et d'utilisation des services d'urgence parmi les patients ayant une culture sanitaire limitée. Cela peut conduire à des milliards de dollars de dépenses dans des coûts de santé, dépenses qui pourraient être évitées. La culture sanitaire (traduction pratiquée au Canada du terme *health literacy*), est définie comme *"un concept global aidant l'individu à se comporter de manière responsable en matière de santé à l'intérieur et à l'extérieur du système de santé, avec le concours de son environnement social ; cette culture permet aussi à la population d'influencer les sphères sociale et politique de telle manière qu'elles promeuvent des comportements salutogènes."* [Nielsen-Bohlman, 2004].

Dans [Zeng & Tse, 2006], les auteurs déclarent que pendant une communication directe entre le patient et son médecin, la nature interactive d'une telle rencontre permet une négociation

des idées à un niveau de discours commun. Par exemple, la personne peut demander des clarifications quand le concept n'est pas très bien assimilé, ou la répétition des points importants. Cependant, en réalité, ce genre de discours idéal n'a pas toujours lieu entre le patient et son médecin, et le patient aura tendance à se renseigner par d'autres moyens, parmi lesquels Internet prend aujourd'hui une place importante. Cependant, la nature même de l'interaction entre les humains et les ordinateurs ne permet pas, contrairement à la communication directe, d'adapter le discours afin de faciliter la compréhension. Par ailleurs, les médias "statiques" comme les brochures, les magazines, les journaux et la télévision, que plusieurs personnes utilisent pour trouver l'information médicale, sont basés sur le principe de "taille unique" ou "plus petit dénominateur commun", et sur la connaissance et le jugement des auteurs. Durant l'une de ces interactions, que ce soit avec un médecin, un ordinateur ou un document, les patients peuvent "remplacer" d'une manière correcte ou incorrecte les termes qu'ils ne comprennent pas en utilisant leurs propres connaissances, leur expérience et leurs préférences. Dans le meilleur des cas, le patient va déduire correctement le sens du terme et va donc enrichir sa connaissance. Dans le pire des cas, il va mal interpréter le terme, lui donnant un sens différent de celui voulu par le médecin. Dans [Rubin, 2003], l'auteur raconte l'histoire d'un patient qui a donné un sens géographique au terme de son médecin *"traitement local"* au lieu du sens anatomique. Inversement, les patients peuvent utiliser des termes techniques, mais les associent à des concepts différents ou incomplets (ex., *"dépression"* pour *"tristesse"*). Les patients ne pouvant pas "remplir les vides" (à cause d'une culture sanitaire insuffisante) vont se sentir perdus, et probablement être plus stressés par leur situation.

Ainsi, le discours des patients peut comprendre une combinaison de la terminologie médicale et des expressions du langage de tous les jours, avec plusieurs interprétations possibles basées sur des facteurs individuel, contextuel, sociétal et culturel.

1.5 La recherche d'information médicale sur le Web

Le Web est l'une des ressources principales utilisées par les usagers pour rechercher de l'information médicale. A cause de (et malgré) la grande quantité d'information disponible sur le Web, les usagers rencontrent souvent des obstacles dans la recherche d'information. Les études ont identifié certains de ces obstacles, qui comprennent la soumission de requêtes mal formées, la disparité dans la terminologie et les termes ayant un sens trop large ou trop étroit, ou en dehors du domaine. De telles requêtes mènent à des résultats incomplets, non pertinents ou inexistants. La plupart des professionnels de santé reconnaissent que les patients et les non-spécialistes ne sont pas familiers avec la terminologie médicale. Ainsi, utilisant leurs propres termes (de nature informelle) au lieu des termes professionnels (savants et formels), ils vont rechercher *"perte de cheveux"* au lieu de *"alopécie"*, et *"douleur au sein"* au lieu de *"mastodynie"*.

Une étude aux États-Unis a essayé de voir si reformuler les requêtes des usagers avec la terminologie provenant du métathésaurus UMLS amenait à de meilleurs résultats au niveau de la recherche d'information [Plovnick & Zeng, 2004]. Les résultats ont montré que 42% des recherches utilisant les requêtes reformulées amènent à des résultats meilleurs que ceux obtenus en utilisant les requêtes originales, 19% amènent à des résultats plus mauvais, et les résultats pour le reste (39%) sont inchangés. Ils ont pu identifier des termes ambigus, des expansions d'acronymes comme facteurs des différences de performance. Ils ont noté une tendance vers une précision accrue en fournissant des substitutions pour les termes des patients, les abréviations et les acronymes. Ils ont estimé que la reformulation des requêtes avec la terminologie professionnelle pourrait être une méthode prometteuse pour améliorer la recherche d'information réalisée par

les patients.

Une équipe de chercheurs aux États-Unis a collecté et analysé les termes utilisés par les patients et ceux utilisés par les médecins pour évaluer quantitativement leurs différences [Zeng et al., 2001]. Ils ont aussi analysé la performance de la recherche d'information en utilisant les termes spécialisés et ceux du grand public. Pour cela, ils ont collecté un ensemble de termes a partir des requêtes soumises par les usagers sur la page *Find-A-Doctor* du site Web de *Brigham and Women's Hospital* (BWH). L'objectif des requêtes était de trouver des médecins avec des intérêts cliniques spécifiques. Ils ont analysé toutes les données provenant de la section "intérêts cliniques". Les résultats ont montré que la terminologie utilisée par les patients diffère de celle utilisée par les médecins sur plusieurs aspects comprenant le taux des fautes d'orthographe (13.9% pour les patients vs 7.8% pour les médecins), le taux d'alignement vers UMLS (Unified Medical Language System) (62.3% vs 74.9%) et la pertinence des résultats de requêtes (56.5% vs 79.5%).

Une seconde recherche aux États-Unis a étudié la nature des échecs de recherche d'information sur deux sites Web de la NLM (National Library of Medicine) destinés au grand public : *ClinicalTrials.gov* et *MEDLINEplus* [McCray & Tse, 2003]. Le but de l'étude était d'analyser et de classer les requêtes qui ne mènent à aucun résultat. Les résultats indiquent que pour ces requêtes vides il y a trois classes de phénomènes pouvant être la cause de ces échecs : des problèmes dans la couverture du domaine, la formulation des requêtes et les fonctionnalités du système. Chacun de ces phénomènes est potentiellement susceptible d'une amélioration qui pourrait augmenter l'accès des usagers à l'information disponible.

Le Web est une source très abondante d'information pour les patients, qui restera partiellement inaccessible sans une stratégie efficace d'exploitation. Ceci peut augmenter la confusion et même l'anxiété des patients, à cause de la non-compréhension de leur maladie. Avec la généralisation d'Internet, il y a un besoin crucial de réfléchir à des méthodes d'aide à la recherche et à la compréhension de l'information médicale. Anticiper ces recherches peut être d'un grand bénéfice, surtout dans la perspective attendue de l'explosion des données sur Internet.

Chapitre 2

Les ontologies dans le domaine médical

2.1 Introduction

Les ontologies sont devenues un moyen indispensable pour représenter et exploiter les données et les connaissances d'un domaine, et plus particulièrement celles du domaine médical. Les données et les connaissances médicales ont été structurées au moyen de taxonomies et de classifications longtemps avant que les ontologies ne deviennent indispensables pour le Web sémantique. Il existe plus de 100 classifications médicales (ex., ICD10, MeSH, SNOMED), ce qui rend difficile l'exploitation des données codées par l'une ou l'autre. Dans l'absence d'une ontologie de la santé - encore à venir, l'initiative UMLS (Unified Medical Language System) essaye de fournir un accès unifié à ces classifications.

L'objectif de ce chapitre est de fournir les notions de base pour appréhender la notion d'ontologie. Malheureusement, la communauté scientifique n'est pas arrivée à un consensus pour définir ce qu'est une ontologie. Ce qui est paradoxal, surtout si l'on sait que l'une des principales caractéristiques d'une ontologie est de représenter un consensus d'une communauté sur un domaine particulier. Cependant, il n'est pas nécessaire d'entrer dans le débat des spécialistes sur les ontologies pour comprendre leurs principales caractéristiques, et donc être capable de les utiliser d'une manière pertinente et efficace dans plusieurs tâches.

2.2 Ontologie : épistémologie et définitions

Dans cette section, nous allons présenter tout d'abord un exemple concret d'ontologie dans le but d'introduire les notions de base liées aux ontologies. Nous présenterons ensuite un petit historique de l'origine des ontologies, de leur début en philosophie à leur introduction dans le domaine de l'informatique.

2.2.1 Ontologie du cerveau et fonctions cognitives : exemple et notions de base

Cet exemple est tiré du projet français BC^3 (Bases de Connaissances Cœur-Cerveau) [Bonnevay & Lamure, 2003]. L'objectif du projet, qui s'est déroulé de 2000 à 2003, était l'extraction et la représentation des connaissances concernant deux organes vitaux, le cerveau et le cœur, dans les domaines anatomique et fonctionnel. Le projet engageait plusieurs équipes de la région Rhône-Alpes et le Centre de Neuropsychologie de l'hôpital de la Pitié-Salpêtrière (Paris).

Pour le cas spécifique du cerveau, un des enjeux était la compréhension des mécanismes cérébraux chez l'homme afin d'améliorer la prise en charge des patients cérébro-lésés. Cette compréhension passe notamment par la réalisation d'une base de connaissances qui rend disponibles les connaissances de nature anatomique et fonctionnelle sur le cerveau, ainsi que les liens anatomo-fonctionnels entre les zones anatomiques du cerveau et les fonctions cognitives [Diallo, 2006]. Ces connaissances ont été formalisées à travers une ontologie anatomique, une ontologie fonctionnelle et une ontologie des tests cognitifs ; ces trois ontologies et les relations entre elles représentaient les connaissances anatomo-fonctionnelles du cerveau. De manière générale, une fonction cognitive (ex. *la mémoire, la langue*) est évaluée par un ensemble de tests cognitifs (ex. *test d'agilité verbale, test de dénomination d'image*). Une région ou une zone particulière du cerveau participe à l'exécution d'une fonction cognitive donnée. Par exemple, *le lobe temporal* est impliqué dans *la mémoire*. La figure 2.1 représente un exemple de l'ontologie construite dans le projet BC³. Il illustre deux notions de base des ontologies : les concepts et les relations. Il montre aussi l'utilisation d'une relation particulière *Partie_de* qui relie une entité à ses composants.

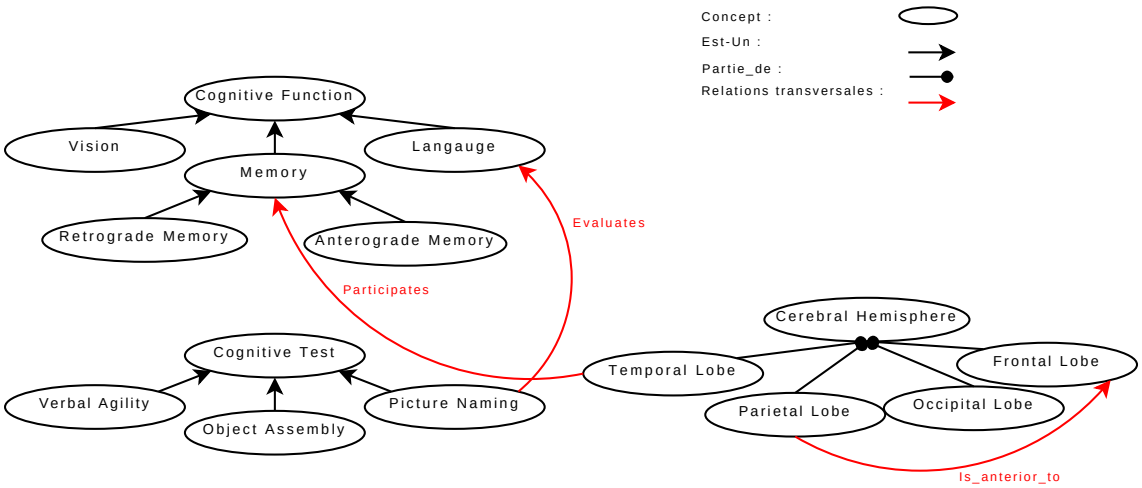


FIGURE 2.1 – Un exemple d'une structure conceptuelle dans le domaine cérébral

Dans cet exemple, il n'existe pas de définitions associées aux concepts. D'une manière générale, l'objectif d'une ontologie étant d'établir un consensus au sein d'une communauté dans un domaine particulier (général ou spécifique), il est nécessaire, durant l'étape de construction de l'ontologie, de fournir des définitions aux concepts, au moins en langage naturel. Ces définitions peuvent révéler des désaccords ou des conflits dans la compréhension des concepts au sein d'un groupe, et amener ce dernier à chercher à atteindre un consensus.

Associer des définitions formelles aux concepts et aux relations nécessite l'utilisation d'un langage formel. *La logique du premier-ordre* est appropriée pour une telle tâche, et a été effectivement utilisée dans le premier travail de formalisation des ontologies, dans l'environnement KIF/Ontolingua [Gruber, 1993]. Actuellement, *les logiques de description* (LD), un sous-ensemble de la logique du premier-ordre, est le formalisme le plus utilisé pour représenter formellement les ontologies. Le langage de représentation des connaissances OWL *(Web Ontology Language)*, l'actuelle recommandation du W3C pour la représentation des ontologies, est basé sur les logiques de description. Le premier travail en médecine qui a utilisé les LD est le système GALEN [Rector & Nowlan, 1994]. Il utilise une LD spécifique appelée GRAIL, dont l'auteur, *I. Horrocks*, est l'un des fondateurs du langage OWL.

2.2.2 Origines et définitions

Les catégories d'Aristote peuvent être considérées comme la première tentative de construire une ontologie de ce qui existe. Aristote a identifié et nommé dix catégories[1] qui servent à classer n'importe quel objet ou être vivant. Ces catégories peuvent nous sembler curieuses aujourd'hui, mais à l'époque ce travail était original et avec le recul on peut le juger remarquable. Elles étaient établies sous forme de liste plate, mais cinq siècles plus tard, Porphyre (234 ?-305 ?) les a organisées en une structure d'arbre et a fourni les principes de base pour différencier les nœuds pères des nœuds fils, ainsi que les nœuds du même niveau (les nœuds frères) (voir Fig 2.2[2]). Ces principes consistent à identifier des ensembles de caractéristiques qui distinguent deux nœuds proches. Ces principes sont connus sous le nom de *principes différentiels* et sont la base de plusieurs approches contemporaines de construction d'ontologie [Roche, 2003, Troncy & Isaac, 2002, Bachimont, 2000].

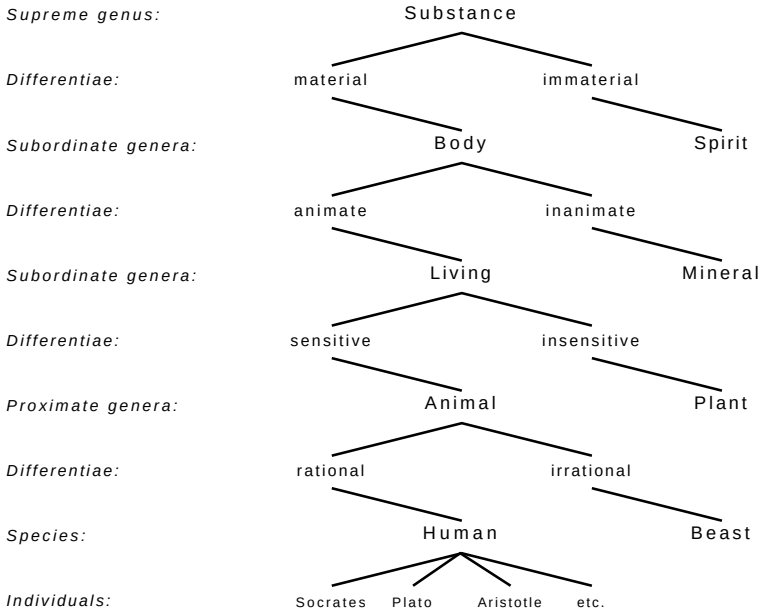


FIGURE 2.2 – Arbre de Porphyre

Le mot latin *Ontologia* est apparu pour la première fois dans un texte en 1613. Le mot anglais *ontology* est apparu pour la première fois dans *An Universal Etymological English Dictionary* par Nathaniel Bailey en 1721 : *"An account of being in the abstract"*. Il est apparu pour la première fois dans un titre de livre en 1733 [3]. Il faisait référence au sens philosophique du terme.
Barry Smith, un philosophe qui travaille sur les ontologies biomédicales, a donné la définition suivante au terme *ontology* : "L'ontologie comme branche de la philosophie est la science qui étudie ce que sont les objets, leurs types, leurs structures, leurs propriétés et les relations qui peuvent exister entre eux dans le domaine de *la réalité*... En d'autres termes, elle cherche à établir une classification des entités. Chaque domaine scientifique aura sa propre ontologie définie par le vocabulaire du domaine et par les formules canoniques de ses théories" [Smith, 2003].

1. Les catégories d'Aristote : la Substance, la Quantité, la Relation, la Qualité, l'Action, la Passion, le Lieu, le Temps, la Situation, la Manière d'être ou de posséder.
2. `http://faculty.washington.edu/smcohen/433/PorphyryTree.html`
3. *"Notes on the history of ontology"* `http://www.formalontology.it/history.htm`

Les travaux du Laboratoire des Systèmes de Connaissance (KSL) de Stanford, vers la fin des années 80, sont à l'origine de l'utilisation du mot *ontologie* dans le domaine de l'informatique , et aujourd'hui largement diffusé à travers son utilisation dans le Web sémantique. Leur but était d'apporter un moyen précis et formel pour définir les bases de connaissance, et de développer une méthodologie pour leur partage et leur réutilisation [Gruber, 1991, Gruber, 1993]. Ces travaux ont donné naissance au langage KIF (*Knowledge Interchange Format*) ainsi qu'au projet Ontolingua et le langage éponyme.

La première définition donnée dans le domaine de l'informatique provient du group sus-cité : *"Une ontologie définit les termes et les relations qui renferment le vocabulaire d'un domaine donné, ainsi que les règles utilisées pour les combiner pour définir des extensions au vocabulaire"* [Neches et al., 1991]. La définition la plus citée des ontologies est celle donnée par Gruber : *"une ontologie est une spécification explicite d'une conceptualisation"* [Gruber, 1993].

Actuellement, il n'existe pas une définition commune à la communauté scientifique. Parmi les différentes définitions données d'une ontologie, il semble qu'il y ait un accord sur le fait qu'une ontologie soit la donnée des concepts d'un domaine, des relations qui les structurent (Est_Un et autres types de relations), du vocabulaire utilisé pour en parler et de leurs définitions (formelles et/ou informelles). Les ontologies sont le support d'une représentation conceptuelle du monde ou d'un domaine particulier. Les relations hiérarchiques *Est_ Un* définissent la structure hiérarchique de l'ontologie (structure d'arbre).

Une liste plus complète des définitions données au terme ontologie peut être trouvée dans [Corcho et al., 2003]. Parmi les plus importantes sont celles basées sur la définition de Gruber. Borst a légèrement modifié cette dernière dans [Borst, 1997] : *"les ontologies sont définies comme une spécification formelle d'une conceptualisation partagée"*. Les définitions de Gruber et Borst ont été expliquées par Studer et ses collègues : *"La conceptualisation renvoie à un modèle abstrait d'un phénomène particulier dans la réalité, par l'identification des concepts pertinents de ce phénomène. "Explicite" veut dire que les types des concepts utilisés ainsi que les contraintes de leur utilisation sont clairement définis. "Formelle" veut dire que l'ontologie doit pouvoir être manipulée par un ordinateur. "Partagée" reflète la notion qu'une ontologie représente une connaissance consensuelle, c'est-à-dire, acceptée par une communauté"* [Studer et al., 1998]. En 1995, Guarino et ses collègues ont collecté et analysé sept définitions d'ontologie en leur apportant leurs interprétations syntaxique et sémantique [Guarino et al., 1995]. Dans cet article, les auteurs proposent de considérer une ontologie comme *"Une théorie logique qui donne une représentation partielle et explicite d'une conceptualisation"*, où *"La conceptualisation est essentiellement une idée du monde qu'une personne ou un groupe de personnes peut avoir"*. Guarino et ses collègues ont défini une méthode de construction d'ontologie en utilisant une théorie de la logique.

La question de *qu'est ce qu'une bonne ontologie* demeure l'objet d'un débat intense. Les auteurs de référence dans ce domaine sont B. Smith [1] et N. Guarino [2]. En France, les travaux sur les ontologies sont centrés autour des terminologies [Charlet, 2002], [Roche, 2005].

2.3 Ressources terminologiques et ontologiques en médecine

2.3.1 Ressources destinées aux professionnels de santé

Depuis la classification des espèces de Linné en 1735, il y a eu une explosion du nombre de classifications dans le domaine biomédical. Ces classifications ont été construites pour répondre

1. `http://ontology.buffalo.edu/smith`
2. `http://www.loa-cnr.it/guarino.html`

à des besoins divers selon des perspectives différentes (biologie, génétique, médecine,...). Par conséquent, il existe une multitude de classifications biomédicales répondant chacune à une catégorie de besoins, mais avec des structures différentes malgré le taux de recouvrement important entre elles.

Avec la prolifération des outils informatiques, le problème de la réutilisation et du partage des données codées avec différents systèmes de classification s'est accentué. Dans cette section nous présentons les principales classifications médicales actuelles et les solutions proposées pour remédier au problème de la réutilisation et du partage des données.

ICD

ICD (International Statistical Classification of Diseases and Related Health Problems), en français, la CIM (Classification Statistique Internationale des Maladies et des Problèmes de Santé Connexes) abrégée en *Classification Internationale des Maladies*, est publiée par l'Organisation Mondiale de la Santé et est utilisée à travers le monde pour enregistrer les causes de morbidité et de mortalité. La classification ICD permet le codage des maladies, des traumatismes et de l'ensemble des motifs de recours aux services de santé. L'objectif principal de cette classification est de *"permettre l'analyse systématique, l'interprétation et la comparaison des données de mortalité et de morbidité recueillies dans différents pays ou régions à des époques différentes"* [WHO, 1995].

SNOMED CT

La SNOMED CT (Systematized NOmenclature of MEDicine - Clinical Terms) [1] est une terminologie clinique qui regroupe la SNOMED Reference Terminology (SNOMED RT) et la Version 3 de la United Kingdom's Clinical Terms (anciennement connue sous le nom de *Read Codes*). Son objectif est de fournir un espace commun pour coder et partager les informations relatives à un patient (histoire, maladie, traitement, etc.). Elle est utilisée dans plusieurs tâches : l'indexation de documents cliniques, l'aide à la décision clinique, l'indexation d'images, etc. Une nouvelle organisation internationale, la IHTSDO (Health Terminology Standards Development Organisation) a acquis la propriété de la SNOMED CT depuis le 26 avril 2007 et va être responsable de son développement et de ses révisions. La terminologie est organisée en une hiérarchie de 18 catégories de premier niveau : les procédures, les entités observables, les structures du corps humain, les événements, ... Ces entités, dites majeures, sont regroupées autour d'une racine appelée Top. Dans sa version de Juillet 2006, SNOMED CT totalise plus de 300 000 concepts et 770 000 descriptions en anglais.

MeSH

Le MeSH est un thésaurus médical conçu par la NLM (National Library of Medicine) aux États-Unis. Il est utilisé en particulier par PubMed (l'interface de la NLM à la base de données MEDLINE [2]) pour l'indexation et la recherche de publications scientifiques. Il compte 24 767 descripteurs dans sa version de 2008. Les descripteurs MeSH sont organisés en 16 catégories : la catégorie A pour les termes anatomiques, la catégorie B pour les organismes, la catégorie C pour les maladies, etc. Chaque catégorie est subdivisée en sous-catégories. A l'intérieur de chaque catégorie, les descripteurs sont structurés hiérarchiquement, du plus général au plus spécifique.

1. `http://www.snomed.org/snomedct/index.html`
2. MEDLINE est une base de données bibliographiques qui couvre tous les domaines médicaux de l'année 1966 à nos jours : plus de 11 millions de références issues de 4 300 périodiques, principalement en langue anglaise.

Le MeSH est traduit en français par l'INSERM et sert aussi de thésaurus au site CISMeF et au site HON que nous allons présenter plus loin.

FMA

La FMA (Foundational Model of Anatomy) est une ontologie de référence dans le domaine de l'anatomie. Elle vise à représenter les entités anatomiques et les relations nécessaires pour la modélisation symbolique de la structure phénotypique du corps humain sous une forme qui soit compréhensible par l'homme et qui soit également traitable par une machine [Rosse & Mejino, 2003]. Les entités anatomiques sont représentées dans FMA, allant des macromolécules biologiques aux cellules, tissus, organes, systèmes d'organes, les majeures parties du corps, y compris le corps entier. Elle contient actuellement autour de 75 000 entités (concepts) anatomiques et plus de 120 000 termes.

DOLCE

L'ontologie DOLCE [1] (Descriptive Ontology for Linguistic and Cognitive Engineering), élaborée par l'équipe de Nicola Guarino (LOA, Trento, Italie), est une ontologie de haut niveau. Elle est l'un des résultats du projet européen WonderWeb Fundation Ontologies Library. Son objectif est d'être utilisée pour concevoir des ontologies de domaine plus spécifiques.

NCI Metathesaurus

Le métathésaurus du NCI (National Cancer Institute) résulte d'un projet de recherche interne au NCI appelé SIS (Science Information System) [NCI, 2006]. Ce projet avait pour objectif de fournir des outils de gestion de connaissances à la direction du NCI. Ce travail a révélé que disposer d'une terminologie du cancer était une condition *sine qua non* pour la réalisation des tâches liées à la gestion des connaissances. Le métathésaurus du NCI était construit pour répondre aux besoins du projet SIS, mais il a survécu à la fin du projet. Par la suite, le centre de la bioinformatique du NCI (NCICB) et l'office de communication du NCI (NCIBO) ont collaboré pour créer un ensemble de services et de ressources qui répondent aux besoins du NCI en terminologie du cancer. Ils ont ainsi initié le projet EVS (Entreprise Vocabulary Service) autour du métathésaurus du NCI. Les principaux résultats du projet EVS sont :

- **Le thésaurus du NCI** [2] **:** c'est une terminologie publique basée sur la logique des descriptions. Le thésaurus fournit une large description textuelle et ontologique de plus de 50 000 concepts.
- **Le métathésaurus du NCI** [3] **:** c'est une base de données de terminologies biomédicales. Il est basé sur le métathésaurus UMLS mais plus centré sur le cancer. Il contient plus de 60 terminologies distinctes et un nombre croissant de vocabulaires du cancer.
- **Infrastructure informatique** [4] **:** c'est un ensemble de logiciels pour créer, gérer et mettre à jour les données de la terminologie. Il garantit également l'accès à la terminologie à travers plusieurs serveurs.

1. `http://www.loa-cnr.it/DOLCE.html`
2. `http://nciterms.nci.nih.gov`
3. `http://ncimeta.nci.nih.gov`
4. `http://gforge.nci.nih.gov/softwaremap/trove_list.php?form_cat=434`

Gene Ontology

Le projet *Gene Ontology* (GO) [1] est le fruit d'une collaboration internationale qui vise à fournir des vocabulaires structurés pour l'annotation des gènes et leurs produits dans différents organismes. Le projet a débuté en 1998 avec une collaboration autour de trois bases de données d'*organismes modèles* : *FlyBase* (Drosophila), *the Saccharomyces Genome Database* (SGD) et *the Mouse Genome Database* (MGD). Depuis, le consortium GO s'est élargi pour regrouper plusieurs bases de données, dont plusieurs répertoires internationaux de gènes de plantes, d'animaux et micro-organismes.

Le projet GO a développé trois ontologies qui décrivent les produits des gènes de plusieurs espèces d'une manière indépendante en termes des processus biologiques associés, des composants cellulaires et des fonctions moléculaires. Le projet assure également trois tâches : premièrement, le développement et la maintenance des trois ontologies, deuxièmement, l'annotation des produits des gènes dans les bases de données associées, et troisièmement, le développement d'outils qui facilitent la création, la maintenance et l'utilisation des ontologies.

UMLS

Dans le but d'améliorer l'accès à l'information médicale à partir de différentes sources, la NLM (National Library of Medicine) a lancé, en 1986, le projet UMLS (Unified Medical Language System) [2]. Comme son nom l'indique, le projet vise à fournir un point d'entrée unique pour n'importe quelle entité médicale présente dans la littérature scientifique. La partie principale d'UMLS est un métathésaurus constitué d'environ 1 885 896 concepts (version 2008AB), et qui croît régulièrement de près de 200 000 concepts par an. Cette taille le rend difficile à gérer et il est presque impossible de garantir sa cohérence. Les concepts du métathésaurus sont principalement liés par des relations de type *Est_ Un*. Cependant, il existe d'autres types de relations comme *associated_ with* et *occurs_ in*. La version 2008AB contient actuellement 54 relations différentes. Chaque concept d'UMLS est relié aux *mêmes* concepts dans d'autres classifications. UMLS regroupe dans sa version actuelle 2009AA 126 classifications. Chaque concept UMLS a un identifiant unique, le CUI (Concept Unique Identifier). A chaque concept est associé un ensemble de termes dans différentes langues. La langue principale est l'anglais et les autres langues sont pour l'instant peu représentées, à l'exception de l'espagnol. Chaque CUI a dans chaque langue un terme préféré unique appelé SUI (String Unique Identifier). Chaque SUI est lié à un ou plusieurs termes selon ses différentes variations lexicales, qui sont les LUI (Lexique Unique Identifier).

OBO

Un des moyens pour réaliser l'intégration des données est l'utilisation de ressources terminologiques ou ontologiques pour annoter les données. L'absence d'une base théorique solide pour la construction et la formalisation des ontologies a amené à une prolifération d'ontologies, ce qui paradoxalement constitue un obstacle pour l'intégration. En 2001, *Ashburner* et *Lewis* ont créé le consortium OBO (Open Biomedical Ontologies) [3] pour définir des stratégies afin de remédier à ces problèmes. OBO est une organisation qui vise à regrouper les développeurs des ontologies biomédicales [Smith et al., 2007]. OBO a établi les principes de base pour la construction et le partage des ontologies *de bonne qualité*. Selon ces principes, les ontologies doivent

1. `http://www.geneontology.org`
2. `http://umlsks.nlm.nih.gov`
3. `http://obofoundry.org`

être ouvertes ; leur utilisation et celle des données qu'elles décrivent dans de nouvelles applications doivent être possibles sans aucune contrainte et sous aucune licence. Elles doivent être réceptives aux modifications résultant des débats des scientifiques. Elles doivent être orthogonales (hiérarchiques) afin de faciliter l'ajout des annotations et pour bénéficier des avantages de la programmation modulaire. Elles doivent être décrites dans une syntaxe formelle et bien définie, et enfin, elles doivent partager le même espace d'identificateurs. OBO regroupe actuellement plus de 60 ontologies et est soutenue par le NIH Roadmap National Center for Biomedical Ontology (NCBO) à travers son BioPortal [Rubin et al., 2006]. Un groupe de développeurs des ontologies OBO a initié la Fondation OBO, une expérience collaborative basée sur l'acceptation volontaire par ses participants d'un ensemble de principes évolutifs [1] qui s'ajoutent à ceux du consortium OBO initial en exigeant, en plus, que les ontologies (i) soient développées d'une manière collaborative, (ii) doivent utiliser des relations communes qui soient définies d'une manière non-ambiguë, (iii) doivent fournir des procédures pour collecter les retours des utilisateurs et pour identifier les différentes versions successives et (iv) doivent posséder une délimitation claire du domaine étudié (de sorte qu'une ontologie des composants cellulaires n'inclut pas des termes de type *base de données* ou *nombre entier*). Depuis la création de la fondation OBO, des ontologies comme GO et FMA ont été révisées [Rosse & Mejino, 2007], et d'autres ont été créées [Haendel et al., 2008, Leontis et al., 2006].

CISMeF

En France, l'équipe CISMeF du Centre Hospitalier Universitaire de Rouen a initié le projet CISMeF (Catalogue et Index des Sites MEdicaux Francophones) [2], qui a débuté en même temps que le site Web du CHU en février 1995. Ce catalogue indexe les principaux sites et documents francophones de qualité médicale contrôlée. En décembre 2007, il a dépassé les 41 300 ressources indexées avec une moyenne de 80 nouvelles ressources par semaine. Cette liste de sites contient un classement thématique, en particulier des spécialités médicales, un classement alphabétique, et un accès par type de ressources. Depuis juin 2000, l'outil associé, Doc'CISMeF, permet d'effectuer des recherches dans le catalogue de ressources, et offre des possibilités de recherches plus étendues [Darmoni et al., 2002]. CISMeF utilise deux outils standards pour organiser l'information : le thésaurus MeSH (Medical Subject Headings), utilisé notamment par la base de données bibliographique Medline, et le format de métadonnées du Dublin Core.

HONsélect

HONsélect [3] est un outil de recherche multilingue qui intègre des ressources Web hétérogènes. Il intègre des bases de données hétérogènes et d'autres ressources médicales disponibles sur le Web telles que le MeSH, MEDLINE pour les publications scientifiques, HONmédia pour les images et vidéos, NewsPage pour l'actualité quotidienne et MedHunt pour les ressources Web générales. Cet outil offre à l'utilisateur pour un terme médical donné une sélection de résultats dans chaque base de données mais également un choix de recherches intéressantes, déterminé automatiquement et avec des liens directs vers des articles scientifiques.

1. disponible à : `http://obofoundry.org`
2. `http://www.cismef.org`
3. `http://www.hon.ch/Project/HONselect_f.html`

2.3.2 Ressources destinées aux usagers de santé

Les usagers de santé utilisent généralement un sous-ensemble du vocabulaire du langage général pour décrire des concepts médicaux. Cependant, les termes empruntés au langage spécialisé sont parfois utilisés d'une manière peu précise, engendrant des lacunes de communication :

> Le vocabulaire médical et le vocabulaire familier diffèrent en contenu et en étendue. ... Les études médicales sont largement concernées par l'identification et la classification des processus biologiques et ceci nécessite inévitablement des vocabulaires spécialisés. ... le résultat est que la plupart de ce que disent les médecins est incompréhensible pour leurs patients. [Thompson, 1984, p. 120]

Les usagers de santé qui consacrent du temps pour s'informer sur un domaine médical particulier (ex. les patients ou leurs familles) et ceux exposés depuis une longue période à une maladie particulière peuvent constituer une exception et être considérés comme spécialistes mais dans un domaine très restreint.

Actuellement, il existe peu de listes de termes qui concernent les usagers de santé : *"A ce jour, les efforts pour organiser les mots... ont été basés sur les besoins des professionnels de santé, qui négligent les besoins des patients"* [Marshall, 2000, p. 1082]. Tony Tse a établi plusieurs facteurs qui peuvent expliquer le manque de recherche sur le vocabulaire des usagers de santé [Tse, 2003]. Les recherches existantes se sont concentrées sur l'évaluation de la compréhension des termes médicaux par des non-professionnels, comme montré dans la section 1.4. Les recherches précédentes étaient essentiellement focalisées sur *les termes* plutôt que sur *les personnes*. Les termes médicaux ont toujours été considérés comme professionnels par définition, et tout problème de communication entre le médecin et son patient était généralement considéré comme une conséquence de l'incapacité du patient à comprendre des concepts complexes [Segall & Roberts, 1980]. Un autre facteur est que l'on a toujours supposé que les personnes qui ont vraiment besoin d'information médicale vont apprendre les termes médicaux. Bien que cette supposition soit vraie pour quelques personnes, elle ignore les obstacles de compréhension tels que l'accessibilité à l'information, le niveau de culture sanitaire et les problèmes culturels.

Identifier le vocabulaire des usagers de santé est une étape essentielle dans la conception de systèmes d'aide à la recherche et à la compréhension de l'information médicale. Cette identification peut également contribuer à la compréhension des schémas conceptuels des usagers de santé afin d'améliorer les campagnes de prévention et les actions d'éducation des patients.

L'initiative CHV (Consumer Health Vocabulary), en français *Vocabulaire des Usagers de Santé* fait partie des initiatives les plus importantes de construction de terminologies destinées aux patients. Elle a été lancée conjointement par la Harvard Medical School et la National Library of Medicine [Zeng et al., 2001]. L'objectif de cette initiative, toujours en cours, est de construire une "première génération" Open Source d'un vocabulaire pour les usagers de santé, basé sur UMLS. Ils ont développé une application Web, appelée VocabTool, pour permettre la sélection et l'annotation des termes candidats extraits d'un corpus de textes [Crowell et al., 2005]. L'analyse contextuelle est utilisée pour discerner le sens des termes et pour atteindre un consensus parmi plusieurs relecteurs provenant de différentes disciplines. Ils ont analysé 12 millions de requêtes portant sur la santé et extrait 90 000 concepts, qui sont en cours d'analyse.

Une équipe de chercheurs en Allemagne a initié un projet pour créer une nouvelle base de données lexicales appelée MedicalWordNet (MWN) [Fellbaum et al., 2006]. Elle comprend les termes médicalement pertinents utilisés et intelligibles par des non-experts. Ces termes sont obtenus à partir d'un corpus de phrases en langage naturel qui est conçu pour fournir des contextes médicalement validés pour les termes du MWN. Le corpus provient essentiellement des sites d'information de santé destinés au grand public, et comprend deux sous-corpus appelés *Medical*

FactNet et *Medical BeliefNet*. Le premier se compose d'énoncés évalués comme corrects (*"statements accredited as true"*) sur la base d'un processus rigoureux de validation, tandis que le second est composé d'énoncés que les non-experts pensent corrects. Cette initiative n'a pas encore été testée sur des grands corpus et elle semble inactive depuis 2005.

Une autre expérience de construction d'un glossaire des termes médicaux scientifiques et populaires en 9 langues (anglais, français, allemand, néerlandais, espagnol, portugais, italien, grec et danois) a été menée en Belgique [1]. Ce travail a été motivé par la directive 92/27/EEC de la communauté européenne. Cette dernière a stipulé l'inclusion dans chaque paquet de médicaments d'une notice d'information complète écrite dans un langage compréhensible, dans les états membres de la communauté européenne à partir du janvier 1994. Ils ont construit un dictionnaire anglais de 1 830 termes médicaux fréquemment utilisés dans l'écriture des informations concernant les médicaments, et neuf glossaires de termes médicaux populaires et scientifiques, chacun dans l'une des neuf langues officielles de l'Union Européenne.

L'équipe du CISMeF a lancé l'initiative CISMeF-Patient qui est l'équivalent français de MEDLINEplus. Il permet d'interroger un catalogue de 4 974 ressources (au 25/06/2007) spécialement destinées au grand public : brochures, listes de diffusion, associations de patient. Plusieurs termes propres aux patients ont été pris en compte par le moteur de recherche Doc'CISMeF pour améliorer les performances de recherche des patients.

2.4 Formalismes pour la représentation des connaissances

Plusieurs travaux de recherche se sont intéressés à la représentation formelle des connaissances afin qu'une machine puisse les coder, les traiter et faire des opérations dessus (raisonnement, résolution de problèmes particuliers, mise à jour,...) [Sowa, 1984, Nardi & Brachman, 2002]. Les approches de représentation des connaissances développées aux alentours des années 70 sont généralement divisées en deux catégories : les formalismes basés sur la logique et ceux qui ne le sont pas. Les premiers sont nés de l'intuition que le calcul des prédicats peut être utilisé d'une manière non-ambiguë pour représenter des faits du monde réel. Les autres sont basés sur des notions plus cognitives, par exemple, les réseaux sémantiques et les représentations à base de règles, inspirées des expériences sur le fonctionnement de la mémoire humaine et l'exécution de certaines tâches comme la résolution mathématique d'un puzzle. Dans cette section, nous allons introduire brièvement deux formalismes très utilisés pour la représentation des connaissances : les graphes conceptuels et les logiques de description.

2.4.1 Les graphes conceptuels

Les graphes conceptuels sont un formalisme de représentation des connaissances qui appartient à la famille des réseaux sémantiques (détaillée au chapitre 6). Ils ont été introduits par *John F. Sowa* sur la base des travaux de *Charles Sanders Peirce*. Les aspects formels des graphes conceptuels sont inspirés de deux bases mathématiques, la logique du premier ordre et la théorie des graphes. Les graphes conceptuels sont utilisés dans plusieurs domaines, tels que le traitement automatique des langues, les systèmes à base de connaissances, l'ingénierie des connaissances et la construction des bases de données, ...

Un graphe conceptuel est un graphe ou un réseau avec deux types de nœuds : les concepts et les relations. Les nœuds sont connectés par des arcs qui sont toujours dirigés. Un arc entrant

1. `http://users.ugent.be/~rvdstich/eugloss/information.html`

relie un concept à une relation conceptuelle, et un arc sortant relie une relation conceptuelle à un concept.

[CONCEPT]→(RELATION)→[CONCEPT]

2.4.2 Les logiques de description

Les logiques de description (LD) sont le nom le plus récent [1] donné à une famille de formalismes de représentation des connaissances qui représente les connaissances d'un domaine d'application par la définition en premier lieu des concepts pertinents de ce domaine, et par la suite l'utilisation de ces concepts pour spécifier les propriétés des objets et des individus qui font partie du domaine. Comme l'indique son nom, une des caractéristiques des logiques de description, qui les distingue de leurs prédécesseurs, est leur dotation d'une sémantique formelle basée sur la logique. Une autre caractéristique est l'accent mis sur le raisonnement, qui est considéré comme un service central pour ce type de formalismes. Le raisonnement permet de déduire des connaissances implicites à partir de connaissances explicitement présentes dans la base de connaissances [Baader & Nutt, 2002]. Les logiques de description exploitent, en général, des sous-ensembles décidables de la logique du premier ordre.

Les logiques de description ont été créées dans le but de surmonter les ambiguïtés des anciens systèmes à base de réseaux sémantiques et de frames. Les premiers systèmes développés à partir de KL-ONE, lui-même jamais implémenté à cause de sa complexité, sont NACK, LOOM et CLASSIC [Brachman & Schmolze, 1985, Baader et al., 2003]. Ces premiers systèmes étaient basés sur une catégorie d'algorithmes de vérification de subsomption de type normalisation/-comparaison qui avaient comme inconvénient leur manque d'expressivité, un temps de calcul souvent polynomial et une indécidabilité dans certains cas. Dans les années 90, une nouvelle classe d'algorithmes apparaît : les algorithmes de vérification de satisfiabilité à base de tableaux. Ces algorithmes raisonnent sur des logiques de description dites expressives, mais en temps exponentiel. Cette nouvelle famille d'algorithmes a montré un potentiel très intéressant et a été utilisée dans de nouvelles applications telles que le Web sémantique.

Un système de représentation de connaissances à base de logiques de description fournit des moyens pour construire des bases de connaissance, effectuer des raisonnements sur leur contenu et les manipuler. Une base de connaissance comprend deux parties, la *TBox* et la *ABox* [Baader & Nutt, 2002] :

- La TBox introduit la *terminologie*, i.e., le vocabulaire du domaine de l'application, décrit les connaissances générales d'un domaine et contient les déclarations des primitives conceptuelles organisées en concepts et rôles (relations).
- La ABox contient des *assertions* sur les individus et les instances de concepts. Plusieurs ABox peuvent être associées à une même TBox ; chacune représente une configuration constituée d'individus, et utilise les concepts et les rôles de la TBox pour l'exprimer.

Un système à base de LD ne regroupe pas seulement les terminologies et les assertions, mais offre aussi des services pour raisonner. Les tâches du raisonnement appliquées à une terminologie ont pour objectif de déterminer si une description est *satisfiable* (i.e., non-contradictoire), ou si une description est plus générale qu'une autre, c'est-à-dire, si la première *subsume* la deuxième. Les problèmes importants d'une ABox sont de trouver si l'ensemble de ses assertions est *consistant*, et si l'ensemble des assertions de la ABox implique qu'un individu particulier est une *instance* d'un concept donné. Les tests de satisfiabilité des descriptions et de consistance des

1. Précédemment les noms utilisés sont les langages terminologiques de représentation des connaissances, les langages conceptuels, les langages de subsomption des termes, et les langages KL-ONE-LIKE

ensembles d'assertions sont nécessaires pour déterminer si une base de connaissance est cohérente ou pas. Avec les tests de subsomption, on peut organiser les concepts d'un domaine en une hiérarchie selon leur degré de généralité. Une description de concept peut être également conçue comme une requête décrivant l'ensemble des objets intéressants pour une tâche particulière. Par conséquent, avec les tests d'instances, on peut retrouver les individus qui satisfont une requête.

2.5 Méthodes de construction d'ontologies

Peu de groupes de recherche proposent des méthodologies pour la construction des ontologies. Cependant, l'ingénierie ontologique étant une discipline relativement immature, chaque groupe de travail emploie sa propre méthodologie. Dans cette section, nous nous sommes basés sur l'article de [Lopez, 1999] pour présenter les méthodologies les plus importantes de construction d'ontologies.

2.5.1 La méthodologie de Uschold et King

Cette méthodologie est basée sur l'expérience de développement de l'*Enterprise Ontology*, une ontologie pour la modélisation des procédés de l'entreprise [Uschold & King, 1995]. Cette méthodologie fournit des principes pour développer des ontologies, qui sont :

1. **Identification de l'objectif**. Définir clairement l'objectif de l'ontologie et les scénarios de son utilisation.
2. **Construction de l'ontologie**, qui est divisée en plusieurs étapes :
   - Identification des concepts et des relations clés du domaine concerné.
   - Production de définitions précises non ambiguës de ces concepts et de ces relations.
   - Identification des termes qui désignent ces concepts et ces relations.
   - Codage de l'ontologie en la représentant explicitement dans un langage formel.
   - Intégration des ontologies existantes.
3. **Evaluation de l'ontologie**. L'évaluation a pour objectif de produire des jugements techniques sur l'ontologie, son environnement logiciel et la documentation produite, en suivant un cahier de spécifications, des questions de compétence ou le monde réel.
4. **Documentation de l'ontologie**, selon des protocoles préalablement établis, qui peuvent différer selon le type et l'objectif de l'ontologie.

2.5.2 La méthodologie de Grüninger and Fox

Cette méthodologie est basée sur l'expérience de construction de l'ontologie du projet TOVE [Gruninger & Fox, 1994] dans le domaine des procédés commerciaux et la modélisation des activités de l'entreprise. Il s'agit essentiellement de construire un modèle logique de la connaissance au moyen de l'ontologie. Ce modèle n'est pas construit directement. Premièrement, une description informelle est établie sur la base des spécifications de l'ontologie à construire. Par la suite, cette description est formalisée. Les étapes proposées sont les suivantes :

1. **Spécification des scénarios**. Selon les auteurs, le développement d'une ontologie est motivé par des scénarios d'usage qui constituent des problèmes ou des exemples d'usage auxquels l'application qui va utiliser l'ontologie doit répondre. Ces scénarios peuvent donner une description informelle de la sémantique des objets et des relations à inclure dans l'ontologie.

2. **Formulation informelle des questions de compétence**. L'ontologie doit être capable par la suite de représenter ces questions au moyen de sa terminologie, et être également capable d'y répondre en utilisant des axiomes et des définitions.
3. **Spécification de la terminologie de l'ontologie dans un langage formel**. Au départ, les questions de compétences précédemment identifiées vont déterminer une liste de termes qui va être utilisée pour spécifier une terminologie dans un langage formel.
4. **Formulation formelle des questions de compétence en utilisant la terminologie de l'ontologie**.
5. **Spécification des axiomes et des définitions de termes de l'ontologie dans un langage formel**. Si les axiomes proposés sont insuffisants pour représenter les questions de compétences et caractériser des solutions à ces questions, un processus itératif d'ajout d'axiomes et de définitions est déclenché jusqu'à atteinte de l'objectif fixé.

2.5.3 METHONTOLOGY

Cette méthodologie a été développée au laboratoire d'Intelligence Artificielle de l'Université Polytechnique de Madrid [Fernandez et al., 1997, Fernandez-Lopez et al., 1999]. Cette méthodologie intègre la construction d'ontologies dans un processus de gestion de projet complet, comprenant aussi bien les étapes de spécification des besoins et de planification que celles, par exemple, de la réalisation, de la maintenance et de la documentation. La construction d'une ontologie à l'aide de cette méthodologie est divisée en dix étapes :

1. collecter l'ensemble des termes qui seront inclus dans l'ontologie, préciser leur définition en langage naturel, identifier leurs synonymes et leurs acronymes ;
2. construire des taxonomies de concepts pour les classer ;
3. identifier les différentes relations binaires entre les concepts à inclure dans l'ontologie ;
4. construire le dictionnaire de concepts, qui inclut, pour chaque concept, ses attributs d'instance, ses attributs de classe et ses relations ;
5. décrire en détail chaque relation binaire qui apparaît dans le dictionnaire de concepts ;
6. décrire en détail chaque attribut d'instance qui apparaît dans le dictionnaire de concepts ;
7. décrire en détail chaque attribut de classe qui apparaît dans le dictionnaire de concepts ;
8. décrire en détail chaque constante (les constantes donnent des informations sur le domaine de connaissances) ;
9. décrire les axiomes formels ;
10. décrire les règles utilisées pour contraindre le contrôle et pour inférer des valeurs aux attributs.

Des outils tels que WebODE [Arpirez et al., 2001] et ODE [Blazquez et al., 1998] fournissent des environnements qui implantent cette méthodologie.

2.5.4 ARCHONTE

B. Bachimont s'est basé sur la sémantique différentielle pour proposer la méthodologie ARCHONTE (ARCHitecture for ONTological Elaborating) [Bachimont et al., 2002]. Selon cette méthodologie, la construction d'une ontologie passe par trois étapes principales :

1. choisir les termes pertinents du domaine et normaliser leurs sens puis justifier la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père ;
2. formaliser les connaissances, ce qui implique par exemple d'ajouter des propriétés à des concepts, des axiomes, de contraindre les domaines d'une relation,etc. ;
3. représenter l'ontologie dans un langage formel de représentation des connaissances.

2.6 Langages pour exploiter les ontologies

Plusieurs langages, basés sur la syntaxe de XML, ont été développés pour l'utilisation des ontologies dans le cadre du Web sémantique. Cette section introduit les langages les plus importants : RDF/RDF(S), OIL et ses successeurs DAML+OIL et OWL.

2.6.1 RDF et RDF(S)

RDF (Resource Description Framework), développé par le W3C (the World Wide Web Consortium) pour décrire des ressources du Web, est un modèle de graphes pour décrire les (méta-)données en permettant leur traitement automatisé [Lassila & Swick, 1999]. Le modèle de données de RDF est équivalent au formalisme des réseaux sémantiques. Il est constitué de trois types d'objets : *les ressources* sont décrites par des expressions RDF et sont toujours identifiées par des URIs (Uniform Resource Identifier) ; *les propriétés* définissent des aspects spécifiques, des caractéristiques, des attributs ou des relations utilisés pour décrire une ressource ; et enfin *les déclarations* attribuent une valeur à une propriété d'une ressource spécifique (cette valeur peut être un autre élément RDF).

Le modèle de données RDF n'offre pas de mécanismes pour définir les relations entre les propriétés (attributs) et les ressources. RDF(S) fournit un ensemble de primitives simples, mais puissantes, pour la structuration de la connaissance d'un domaine en classes et sous-classes, propriétés et sous-propriétés, avec la possibilité de restreindre leur domaine d'origine (rdf :domain) et leur domaine d'arrivée (rdf :range).

RDF(S) est indiqué pour la description de ressources ; cependant il présente assez rapidement des limites lorsqu'il s'agit de son utilisation comme langage de représentation d'ontologies ayant de fortes contraintes [McBride, 2004].

2.6.2 OIL

OIL (Ontology Interchange Language), développé dans le cadre du projet OntoKnowledge [1], permet l'interopérabilité sémantique entre les ressources Web. Sa syntaxe et sa sémantique sont basées sur des propositions existantes (OKBC, XOL et RDF(S)), fournissant des primitives de modélisation comme celles utilisées dans des approches basées sur les frames et l'ingénierie ontologique (les concepts, les taxonomies de concepts, les relations,...), des sémantiques formelles et des procédures de raisonnement inspirées des approches des LD. OIL possèdent les couches suivantes : *le noyau OIL* (Core OIL) qui regroupe les primitives OIL qui possèdent une correspondance directe avec les primitives de RDF(S) ; *OIL standard* (standard OIL) est le modèle complet de OIL qui utilise un nombre plus grand de primitives que celles définies dans RDF(S) ; *les instances OIL* (OIL instance) ajoutent au modèle précédent des instances de concepts et de rôles ; et *OIL "lourd"* (heavy OIL) est la couche qui contient les extensions futures de OIL.

1. www.ontoknowledge.org/OIL

2.6.3 DAML+OIL

DAML+OIL (DARPA Agent Markup Language + OIL) a été développé par un effort commun entre les États-Unis et l'EUROPE (IST) dans le cadre de DAML, un projet de DARPA qui a pour objectif de permettre l'interopérabilité sémantique en XML [Horrocks & van Harmelen, 2001]. Par conséquent, DAML+OIL partage le même objectif que OIL. DAML+OIL est construit sur la base de RDF(S) et OIL. Il possède plus de capacités de raisonnement. OILEd, OntoEdit, Protégé, et WebODE sont des outils qui permettent d'éditer des ontologies en DAML+OIL.

2.6.4 OWL

En 2001, le W3C a formé un groupe de travail appelé *Web-Ontology* (WebOnt). L'objectif de ce groupe était de définir un nouveau langage d'ontologie pour le Web sémantique. OWL (Web Ontology Langage) a été le résultat des travaux de ce groupe. La sémantique du langage peut être définie via une transformation en logiques de description, dont ce langage est inspiré. On peut utiliser les outils de raisonnement développés pour les LD afin d'effectuer des inférences sur OWL. OWL est devenu une recommandation (i.e., un standard) du W3C en 2004 [McGuinness & van Harmelen, 2004].

Une particularité de OWL est qu'il est divisé en trois sous-langages avec des expressivités ascendantes et compatibles : OWL Lite, OWL DL et OWL Full, destinés à différents groupes d'utilisateurs :

OWL Lite est destiné aux utilisateurs qui ont besoin d'exprimer une classification sous une forme hiérarchique et des contraintes simples (contraintes de cardinalité de type 0 ou 1). Ce langage est particulièrement indiqué pour représenter des taxonomies.

OWL DL est destiné aux utilisateurs qui désirent une expressivité maximale sans perte de complétude computationnelle (garantie de calcul de tous les axiomes) et de décidabilité (tous les calculs s'achèvent en un temps fini) des systèmes de raisonnement. OWL DL est basé sur les logiques de description. OWL DL ajoute des restrictions telles que la séparation des types (une classe ne peut pas être à la fois un individu ou une propriété et une propriété ne peut pas être à la fois une classe ou un individu). OWL DL a l'avantage de fournir un support pour les inférences, mais la compatibilité totale avec RDF/RDF(S) est perdue.

OWL Full est destiné aux utilisateurs qui désirent des capacités d'expressivité maximale. Il contient tous les éléments disponibles de OWL et il permet de les combiner avec RDF et RDF Schema. OWL Full est totalement compatible avec RDF. Sa grande expressivité le rend indécidable.

2.6.5 SKOS

SKOS (Simple Knowledge Organisation System) [1] désigne une famille de langages formels permettant une représentation standard des thésaurus, classifications ou tout autre type de vocabulaire contrôlé et structuré. SKOS est construit sur la base du langage RDF, dont il est une application, et son principal objectif est de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du Web sémantique. SKOS est actuellement développé dans le cadre du W3C et n'a pas encore le statut de *Recommendation Candidate* [2].

Les différents composants de SKOS sont :

1. `http://www.w3.org/2004/02/skos/`
2. `http://fr.wikipedia.org/wiki/SKOS`

1. **le noyau SKOS** (SKOS Core), définit les classes et les propriétés suffisantes pour la représentation standard des thésaurus. Les objets primitifs dans SKOS ne sont pas des termes, mais des concepts abstraits dont les termes sont des propriétés. Un concept SKOS est défini comme une ressource RDF et identifié par une URI. Chaque concept peut avoir des propriétés RDF, comme par exemple un terme préférentiel par langue, des synonymes avec une spécification possible de la langue, des définitions, des concepts reliés, etc. ;
2. **le SKOS Mapping**, qui définit un vocabulaire pour exprimer des correspondances (alignements exacts ou correspondances approximatives) entre concepts provenant de schémas différents ;
3. **les extensions de SKOS**, qui devront permettre de déclarer des relations entre concepts avec une sémantique plus précise que les relations définies dans SKOS Core, par exemple des relations tout-partie ou classe-instance, ou de préciser certains attributs (note éditoriale, note historique, acronyme, etc.).

2.7 Outils de développement des ontologies

Ces dernière années, il y a eu un grand nombre d'environnements et d'outils de construction d'ontologies. Ces outils visent à fournir un support pour le processus de développement des ontologies et pour leurs usages par la suite. Dans cette section, nous présentons brièvement les environnements les plus importants de développement d'ontologies.

L'environnement *Ontolingua Server* a été le premier outil de développement d'ontologies créé [Farquhar et al., 1996]. Il a été développé au Laboratoire KSL (Knowledge Systems Laboratory) à l'Université de Stanford. Cet environnement a été construit pour la première fois au début des années 90 dans le but de faciliter le développement des ontologies Ontolingua dans le cadre d'une application Web. Initialement, le module principal de l'outil était un éditeur d'ontologies, ensuite d'autres modules ont été ajoutés à l'environnement, tels que *Webster* un résolveur d'équation, *Chimaera* un outil de fusion d'ontologies, etc. L'éditeur d'ontologies fournit aussi des traducteurs pour d'autres langages, tels que LOOM, Prolog, CLIPS, etc.

A la même période, *OntoSaurus* a été développé par l'Institut des Sciences de l'Information (ISI) à l'Université de South California [Swartout et al., 1997]. OntoSaurus est constitué de deux modules : un serveur d'ontologie, qui utilise LOOM comme système de représentation des connaissances, et un navigateur Web pour les ontologies LOOM. Des traducteurs de LOOM vers Ontolingua, KIF, KRSS et C++ sont également disponibles.

En 1997, le Knowledge Media Institute (KMI) de l'Open University a développé *Tadzebao* et *WebOnto* [Domingue, 1998]. WebOnto est un éditeur d'ontologies en OCML (Operational Conceptual Modelling Language). Son principal avantage par rapport aux autres outils est qu'il permet l'édition collaborative d'ontologies avec un module de discussion synchrone ou assynchrone entre les différents développeurs d'ontologies.

La principale similarité entre les outils mentionnés précédemment est qu'ils ont une relation forte avec un langage spécifique (Ontolingua, LOOM et OCML respectivement). Ils étaient créés dès le départ pour faciliter la création, la navigation et l'édition des ontologies dans ces langages. En plus, ils étaient uniquement destinés à des activités de recherche et la majorité d'eux ont été développés comme des outils isolés qui n'offraient pas beaucoup de possibilités d'extension.

Au cours des dernières années, une nouvelle génération d'environnements d'ingénierie ontologique a vu le jour. Le critère de construction de ces environnements est beaucoup plus ambitieux que celui des outils précédents. Ils ont été créés pour intégrer des ontologies dans des systèmes d'information effectifs. Ils constituent soient des environnements intégrés robustes ou des modules

qui fournissent un support technologique aux différentes activités d'un cycle de vie d'ontologie. Ils sont extensibles et possèdent une architecture modulaire où de nouveaux modules peuvent être facilement intégrés. Les modèles de connaissance dans ces environnements sont indépendants des langages ontologiques. Parmi ces environnements, on peut citer *Protégé*, *WebODE* et *OntoEdit*.

Protégé a été développé par le Stanford Medical Informatics (SMI) de l'Université de Stanford. C'est une application open source, standalone (i.e., une application à part entière) avec une architecture extensible. Le noyau de cette application est un éditeur d'ontologies, qui contient une librairie de plugins qui ajoute des fonctionnalités supplémentaires à l'environnement. Actuellement, des plugins sont disponibles pour l'importation/exportation de langages ontologiques (FLogic, Jess, OIL, XML, Prolog, RDF(S), OWL, XML Schema), l'accès OKBC (Open Knowledge Based Connectivity), la création et l'exécution de contraintes (PAL), la fusion des ontologies (PROMPT), etc. Protégé continue à s'enrichir régulièrement de l'apport de la communauté des utilisateurs et des développeurs.

WebODE [Arpirez et al., 2001] est le successeur de *ODE* (Ontology Design Environment) [Blazquez et al., 1998], et a été développé au Laboratoire d'Intelligence Artificielle de L'Université Technique de Madrid (UPM). Il est également caractérisé par son architecture extensible, mais il n'est pas utilisé comme une application standalone, mais comme un serveur Web avec une interface Web. Le noyau de cet environnement est le service d'accès à l'ontologie, qui est utilisé par tous les autres services et applications ajoutés au serveur. Il existe plusieurs services pour : l'importation/exportation de langages ontologiques (XML, RDF(S), OIL, DAML+OIL, CARIN, FLogic, Jess, Prolog), l'édition d'axiomes avec *WebODE Axiom Builder* (WAB) [Corcho et al., 2002], la documentation, l'évaluation et la fusion d'ontologies. Les ontologies WebODE sont stockées dans une base de données relationnelle. Finalement, WebODE couvre la majorité des activités du processus de développement d'ontologies proposé par METHONTOLOGY.

OntoEdit [Sure et al., 2002b] a été développé au Knowledge Management Group (AIFB) de l'Université de Karlsruhe. Il est similaire aux outils précédents : c'est un environnement extensible et flexible, basé sur une architecture modulaire à base de plugins, qui fournit des fonctionnalités pour l'édition et la navigation dans des ontologies. Il inclut des plugins pour réaliser des inférences en utilisant *Ontobroker* [Decker et al., 1999], des importations et des exportations d'ontologies dans différents formats (FLogic, XML, RDF(S), DAML+OIL), etc. Deux versions d'OntoEdit sont disponibles : OntoEdit Free et OntoEdit Professionnel. Récemment, l'outil KAON (Karlsruhe Ontology) a été développé comme successeur d'OntoEdit.

Finalement, avec l'émergence du Web sémantique, les outils pour le développement des ontologies en DAML+OIL et RDF(S) ont proliféré. Les outils mentionnés précédemment permettent l'importation et l'exportation des ontologies en DAML+OIL et RDF(S). Il existe aussi plusieurs outils isolés qui permettent la création des ontologies DAML+OIL ; les plus représentatifs sont : OILEd (un outil basé sur les LD) et DUET (un plugin pour Rational Rose basé sur UML).

OILEd [Bechhofer et al., 2001] a été initialement développé comme un éditeur d'ontologies OIL, dans le cadre du projet européen *IST On-To-Knowledge*. Cependant, OILEd a évolué et il est devenu un éditeur d'ontologies DAML+OIL. Les utilisateurs de OILEd peuvent se connecter à FaCT, un moteur d'inférence qui fournit des fonctionnalités de tests de consistance et de classification automatique de concepts. OILEd fournit également plusieurs options de documentation (HTML, visualisation graphique d'ontologies, etc.).

DUET [Kogut et al., 2002] a été développé par *AT&T Government Solutions Advanced Systems Group*. Il offre un environnement de visualisation UML et d'édition pour le langage DAML+OIL, qui est intégré comme un plugin à la suite Rational Rose. Cet outil n'est pas destiné aux ingénieurs de connaissance mais pour les concepteurs de bases de données et de

systèmes d'information, qui peuvent modéliser leurs ontologies en UML et les traduire ensuite en DAML+OIL. L'outil *VOM* (Medius Visual Ontology Modeller) a été également créé comme plugin à la suite Rational Rose [Kendall et al., 2002].

Plusieurs autres outils existent, avec des objectifs différents : des outils spécialisés dans la fusion d'ontologies (Chimaera, Protégé-PROMPT), la traduction des ontologies dans différents formats (Ontomorph), l'annotation à base d'ontologies de pages Web (COHSE, OntoMat, SHOE Knowledge Annotator), l'évaluation des ontologies (OntoAnalyser, ONE-T, ODEClean), etc.

2.8 Modélisation des connaissances à partir de textes

Une des approches existantes de conception d'ontologies consiste en la construction d'ontologies à partir de textes. Des travaux ont déjà été réalisés dans ce sens [Aussenac-Gilles et al., 2003]. L'hypothèse sous-jacente est que les connaissances, et donc les concepts permettant de modéliser un domaine donné, sont contenus dans des corpus de textes représentatifs du domaine. Bachimont [Bachimont, 2000] adopte ce point de vue : *"le travail de modélisation doit s'effectuer à partir de documents attestés dans la pratique d'un domaine et rassemblés en un corpus"*.

Comme mentionné précédemment, il existe une multitude d'outils d'édition d'ontologies : le sondage de *XML.COM* des éditeurs d'ontologies [1] en a recensé 94 en 2004. Cependant, seuls 25 outils d'entre eux offrent un support lexical, i.e., fournir plusieurs lexicalisations d'un concept en différents langages, et seuls 9 d'entre eux permettent l'extraction d'information à partir de textes. En plus, plusieurs de ces outils ne sont pas disponibles gratuitement et, par conséquent, leur capacité d'extraire de l'information à partir de textes ne peut être testée. Dans ce qui suit, nous décrivons deux éditeurs d'ontologies qui incluent des fonctionnalités de modélisation de connaissances à partir de textes : KAON et Terminae.

2.8.1 Text-To-Onto et KAON

KAON (Karlsuhe Ontology and Semantic Web) est un environnement Open Source modulaire, basé sur Java, destiné à la conception, au développement et à la gestion d'ontologies. Il a été développé au Knowledge Management Group (AIFB) de l'Université de Karlsruhe [2]. Ce logiciel permet la construction d'ontologies à partir de textes, grâce au module *Text-To-Onto* qui inclut l'extraction de termes, l'extraction d'associations de concepts et le nettoyage de l'ontologie (i.e., supprimer les concepts candidats non pertinents). L'extraction des termes produit une liste de termes simples et complexes avec leurs fréquences et plusieurs autres indicateurs statistiques.

2.8.2 TERMINAE

Terminae est un outil destiné à aider les utilisateurs dans la construction d'ontologies et de terminologies à partir de corpus de textes. Il a été développé depuis 1997 par *B. Biébow* et *S. Szulman* au LIPN de l'Université Paris-Nord II [Aussenac-Gilles et al., 2003]. Il utilise les résultats de plusieurs outils d'ingénierie linguistique :

- LEXTER : extraction des termes candidats.
- SYNTEX : extraction d'une liste de noms et syntagmes nominaux, structurée par des relations de dépendance syntaxique.
- LINGUAE : extraction de patterns récurrents de termes.

1. `http://www.xml.com/pub/a/2004/07/14/onto.html`
2. `http://kaon.semanticweb.org/`

En fonction des résultats donnés par ces outils, les termes candidats pertinents peuvent être sélectionnés et décrits. Dans l'étape suivante, le sens des termes est normalisé et spécifié ; les occurrences des termes candidats dans le corpus peuvent être consultées pour vérifier s'ils sont polysémiques et pour préciser leurs définitions. Finalement, les concepts peuvent être formalisés et insérés dans l'ontologie.

2.9 Conclusion

Dans ce chapitre nous avons fait le tour des notions les plus importantes autour des ontologies. Cette vue d'ensemble a dévoilé une dichotomie majeure. D'une part, les ontologies doivent permettre de résoudre les problèmes d'hétérogénéité des données liés à l'existence de plusieurs langues et plusieurs cultures. D'autre part, elles se basent sur quelques suppositions théoriques qui sont incompatibles avec ce point de vue, spécialement la notion du consensus et l'universalité des concepts qu'elles définissent. Néanmoins, les ontologies jouent un rôle très important dans le développement du Web sémantique. Elles sont utilisées dans plusieurs domaines tels que la RI afin de pallier les insuffisances des modèles basés sur les mots-clés, le commerce électronique (où elles offrent un vocabulaire partagé) et la gestion des connaissances (accès intelligent à l'information, partage de connaissances, etc.).

Chapitre 3

Recherche d'information, concepts de base et principaux modèles

3.1 Introduction

Dans ce chapitre nous introduisons les concepts de base de la recherche d'information (RI) et les principaux modèles utilisés par les systèmes de recherche d'information (SRI).

Plusieurs définitions ont été données à la RI. Officiellement, c'est toute action, méthode ou procédure ayant pour objet d'extraire d'un ensemble de documents les informations voulues (d'après l'AFNOR, 1979). Dans un sens plus large, toute opération (ou ensemble d'opérations) ayant pour objet la recherche, la collecte et l'exploitation d'informations en réponse à une question sur un sujet précis [1]. Les SRI servent d'interface entre une source contenant des quantités considérables de documents et des utilisateurs cherchant, via des requêtes, des informations susceptibles de se trouver dans cette collection. Les SRI intègrent un ensemble de techniques permettant de sélectionner ces informations. Elles peuvent être résumées en quatre fonctions, qui sont le stockage de l'information, l'organisation de ces informations, la recherche d'informations en réponse à des requêtes utilisateurs et la restitution des informations pertinentes pour ces requêtes. La dernière fonction est celle qui est visible pour l'utilisateur [Baziz, 2005].

Avec l'invention des ordinateurs, le besoin de stocker et de trouver de grandes quantités d'information a été vite ressenti. En 1945, *Vannevar Bush* a publié un article très novateur *"As We May Think"* qui a donné naissance à l'idée d'un accès automatique à une large quantité de connaissances [Bush, 1945]. Dans les années 50, inspirés de cette idée, plusieurs travaux sur la recherche automatique de documents ont vu le jour. Une des méthodes des plus remarquables à cette époque et qui a marqué la naissance de la recherche d'information est celle de *H.P. Luhn*, en 1957, qui a proposé le système d'indexation KWIC, qui sélectionnait les index selon la fréquence des mots dans les documents et filtrait des mots vides de sens en employant des listes de mots "vides", comme "le", "la", "un", ... [Luhn, 1957].

Dans les années 60, plusieurs avancées dans le domaine de la RI ont été réalisées. Les plus notables sont le développement du système SMART par *Gerard Salton* et ses étudiants [Salton, 1971], et les évaluations *Cranfield* réalisées par *Cyril Cleverdon* et son groupe [Cleverdon, 1967]. Les tests de Cranfield font partie d'une méthodologie d'évaluation des SRI qui est toujours utilisée aujourd'hui. Le système SMART couplé avec cette méthode d'évaluation a permis aux chercheurs de réaliser plusieurs expérimentations et depuis, les progrès dans la RI ont été très rapides [Singhal, 2001].

1. `http://www.uhb.fr/urfist/Supports/RechInfoInit/RechInfo3Problematique.html`

Les années 70 et 80 ont connu plusieurs avancées basées sur les développements des années 60. Plusieurs modèles de RI ont vu le jour. Ces nouveaux modèles et techniques se sont avérés efficaces sur des petites collections de textes (plusieurs milliers de documents), les seules disponibles à l'époque pour les chercheurs. Cependant à cause de ce fait, la question de savoir si ces modèles resteront performants sur des larges collections de documents demeure sans réponse. En 1992, la création de la conférence de recherche de textes TREC (Text Retrieval Conference), une série de conférences d'évaluation des SRI, soutenue conjointement par le National Institute of Standards and Technology (NIST) et par l'Advanced Research and Development Activity (ARDA) Center du Département de la Défense des États-Unis, a remédié à ce problème. Le but de ce programme est d'encourager les travaux dans le domaine de la recherche d'information en fournissant l'infrastructure nécessaire à une évaluation objective à grande échelle des méthodologies de recherche textuelle et l'accroissement de la rapidité du transfert des technologies.

Avec les grandes collections de textes diponibles sous TREC, plusieurs techniques anciennes ont été modifiées et plusieurs nouvelles sont apparues (et continuent d'apparaître), qui visent à améliorer la RI sur de large collection de documents. TREC s'est aussi divisée en plusieurs branches de la RI, comme la recherche sur la parole, la recherche d'information dans d'autres langues que l'anglais, le filtrage de l'information, l'interaction utilisateur-SRI, etc. Les algorithmes développés en RI ont été les premiers utilisés pour faire des recherches sur le *World Wide Web* entre 1996 et 1998. La recherche sur le Web s'est développée par la suite pour des systèmes qui prennent en compte des liens croisés entre les différentes pages Web.

3.2 Concepts de base de la Recherche d'Information

Collection de documents : La collection de documents est l'ensemble de documents exploitables et accessibles par le SRI. Le Web constitue actuellement la plus grande collection de documents mais reste partiellement exploitée par les moteurs de recherche.

Document : Le document représente une unité (item) de la collection. Il peut s'agir d'un document entier ou d'une partie de ce document. Le document peut être de plusieurs types : textuel, multimédia, vidéo, etc. Dans ce mémoire, un document représente des informations textuelles.

Besoin d'information : Le besoin d'information représente le besoin de l'utilisateur en information. Peter Ingwersen a classé le besoin d'information en trois catégories : i) le besoin vérificatif (l'utilisateur possède les données qui lui permettent d'accéder à l'information, ce besoin est stable et ne change pas avec le temps), ii) le besoin thématique connu (l'utilisateur cherche de nouvelles informations sur un domaine ou une thématique qu'il connaît déjà, dans ce cas il se peut que le besoin soit exprimé d'une manière incomplète), iii) le besoin thématique inconnu (l'utilisateur cherche de nouvelles informations dans des domaines qu'il connaît pas, et le besoin est toujours exprimé d'une manière incomplète).

Requête : La requête constitue l'expression du besoin d'information de l'utilisateur. La requête peut être exprimée en langage naturel, booléen ou graphique ou par des mots-clés.

Pertinence : La pertinence est définie par le degré ou la mesure de correspondance ou d'utilité qui existe entre un document et une requête ou un besoin d'information tel qu'il est exprimé par

l'utilisateur. Cette mesure est difficile à évaluer et elle est souvent donnée en termes de rappel et précision.

Représentation des documents : Ce processus, appelé ***indexation***, consiste à représenter les documents sous une forme exploitable par le SRI. Il s'agit d'identifier *les descripteurs* les plus significatifs dans chaque document et dans chaque requête, et de leur associer des poids pour représenter leur degré de représentativité. Les descripteurs sont en général les termes présents dans le document, souvent transformés par opérations telles que *la lemmatisation*, qui consiste à associer au mot sa forme canonique (l'infinitif pour les verbes, le masculin singulier pour les noms, ...).

L'indexation peut être de trois types, selon le niveau d'automatisation de la représentation des documents :

1. l'indexation manuelle : les documentalistes identifient manuellement l'ensemble des descripteurs du document ;
2. l'indexation automatique : l'identification des descripteurs est entièrement automatisée ;
3. l'indexation semi-automatique : le spécialiste de l'indexation sélectionne manuellement une liste de descripteurs significatifs à partir d'une liste extraite des documents d'une manière automatique.

3.3 Architecture générale d'un système de recherche d'information

Le but d'un SRI est de permettre à un utilisateur d'obtenir des informations à partir d'une ressource de connaissances dans le but de l'aider à résoudre un problème (helps him in problem management) [Robertson, 1981].

Ces objectifs de RI ont été décrits dans des modèles du type montré sur les figures 3.1 et 3.2 [Belkin & Croft, 1992, Baziz, 2005]. Dans ce modèle, une personne avec des objectifs particuliers, par exemple la réalisation d'une tâche au travail, constate que ses objectifs ne peuvent pas être atteints à cause de son manque de connaissances. Cette situation problématique va engendrer un besoin d'information qui va pousser la personne à chercher activement des informations (ex. soumettre une requête à un SRI). La requête est l'expression du besoin d'information. Comme il est souvent difficile d'exprimer ce besoin, une requête dans un SRI est toujours considérée comme incomplète et approximative. Pour apparier une requête et un ensemble de documents, les deux doivent être indexés selon l'un des principes montrés précédemment. Vient ensuite le processus de recherche qui est le noyau d'un SRI. Il comprend la fonction qui permet d'associer à une requête l'ensemble des documents pertinents à restituer. Ce processus est étroitement lié au modèle d'indexation et est considéré comme le point essentiel qui différencie un SRI d'un autre. Les trois modèles classiques de RI sont les modèles booléen, vectoriel et probabiliste. Dans le modèle booléen, les documents et les requêtes sont représentés comme un ensemble de termes. Le modèle est aussi dit *ensembliste*. Dans le modèle vectoriel, les documents et les requêtes sont représentés comme des vecteurs dans un espace multi-dimensionnel. Le modèle est aussi dit *algébrique*. Dans le modèle probabiliste, la modélisation des documents et des requêtes est basée sur la théorie des probabilités. De ce fait, le modèle est dit *probabiliste* [Baeza-Yates & Ribeiro-Neto, 1999].

A travers les années, des paradigmes alternatifs de modélisation pour chaque type de modèle ont vu le jour (i.e., ensembliste, algébrique et probabiliste). Pour les modèles ensemblistes, on peut distinguer les modèles booléens étendu et flou. En ce qui concerne les modèles algébriques

alternatifs, on peut citer le modèle vectoriel généralisé, l'indexation sémantique latente et les modèles de réseaux de neurones. Pour les modèles probabilistes alternatifs, on distingue le réseau d'inférence et le réseau de croyance. Ces modèles seront brièvement introduits dans la prochaine section.

Le processus de recherche peut contenir des étapes supplémentaires qui visent à améliorer les performances du SRI telles que *la reformulation automatique des requêtes.*

Les trois principales fonctions d'un SRI, à savoir l'indexation, la recherche et la reformulation des requêtes, peuvent utiliser une ressource externe telle qu'une terminologie ou une ontologie. Dans le dernier cas, le processus est appelé *la recherche d'information guidée par une ontologie*, que nous détaillerons plus loin.

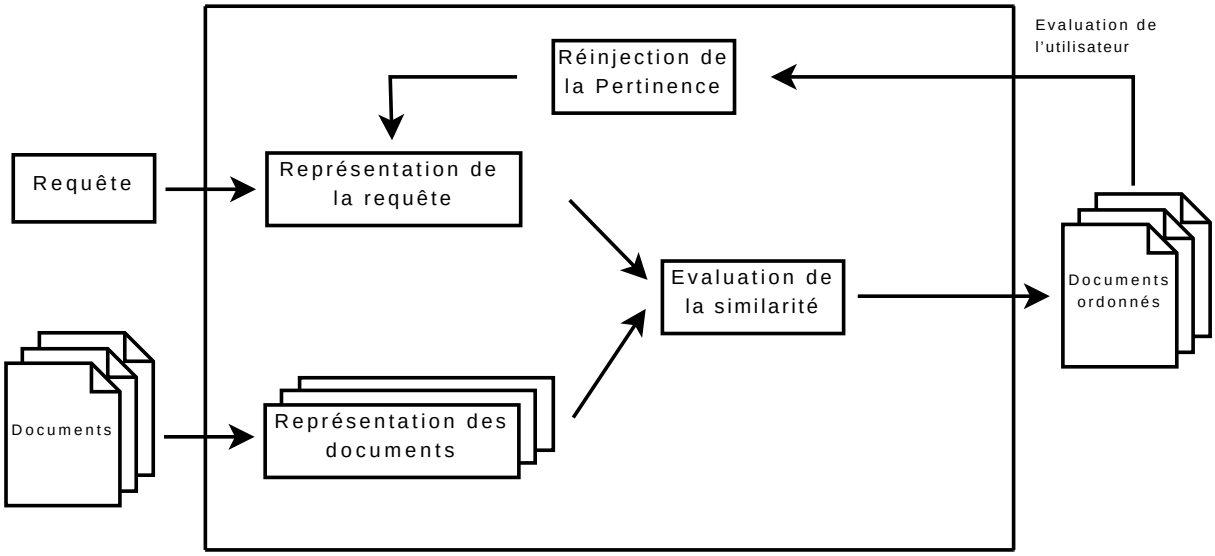


FIGURE 3.1 – Système de recherche d'information classique

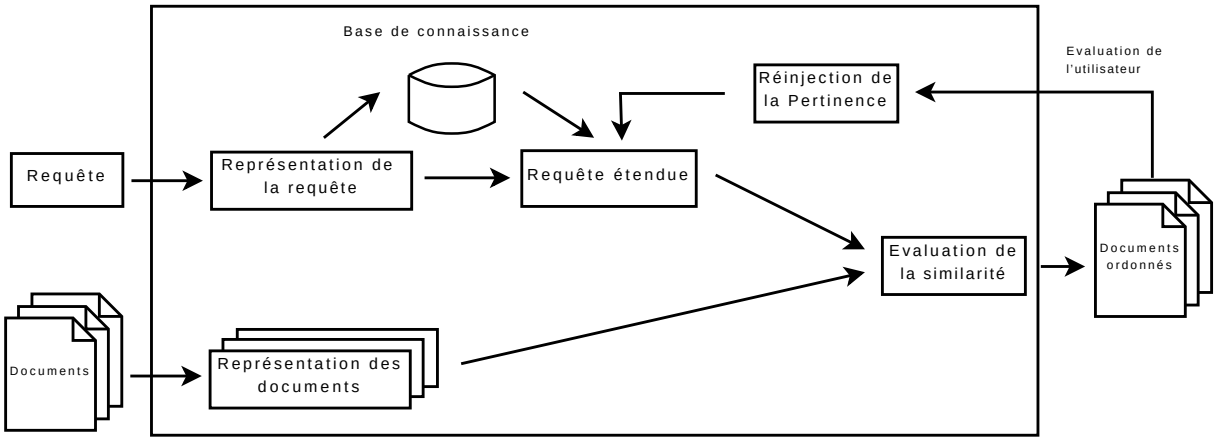


FIGURE 3.2 – Système de recherche d'information intelligent

3.4 Les modèles de recherche d'information

"Un modèle de RI est un ensemble de techniques et un algorithme qui visent à classer les documents selon leur pertinence par rapport à une requête donnée. Plus formellement, un modèle de RI est un quadruplet $[D, Q, F, R(q_i, d_j)]$, où D est un ensemble de vues logiques de documents (listes de descripteurs de documents), Q est un ensemble de requêtes d'utilisateurs, F est un framework pour modéliser les documents et les requêtes, et $R(q_i, d_j)$ est une fonction de classement qui associe une valeur numérique à la requête q_i et le document d_j" [Baeza-Yates & Ribeiro-Neto, 1999].

Un modèle de RI doit donc fournir un cadre théorique précis pour évaluer la pertinence d'un document à une requête particulière à partir de leurs représentations respectives. On peut classer les modèles de RI en trois catégories. Les modèles booléens, les modèles statistiques et les modèles basés sur des ressources de connaissance externes. Nous présentons dans ce qui suit les principaux modèles issus de chaque catégorie.

3.4.1 Les modèles booléens

Le modèle booléen est historiquement le premier modèle et le plus utilisé dans l'implémentation des SRI. Il est basé sur la théorie des ensembles ainsi que l'algèbre booléenne. Dans ce modèle, les documents sont considérés comme un ensemble de termes d'indexation, et les requêtes comme des expressions booléennes entre ces termes. Plus formellement, on peut le schématiser comme suit :

D : l'ensemble des termes d'indexation présents dans le document
chaque terme est soit présent (1) ou absent (0)

Q : une expression booléenne entre les termes d'indexation
les termes sont reliés par des opérateurs logiques $ET(\wedge)$, $OU(\vee)$, $NON(\neg)$

F : une algèbre booléenne entre l'ensemble des termes et l'ensemble des documents

R : un document est considéré comme pertinent pour une requête s'il satisfait l'expression de la requête.

Même si ce modèle a connu un grand succès, il présente un certain nombre d'inconvénients :

- l'absence d'une fonction de classement des résultats selon leur pertinence,
- l'utilisateur doit avoir une bonne connaissance du domaine pour formuler de bonnes requêtes,
- tous les termes de la requête ont la même importance ; aucune fonction n'est disponible pour affecter des poids aux termes de la requête,
- tous les documents pertinents et qui ne correspondent pas parfaitement à la requête sont écartés du résultat.

Pour remédier à ces inconvénients, plusieurs améliorations au modèle de base ont été proposées. Salton et ses collègues ont proposé le modèle booléen étendu [Salton et al., 1983b]. Ils ont proposé de pondérer les termes des documents et des requêtes pour refléter leur importance, en utilisant la méthode dite *P-norm*.

En 1993, Lee et ses collègues ont proposé une autre extension au modèle booléen classique, basée sur la théorie des ensembles flous [Lee et al., 1993]. Ils ont remplacé les opérateurs binaires dans le modèle booléen classique par des opérateurs flous appelés *opérateurs de compensation positifs*. Dans ce type de modèle, un élément possède un degré variable d'appartenance à un ensemble, au lieu du choix traditionnel entre *1* pour dire *appartient* et *0* pour dire *n'appartient pas*.

3.4.2 Les modèles vectoriels

Le modèle vectoriel représente les documents et les requêtes de l'utilisateur dans un espace multidimensionnel de dimension N (N étant le nombre de termes d'indexation de l'ensemble des documents) [Salton & McGill, 1986]. Chaque composante du vecteur correspond à un mot-clé, un terme ou un concept dans le document. La valeur qui lui est affectée reflète l'importance de l'item dans la description du contenu du document. Les documents et les requêtes sont comparés en comparant leurs vecteurs, en utilisant, par exemple, le produit scalaire ou la mesure du cosinus. La pondération des composantes de la requête est soit la même que celle utilisée pour les documents, soit donnée par l'utilisateur lors de sa formulation.

D'une manière plus formelle, l'index d'un document d_j est le vecteur $\vec{d_j} = (w_{1j}, w_{2j}, ..., w_{nj})$, où $w_{ij} \in [0,1]$ dénote le poids du terme t_i dans le document d_j. Le poids d'un terme dénote son intérêt dans le document. Une requête est également représentée par un vecteur $\vec{q} = (w_{1q}, w_{2q}, ..., w_{nq})$, où $w_{iq} \in [0,1]$ est le poids du terme t_i dans la requête q. La figure 3.3 montre un exemple d'espace vectoriel composé de trois termes t_1, t_2 et t_3. Les index de deux documents D_1 et D_2 et une requête Q sont représentés dans cet espace.

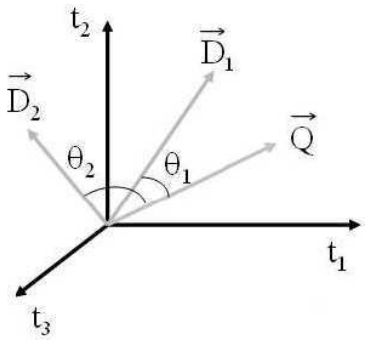


FIGURE 3.3 – Représentation vectorielle de deux documents et d'une requête

La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents. Une mesure classique est le cosinus de l'angle formé par les deux vecteurs :

$$\text{correspondance}(d_j,q) = cos(d_j,\vec{q})$$

où $cos(\vec{d_j}, \vec{q})$ est le cosinus de l'angle formé par les vecteurs $\vec{d_j}$ et $\vec{q}$:

$$cos(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{\| \vec{d_j} \| \cdot \| \vec{q} \|} = \frac{\sum_{i=1}^{n} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{n} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{n} w_{iq}^2}}$$

Plus deux vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. Dans l'exemple de la figure 3.3, le document D_1 est plus similaire à la requête que le document D_2. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne satisfont la requête que partiellement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante. La formulation de requêtes sous la forme d'expressions logiques comme *"documents concernant le sport, mais **pas** le football"* n'est toutefois pas possible dans le modèle vectoriel [Martinet, 2004].

Les modèles vectoriels font l'hypothèse que chaque terme est une "dimension" au sens des espaces vectoriels, alors que l'utilisation des termes n'est pas orthogonale. De plus, ils utilisent

les termes comme unité de base de la recherche d'information, sans tenir compte des problèmes de polysémie. Même si le modèle vectoriel est critiqué, il est l'un des modèles de RI classique les plus étudiés et les mieux acceptés.

3.4.3 Les modèles probabilistes

Le modèle probabiliste a été introduit en 1976 par *Stephen Robertson* et *Karen Sparck Jones* [Robertson, 1977]. Ce modèle essaye de modéliser le problème de RI en utilisant les lois des probabilités. Il est basé sur l'hypothèse suivante : Étant donnée une requête q et un document d_j dans la collection, le modèle probabiliste essaye d'estimer la probabilité que l'utilisateur juge le document d_j pertinent. Le modèle suppose que cette probabilité de pertinence dépend uniquement des représentations de la requête et du document. En plus, le modèle suppose qu'il existe un sous-ensemble de documents que l'utilisateur considère comme l'ensemble idéal de réponses à la requête q. Ce sous-ensemble est appelé l'ensemble R et devrait maximiser la probabilité de pertinence par rapport à l'utilisateur. Les documents dans l'ensemble R sont considérés comme pertinents à la requête. Les documents qui n'y appartiennent pas sont considérés comme non pertinents.

Cette hypothèse est assez ambiguë parce qu'elle ne montre pas explicitement comment calculer ces probabilités de pertinence.

Définition *Pour le modèle probabiliste, les variables des poids des termes d'indexation sont toutes binaires i.e.,* $w_{ij} \in \{0,1\}$, $w_{iq} \in \{0,1\}$. *Une requête* q *est un sous-ensemble de termes d'indexation. Soit* R *l'ensemble des documents connu pour être pertinent et soit* $\overline{R}$ *le complément du sous-ensemble* R *(i.e., l'ensemble des documents non pertinents). Soient* $P(R \mid \vec{d_j})$ *la probabilité que le document* d_j *soit pertinent pour la requête* q *et* $P(\overline{R} \mid \vec{d_j})$ *la probabilité que le document* d_j *soit non-pertinent à la requête* q. *La similarité* $sim(d_j, q)$ *du document* d_j *et de la requête* q *est définie ainsi :*

$$sim(d_j, q) = \frac{P(R \mid \vec{d_j})}{P(\overline{R} \mid \vec{d_j})}$$

En utilisant la règle de Bayes,

$$sim(d_j, q) = \frac{P(\vec{d_j} \mid R) \cdot P(R)}{P(\vec{d_j} \mid \overline{R}) \cdot P(\overline{R})}$$

$P(\vec{d_j} \mid R)$ représente la probabilité de sélectionner au hasard le document d_j de la collection R des documents pertinents. $P(R)$ est la probabilité qu'un document sélectionné au hasard dans la collection entière soit pertinent. Le sens des probabilités $P(\vec{d_j} \mid \overline{R})$ et $P(\overline{R})$ est analogue et complémentaire.
Comme $P(R)$ et $P(\overline{R})$ sont similaires à tous les documents de la collection, on peut écrire :

$$sim(d_j, q) \sim \frac{P(\vec{d_j} \mid R)}{P(\vec{d_j} \mid \overline{R})}$$

En supposant l'indépendance des termes d'indexation,

$$sim(d_j, q) \sim \frac{(\prod_{g_i(\vec{d_j})=1} P(k_i \mid R)) \cdot (\prod_{g_i(\vec{d_j})=0} P(\overline{k_i} \mid R))}{(\prod_{g_i(\vec{d_j})=1} P(k_i \mid \overline{R})) \cdot (\prod_{g_i(\vec{d_j})=0} P(\overline{k_i} \mid \overline{R}))}$$

$P(k_i \mid R)$ est la probabilité que le terme d'indexation k_i soit présent dans un document sélectionné au hasard de l'ensemble R. $P(\overline{k_i} \mid R)$ est la probabilité que le terme k_i ne soit pas présent dans un document sélectionné au hasard de l'ensemble R. Les probabilités associées à l'ensemble $\overline{R}$ ont un sens analogue au précédent.

En introduisant les logarithmes, en rappelant que $P(k_i \mid R) + P(\overline{k_i} \mid R) = 1$, et en ignorant les facteurs qui sont constants pour tous les documents dans le contexte d'une même requête, on peut finalement écrire :

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{iq} \cdot w_{ij} \cdot (log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})})$$

qui est l'expression clé pour le calcul de pertinences dans le modèle probabiliste.

Comme on ne connaît pas l'ensemble R au départ, il est nécessaire de proposer une méthode pour calculer les probabilités $P(k_i \mid R)$ et $P(k_i \mid \overline{R})$. Il existe plusieurs méthodes pour réaliser ce calcul. Cette fonction d'estimation est la partie principale du modèle et c'est celle qui différencie un modèle probabiliste d'un autre. L'idée initiale des modèles probabilistes a été proposée par Maron et Kuhns dans un article publié en 1960 [Maron & Kuhns, 1960]. Depuis, plusieurs modèles probabilistes ont été proposés, chacun basé sur une technique différente d'estimation des probabilités [Callan et al., 1992, Robertson et al., 1995].

L'avantage principal du modèle probabiliste, en théorie, est que les documents sont classés par ordre décroissant de leur probabilité de pertinence. Les inconvénients comprennent : (i) le besoin initial de séparer les documents en deux ensembles, les documents pertinents et les documents non-pertinents ; (ii) le fait que la méthode ne prenne pas en compte la fréquence d'apparition d'un terme dans un document (i.e., tous les poids sont binaires) ; et (iii) la supposition que les termes d'indexation sont indépendants.

3.4.4 Brève comparaison des modèles classiques de RI

En général, le modèle booléen est considéré comme la méthode classique de RI la moins performante. Son problème majeur est son incapacité de reconnaître les correspondances partielles, ce qui conduit souvent à des performances pauvres du système. Cependant, il existe un débat entre les chercheurs pour dire si le modèle vectoriel dépasse en performances le modèle probabiliste. *Croft* a réalisé quelques expérimentations et a fini par conclure que le modèle probabiliste fournit de meilleurs résultats que ceux fournis par le modèle vectoriel. Cependant, les expériences conduites ensuite par *Salton* et *Buckley* réfutent ce constat. A travers différentes mesures, ils ont montré que le modèle vectoriel dépasse les performances du probabiliste pour les collections de documents générales. Ce point de vue semble être également le plus partagé parmi la communauté de chercheurs.

3.5 Modèles ensemblistes alternatifs

Dans cette section, nous allons brièvement introduire deux modèles ensemblistes alternatifs, à savoir, le modèle basé sur les ensembles flous et le modèle booléen étendu.

3.5.1 Le modèle basé sur les ensembles flous

Représenter les documents et les requêtes à travers des ensembles de mots-clés conduit à des descriptions qui sont partiellement liées à la sémantique réelle du contenu des ressources. Par

conséquent, la correspondance d'un document avec une requête est approximative (ou vague). Cette dernière constatation peut être modélisée en considérant que chaque requête définit un ensemble *flou* et que chaque document possède un degré d'appartenance à cet ensemble (généralement inférieur à 1). Cette interprétation du processus de RI (en termes de concepts de la théorie des ensembles flous) est à l'origine des différents modèles de RI basés sur cette théorie proposée par *Zadeh* en 1965 [Zadeh, 1965].

Sans donner les formules utilisés dans le calcul du degré d'appartenance dans ce type de modèle, le principe est le suivant : une requête est représentée par un arbre, un nœud avec l'opérateur OR (resp. AND) est évalué en prenant le maximum (resp. minimum) sur les valeurs de ses fils, ce qui correspond à l'union (resp. intersection) floue des sous-ensembles flous correspondant à ses fils. Ce modèle permet d'obtenir un score de pertinence pour un document dans l'intervalle $[0, 1]$, ce qui permet cette fois de classer les documents. Des extensions ont été proposées à ces modèles [Bordogna & Pasi, 2000, Boughanem et al., 2005] notamment pour améliorer le classement (ranking) des documents sélectionnés.

3.5.2 Le modèle booléen étendu

La recherche booléenne est la méthode de RI la plus simple. Cependant, comme elle n'offre aucun moyen pour affecter des poids aux termes selon leur importance dans la requête ou dans le document, et aucun classement des résultats de la recherche n'est donné. Par conséquent, la taille du résultat peut être très grande ou au contraire très petite [Baeza-Yates & Ribeiro-Neto, 1999]. A cause de ces problèmes, les SRI modernes n'utilisent plus ce modèle. En fait, la plupart de ces systèmes adoptent dans leurs noyaux une forme de la recherche vectorielle. Les raisons viennent du fait que le modèle vectoriel est simple, rapide et offre de meilleures performances de recherche. Une alternative du modèle booléen a été par conséquent de l'étendre avec des fonctionnalités de correspondance partielle et l'assignation de poids aux termes. Cette stratégie permet de combiner des formulations de requêtes booléennes avec des stratégies de recherche vectorielle. Le premier modèle booléen étendu a été proposé par *Salton*, *Fox* et *Wu* en 1983 [Salton et al., 1983b].

3.6 Modèles algébriques alternatifs

3.6.1 Indexation Sémantique Latente (LSI)

Plusieurs techniques statistiques et d'intelligence artificielle ont été utilisées pour étendre le modèle vectoriel afin de répondre aux deux problèmes fréquents en RI causés par la synonymie et la polysémie. l'une de ces méthodes est celle de ***l'indexation sémantique latente*** (Latent Semantic Indexing) [Deerwester et al., 1990, Furnas et al., 1998]. Dans cette technique, les associations entre les termes et les documents sont calculées et utilisées dans le processus de recherche. L'hypothèse est qu'il existe une certaine structure "latente" dans l'usage d'un terme à travers les différents documents et que des techniques statistiques peuvent être utilisées pour estimer cette structure. Un avantage de cette approche est que l'on peut trouver des documents pertinents à une requête même s'ils n'ont aucun terme en commun. La différence entre le LSI et le modèle vectoriel est que le LSI représente les termes et les documents dans un espace dimensionnel réduit par rapport à celui du modèle vectoriel. Le LSI utilise une technique de réduction de dimension appelée *décomposition en valeurs singulières* (Singular Value Decomposition) ou SVD. L'inconvénient de cette méthode est qu'elle est sensible à la quantité et à la qualité des données traitées. En plus, la complexité de l'algorithme SVD est très coûteuse.

3.6.2 Le modèle connexionniste

Le modèle connexionniste est basée sur une architecture en réseau de neurones. Un réseau de neurones est une représentation graphique très simplifiée du réseau neuronal du cerveau humain. Les nœuds dans le graphe représentent les unités de calcul et les arcs jouent le rôle des connexions synaptiques. Pour simuler le fait que la puissance d'une connexion synaptique change à travers le temps, un poids numérique est assigné à chaque arc dans le réseau de neurones. A chaque instant, l'état d'un nœud est défini par son niveau d'activation (qui est une fonction de son état initial et des signaux qu'il reçoit comme entrée). En fonction de son niveau d'activation, un nœud A peut envoyer un signal à son voisin le nœud B. La puissance de ce signal au nœud B dépend du poids associé à l'arc entre les nœuds A et B.

Ce modèle sera détaillé d'une manière plus formelle au chapitre 6.

3.7 Modèles probabilistes alternatifs

Les réseaux bayésiens sont de plus en plus étudiés comme alternative aux modèles probabilistes de RI parce qu'ils fournissent un formalisme propre pour combiner différentes sources de données (requêtes passées, cycles de feedback passés, différentes formulations de requêtes) pour servir de support au processus de classement d'un document donné. Cette combinaison de données peut être utilisée pour améliorer les performances de la RI. Nous décrivons brièvement la notion de réseaux bayésiens dans la section suivante.

3.7.1 Les réseaux bayésiens

Les réseaux bayésiens sont des graphes dirigés acycliques [Pearl, 1988] dans lesquels les nœuds représentent des variables aléatoires, les arcs décrivent des relations causales entre ces variables, et la puissance de ces relations est exprimée par des probabilités conditionnelles. Une relation causale est représentée dans le graphe par un lien dirigé du nœud parent au nœud fils.

Soit x_i un nœud dans le réseau bayésien G et Γ_{x_i} l'ensemble des nœuds parents du nœud x_i. L'influence de l'ensemble Γ_{x_i} sur x_i peut être spécifiée par un ensemble de fonctions $F_i(x_i, \Gamma_{x_i})$ qui satisfait :

$$\sum_{\forall x_i} F_i(x_i, \Gamma_{x_i}) = 1$$

$$0 \leq F_i(x_i, \Gamma_{x_i}) \leq 1$$

où x_i désigne également les états des variables aléatoires associées au nœud x_i. Cette spécification est complète et consistante car le produit $\prod_{\forall x_i} F_i(x_i, \Gamma_{x_i})$ constitue une distribution de probabilité jointe pour les nœuds dans G.

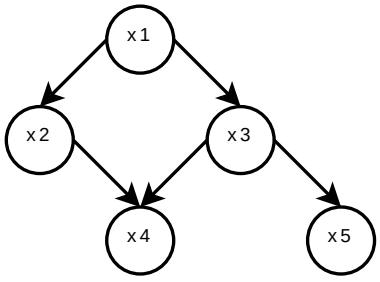


FIGURE 3.4 – Un exemple de réseau bayésien

La figure 3.4 illustre un exemple de réseau bayésien pour une distribution de probabilité jointe $P(x_1, x_2, x_3, x_4, x_5)$. Dans ce cas, les dépendances déclarées dans le réseau permettent l'expression de la distribution de la probabilité jointe en termes de probabilités conditionnelles locales comme suit :

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_3)$$

La probabilité $P(x_1)$ est appelée la probabilité *préalable* pour le réseau et peut être utilisée pour modéliser la connaissance antérieure sur la sémantique de l'application.

3.8 Modèles à base de ressources sémantiques externes

La majorité des systèmes de recherche d'information actuels se base sur l'approche de *sac-de-mots* pour représenter le contenu des documents. Dans ce cas, un document est représenté par l'ensemble des mots clés qu'il contient. Cette représentation ne prend pas en compte les relations sémantiques qui peuvent exister entre les mots, comme la synonymie, ou les phénomènes linguistiques comme la polysémie ou l'homonymie. Ceci peut être source de silence (plusieurs documents pertinents non retournés), ou de bruit (plusieurs documents non pertinents retournés).

Ces dernières années, beaucoup de travaux ont souligné l'insuffisance de cette représentation basée sur des mots simples [Khan, 2000, Guarino et al., 1999]. Une solution consiste à intégrer au processus de recherche des mécanismes qui permettront de prendre en compte la sémantique des textes dans le processus de recherche d'information. L'une des pistes exploitées est l'utilisation de ressources linguistiques ou sémantiques externes, comme les thésaurus ou les ontologies, dans l'indexation et la recherche des documents.

3.8.1 Utilisation des thésaurus en RI

Parmi les ressources utilisées pour améliorer la recherche d'information se trouve les thésaurus. Ils ont été développés à l'origine pour cet objectif, à savoir, regrouper le vocabulaire utilisé par les indexeurs et les utilisateurs qui cherchent de l'information [Bernier & Heumann, 1957, Vickery, 1960].

Parmi les travaux les plus notables dans ce domaine on peut citer ceux de Sparck Jones [Jones & Needham, 1968, Sparck Jones & Jackson, 1970], de Salton sur la construction automatique de thésaurus et l'expansion de requêtes [Salton, 1968, Salton et al., 1983b] et de Van Rijsbergen et ses collègues sur la co-occurrence des termes [Rijsbergen et al., 1981]. Ces expériences ont montré que sous certaines conditions l'utilisation des thésaurus apportait une amélioration des performances de la RI.

Les thésaurus ont été également exploités dans des tâches de désambiguïsation. La désambiguïsation essaye de résoudre les problèmes liés à la polysémie des mots. Le thésaurus est utilisé dans cette tâche pour déterminer lequel parmi les sens d'un mot donné est invoqué dans un contexte spécifique. Parmi ces travaux, citons l'exemple de Yarowsky [Yarowsky, 1992] qui a utilisé le thésaurus Roget, tandis que Voorhees [Voorhees, 1993], Richardson et Smeaton [Richardson & Smeaton, 1995], Gonzalo et ses collègues [Gonzalo et al., 1998] ont utilisé WordNet.

3.8.2 Utilisation des ontologies en RI

L'utilisation des ontologies pour dépasser les limites de la RI à base de mots-clés a été mise en avant comme l'une des principales motivations du Web sémantique depuis son émergence vers

la fin des années 90 [Berners-Lee et al., 2000]. Les ontologies offrent un espace sémantique pour décrire les documents et les requêtes à travers les concepts qu'elles regroupent et les relations qui les relient. Elles sont plus riches que les thésaurus, et en plus de la hiérarchie de concepts basée sur des relations d'hyperonymie/hyponymie (Is-A) ou de méronymie (Part-Of), elles peuvent contenir tout type de relation jugé utile ainsi que des contraintes portant sur le domaine concerné.

Les premiers travaux datent des années 1990 [Guarino et al., 1999, McGuinness, 1998, Fensel et al., 1999, Heflin et al., 2003]. Les ontologies offrent un espace conceptuel sur lequel les systèmes s'appuient pour saisir une partie de la sémantique présente dans les documents et les requêtes. Cette sémantique vient de l'utilisation des représentants des concepts (termes) de l'ontologie comme vocabulaire de référence. Comparé à un espace de représentation basé sur les mots-clés, un index basé sur les ontologies présente deux avantages [Kabel et al., 2004] :

1. la présentation des résultats de recherche peut s'effectuer suivant les catégories présentes dans l'ontologie ;
2. la formulation et le raffinement des requêtes peuvent reposer sur un vocabulaire structuré fourni par l'ontologie.

Un des problèmes à surmonter dans cette approche est l'identification des concepts dans les ressources à indexer (concept mapping). Dans la majorité des travaux existants, le processus de détection de concepts est un processus semi-automatique. Une phase supplémentaire est souvent nécessaire pour corriger les concepts trouvés par un algorithme de désambiguïsation. Ceci est dû notamment aux problèmes de synonymie et de polysémie présents dans le langage naturel.

Plusieurs SRI basés sur une ontologie ont vu le jour. Le système *OntoSeek* de Guarino [Guarino et al., 1999] fut l'une des premières expériences d'utilisation des ontologies pour la RI. Il utilise l'ontologie SENSUS, qui est une extension et une réorganisation de WordNet, pour décrire les requêtes des utilisateurs et les ressources qui concernent des pages jaunes et des catalogues de produits en ligne. L'utilisation de l'ontologie pour la RI dans ce domaine a montré une amélioration dans les performances du système. *CORESE* (COnceptual REsource Search Engine) [Corby et al., 2006] est un moteur RDF(S) basé sur les graphes conceptuels. Il permet de charger des schémas RDFS et des annotations RDF dans le formalisme des graphes conceptuels. Il permet ensuite d'interroger la base d'annotations ainsi créée en utilisant l'opérateur de projection des graphes conceptuels. Le résultat obtenu est traduit en RDF pour être retourné en réponse à la requête. Roussey et ses collègues [Roussey et al., 2001] présentent une approche basée sur l'utilisation des graphes conceptuels pour la description sémantique de documents multilingues.

Un autre avantage d'utiliser les ontologies dans la RI est la réduction de temps de recherche. Ce point a été démontré par le système *ONTOWEB* [Kim, 2005]. Ce système permet une recherche sémantique sur les sites des organisations internationales comme par exemple la banque mondiale. On peut également citer dans le cadre du Web sémantique le système *FindUR* [McGuinness, 1998], l'initiative $(KA)^2$ de Benjamins et Fensel [Benjamins & Fensel, 1998] et la plateforme *KIM* [Kiryakov et al., 2003, Popov et al., 2003].

L'utilisation des ontologies dans la RI est considérée comme une alternative prometteuse pour améliorer les performances de la RI et le temps de réponse du système [Sure et al., 2002a, Guarino et al., 1999, McGuinness, 1998].

3.9 La mesure TF*IDF en Recherche d'information

La mesure *tf*idf* [Salton et al., 1983a] est une mesure utilisée pour évaluer l'importance d'un terme dans un document ou une collection de documents. Cette importance augmente proportionnellement aux occurrences du terme dans le document mais est contrebalancée par la fréquence

du terme dans la collection des documents. En effet, la seule mesure du nombre d'occurrences d'un terme dans une collection ne permet pas de connaître sa spécificité. Un terme qui est commun à beaucoup de documents est moins utile pour la recherche qu'un terme qui n'apparaît que dans peu de documents. En bref, le *tf* mesure l'importance d'un terme dans un document tandis que l'*idf* mesure sa spécificité dans toute la collection. La formule du *tf*idf* s'écrit :

$$w_{i,j} = tf_{i,j} \log \frac{N}{df_i}$$

où :

1. $w_{i,j}$ est le poids du terme i dans le document j ;
2. $tf_{i,j}$ est la fréquence d'apparition du terme i dans le document j ;
3. $\log \frac{N}{df_i}$ est l'*idf*, df_i est le nombre de documents contenant i et N est le nombre de documents de la collection.

3.10 Évaluation des systèmes de recherche d'information

L'évaluation des performances de la RI dépend de la tâche de recherche à évaluer. Par exemple, la tâche de recherche peut consister en une simple séquence (i.e., l'utilisateur soumet une requête et reçoit un ensemble de réponses) ou une session interactive entière (i.e., l'utilisateur détermine son besoin d'information à travers une série d'étapes interactives avec le système). Ces deux modes de recherche sont deux processus différents et par conséquent leurs méthodes d'évaluations seront différentes. En effet, dans une session interactive, l'effort fourni par l'utilisateur, les caractéristiques de l'interface du système, l'assistance fournie par le système et la durée de la session sont des facteurs importants qui doivent être mesurés et évalués. Dans une recherche simple, l'aspect le plus important à observer est la qualité de l'ensemble de réponses qui est généré.

Les suggestions de mesures et les techniques d'évaluation des systèmes de recherche d'information se sont multipliées depuis une vingtaine d'années, dans la lignée des projets de la **DARPA** [1] dont le plus connu est le programme **TREC**.

Des mesures de rappel et de précision, que nous allons détailler dans ce qui suit, servent à évaluer les deux objectifs essentiels d'un SRI, à savoir :

- Retrouver tous les documents pertinents,
- Rejeter tous les documents non pertinents.

3.10.1 Les mesures de Rappel/Précision

Considérons un exemple d'une requête I (sur une collection de test de référence) et son ensemble R de documents pertinents. Soit $|R|$ le nombre de documents dans cet ensemble. Supposons qu'une stratégie de recherche donnée (qui a été évaluée) génère un ensemble A de réponses à la requête I. Soit $|A|$ le nombre de documents dans cet ensemble. La valeur $|R_a|$ désigne le nombre de documents dans l'intersection des deux ensembles R et A. La figure 3.5 illustre ces ensembles.

1. Defense Advanced Research Projects Agency

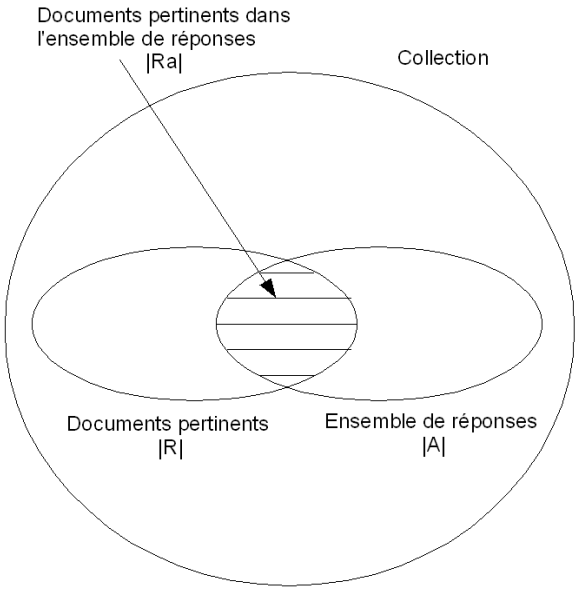


FIGURE 3.5 – Les mesures de rappel et de précision pour un exemple de requête

Les mesures de rappel et de précision sont définies comme suit :

Le rappel est le rapport entre les documents retrouvés pertinents et l'ensemble des documents pertinents de la base :

$$Rappel = \frac{|R_a|}{|R|}$$

La précision le rapport entre l'ensemble des documents sélectionnés pertinents et l'ensemble des documents sélectionnés :

$$Precision = \frac{|R_a|}{|A|}$$

3.10.2 La courbe de Précision-Rappel

Les mesures de rappel et de précision supposent que tous les documents dans l'ensemble des réponses A ont été examinés. Cependant, l'utilisateur ne reçoit pas l'ensemble des réponses en bloc. Les documents dans l'ensemble A sont plutôt classés par leur degré de pertinence. L'utilisateur par la suite examine cette liste ordonnée en commençant par les documents les mieux classés. Dans cette situation, les mesures de rappel et de précision varient au fur et à mesure que l'utilisateur examine cet ensemble de réponses A. Par conséquent, une évaluation propre exige de tracer une courbe de rappel/précision.

Comme précédemment, prenons une collection de référence et son ensemble d'exemples de requêtes. Étudions un exemple d'une requête donnée q. Un ensemble R_q qui contient les documents pertinents pour la requête q a été défini. L'ensemble R_q est composé des documents suivants :

$$R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

Selon un groupe de spécialistes, il existe une dizaine de documents pertinents pour la requête q dans la collection de référence. Considérons maintenant un algorithme de RI qui retourne pour la requête q le classement suivant des documents dans l'ensemble de réponses :

1.d_{123}•	6.d_9•	11.d_{38}
2.d_{84}	7.d_{511}	12.d_{48}
3.d_{56}•	8.d_{129}	13.d_{250}
4.d_6	9.d_{187}	14.d_{113}
5.d_8	10.d_{25}•	15.d_3•

TABLE 3.1 – Classement des documents pour la requête q

Les documents pertinents pour la requête q sont marqués par un point dans le tableau après le numéro du document. Si nous examinons ce classement en partant du premier document, nous faisons plusieurs remarques. Premièrement, le document d_{123} qui est classé en numéro 1 est pertinent. De plus, ce document correspond à 10% de tous les documents pertinents dans l'ensemble R_q. Donc, nous pouvons dire que nous avons une précison de 100% à 10% de rappel. Deuxièmement, le document d_{56} qui est classé en numéro 3 est le prochain document pertinent. A ce stade, nous pouvons dire que nous avons une précision d'environ 66% (2 documents sur 3 sont pertinents) à 20% de rappel (2 documents sur les dix pertinents pour la requête ont été trouvés). En procédant de cette manière, nous pouvons tracer une courbe de rappel/précision, illustrée dans la figure 3.6. La précision à des niveaux de rappel supérieurs à 50% tombe à 0 parce que seuls 5 des documents pertinents ont été trouvés. Cette courbe de rappel/précision est généralement basée sur 11 (au lieu de 10) niveaux de rappel qui sont 0%, 10%,. . . ,100%. Pour le niveau de rappel à 0%, la précision est obtenue par une procédure d'interpolation.

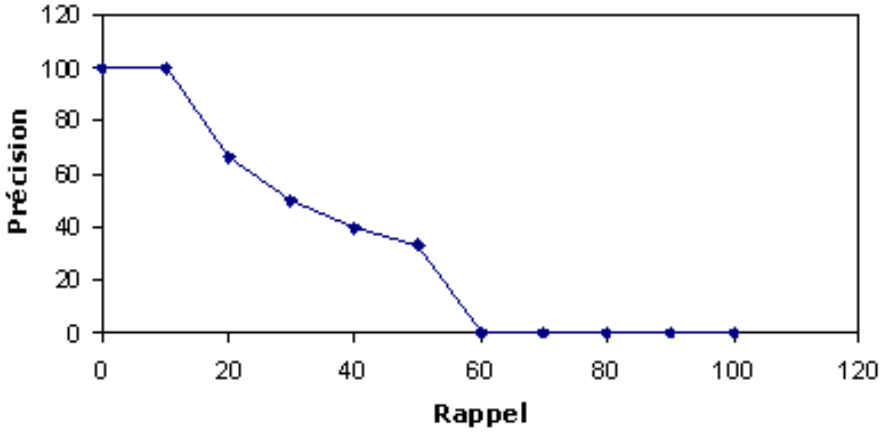


FIGURE 3.6 – La précision à 11 niveaux de rappel

3.10.3 Les mesures combinées

Les mesures de rappel et de précision, malgré leur popularité, ne constituent pas les mesures les plus appropriées pour évaluer les performances d'un SRI. Des mesures alternatives ont été proposées, parmi lesquelles on peut citer :

La mesure F (F-measure) et la mesure E (E-measure) : Van Rijsbergen [Rijsbergen, 1979] a proposé de combiner le rappel et la précision dans la F-measure et la E-measure, qui se calculent comme suit :

$$F(j) = \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{p(j)}}$$

et

$$E(j) = 1 - F(j)$$

où $r(j)$ est le rappel au j^{ieme} document dans le classement, $p(j)$ est la précision pour ce même document et b est le paramètre spécifié par l'utilisateur et qui reflète l'importance relative du rappel et de la précision. Des valeurs de b supérieures à 1 indiquent que l'utilisateur est plus intéressé par la précision que le rappel tandis que des valeurs inférieures à 1 indiquent l'inverse.

3.11 Conclusion

Dans ce chapitre, nous avons fait le tour des principales notions de la RI. Nous avons également introduit brièvement les principaux modèles utilisés par les SRI. Nous avons présenté d'une manière succincte l'utilisation des ontologies dans la RI et les principaux travaux qui ont marqué cette discipline.

Plusieurs travaux se dirigent maintenant vers l'intégration de ressources sémantiques dans le processus de RI. Ces techniques visent à s'affranchir des limites des SRI traditionnels, à savoir la production d'un nombre important de réponses non pertinentes. Les travaux actuels portent sur deux niveaux : les documents et les requêtes. Pour les requêtes, parmi les tentatives les plus marquantes on retrouve :

1. la réinjection de pertinence ou reformulation de requête (relevance feedback), qui vise à étendre la portée de la recherche en intégrant des termes issus soit des documents pertinents, soit des documents en tête de la liste de réponse trouvée automatiquement.
2. L'expansion de requête vise à renforcer l'expression de la requête de l'utilisateur (qui est souvent très courte, environ 2 à 3 mots [1], ou mal exprimée) par l'intégration d'autres termes reliés. Ces termes reliés peuvent être obtenus soit en exploitant un thésaurus, soit en utilisant un calcul basé sur des co-occurences de termes.

Le projet du Web sémantique essaye de proposer des solutions aux problèmes de la RI classique en utilisant des outils sémantiques et des techniques d'indexation conceptuelle. Il y a encore beaucoup de travail à faire mais ces techniques ont prouvé qu'elles constituaient une piste prometteuse pour améliorer les performances des SRI.

1. Étude menée par AOL, 2006

Deuxième partie

Construction d'une ontologie du cancer du sein

Chapitre 4

Construction du corpus

4.1 Construction des ontologies à partir d'un corpus de textes

Les premières initiatives de construction d'ontologies, telles que : METHONTOLOGY [Fernandez et al., 1997] et On-To-Knowledge [Sure et al., 2002b], se sont appuyées sur des méthodologies établies selon les mêmes critères que le génie logiciel. L'inconvénient de ces méthodes est qu'elles fournissent peu d'aide sur les modalités d'extraction des connaissances et sur la manière de déterminer le contenu de l'ontologie. Depuis, de nouvelles approches ont vu le jour, qui s'intéressent à l'expression des connaissances et donc à la langue comme vecteur principal de transmission de connaissances. Ces méthodes s'articulent autour de l'extraction des connaissances à partir de textes. Les premiers travaux qui se sont intéressés à la construction d'ontologies à partir de textes datent de 1995. Ils s'appuient sur des outils de TAL (Traitement Automatique de la Langue) pour déboucher sur des propositions méthodologiques [Bourigault & Aussenac-Gilles, 2003, Szulman et al., 1999]. Une étape primordiale dans ces méthodes est la construction d'un corpus approprié qui deviendra la source privilégiée permettant de caractériser les notions utiles à la modélisation ontologique et le contenu sémantique qui leur correspond.

Plusieurs définitions ont été données au terme "corpus textuel". Nous citons celle donnée dans [Condamines, 2003] : *"Collection de textes (éventuellement un seul texte) constituée à partir de critères linguistiques ou extra-linguistiques pour évaluer une hypothèse linguistique ou répondre à un besoin applicatif"*. Dans le cadre de notre travail, notre besoin est applicatif. Le corpus constitue la source de termes ou expressions utilisés par les médiateurs de santé ou le grand public pour parler du cancer du sein. Ce besoin applicatif s'inscrit dans le cadre des travaux en terminologie. La terminologie est traditionnellement associée à la catégorisation des termes utilisés dans le discours d'un domaine spécifique, également appelé *langue de spécialité* (language for special purposes). Dans les années 70, *Alain Rey* a proposé de faire la distinction entre les volets appliqué et théorique de la terminologie et de les étiqueter respectivement terminographie et terminologie. La terminographie regroupe les diverses activités d'acquisition, de compilation et de gestion des termes. La terminologie se penche sur les questions fondamentales que soulève l'étude des termes et propose un cadre conceptuel pour les appréhender [L'Homme, 2004]. Des protocoles et des méthodologies ont été développés pour la discipline de la terminographie, par les académiciens (e.g., [Cabré, 1999, Picht & Draskau, 1985]) et les organismes de standardisation, tels que ISO/TC37 (1952) et ISO 704 (1987). Ces méthodes se sont concentrées sur la communication entre les spécialistes d'un domaine donné. Cependant, comme *Bowker* l'a observé (2001), *"La communication en langues de spécialité ... devient de moins en moins limitée aux*

spécialistes du domaine : elle . . . envoie son message au grand public à travers les grands médias" [Cabré, 1999, p. 590].

Par ailleurs, La lexicologie est l'étude des mots du langage général et de leurs usages. *Cabré* fait la distinction entre la lexicologie et la terminologie comme suit :

> La lexicologie s'intéresse aux mots dans le but de représenter les compétences lexicales des personnes. La terminologie . . . s'intéresse aux termes dans le but d'établir une référence aux concepts du monde réel. . . . les objectifs de la terminologie sont clairement définis par rapport à ceux de la lexicologie descriptive, parce que la terminologie n'essaye pas de fournir une explication des connaissances que les experts ont des termes . . . . Plutôt, la terminologie . . . vise à identifier et à nommer les concepts propres à un domaine spécifique. (p. 36-37)

L'hypothèse adoptée dans cette thèse est que le vocabulaire médical utilisé par les usagers de santé se rapproche plus de la terminologie que de la lexicologie. En conséquence, les méthodologies propres à la terminographie ont été utilisées comme un point de départ pour explorer les termes utilisés par les usagers de santé. Ce travail peut être utile pour des recherches ultérieures dans l'exploration de l'usage du vocabulaire non-professionnel dans des domaines de spécialité.

Pour réaliser ces objectifs, deux ensembles de documents issus du Web ont servi comme sources pour la construction de deux types de corpus :

1. Corpus destiné aux usagers de santé (que nous appelons corpus médiateur [1]), qui consiste en des documents écrits par des professionnels de santé, des journalistes, des communicateurs dans le domaine de la santé ou même des patients avec une longue expérience de la maladie.
2. Corpus des usagers de santé, qui consiste en des messages de discussion écrits par les usagers de deux forums de patients sur le cancer du sein.

4.2 Le Web comme source de corpus

Le Web offre un accès à une énorme quantité de documents, sous plusieurs formes (textes, images, vidéos ou sons) et dans différentes langues. Ces documents sont dans leur majorité gratuits, faciles d'accès et dans un format électronique. Cette richesse a poussé plusieurs travaux de recherche à s'intéresser à la construction automatique de corpus à partir du Web [Bernhard, 2006, Kilgarriff & Grefenstette, 2003]. Le Web permet de constituer rapidement des corpus pour différentes applications dans différents domaines : littérature, articles scientifiques ou de vulgarisation, journaux, blogs ou forums de discussion,. . . [Baroni & Bernardini, 2004, Ghani et al., 2001, Naets, 2005].

Néanmoins, l'utilisation du Web comme source de corpus pose plusieurs problèmes. Les plus importants sont d'ordre juridique et technique :

– **Problèmes juridiques :** Ces problèmes sont liés aux questions juridiques de droit d'auteur. Dans [Kilgarriff & Grefenstette, 2003], les auteurs éludent le problème en argumentant que le fait d'utiliser une sous-partie du Web comme corpus est moins grave que ce que font les moteurs de recherche commerciaux, qui sauvegardent des index et des pages en cache immenses. Certaines initiatives commencent à apparaître, comme les licences *Creative Commons* pour les œuvres intellectuelles [2], qui visent à encadrer juridiquement la diffusion

1. Mediator corpus : appellation utilisée par Tony Tse en référence aux médiateurs de santé, qui sont toute personne professionnelle qui écrit des informations destinées aux patients ou au grand public.

2. `http://fr.creativecommons.org`

d'œuvres en ligne. Cette initiative est désormais relayée par le moteur de recherche Yahoo via le site `http://search.yahoo.com/cc` qui permet la collecte de documents accompagnés d'une licence Creative Commons. D'une manière générale, pour un corpus constitué à partir d'Internet, seule la liste des URLs des documents constituant le corpus est libre de droits et donc distribuable, à l'exclusion des documents eux-mêmes.

– **Problèmes techniques :** Les données extraites à partir du Web sont parfois fortement bruitées (fautes d'orthographe, langage de *chat*, éléments de navigation, publicités, etc). Il est donc nécessaire d'effectuer un certain nombre de traitements sur les textes pour les rendre utilisables.

Il existe trois approches principales pour l'utilisation du Web comme source de données linguistiques [Baroni & Ueyama, 2006] :

1. Utilisation d'un moteur de recherche pour l'estimation de co-occurrences de termes par le nombre de pages retournées par le moteur de recherche pour des requêtes qui contiennent les termes en question [Turney, 2001, Léon & Millon, 2005]. Cette approche présente de sérieuses limites. En effet, il est impossible d'effectuer des requêtes complexes, utilisant les expressions régulières par exemple.

2. Construction de corpus à partir de requêtes sur des moteurs de recherche [Ghani et al., 2001, Baroni & Bernardini, 2004]. Cette approche utilise des APIs permettant d'interroger automatiquement divers moteurs de recherche, sans passer par un navigateur Web. Ces APIs retournent une liste d'URLs correspondant à une requête. Des traitements sont nécessaires par la suite pour récupérer et nettoyer le contenu des pages pointées par ces URLs.

3. Utilisation de robots de parcours du Web [Vaufreydaz, 2002, Baroni & Ueyama, 2006, Liu & Curran, 2006]. Cette approche sera détaillée dans la section 4.3.2.

4.3 Les sources de documents

4.3.1 Le corpus médiateur

En terminographie, *"La communication scientifique et technique, orale ou écrite est la source de base pour l'extraction des termes"* [Cabré, 1999, p. 118]. De plus, selon *Cabré*, les sources utiles doivent avoir trois qualités : représenter le domaine concerné, contenir des informations actuelles, et enfin la relation avec le domaine concerné doit être explicite.

La construction du corpus médiateur a été manuelle. Les sites et les pages Web utilisés pour collecter les documents de ce corpus ont été sélectionnés en essayant de respecter au maximum les trois qualités citées précédemment. Le moteur de recherche *Google* a été utilisé pour générer la liste des pages Web en rapport avec le cancer du sein, en soumettant la requête *cancer du sein*. Nous avons commencé par les pages Web les plus populaires (les pages Web en haut de la liste des réponses). Les pages ont été sélectionnées selon des critères qualitatifs : la représentativité du domaine, le public ciblé par le site (grand public ou professionnel de santé), l'auteur de la page (professionnel de santé ou non), le langage utilisé (facile ou difficile). Ces critères nous ont permis de sélectionner les pages Web les plus appropriées pour fournir les termes représentatifs utilisés par les médiateurs de santé. Pour représenter les différents types et modalités de communication médicale, des articles de sites gouvernementaux et commerciaux ont également été sélectionnés.

Le corpus médiateur collecté

Nous avons collecté 583 documents en français. Après revue du corpus, nous avons décidé de supprimer 9 documents qui étaient destinés à des professionnels de santé, et qui font partie d'un guide de pratique clinique. Pour chaque document, les attributs suivants ont été enregistrés :

- Identifiant du document : nombre entier qui correspond également au nom du fichier qui contient le document.
- Nom de la source : le site Web ou la page où le document a été extrait.
- Titre du document.
- Date d'extraction.
- Auteur ou distributeur.
- Nature de l'information : définitions, informations générales, actualités médicales, ...
- Nombre de mots : a été obtenu en utilisant le logiciel libre *WordCount* [1]

Le tableau 4.1 résume les principales caractéristiques du corpus médiateur collecté. Le nombre de formes différentes ne tient pas compte de la liste des mots vides, des chiffres ainsi que des caractères qui ne sont pas alphanumériques.

Nombre de documents	574
Taille	3,04 Mo
Nombre total de mots	444 355
Nombre de formes différentes	14 045
Taille moyenne des documents	5,30 Ko
Nombre de sources	109

TABLE 4.1 – Caractéristiques du corpus médiateur

4.3.2 Le corpus des usagers de santé

Dans les sections suivantes, nous détaillons les différentes étapes de construction du corpus des usagers de santé à partir du Web en utilisant un robot de parcours du Web. Cette approche a notamment été utilisée par [Vaufreydaz, 2002, Liu & Curran, 2006]. L'ensemble des fonctionnalités décrites a été implémenté en Java.

Collecte de pages Web

Un robot de parcours du Web (en anglais : Web crawler ou Web spider) est un logiciel qui télécharge automatiquement les pages Web pour une analyse ou une exploration. Le robot développé pour ce travail est simple et conçu pour télécharger automatiquement les pages Web de deux forums de patients atteints du cancer du sein. Ce robot télécharge ces pages à partir d'une liste d'URLs. Il se limite exclusivement à la page indiquée par l'URL et ne parcourt pas les pages pointées par les hyperliens présents dans la page. Nous avons fait ce choix pour bien contrôler le téléchargement et le limiter aux pages qui contiennent les messages de patients. Le forum est composé de plusieurs thèmes de discussion. Chaque thème contient plusieurs pages de discussion entre les patients. Pour chaque thème, nous avons collecté la page racine. En étudiant la structure du site, nous avons généré le reste des URLs qui correspondent aux pages plus profondes du même thème en utilisant les expressions régulières.

Par cette procédure nous avons pu collecter 477 pages Web du forum `Essentielles.net` et 1 985 pages Web du forum de La Ligue Contre le Cancer. Les documents collectés ne constituent

1. `http://www.tawbaware.com/wc.htm`

pas encore un corpus exploitable. Ils contiennent encore beaucoup d'informations inutiles pour notre application. Dans la prochaine section, nous détaillons le processus d'extraction du contenu des pages collectées.

```
1  <html>
2   <head>
3    ...
4   </head>
5   <body>
6     ...
7    <a href="http://assoc...>Les phosphatases</a>
8    <p class="author">
9    <img src="effets-secondaires...>
10   <a href="http://assoc..." >tatiedaniele</a>
11   le Sam 7 Mar - 20:24
12   </p>
13   <div class="clearfix"></div>
14   <div style="display: none;"></div>
15   <div class="content clearfix"><!-- google_ad_section_start -->
16   Comme d'hab, j'ai eu ma prise de sang ce matin pour ma "cure". Comme d'
         hab, certains blancs sont en dessous de la norme, mais pas de soucis
         infectieux. Les phosphatases alcalines ont monté. Ils sont dans la
         norme mais sont passé au-dessus de 240 alors qu'avant ils étaient
         entre 150 et 180.<br>Qui peut me dire d'où cela pourrait venir? Merci
17    ...
18   <!-- google_ad_section_end -->
19   </div>
20   </div>
21   <div class="postprofile" id="profile7567"><!-- div class="online2">
22    ...
23  </body>
24 </html>
```

Listing 4.1 – exemple d'une partie de page Web collectée

Extraction du contenu des pages

Un document HTML contient du texte formaté par des balises comme <p> par exemple, qui marque un début de paragraphe. Ces balises définissent la structure et la forme de la page HTML. Pour extraire le contenu des documents collectés dans la phase précédente, nous avons développé un parseur qui se base sur la structure des pages HTML pour localiser les parties à extraire et qui correspondent aux messages des utilisateurs du forum. Le Listing 4.1 montre un exemple d'une page Web[1] à traiter. La partie qui nous intéresse correspond au paragraphe qui commence à la 16ème ligne. La structure stable des pages nous a permis de définir un chemin de balises qui a été utilisé par le parseur pour localiser le texte à extraire. L'API *HTMLParser* a été utilisée pour réaliser cette procédure. Chaque message a été sauvegardé dans un fichier à part. Nous avons également extrait les pseudonymes des utilisateurs du forum pour établir des statistiques par la suite en utilisant l'API *JExcelAPI*. Sur l'exemple de la page Web, cela consiste à extraire *tatiedaniele*.

1. Des parties de la page Web ont été supprimées par souci d'espace.

Lors de la construction de ce corpus, des problèmes éthiques liés à l'utilisation des informations issues des communautés en ligne ont été soulevés. Reconnaissant que plusieurs chercheurs collectent des données issues de ces sources, *Eysenbach* et *Till* ont proposé plusieurs catégories pour l'utilisation éthique de ce type de données [Eysenbach & Till, 2001]. Basé sur leur classification, cette étude est de type *"analyse passive"*, car elle n'est ni intrusive ni rétrospective et protège de ce fait la vie privée des personnes. De plus, nous avons nettoyé les messages des adresses e-mail personnelles qui peuvent être données par certains patients lors de leurs discussions.

Le corpus des usagers de santé collecté

Nous avons collecté un corpus issu de deux forums de patients atteints du cancer du sein : `Essentielles.net` et le forum de La Ligue Contre le Cancer. Le tableau 4.2 résume les caractéristiques de ce corpus.

`Essentielles.net`	
Nombre de pages Web collectées	477
Taille après traitement	3,41 Mo
Nombre de messages	1 911
Nombre de participants	564
Nombre de formes différentes	23 814
La Ligue Contre le Cancer	
Nombre de pages Web collectés	1 985
Taille après traitement	4,02 Mo
Nombre de messages	7 932
Nombre de participants	2 008
Nombre de formes différentes	19 653

TABLE 4.2 – Caractéristiques du corpus des usagers de santé

Ces corpus ont été utilisés comme source de termes et de connaissances pour la construction de l'ontologie. Dans le prochain chapitre, nous allons décrire les différentes étapes de la construction de cette ontologie.

Chapitre 5

Traitement du corpus et conceptualisation

5.1 Introduction

Les usagers de santé rencontrent souvent des problèmes pour trouver, comprendre et interpréter l'information médicale, à cause de leur manque de connaissances dans ce domaine. Idéalement, des terminologies d'usagers de santé refléteront les différentes manières que les usagers de santé utiliseront pour exprimer des concepts médicaux. Ces terminologies pourront ainsi servir à limiter les difficultés rencontrées par ce public. Cependant, malgré les recherches récentes sur les différences entre le langage médical et le langage grand public, peu d'efforts ont été fournis pour développer et évaluer ce type de ressources destinées aux usagers de santé.

Ce chapitre détaille les étapes de construction d'une ontologie du cancer du sein à partir de ressources textuelles. Les corpus construits lors de l'étape précédente sont utilisés pour l'acquisition des connaissances nécessaires au processus de construction. La méthode suivie est fondée sur les principes de la sémantique différentielle. Elle s'inspire des travaux de C. Roche et B. Bachimont sur la construction des ontologies [Roche, 2003, Bachimont et al., 2002]. L'ontologie résultante a été analysée sur plusieurs niveaux : termes, termes-concepts, concepts et relations. L'ensemble des résultats a montré un certain nombre de caractéristiques des concepts et des termes utilisés par les usagers de santé.

5.2 Extraction des n-grammes du corpus

L'extraction des termes est *"la tâche de détecter automatiquement, à partir de corpus textuels, des unités lexicales qui désignent des concepts dans des domaines de thématique restreinte"* [Vivaldi et al., 2001, p. 515]. Ce problème est difficile et plusieurs techniques ont été proposées pour le résoudre, plus spécialement des méthodes statistiques et linguistiques [Vivaldi et al., 2001, Rousselot & Frath, 2000]. Ces dernières comprennent des approches de traitement de corpus par analyse syntaxique et distributionnelle ou par des patrons lexico-synatxiques. Pour que ces approches d'extraction automatique soient efficaces, une connaissance linguistique préalable spécifique au domaine est nécessaire. Que ce soit un système qui utilise des mécanismes linguistiques, probabilistes ou combinés, les caractéristiques des termes spécifiques au domaine doivent être fournies implicitement (i.e., par la qualité du corpus) ou explicitement (e.g., liste de termes de base ou patrons lexico-syntaxiques spécifiques au domaine). Une langue de spécialité bien étudiée peut être un candidat approprié pour de tels systèmes. Cependant,

nous pensons que nous ne disposons pas d'assez d'information sur le vocabulaire médical utilisé par les usagers de santé pour profiter de ces approches. Par conséquent, nous avons choisi d'extraire les n-grammes du corpus en nous basant sur la méthode des segments répétés qui consiste à repérer les séquences de mots répétées dans le corpus [Lebart & Salem, 1994].

Les résultats obtenus ont été améliorés par l'utilisation de filtres [Rousselot, 2004] identifiant les mots qui indiquent des frontières de termes, en complément des signes de ponctuation délimiteurs de séquences (filtre "coupant" : verbes courants, adverbes, pronoms relatifs, conjonctions) et ceux qui ne peuvent se trouver aux bornes d'un terme (articles, prépositions). Les redondances sont ensuite supprimées en utilisant le mécanisme de l'intersection lexicale (également appelé contrainte d'autonomie par [Drouin, 2004]) : par exemple, si l'on trouve *cancer du sein inflammatoire* de fréquence 3 et *cancer du sein* de fréquence 3, *cancer du sein* est considéré comme un sous-segment de *cancer du sein inflammatoire* et est donc supprimé.

Les n-grammes sont classés par ordre de fréquence décroissante. L'extraction a été faite en mode itératif. A chaque itération, les mots les plus productifs et qui ne sont pas représentatifs du domaine ont été ajoutés au filtre. Nous avons gardé pour analyse les expressions régulières avec une fréquence supérieure à 6. A la fin du processus, nous avons obtenu 6 896 termes candidats du corpus médiateur et 11 723 termes candidats du corpus des usagers de santé.

5.3 Analyse des données : Concordancier

5.3.1 La méthode ARCHONTE

Peu de méthodologies proposent réellement de guider l'ingénieur des connaissances dans la construction d'ontologie pour définir les concepts du domaine et les relations qui les structurent. L'essentiel des démarches reposent sur une intuition quant à la manière de modéliser le domaine ou sur l'avis d'un expert. Nous nous sommes basés dans notre travail sur la méthodologie ARCHONTE (ARCHitecture for ONTological Elaborating), proposée par *B. Bachimont*, qui définit des directives précises pour expliciter véritablement les concepts à l'aide du langage et qui s'appuie sur la sémantique différentielle [Bachimont, 2000, Bachimont et al., 2002]. *B. Bachimont* propose de contraindre l'ingénieur des connaissances à un "engagement sémantique", c'est-à-dire à expliciter clairement le sens de chacun des concepts de l'ontologie, en introduisant une "normalisation sémantique" :

> Les primitives nécessaires à la représentation des connaissances doivent être modélisées à partir des données empiriques dont on dispose, à savoir l'expression linguistique des connaissances. Le travail de modélisation doit s'effectuer à partir de documents attestés dans la pratique d'un domaine et rassemblés en un corpus. Le corpus est constitué de documents produits dans le contexte où le problème à résoudre se pose [Bachimont, 2000].

Le corpus textuel dans cette méthodologie est la source privilégiée permettant de caractériser les notions utiles à la modélisation ontologique et le contenu sémantique qui leur est associé. ARCHONTE permet de décrire les variations des sens des termes considérés en contexte. Elle comporte trois étapes :

1. choisir les termes pertinents du domaine et normaliser leur sens, puis justifier la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarité et de différence que chaque concept entretient avec ses concepts frères et son concept père ;
2. formaliser les connaissances, ce qui implique par exemple d'ajouter des propriétés à des concepts, des axiomes, de contraindre les domaines d'une relation, ...

3. opérationnaliser l'ontologie dans un langage de représentation des connaissances.

Normalisation sémantique et engagement sémantique

L'étape de normalisation sémantique a pour objectif de rendre explicite le sens des expressions linguistiques. Il s'agit, par ce processus, d'en faire des primitives du domaine, c'est-à-dire d'identifier les notions élémentaires à partir desquelles l'ensemble des connaissances du domaine sont construites. *B. Bachimont* propose de se baser sur la sémantique différentielle, présentée dans les travaux de *F. Rastier* [Rastier et al., 1994], pour réaliser cette étape. Cette théorie attribue un sens aux termes grâce à la définition de traits sémantiques génériques et spécifiques. Ces traits permettent de définir un cadre interprétatif selon l'objectif de la tâche à réaliser. Ce cadre permet de stabiliser la définition des concepts et par conséquent les normer selon un certain point de vue. Dans la réalisation de notre ontologie, le point de vue est celui de l'usager de santé et non plus du professionnel de santé. En pratique, l'ingénieur des connaissances doit exprimer en langue naturelle les identités (traits sémantiques génériques) et les différences (traits sémantiques spécifiques en opposition les uns avec les autres) que chaque notion entretient avec celles qui lui sont proches. La structuration de ces notions, en fonction des identités et des différences qu'elles partagent avec leurs notions mères et leurs notions sœurs dans un arbre, permet de passer à "l'ontologie différentielle". *B. Bachimont* propose de définir quatre principes fondamentaux, les principes différentiels :

- le principe de communauté avec le père : expliciter en quoi le fils est identique au père qui le subsume,
- le principe de différence avec le père : expliciter en quoi le fils est différent du père qui le subsume,
- le principe de différence avec les frères : expliciter la différence de la notion considérée avec chacune des notions sœurs,
- le principe de communauté avec les frères : il faut expliciter la communauté existant entre la notion considérée et chacune des notions sœurs. Ce principe de communauté doit être différent du principe de communauté existant avec le parent. La communauté entre les notions filles doit permettre de définir des différences mutuellement exclusives entre les notions filles. *B. Bachimont* illustre très clairement ce principe par l'exemple suivant :

  > L'unité parente est "être humain", et les unités filles sont "homme" et "femme". Ces unités partagent le fait d'être des humains. Mais cette propriété ne permet pas de définir en quoi les hommes et les femmes sont différents. On choisit alors comme principe de communauté la sexualité, où l'on peut attribuer à "homme" le trait masculin et à "femme" le trait féminin. Ces deux traits sont mutuellement exclusifs, car ce sont deux valeurs possibles d'une même propriété [Bachimont, 2000].

A la fin de cette étape, on obtient une taxinomie de notions. La signification de chacune s'obtient de manière compositionnelle en parcourant les identités et les différences qui définissent l'ensemble des notions de l'arbre.

Formalisation des connaissances et engagement ontologique

La seconde étape permet de formaliser les connaissances du domaine à représenter. Il s'agit de définir des concepts, et non plus des notions, selon une sémantique formelle et extensionnelle. Les concepts sont exprimés dans un langage de représentation des connaissances et leur sens est décontextualisé. Selon la sémantique extensionnelle, les concepts sont liés à un ensemble de

référents dans le monde, i.e., à un ensemble d'objets du domaine. Cet ensemble est appelé l'extension du concept. A ce stade, des opérations ensemblistes, telles que la réunion ou l'intersection, peuvent être utilisées pour composer de nouveaux sens et donc de nouveaux concepts formels. Cette étape permet également de formaliser les relations qui existent entre les concepts.

Opérationalisation

La dernière étape a pour objectif d'informatiser l'ontologie dans un langage opérationnel de représentation des connaissances (voir section 2.4). Les concepts deviennent alors manipulables par un ordinateur en s'appuyant sur la spécification informatique des opérations mathématiques associées aux concepts. Ces opérations peuvent être de plusieurs sortes en fonction du formalisme de représentation des connaissances choisi. Cette étape a pour résultat une "ontologie computationnelle".

Notre expérimentation de construction d'ontologie a suivi cette méthodologie, avec quelques adaptations à notre application. Cette méthodologie a été développée pour des langues de spécialité bien étudiées. Notre cadre de travail rentre dans ce champ mais avec des considérations plus spécifiques (voir section 4.1).

5.4 Sélection des termes candidats du domaine

5.4.1 Définition des termes du domaine

La liste des termes et expressions extraite des corpus de textes est l'étape de départ dans la construction de l'ontologie. Cette liste contient encore beaucoup de bruit. Il faut donc filtrer et identifier les termes spécifiques au domaine du cancer du sein. Nous appelons les termes qui nous intéressent *les termes du domaine*. Ces termes sont de plusieurs sortes :

- termes spécifiques au domaine de la cancérologie : cancer, tumeur, gène BRCA1, ...
- termes du domaine mais d'utilisation commune à toute la médecine : fièvre, palpation, prélèvement, diagnostic, ...
- termes du domaine polysèmes ayant un sens médical et un sens autre dans la langue générale : plaquette, opération, examen, ...
- termes généraux, mais utilisés par les usagers de santé pour un sens médical : gros bras (pour lymphœdème du bras), bleu (pour hématome), ...
- termes généraux, mais qui sont intéressants pour le cancer du sein et les usagers de santé : perruque, association de patients, soutien psychologique, ...

Nous avons utilisé le parseur développé pour l'extraction du corpus des usagers de santé pour extraire des éléments spécifiques dans les pages Web qui forment le corpus médiateur (titres, sous-titres, expressions en gras, expressions en italique, expressions soulignées). Ces éléments ont été reconnus grâce aux balises HTML : <h1>, <h2>, , ... Cette structure interne des sites Web a permis d'identifier des *termes fondamentaux du domaine*. En effet, l'importance de certains termes peut être présumée au vu de la place qu'ils occupent dans la structure des documents.

5.4.2 Extraction, filtrage et sélection

La liste des termes extraits des corpus contient encore beaucoup de bruit. Par conséquent, nous avons procédé à un filtrage pour isoler ou éliminer un certain nombre de termes :

– éliminer les termes qui contiennent des chiffres mais ne contiennent pas de majuscules : 3ème séance, température à 38 ° , 30mm, ...
– isoler les syntagmes qui contiennent des caractères qui ne font pas partie de l'alphabet : rp+, ro+, mg/j,...
– isoler les termes qui contiennent plusieurs majuscules ou une majuscule et des chiffres : BRCA1, BRCA2, P53, TAXOTERE,...

Nous obtenons une liste réduite de termes candidats. L'identification des termes du domaine doit se faire manuellement. Plusieurs expérimentations d'identification automatique des termes du domaine se sont soldées par des échecs [Bowker, 1996, Lame, 2002]. *Bowker* a d'ailleurs conclu que "la terminographie est une discipline où une analyse manuelle minutieuse ne peut pas être remplacée ; elle peut être facilitée..." [Bowker, 1996, p. 49]. Pour faciliter la tâche de sélection des termes du domaine, nous avons utilisé un concordancier [Bernhard, 2003]. Techniquement, le concordancier permet à l'utilisateur de formuler des requêtes sous forme d'expressions régulières dans le corpus et d'afficher toutes les parties du corpus contenant cette expression. Cette vision d'un terme dans son contexte permet en général de déterminer très rapidement son sens. La figure 5.1 montre un exemple de recherche dans le concordancier. Le terme *"gros bras"* a été soumis au concordancier. On voit que dès la première occurrence on peut conclure que *"gros bras"* veut dire *"lymphœdème du bras"*. Le concordancier est ainsi un très bon outil pour l'aide à la construction d'ontologies.

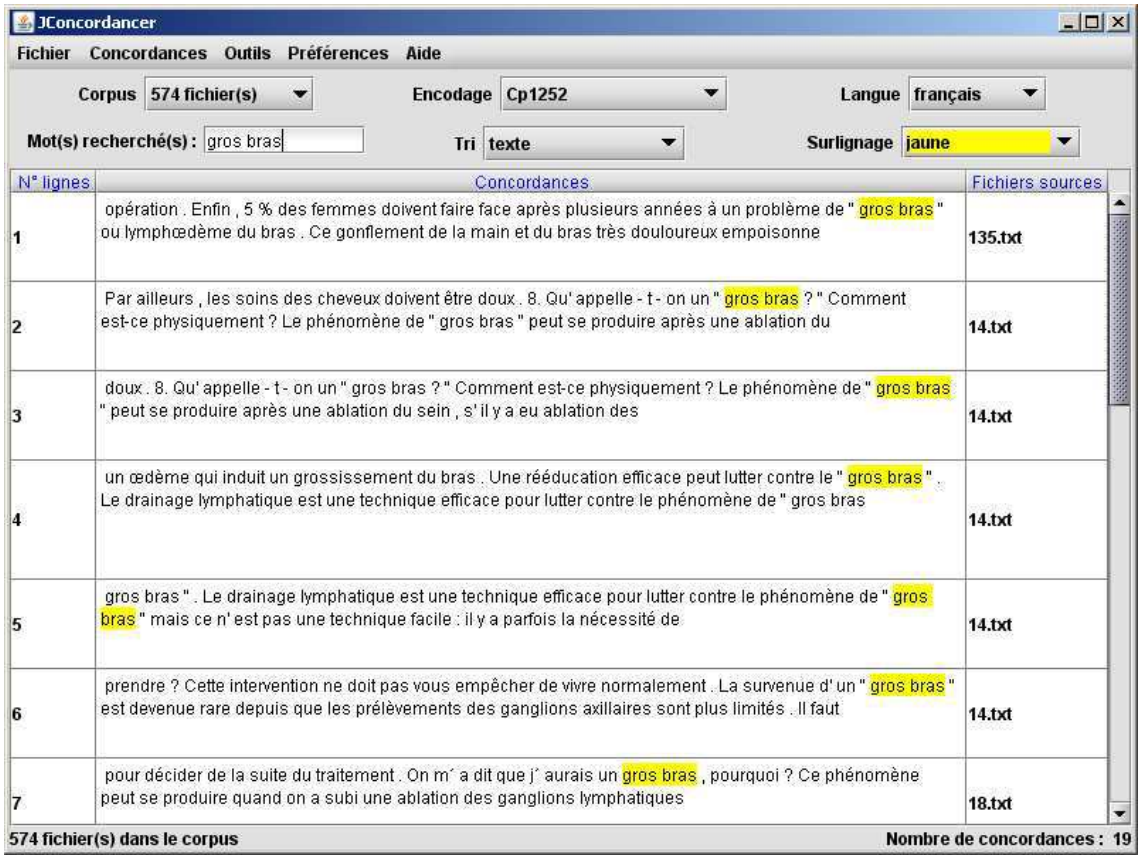


FIGURE 5.1 – Exemple de recherche dans le concordancier

L'analyse des *termes fondamentaux du domaine* (voir section 5.4.1) a permis de repérer les grands axes conceptuels typiques du corpus et donc du domaine. Ces axes sont indexés pour pouvoir les utiliser par la suite dans le classement des autres termes. Le tableau 5.1 montre ces axes et leurs descriptions.

Ces axes permettent de regrouper les termes du domaine dans des groupes conceptuels *larges* et d'adopter ainsi une méthode de construction descendante.

Index	Nom	Description
1	Organe	Cet axe représente les concepts de type organes du corps humain, par exemple : cœur, poumon, rein, ...
2	Fonction	Cet axe représente les concepts de type fonctions des organes du corps humain, par exemple : respiration, purification du sang, digestion, ...
3	Substance	Cet axe représente les concepts de type substance sécrétée par le corps humain, par exemple : salive, urine, sueur, ...
4	Pathologie	Cet axe représente les concepts de type pathologie ou maladie qui touchent l'homme, par exemple : cancer, dépression, diabète, ...
5	Symptôme	Cet axe représente les concepts de type symptôme ou signe qui apparaissent sur l'homme, par exemple : fièvre, fatigue, boule au sein, ...
6	Cause	Cet axe représente les concepts qui peuvent causer des pathologies, par exemple : virus, bactérie, cause génétique, ...
7	Examen	Cet axe représente les concepts de type examen appliqué à l'humain pour diagnostiquer une pathologie ou sa cause, par exemple : analyse du sang, analyse d'urine, prise de tension, ...
8	Traitement	Cet axe représente les concepts de type traitement ou thérapie d'une maladie, par exemple : chimiothérapie, radiothérapie, chirurgie, ...

TABLE 5.1 – Axes conceptuels pour la construction de l'ontologie

L'étape suivante consiste à élaborer la structure hiérarchique de l'ontologie. Dans chaque groupe conceptuel, nous avons étudié les termes pour les placer au sein de l'ontologie. Pour chaque terme, un module de recherche développé affiche dans un tableau l'ensemble des autres termes qui le contiennent. Ces termes peuvent avoir des relations avec le premier et permettent ainsi de créer des sous-hiérarchies conceptuelles. Le tableau 5.2 montre un exemple de ces recherches pour le terme *chimiothérapie*.

Chimiothérapie	chimiothérapie adjuvante, chimiothérapie à haute dose, chimiothérapie préopératoire, médicaments de chimiothérapie, chimiothérapie avant la chirurgie, chimiothérapie fec, chimiothérapie néo-adjuvante, traitements de chimiothérapie, chimiothérapie dite adjuvante, chimiothérapie locale, chimiothérapie renforcée, combinaisons de chimiothérapie, cycle de chimiothérapie, effets secondaires de la chimiothérapie, toxicité de la chimiothérapie, traitement de chimiothérapie, bénéfice de la chimiothérapie.

TABLE 5.2 – Liste des termes qui contiennent le terme *chimiothérapie*

5.4.3 Mise en œuvre des principes différentiels

La mise en œuvre des principes différentiels permet de créer la structure hiérarchique de l'ontologie. Il convient alors de préciser pour chaque concept les principes différentiels qui le définissent. Ces principes permettent de le placer dans la hiérarchie et de justifier cet emplacement. Ces principes servent aussi de documentation à l'ingénieur de connaissance en cas de mise à jour ou de révision de l'ontologie. Par exemple, le concept *cancer du sein intracanalaire* et le concept *cancer du sein intralobulaire* sont des concepts frères. Le principe différentiel entre ces concepts est relatif à l'emplacement de la prolifération des cellules malignes : les canaux galactophoriques pour le premier et les acini situés dans les lobules pour le deuxième.

5.4.4 Création des relations

Le concordancier est un bon support pour la recherche de relations sémantiques à partir de leurs formes lexicales dans le corpus. Une attention particulière a été accordée à cette phase. Les relations sémantiques jouent un rôle important dans notre application. En plus des relations *classiques* telles que *Est_Un*, *Partie_De*, nous avons défini un ensemble de 61 relations qui sont énumérées en annexe A. Ce travail s'est appuyé sur les travaux de *Soergel* et *Slaughter* [Soergel et al., 2004, Slaughter et al., 2006]. Grâce aux relations sémantiques, nous avons pu modéliser une partie des connaissances et des croyances des usagers de santé. Cette partie sera plus détaillée dans la section de discussion 5.8.

5.5 Édition et formalisation : PROTÉGÉ 3.2

A ce stade du travail, nous avons défini l'ensemble des concepts de l'ontologie et les termes du domaine (les synonymes), ainsi que les relations sémantiques qui les relient. Nous avons utilisé le logiciel PROTÉGÉ dans l'étape de formalisation de l'ontologie. Cette étape a consisté en l'édition des concepts de l'ontologie, des termes qui les désignent et des relations qui les relient. Elle permet également d'introduire des axiomes logiques qui définissent le comportement des individus qui constituent les extensions des concepts formels. Comme notre ontologie est destinée à des applications de recherche d'information, nous n'avons pas d'individus au sein de l'ontologie ce pourrait être, par exemple, une liste d'oncologues pour le concept *Medecin_Oncologue*. L'ontologie obtenue a été traduite en langage OWL, qui répond à nos besoins en termes d'expressivité et de maniabilité. L'annexe B montre un extrait de l'ontologie construite.

5.6 Mapping vers UMLS et CHV

Au cours de cette étape, nous avons cherché à relier les concepts de notre ontologie à ceux d'UMLS (voir section 2.3.1) et par conséquent ceux de CHV (voir section 2.3.2). L'intérêt de ce travail est double. D'une part, UMLS est le point d'entrée à plusieurs terminologies médicales et le mapping vers cette ressource facilite le mapping vers plusieurs d'autres terminologies. D'autre part, comme UMLS est une ressource destinée aux professionnels de santé, ce mapping peut être le point de départ d'une étude comparative entre les deux types de ressources.

Le mapping a été effectué manuellement. Pour chaque concept, la liste de ces termes a été traduite en anglais. Ensuite, les termes en question ont été soumis aux serveurs d'UMLS 2008AA [1] et CHV [2]. Seuls les cas de correspondance exacte ont été considérés. Par exemple, le concept

1. `https://login.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login`
2. `www.consumerhealthvocab.org`

Adenopathie_ Tumorale n'a pas de correspondance exacte dans UMLS. Le concept *Adenopathie* est suggéré par le serveur d'UMLS comme correspondance partielle. Ce dernier n'a pas été pris en compte dans le processus de mapping. Le tableau 5.3 montre les résultats de ce mapping :

Type de correspondance	Pourcentage
Correspondance exacte	83%
Correspondance partielle	3%
Aucune correspondance	14%

TABLE 5.3 – Résultats du mapping vers UMLS

En plus de ces résultats, nous avons fait cet ensemble de remarques :

- 5 concepts ont des correspondances multiples dans UMLS. Par exemple, *Cancer de l'ovaire* est désigné par les deux concepts *Ovarian carcinoma* et *Malignant neoplasm of ovary* dans UMLS. Le concept *Sein* peut être également désigné par les deux concepts *Breast* et *Entire breast*.
- 2 paires de concepts ont une correspondance unique dans UMLS. Les concepts *Mammographie* et *Mammogramme* sont alignés vers le concept *Mammography* dans UMLS. La même chose est observée pour la paire de concepts *Primipare* et *Primiparité* et le concept *Primiparity* dans UMLS.

Ces résultats montrent des différences entre les deux terminologies au niveau des termes et au niveau des concepts. Ces différences vont être analysées dans la section de discussion 5.8.

5.7 Résultats obtenus

Nous avons pu identifier 1 287 concepts désignés par 2 783 termes français. L'ensemble des concepts et des termes a été revu par un médecin oncologue au CHU de Grenoble et deux cadres infirmiers du service d'oncologie.

5.8 Discussion des résultats

Cette section décrit les différentes comparaisons que nous avons effectuées sur les termes et les concepts qui proviennent des deux différents corpus : médiateurs et usagers de santé. Nous discutons également les résultats du mapping vers UMLS et CHV. Quatre classes d'analyse ont été conduites :

- Analyse des termes.
- Analyse des concepts.
- Analyse des termes-concepts.
- Analyse des relations.

5.8.1 Analyse des termes

Des recherches actuelles utilisent la longueur des termes comme substitut de la complexité des termes dans les calculs de "*la lisibilité*" (readability) des documents [Gemoets et al., 2004, Rosemblat et al., 2006, Zeng et al., 2007]. Nous avons voulu comparer la longueur des termes qui proviennent des deux corpus ainsi que les termes d'une ontologie de cancer du sein destinée

aux professionnels de santé développée au cours du projet européen INFACE [1]. Pour cela, nous avons développé un outil qui a effectué les calculs suivants :

	Usager de santé	Médiateur	INFACE
Moyenne des caractères par terme	21,5	22,8	27,4
Moyenne des mots par terme	3,1	3,0	3,8

TABLE 5.4 – Longueur des termes

Bien que les termes dans le corpus médiateur soient de plus d'un caractère plus longs que les termes issus du corpus des usagers de santé, ils sont pratiquement identiques en nombre de mots par terme. Les termes issus de l'ontologie destinée aux professionnels contiennent plus de mots et un nombre plus élevé de caractères. Les usagers de santé utilisent généralement des expressions descriptives au lieu du terme médical. Ceci implique l'utilisation de beaucoup de mots, par exemple : *enlever le sein* pour *mastectomie* ou *perte de cheveux* pour *alopécie*. Cependant, cette comparaison présente une limitation. L'ontologie INFACE ne contient pas les mêmes concepts que notre ontologie. Elle a un niveau de granularité beaucoup plus élevé et contient par conséquent des termes très longs qui désignent des concepts très spécifiques, comme par exemple : *reflux de la lymphe tissulaire dans les tissus mammaires*.

5.8.2 Analyse des concepts

Le mapping de l'ontologie vers UMLS a révélé plusieurs choses intéressantes. Plusieurs concepts ont des correspondances multiples dans UMLS. Par exemple, le concept *Cancer de l'ovaire* a été relié aux deux concepts UMLS *Ovarian carcinoma* et *Malignant neoplasm of ovary*. A notre avis, la présence des deux concepts dans UMLS est injustifiée et un rapport soulignant la liste des concepts redondants a été rédigé et envoyé à la NLM dans le cadre de l'*UMLS Annual Report*. D'autres concepts de notre ontologie, au contraire, ont été alignés au même concept UMLS. Par exemple, les concepts *Mammographie* et *Mammogramme* sont alignés vers le concept *Mammography* dans UMLS, UMLS ne faisant pas la différence entre les deux. Dans notre cas, nous les avons séparés car le Mammogramme est le résultat de la Mammographie : une image radiologique pour le premier et une technique d'examen radiologique pour le deuxième. Ces cas montrent des anomalies de conceptualisation dans UMLS, que nous avons également signalées.

5.8.3 Analyse des termes-concepts

Analyse de la variabilité expressive des concepts

Le langage naturel est flexible et fournit plusieurs manières d'exprimer la même notion. Cette "variabilité expressive" peut être modélisée par le nombre de termes qui désignent le même concept [Tse & Soergel, 2003]. Nous avons voulu étudier cette variabilité expressive pour savoir quel est le type de concepts qui a la plus grande variabilité expressive. La moyenne de la variabilité expressive est de 2,16 pour l'ensemble de l'ontologie. La distribution de la fréquence totale des termes par concept est illustrée dans la figure 5.2. La distribution des termes suit une courbe de type Zipf, avec la majorité des concepts qui possèdent un seul terme.

1. Visual Interfaces for Timely retrieval of Patient-Related Information, projet du 5ème PCRD, Sept. 2002 – Août 2004 `http://www.inface.org`.

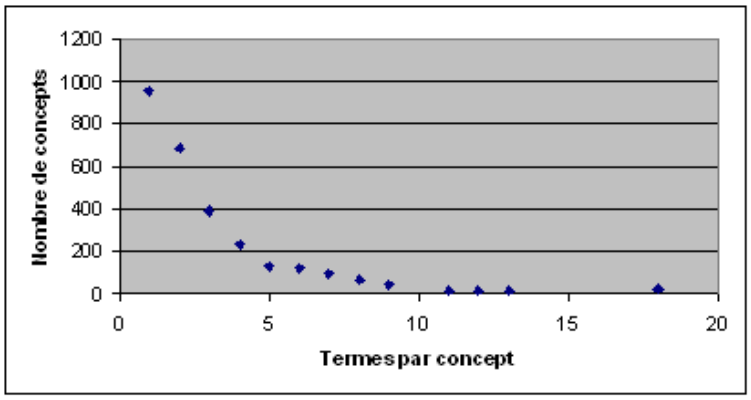


FIGURE 5.2 – Distribution des termes par concept

Pour étudier le type de concepts ayant la plus grande variabilité expressive, nous avons utilisé la classification des descriptions médicales de *Blois* [Blois, 1984] pour classer les concepts dans des groupes sémantiques. La classification de *Blois* (voir Tableau 5.5) a été construite dans le but de représenter les maladies par un ensemble d'attributs sur l'ensemble des classes. Bien que l'ambiguïté existe sur tous les niveaux d'attributs, *Blois* a observé que l'ambiguïté est particulièrement remarquée dans les niveaux supérieurs où "il s'agit des objets de tous les jours ... et des processus de la vie quotidienne ..." (p.61).

Les niveaux de Blois
0 : Patient comme un entier.
-1 : Parties majeures du patient : e.g., abdomen, tête.
-2 : Systèmes physiologiques : e.g., système cardiovasculaire.
-3 : Parties du système, ou organes : e.g., vaisseau, cœur.
-4 : Parties d'organe, ou tissu : e.g., myocarde, moelle osseuse.
-5 : Cellule : e.g., cellule épithéliale, lymphocyte.
-6 : Parties des cellules : e.g., membrane cellulaire, noyau de la cellule.
-7 : Macromolécule : e.g., enzyme, protéine.
-8 : Micromolécule : e.g., glucose, acide ascorbique.
-9 : Atomes ou ions : e.g., sodium, calcium.

TABLE 5.5 – Les niveaux hiérarchiques des descriptions médicales de Blois

Comme les usagers de santé sont plus familiers avec les objets et les processus de tous les jours (niveaux supérieurs) que les cellules et les molécules (niveaux inférieurs), *Blois* a établi que les usagers de santé vont plutôt utiliser les termes de ces niveaux que ceux des niveaux inférieurs. La hiérarchie a été divisée en deux niveaux globaux : "expérience quotidienne", défini comme les niveaux 0 (usager de santé) à -4 (Partie d'organe ou tissu), et "niveau technique", allant du niveau -5 (cellule) au niveau -9 (atomes ou ions). C'est à dire que les cinq niveaux supérieurs concernent les objets et les entités qui peuvent être directement observées ou expérimentées dans la vie de tous les jours. Par exemple, les organes et les tissus sont identifiables par l'observation directe de notre corps (e.g., l'oeil ou la peau) ou des produits animaliers. Par contre, les cellules, les bactéries et les virus, qui ne sont pas directement observables, peuvent cependant être détectés indirectement à travers leurs effets (e.g., lait caillé ou viande avariée). Les associations au niveau de l'expérience quotidienne semblent logiques et simples. Cependant, aux niveaux inférieurs, les

effets peuvent ne pas être notables et la chaîne du raisonnement causal devient plus longue.

Nous avons pris un échantillon au hasard de 100 concepts. 50 concepts avec une variabilité expressive > 5 et 50 autres concepts avec une variabilité expressive ≤ 5. Nous avons rencontré des difficultés pour classer les concepts au sein de cette classification. La classification a été conçue pour classer les maladies mais l'ontologie contient d'autres concepts que des maladies, tels que les traitements, les examens cliniques et les concepts liés aux conditions psychologiques et sociales des patients. Ces concepts ont été classés au sein de la hiérarchie selon leur cause, leur effet ou ce sur quoi ils agissent. Par exemple, le concept *Tumorectomie* a été classé au niveau -4 (celui des parties d'organes) car la tumorectomie est appliquée sur la partie de l'organe touchée par la tumeur. Pareillement, le concept *Antibiotique* a été classé au niveau -5 car les antibiotiques sont administrés contre les bactéries, et les bactéries sont au niveau cellulaire. Le tableau 5.6 montre la distribution des concepts selon les niveaux de classification de *Blois*.

Niveaux	(VE ≤ 5)	(VE > 5)
0	6	4
-1	9	2
-2	4	7
-3	1	9
-4	4	17
-5	2	8
-6	2	0
-7	8	3
-8	9	0
-9	5	0

TABLE 5.6 – Distribution des concepts selon leur niveau de variabilité expressive sur les niveaux de Blois

Les concepts avec une variabilité expressive élevée concernent généralement des concepts du niveau supérieur "expérience quotidienne", plus spécialement le niveau -4. Il s'agit donc d'un niveau de concepts qui peuvent être rencontrés facilement dans la vie de tous les jours mais qui commencent à être à la frontière des concepts techniques. Les concepts concernent généralement des procédés médicaux ou des symptômes. Ils désignent généralement des concepts que l'on commence à connaître quand on côtoie la maladie. Les usagers de santé, quand ils ne connaissent pas le terme médical exact, décrivent le procédé médical (*enlever le sein* au lieu de *mastectomie*) ou le symptôme qu'ils ressentent (*mal de tête* au lieu de *céphalée*). Comme ce type de descriptions n'est ni normalisé ni limité, la production des termes est plus élevée. Par contre, les concepts avec une variabilité expressive faible concernent soient des concepts connus et bien appréhendés par les usagers de santé, comme le nom des organes (e.g., foie, poumon, ...) ou des concepts très spécifiques aux professionnels de santé (e.g., Tamoxifène, Bilirubine, ...).

Nous avons conscience que cette étude présente des limitations et qu'une classification plus fine dévoilerait plus de caractéristiques. Un travail plus approfondi doit être effectué, qui dépasse le cadre de ce travail.

Analyse de recouvrement entre les deux terminologies

Nous avons comparé les deux terminologies issues des corpus médiateurs de santé et grand public pour déterminer le degré de recouvrement conceptuel et terminologique entre les deux. La comparaison a été conduite en deux étapes :

1. Le recouvrement conceptuel : déterminer les concepts qui sont communs aux deux terminologies et les concepts propres à chacune.
2. Le recouvrement terminologique : Pour les concepts communs, déterminer le nombre de termes communs aux deux terminologies.

Les listes des termes et des concepts qui proviennent de chaque corpus ont été comparées. Le tableau 5.7 montre les résultats de cette comparaison :

	Commun	Spécifique usager de santé	Spécifique médiateur
Concepts	1 254	8	25
Termes	2 238	289	182

TABLE 5.7 – Comparaison entre les deux terminologies

Les deux terminologies partagent un grand nombre de concepts. En étudiant de plus près les concepts propres à chaque terminologie, nous avons pu faire ces constatations :

- Les concepts propres à la terminologie des médiateurs de santé concernent généralement des concepts médicaux très spécifiques, surtout d'ordre anatomique. Par exemple : *Tubercules de Montgomery*.
- Les concepts propres à la terminologie des usagers de santé concernent soit des concepts d'ordre social et psychologique, soit des concepts généraux. Par exemple : *Remboursement des soins*, *Produit de beauté*.

En étudiant les termes propres à chaque terminologie, nous avons pu faire les constatations suivantes :

- Termes très spécifiques : comme dans le corpus des médiateurs de santé il y a des descriptions de procédés médicaux, de maladies et du corps humain, les médiateurs de santé utilisent des termes très spécifiques à la médecine que les usagers de santé n'emploient pas entre eux.
- Fautes d'orthographe : les termes médicaux sont parfois difficiles à retenir. Les usagers de santé interprètent parfois mal les mots qu'ils entendent. Ils utilisent par conséquent des homophones et des mots du langage général morphologiquement similaires. Par exemple : *limphocyte* (lymphocyte), *maladie de Hojkin* (maladie de Hodgkin).
- Termes abrégés : les termes médicaux sont généralement longs. Les usagers de santé utilisent des formes abrégées et non standardisées telles que les abréviations et les coupures. Par exemple : *neurochir* (neurochirurgien), *chimio* (chimiothérapie).
- Définition/Description : ignorant le terme spécifique, les usagers de santé peuvent définir le concept ou décrire ses propriétés. Par exemple : *enlever le sein* (mastectomie), *diluant de sang* (anticoagulant).
- Exemples : ne connaissant pas le terme spécifique, les usagers de santé peuvent utiliser un concept particulier de la classe qui représente *un exemple* (concept plus spécifique) du concept en question. par exemple : *aspirine* (antalgique).
- Néologismes : les usagers créent parfois de nouveaux mots, utilisés parfois au sein d'un groupe de patients. Par exemple : *cancerinette* (jeune femme atteinte d'un cancer), *mort-vivant* (patient en fin de vie).

5.8.4 Analyse des relations

Les usagers de santé, en plus de leur problème de terminologie médicale, rencontrent parfois des problèmes pour comprendre comment les concepts médicaux sont reliés entre eux. L'annexe A

illustre l'ensemble des relations créées dans cette ontologie. Parmi les relations créées, on trouve la *Relation_X* qui modélise le fait qu'il existe une relation entre deux concepts mais qui ne peut pas être spécifiée. Cette relation est utilisée pour définir un lien entre des concepts que les usagers de santé relient sans qu'il y ait une justification médicale suffisante derrière. Par exemple, le concept *Pilule_Contraceptive* est reliée au concept *Cancer_Sein* parce qu'un certain nombre d'usagers de santé croient qu'il y a un lien entre les deux, bien que les études scientifiques n'en aient établi aucun. Par conséquent, nous avons choisi de les relier en utilisant *Relation_X*. Nous avons fait de même pour les concepts mal appréhendés par les usagers de santé. Par exemple, le concept *Vaginite* est relié au concept plus général *Maladie_Vagin* par les deux relations *Est_Un* et *Relation_X*. Les usagers de santé associent la vaginite à tous les problèmes du vagin, bien que la vaginite ne représente que l'inflammation de la muqueuse du vagin. L'utilisation de ce type de relation pour modéliser ce type de phénomènes permet de ne pas altérer la structure bien fondée et stable de l'ontologie. Cette relation peut également changer ou disparaître au fil du temps et la modéliser de cette manière facilite la mise à jour de l'ontologie.

5.9 Limitations de la méthodologie

Dans une recherche exploratoire, l'identification des limitations de la procédure suivie est importante pour permettre d'éventuelles améliorations ou pour engager des travaux similaires. Les limitations rencontrées dans cette étude et les mécanismes utilisés pour diminuer leurs effets sont discutés. Le manque de temps et de ressources ne nous ont malheureusement pas permis de les surmonter :

- **Les corpus utilisés :** les textes des corpus constituent le point de départ de notre travail. Cependant, nous devons noter certaines limitations :
  - Pour collecter les termes utilisés par les usagers de santé, nous avons utilisé un corpus de textes issus de deux forums de patients. Par conséquent, nous avons ciblé un profil bien spécifique d'usagers de santé, celui des utilisateurs d'Internet. Ce profil sous-entend que les usagers de santé ont une curiosité plus élevée que les autres et par conséquent possèdent plus de connaissances que les autres. Une autre source du corpus pourrait être les conversations entre patients dans des groupes de paroles ou avec des professionnels de santé.
  - La taille du corpus varie d'un travail à l'autre, il semble difficile de donner des indications précises quant à un nombre de mots optimum. Le corpus des usagers de santé dépasse le corpus des médiateurs en nombre de formes différentes. Cela est sans doute dû à la nature du langage. Un langage académique et normalisé pour le corpus des médiateurs et un langage libre et ressemblant à celui du langage parlé pour celui des usagers. Les travaux existants de construction d'ontologies à partir de textes se sont servis de corpus de textes d'une moyenne de 350 000 mots, et nos corpus dépassent cette moyenne [Moigno et al., 2002, Baneyx, 2007].
- **La méthode d'extraction de termes :** la méthode d'extraction des termes utilisée se base sur des frontières définies de termes. Parmi ces frontières se trouvent les signes de ponctuation. Le corpus des usagers de santé est constitué de messages sur des forums d'utilisateurs. La structure du langage de ces messages ressemble à celle du langage parlé. L'utilisation des signes de ponctuation est peu respectée dans ce type de langage. La méthode d'extraction utilisée s'avère peu adaptée pour ce type de corpus. Il faudra réfléchir à d'autres algorithmes plus adaptés à ce type de ressources.

– **L'implication des usagers de santé :** Par manque de ressources, le processus de travail n'a malheureusement pas inclus des usagers de santé pour valider les étapes de conception de l'ontologie.

5.10 Conclusion

La procédure développée dans ce chapitre a montré quelques-unes des différences qui existent entre la terminologie des usagers de santé et celle des professionnels de santé. Nous avons pu constater que la principale différence réside au niveau des termes. Par conséquent, séparer le développement des ontologies destinées aux usagers de santé de celles des professionnels de santé n'est pas, à notre avis, justifiable. Il serait plus utile de reprendre des ressources destinées aux professionnels de santé et d'y ajouter des termes utilisés par les usagers de santé. Par contre, une réflexion plus profonde est indispensable sur un format de représentation de ce type de ressources. Une annotation du type des termes selon leur *technicité* pourrait faciliter l'utilisation de la même ressource selon le public des utilisateurs cibles.

L'approche de construction par corpus possède plusieurs limitations. La principale limitation est liée à la difficulté d'assigner correctement un sens aux termes repérés dans le corpus. Comme seuls les textes et non pas leurs auteurs sont disponibles, trouver le sens des termes est limité à l'interprétation de l'ingénieur des connaissances de l'intention de chaque auteur. Haas et Hert ont souligné ce problème : *"Même si les mots des utilisateurs peuvent être vus, l'intention derrière leur utilisation, ou ce que réellement veut l'utilisateur (le contenu et le contexte de l'utilisateur), ne peut être connu"* [Haas & Hert, 2002, p. 44]. L'utilisation des messages de patients sur des forums réduit cette limitation en fournissant plus de contexte par rapport aux travaux existants qui s'appuient sur des requêtes d'utilisateurs sur des moteurs de recherche ou des sites Web. Cependant, l'intention réelle peut être mieux cernée en utilisant des approches interactives.

Identifier et interpréter les termes des usagers est une *cible mouvante*, non seulement au niveau de la population, en tenant en compte les facteurs liés à la culture, l'éducation, les statuts sociaux et d'autres encore, mais également au niveau individuel, tels que l'expérience personnelle et l'apprentissage par l'exposition aux concepts professionnels. Cependant, le besoin de combler les lacunes entre la terminologie des professionnels et celle du grand public existe clairement. *"Avec les nouveaux environnements Internet e-santé ... pour permettre la participation des patients, cependant, les mots des patients doivent être traités avec autant de respect que les mots des professionnels de santé"* [Rose et al., 2001]. Les procédures décrites dans cette thèse peuvent être les étapes initiales sur le travail qui concerne "les mots des patients".

Troisième partie

Application de l'ontologie pour la recherche d'information

Chapitre 6

Les techniques de propagation d'activation pour la recherche d'information

6.1 Introduction

Ce chapitre fait le tour des principales applications des techniques de propagation d'activation sur des réseaux sémantiques en RI. L'origine de ces travaux est *la recherche d'information associative*. L'idée derrière cette forme de RI est qu'il est possible de trouver des réponses pertinentes en utilisant des informations *associées* à celles déjà trouvées par l'utilisateur, et estimées pertinentes par lui. Les associations entre les informations peuvent être statiques et déjà existantes avant la session de RI, ou dynamiques et déterminées pendant le temps d'exécution. Dans le premier cas, les associations entre les unités d'information (documents ou parties de documents, termes extraits, termes d'indexation, concepts, etc.) sont créées en amont du SRI et utilisent les relations sémantiques qui existent entre ces unités, telles que par exemple les relations dans un thésaurus ou dans une ontologie entre les termes ou les concepts, les citations bibliographiques entre les documents, ou les similarités statistiques entre les documents ou les termes. Dans le deuxième cas, le système détermine les associations entre les unités d'information à travers l'interaction avec l'utilisateur, par exemple en retrouvant des documents qui sont similaires à ceux jugés pertinents par l'utilisateur (cette technique particulière est appelée *"relevance feedback"* ou *"réinjection de la pertinence"*). Ces techniques ont été largement utilisées par la communauté de RI. Dans ce chapitre, nous nous concentrons sur la première technique, qui est souvent appelée *la recherche associative*.

Dans la recherche d'information associative, les associations entre les différentes unités d'information sont souvent représentées sous forme de réseau, où les unités d'information sont représentées par des nœuds et les associations par des liens connectant ces nœuds. L'algorithme qui consiste à trouver de l'information associée à une information déjà considérée comme pertinente est le plus souvent implémenté par le moyen d'une technique appelée *la propagation d'activation* (spreading activation). Dans ce qui suit, nous décrivons les différents types de techniques de propagation d'activation utilisées dans le contexte de la RI associative, et nous revenons sur les différentes expériences de développement et d'évaluation de SRI basés sur cette technique. Actuellement, cette technique suscite de moins en moins d'intérêt. La principale raison est que la construction d'un réseau d'associations entre les différentes unités d'information nécessite beaucoup de temps, surtout quand le domaine concerné par la collection des documents est très

étendu. La plupart des expériences intéressantes en RI associative ont été réalisées sur des domaines restreints et avec de petites collections de documents. Dans ce cas, les associations sont construites manuellement ou semi-automatiquement. Cependant, de nos jours, on dispose de machines de plus en plus puissantes et leur coût est en constante décroissance, rendant possible la construction automatique des réseaux d'associations pour de grandes collections de documents. On dispose également de plus en plus de ressources sémantiques qui représentent des concepts et leurs relations dans plusieurs domaines donnés.

6.2 Les réseaux sémantiques

Depuis leur introduction par *Quillian* [Quillian, 1968], les réseaux sémantiques ont joué un rôle important dans le domaine de la représentation des connaissances. Selon la définition de Quillian, les réseaux sémantiques expriment les connaissances d'un domaine en termes de concepts, leurs propriétés et les relations hiérarchiques entre ces concepts. Chaque concept est représenté par un nœud et les relations hiérarchiques entre les concepts sont obtenues en connectant les nœuds des concepts appropriés avec les liens *"est-un"* ou *"instance-de"*. Les nœuds du niveau inférieur représentent des individus, tandis que les nœuds de niveaux supérieurs représentent des classes ou des catégories d'individus. Les relations entre concepts sont des relations de type "est-un" s'appliquant entre deux concepts ou bien des relations "instance-de" s'appliquant entre un individu et un concept. Les propriétés sont également représentées par des nœuds. Une propriété est liée à un concept par un lien spécial libellé. Une propriété est attachée au concept approprié le plus haut dans la hiérarchie auquel elle peut être appliquée. Si une propriété est attachée à un nœud, on considère qu'elle peut être appliquée à tous les descendants de ce nœud. Un exemple de réseau sémantique est montré dans la figure 6.1 [Crestani, 1997].

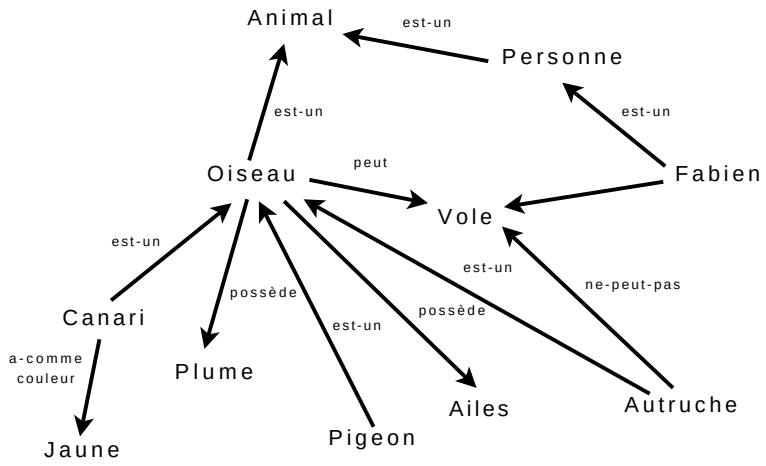


FIGURE 6.1 – Un exemple d'un réseau sémantique

6.3 La recherche d'information associative utilisant la propagation d'activation

Historiquement parlant, la propagation d'activation (PA) n'était pas le premier paradigme utilisé dans la RI associative. Les études sur la recherche d'information associative datent du début des années soixante et ont leur origine dans les études statistiques des associations entre

les termes ou les documents d'une collection. Le *modèle de recherche associatif linéaire* est l'un des premiers modèles utilisés. Ce modèle, qui correspond au modèle de Salton, consiste à élargir la requête d'origine en utilisant des associations obtenues statistiquement entre termes, termes-documents, et documents-documents [Salton, 1968]. Cette technique est basée sur l'hypothèse qu'il existe des relations qui peuvent être déterminées statistiquement entre les termes, entre les documents et entre les termes et les documents. Ces relations ou associations peuvent être représentées sous la forme d'une matrice de similarité. L'estimation quantitative des similarités entre les termes peut être obtenue, par exemple, par l'analyse statistique de la co-occurrence des termes dans les documents. Les similarités entre les documents peuvent être obtenues en évaluant, par exemple, les similarités dans la répartition des termes dans les documents, ou au moyen des co-citations et autres indicateurs bibliographiques. Il y a plusieurs hypothèses fortes sous-jacentes à ce type de modèles et plusieurs études ont montré qu'il est difficile de générer des méthodes d'expansion de requêtes valides pour des ensembles variés de documents [Preece, 1981, Salton & Buckley, 1988]. Ce constat est fondé sur plusieurs observations. Premièrement, les similarités statistiques calculées à partir de certains documents ou de certains termes, peuvent être uniquement valables localement pour certains domaines et pour certaines applications. Deuxièmement, la plupart des méthodes de calcul des degrés d'association entre les documents sont basées sur l'hypothèse que les termes ou les documents sont indépendants les uns des autres (sans corrélation). Une telle hypothèse n'est plus acceptée dans les nouvelles directions de recherche dans le domaine de la RI.

Récemment, ces modèles de recherche associative ont été corrigés en utilisant *le modèle de propagation d'activation (PA)* qui s'inspire des mécanismes de fonctionnement de la mémoire humaine. Ce modèle a son origine dans des études philosophiques [Rumelhart & Norman, 1983] et a été introduit en informatique dans le domaine de l'intelligence artificielle pour fournir un cadre de calcul aux réseaux sémantiques. Son utilisation a été étendue à plusieurs domaines tels que les sciences cognitives, les bases de données, la psychologie, la biologie, et plus récemment, la RI. Le modèle de PA de base a été sujet à plusieurs améliorations afin d'augmenter son efficacité dans plusieurs domaines d'application ; par ailleurs, la façon dont il est utilisé dans le domaine de la RI est différente de celle initialement utilisée dans le domaine de la psychologie. Dans les prochaines sections, nous décrivons en détail ce modèle.

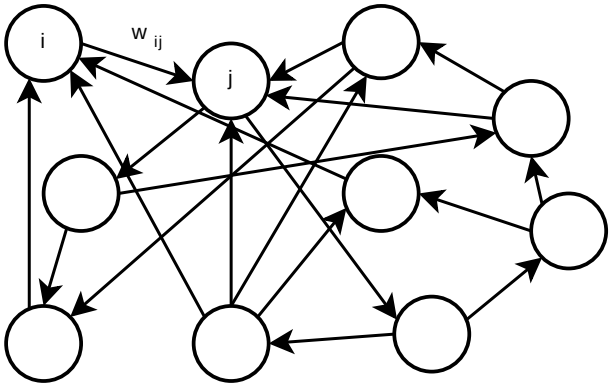


FIGURE 6.2 – Structure en réseau d'un modèle de propagation d'activation

6.3.1 Le modèle simple de propagation d'activation

Le modèle de PA dans sa forme de base est assez simple. Il repose sur une technique de calcul appliquée à une structure de données en réseau. Cette structure consiste en un ensemble

de nœuds connectés par des liens, comme montré dans la figure 6.2. Les nœuds représentent les objets ou les propriétés d'objets du "monde réel". Ils sont généralement étiquetés par les noms des objets qu'ils représentent. Les liens représentent les relations entre les nœuds et ils peuvent être étiquetés et/ou avoir un "poids" qui est une valeur numérique qui représente *la force* de la relation entre les deux nœuds. Un lien est généralement dirigé, et possède selon cette direction une étiquette et un poids donnés. Cette structure est très similaire à celle d'un réseau sémantique mais est plus générale que celle décrite dans la section 6.2.

Les étapes de calcul sont définies par une séquence d'itérations comme celle décrite sur le schéma de la figure 6.3. Chaque itération est suivie par une autre jusqu'à ce qu'elles soient arrêtées par l'utilisateur ou qu'elles déclenchent une condition d'arrêt. Une itération consiste en :

1. une ou plusieurs *pulsations* (pulses) ;
2. vérification des conditions d'arrêt.

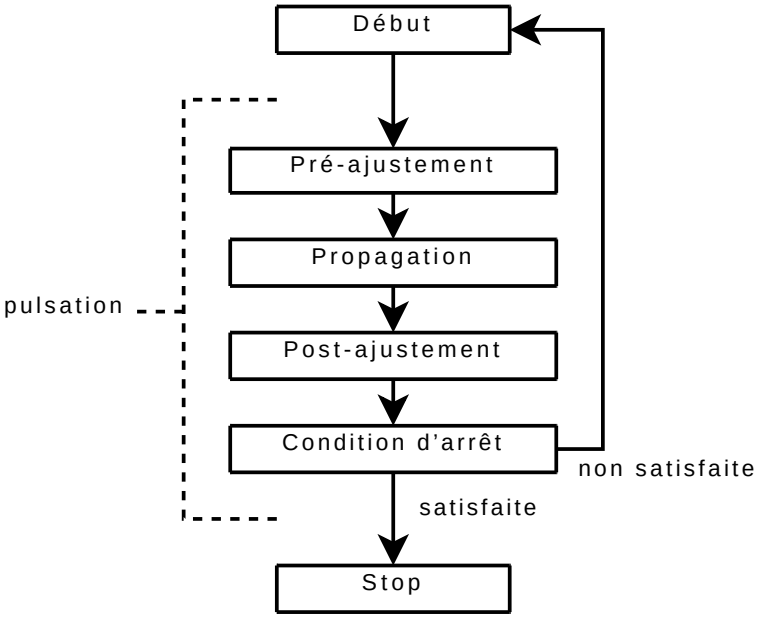


FIGURE 6.3 – Le modèle de propagation d'activation

Ce qui distingue un modèle simple de PA d'autres modèles plus complexes est la séquence d'actions qui composent une pulsation. Une pulsation est constituée de trois phases :

1. le pré-ajustement ;
2. la propagation ;
3. le post-ajustement.

Dans les phases de pré-ajustement et de post-ajustement, qui sont optionnelles, une forme d'atténuation de l'activation peut être appliquée aux nœuds activés. Ces phases sont utilisées pour éviter la rétention de l'activation des pulsations précédentes, ce qui permet de contrôler l'activation des nœuds ainsi que l'activation du réseau entier. Ces phases implémentent une forme de "perte d'intérêt" aux nœuds qui ne sont pas continuellement activés.

La phase de propagation consiste en un nombre de passages de pulsations d'activation d'un nœud à tous ceux qui lui sont connectés. Il existe plusieurs moyens pour propager une activation à travers un réseau. Dans sa forme la plus simple, pour une seule unité, la PA consiste premièrement

à calculer la valeur d'entrée en utilisant cette formule :

$$I_j = \sum_i O_i w_{ij}$$

où :
I_j est la somme des entrées du nœud j ;
O_i est la sortie du nœud i connecté au nœud j ;
w_{ij} est le poids associé au lien entre le nœud i et le nœud j.

Les entrées et les poids sont généralement des nombres réels mais ils peuvent être également, selon les besoins spécifiques de l'application, des valeurs binaires (0 ou 1), des valeurs excitatrices ou inhibitrices (+1 ou -1), ou des nombres réels reflétant la force de la relation entre les nœuds.

Après avoir déterminé la valeur d'entrée d'un nœud, il faut déterminer sa valeur de sortie. Le type numérique de la valeur de sortie est également déterminé par les besoins spécifiques de l'application. Les deux types les plus utilisés sont le type binaire actif/non-actif (0 ou 1) et le type réel. Dans les modèles de PA, on ne distingue généralement pas entre la valeur d'activation d'un nœud et sa valeur de sortie. Le niveau d'activation d'un nœud est sa valeur de sortie. Cette valeur est généralement calculée comme une fonction de la valeur d'entrée :

$$O_j = f(I_j)$$

Il existe plusieurs fonctions qui peuvent être utilisées dans l'évaluation de la valeur de sortie. Quelques exemples sont montrés dans la figure 6.4. La fonction la plus utilisée dans le modèle de base de PA est la fonction *seuil*. Elle est utilisée pour déterminer si le nœud j doit être considéré comme actif ou non. L'application de la fonction seuil à la formule précédente dans le cas des valeurs binaires donne le résultat suivant :

$$O_j = \begin{cases} 0 & I_j < k_j \\ 1 & I_j > k_j \end{cases}$$

où k_j est la valeur seuil pour le nœud j. La valeur seuil de la fonction d'activation dépend de l'application et peut varier d'un nœud à un autre, d'où l'utilisation de la notation k_j.

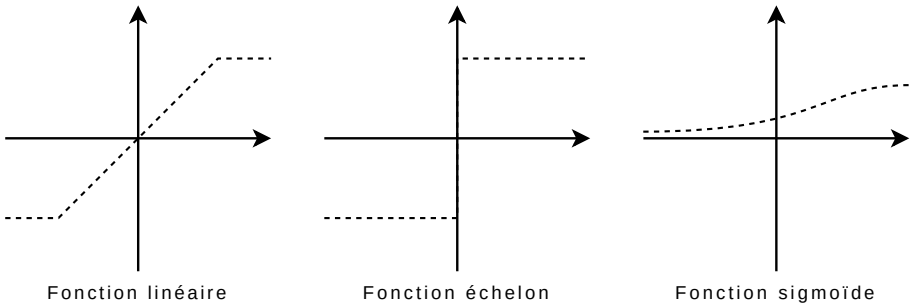


FIGURE 6.4 – Quelques exemples des fonctions d'activation les plus utilisées

Après le calcul de la valeur de sortie d'un nœud, cette valeur est propagée à tous ses nœuds voisins. Généralement, on envoie la même valeur à tous les nœuds. Pulsation après pulsation, l'activation se propage à travers le réseau, atteignant des nœuds éloignés de ceux activés initialement. Après un certain nombre de pulsations la condition d'arrêt est vérifiée. Si la condition est satisfaite, le processus de PA est arrêté, sinon, il enchaîne une autre itération de pulsations. La

PA est donc un processus itératif qui consiste en un ensemble de pulsations et de vérifications de la condition d'arrêt.

Le résultat de la PA sont les niveaux d'activation atteints par les différents nœuds à la fin du processus. L'interprétation du niveau d'activation de chaque nœud dépend de l'application et, en particulier, des caractéristiques de l'objet modélisé par ce nœud.

6.3.2 La propagation d'activation par contraintes

Le modèle simple de PA présente quelques inconvénients :
- A moins qu'elle ne soit soigneusement contrôlée par les phases de pré-ajustement et de post-ajustement, la propagation finit souvent par atteindre tout le réseau ;
- l'absence d'une utilisation efficace des informations fournies par les labels des associations (la sémantique des associations) ;
- la difficulté d'implémenter une forme d'inférence basée sur la sémantique des associations.

Ces problèmes peuvent trouver des solutions en tenant compte au cours du processus de la PA des différentes significations des relations entre les nœuds. Ceci peut être obtenu en utilisant les informations fournies par les labels des associations et en traitant différemment les liens selon leur sémantique. De cette manière, il est possible d'implémenter des heuristiques ou de propager l'activation selon des règles d'inférence. Un des moyens pour réaliser cela est d'imposer des contraintes au processus de propagation. Nous illustrons dans ce qui suit les contraintes les plus utilisées dans le modèle de PA.

La contrainte de distance

La PA doit cesser quand elle atteint des nœuds qui sont éloignés des nœuds initialement activés en termes de nombre de liens parcourus pour les atteindre. Ceci repose sur l'hypothèse que la force d'une relation entre deux nœuds décroît avec leur distance sémantique. Les relations peuvent être classées par rapport à leurs distances en termes de liens. Les relations entre deux nœuds directement connectés sont appelées des relations de premier ordre. Les relations entre deux nœuds connectés par l'intermédiaire d'un autre nœud sont appelées des relations de second ordre, et ainsi de suite.

La contrainte de connectivité

La PA doit cesser aux nœuds avec une grande connectivité, c'est-à-dire, les nœuds connectés à un grand nombre de nœuds. L'objectif de cette contrainte est d'éviter une large propagation due à des nœuds avec un sens sémantique étendu et donc connectés à plusieurs autres nœuds.

La contrainte de chemin

L'activation doit se propager en utilisant des chemins spécifiques qui reflètent les règles d'inférence propres à l'application. Les poids ou les étiquettes des liens peuvent être utilisés pour diriger la propagation selon des chemins spécifiques et éviter des chemins moins significatifs pour l'application.

La contrainte d'activation

En utilisant une fonction seuil au niveau du nœud, il est possible de contrôler la PA dans le réseau. On peut, par exemple, changer la valeur du seuil en fonction du niveau d'activation atteint

par le réseau entier à chaque pulsation. Il est également possible d'affecter différentes valeurs de seuil à chaque nœud ou groupe de nœuds. Cette technique peut augmenter considérablement le temps de calcul mais permet d'implémenter des règles d'inférence complexes.

Selon le modèle simple de PA montré dans la figure 6.3, ces contraintes peuvent être intégrées dans la phase de pré-ajustement (cas de la contrainte de distance, de connectivité et du chemin) ou durant la phase de post-ajustement (cas de la contrainte d'activation).

Un bon exemple de l'utilisation de cette technique dans le domaine de la RI est illustré dans [Kjeldsen & Cohen, 1987, Cohen & Kjeldsen, 1987]. Nous allons décrire plus en détail ce modèle dans la section 6.4.2, qui traite le système appelé GRANT.

6.3.3 La propagation d'activation avec feedback

Un autre moyen pour améliorer les performances du modèle de PA est d'utiliser un feedback provenant d'une source externe. Dans ce cas, une évaluation du niveau d'activation d'un ensemble de nœuds ou du réseau entier est difficile à obtenir automatiquement. Le feedback provient alors d'un processus externe ou de l'utilisateur du système. L'utilisateur évalue le niveau d'activation atteint par quelques nœuds et le modifie selon ses besoins. En plus, l'utilisateur peut désigner des chemins pour la PA différents de ceux désignés par la contrainte du chemin. Le processus de PA s'adapte alors aux besoins spécifiques de l'utilisateur.

Ce type de modèle est très utile dans des applications qui nécessitent un grand nombre de règles d'inférence à représenter sous forme de contraintes. Le feedback de l'utilisateur peut être soit utilisé durant la phase de pré-ajustement, ainsi l'utilisateur peut diriger la PA d'une pulsation, ou durant la phase de post-ajustement permettant ainsi à l'utilisateur d'évaluer le résultat de la PA d'une pulsation et de diriger la prochaine par rapport à cette évaluation. Quelques exemples d'utilisation de ce modèle sont donnés dans la section suivante.

6.4 L'utilisation de la propagation d'activation en RI

Les techniques de PA utilisées en RI sont basées sur l'existence de réseaux qui spécifient des relations entre des termes ou des documents. Les nœuds peuvent correspondre à des termes, des documents, des articles, des journaux ou des auteurs, etc. Il n'y a pas d'homogénéité dans le réseau. Un nœud peut représenter tout objet que l'on veut modéliser. Un lien représente une association entre un nœud et un autre nœud, comme par exemple, un auteur et un document qu'il a écrit ou un document et un autre qu'il cite. Un exemple de représentation d'une partie d'une collection de documents est montré dans la figure 6.5. L'ensemble des nœuds et des liens est déterminé par les données disponibles et l'objectif de l'application. Les relations entre les nœuds peuvent être exprimées par une paire de liens. Le lien *écriture*, par exemple, peut être exprimé par les deux liens *"écrit"* et *"écrit par"*. Ces deux liens connectent les mêmes nœuds, mais leurs directions sont inversées. Des règles de traitement spécifiques peuvent inhiber l'activation sur l'une des directions, les utiliser de différentes manières, ou leur associer différents poids avec différentes directions.

Étant donné cette structure de réseau, l'activation commence par donner un certain niveau d'activation à quelques nœuds de départ. Ces nœuds sont généralement identifiés par le résultat initial d'une requête, ou par les documents ou les termes trouvés par une opération précédente de recherche. L'activation atteint au début les nœuds voisins des nœuds initialement activés, et continue ainsi de se propager sur le réseau. Le niveau d'activation d'un nœud est calculé en utilisant l'une des fonctions spécifiées dans la section 6.3.1. Le processus se termine quand

une condition d'arrêt est rencontrée. Le niveau d'activation des documents reflète alors leur pertinence par rapport à la requête.

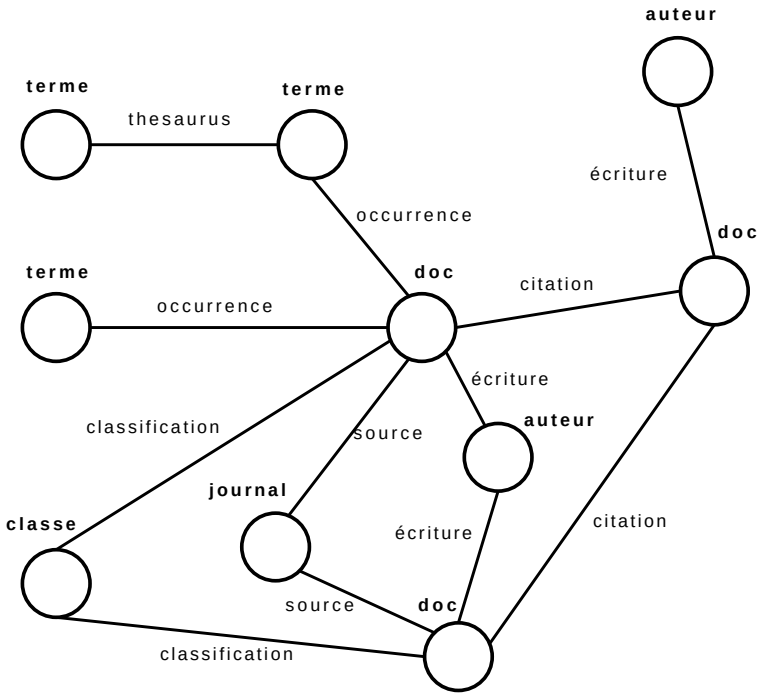


FIGURE 6.5 – Un exemple de représentation d'une partie d'une collection de documents

La plupart des techniques de PA utilisées en RI diffèrent de celle du modèle simple sur plusieurs aspects :

- Le niveau d'activation d'un nœud atteint par la PA est déterminé par le niveau d'activation initial et le type des nœuds et des liens traversés par l'activation avant de l'atteindre.
- Les contraintes de distance sont largement utilisées en arrêtant ou en dégradant l'activation à une distance spécifique du nœud original.
- Les nœuds avec une grande connectivité peuvent recevoir un traitement spécial dans le processus de PA, dans le but d'éviter de propager une large activation dans le réseau.

Cependant, la pertinence du processus dépend fortement de l'existence d'un réseau représentatif. Le problème de construction d'un réseau qui représente les relations nécessaires et utiles (en termes des objectifs de la RI) a toujours été le point critique des nombreuses tentatives de l'utilisation des techniques de PA en RI. Ces réseaux sont difficiles à construire et à maintenir. Leur construction nécessite une connaissance approfondie des domaines concernés par l'application, que seuls les experts peuvent fournir. En plus, leur construction est très coûteuse et nécessite un temps considérable. Ces problèmes sont la principale raison de l'intérêt croissant porté aux techniques de construction automatique de réseaux sémantiques et, en particulier, les techniques d'apprentissage automatique en RI.

Dans ce qui suit, nous présentons les principaux travaux réalisés dans le domaine de la propagation d'activation en RI.

6.4.1 Les premières étapes : les travaux de Preece et Shoval

Les travaux de *Preece*, décrits en détail dans [Preece, 1981], peuvent être considérés comme l'une des premières tentatives d'utilisation de la recherche associative par des techniques de PA. Dans sa thèse, il a examiné en profondeur l'approche de PA en RI associative. Il affirme que la plupart des techniques classiques de RI peuvent être simulées avec différentes techniques de PA sur un réseau de représentation de la collection des documents. Cette séparation entre la structure des données et le traitement des données peut être considérée comme la première tentative de modélisation conceptuelle dans des applications de RI. En combinant différentes structures de réseaux de données avec différentes techniques de calcul et en utilisant plusieurs formes de *poids* pour la recherche associative, il a montré comment il est possible d'implémenter le modèle booléen et le modèle vectoriel. En plus, il a montré comment, en utilisant la réinjection de pertinence, la PA peut être utilisée pour la classification automatique, l'indexation, et pour la construction de concepts. C'est spécialement pour les deux dernières applications que la PA a montré son plus grand potentiel. Ces idées viennent vraisemblablement de l'apparition des premiers papiers sur les réseaux de neurones et il est facile de voir l'intention d'étendre le paradigme de PA dans cette direction. Cependant, la capacité de calcul des machines vers la fin des années 70 a limité *Preece* dans ses recherches, et ne lui a pas permis d'utiliser les techniques de PA avec un paradigme d'apprentissage automatique. En effet, toutes les expériences qu'il avait effectuées pour tester son modèle ont été réalisées en utilisant une petite collection de documents, qui ne pouvait être comparée à la taille des collections utilisées dans des applications *"réelles"*. Un inconvénient majeur de ce modèle est la construction manuelle des réseaux sémantiques pour représenter la collection des documents. Cet inconvénient est commun à la plupart des applications présentées dans ce chapitre.

Le travail de *P. Shoval* [Shoval, 1981], développé en parallèle avec celui de *Preece*, est une tentative pour implémenter l'expansion interactive des requêtes en utilisant la PA. La base de connaissances utilisée par *Shoval* est un réseau sémantique basé sur un thésaurus. Les types de liens utilisés sont ceux utilisés traditionnellement dans un thésaurus, exprimant les relations hiérarchiques, les relations de synonymie et les relations d'apparentement. En plus de ces relations, deux types de liens ont été ajoutés : les liens générateurs, qui combinent les termes sources/-génériques à un concept multi-termes (ex. les termes "information" et "system" sont reliés au terme "information system"), et les liens du modèle, qui sont utilisés pour étendre la connaissance propre à un domaine particulier sur un concept (ex. "business" est lié à "oraganizational area"). La technique de calcul est une forme de PA avec feedback, où le feedback provient de l'utilisateur via un processus interactif. Les étapes d'un processus de PA sont :

1. Le système utilise les termes de l'utilisateur pour leur expansion sur le réseau sémantique.
2. Les nouveaux termes sont comparés entre eux. Si une intersection existe (cas de polysémie par exemple), il est d'abord vérifié que cette intersection inclut de nouveaux termes d'entrée. Si c'est le cas, il est demandé à l'utilisateur de juger si les nouveaux termes sont adéquats à sa requête ou non.
3. Si l'utilisateur rejette ces termes, la recherche dans l'une ou l'autre des directions est arrêtée et les termes sont marqués pour qu'ils ne puissent pas être atteints lors des prochaines pulsations, sinon le terme est à nouveau étendu.
4. Tant qu'il y a d'autres intersections le processus continue comme en 2.
5. Finalement, le système collecte les termes suggérés et les affiche à l'utilisateur qui pour chaque terme a le choix de l'accepter (et dans ce cas il peut être utilisé pour étendre plus les termes), de le rejeter (cas pour lequel le système va essayer de trouver des termes

alternatifs à étendre), ou de demander pourquoi il a été suggéré (dans ce cas le système affiche le chemin emprunté pour atteindre ce terme à partir des termes initiaux).

Ce système a l'avantage d'utiliser des thésaurus qui peuvent être génériques, et donc indépendants d'un domaine particulier. Cependant, la technique de PA utilisée est assez simpliste et nécessite un retour constant de la part de l'utilisateur. L'utilisateur est sollicité pour chaque intersection trouvée, ce qui peut vite devenir une tâche lourde à gérer.

6.4.2 Le système GRANT

Le système GRANT du *P.R. Cohen* et *R. Kjeldsen* est l'un des premiers systèmes à utiliser la PA par contraintes dans la RI [Cohen & Kjeldsen, 1987, Kjeldsen & Cohen, 1987]. Dans GRANT, la connaissance sur les propositions de recherche et les agences potentielles de financement est organisée en utilisant un réseau sémantique. Les thèmes de recherche et les agences sont connectés en utilisant plusieurs types de liens. Une requête peut concerner un ou plusieurs sujets de recherche, ou une ou plusieurs agences de financement. La recherche est réalisée par la technique de PA par contraintes, en utilisant presque tous les types de contraintes décrits dans la section 6.3.2, en particulier, les contraintes de chemin.

D'un point de vue heuristique, le système GRANT peut être considéré comme un système d'inférence qui applique itérativement un seul schéma d'inférence :

$$\text{Si } x \text{ et } R(x, y) \rightarrow y$$

où $R(x, y)$ est un chemin connectant les deux nœuds x et y et qui peut contenir un ou plusieurs liens. Pour l'application en question, cela est équivalent à la règle d'inférence de type : "si une agence de financement est intéressée par le thème de recherche x, et qu'il existe une relation entre le thème de recherche x et le thème de recherche y, alors l'agence est probablement intéressée par le thème y". Les contraintes utilisées pour la PA permettent d'établir des préférences pour certains chemins en leur attribuant des valeurs positives, et d'en éviter certains en leur attribuant des valeurs négatives. Le mécanisme d'évaluation des chemins permet de classer les nœuds trouvés.

L'utilisation des contraintes pour la PA et des règles de *"préférence"* pour des chemins particuliers a permis au système d'obtenir des résultats très intéressants. En pratique, les auteurs ont démontré que l'utilisation de la PA par contraintes pour cette application donnait des valeurs raisonnables de rappel et précision. Ces valeurs étaient meilleures que celles données par une recherche à base de mots-clés. Cette technique s'est avérée également intéressante pour *"les cas difficiles"* où même un expert humain aurait des difficultés à donner des réponses.

Développer un système comme GRANT nécessiterait un grand effort d'ingénierie de connaissance pour construire le réseau sémantique. Cet effort consiste en une analyse profonde du domaine concerné par l'application afin de déterminer les concepts et les relations appropriés ainsi que les *"préférences"* à attribuer à certains chemins sur le réseau.

6.4.3 Le système I³R

L'objectif principal de *W.B. Croft*, *T.J. Lucia*, *J. Crigean* et *P. Willet*, en développant le système I³R, était principalement d'étudier la possibilité de retrouver des documents par une "une inférence plausible" [Croft et al., 1989]. Le système a été développé pour accomplir des tâches de recherche intermédiaire. Il accomplit ces tâches en utilisant la connaissance du domaine pour affiner la description des requêtes, initier des stratégies de recherche appropriées, assister les utilisateurs dans l'évaluation des résultats de leurs recherches, et à reformuler leurs requêtes. Dans sa version initiale, la connaissance du domaine était représentée en utilisant un arbre de

ET/OU de concepts. Les documents ont été représentés par des descripteurs de termes uniques. Le système utilise la connaissance du domaine pour inférer des concepts qui sont reliés à ceux mentionnés dans la requête. Le mécanisme d'inférence utilise une forme de "propagation de certitude" sur le réseau des concepts. Dans une version plus récente, la représentation de la connaissance ressemblait plus dans sa structure à un réseau sémantique, comme celui montré dans la figure 6.6. Dans ce type de structure, il n'est pas nécessaire de faire la distinction entre les concepts, les termes et les documents. Ils sont tous considérés comme des nœuds et ils agissent de la même manière que ceux décrits dans la section 6.3.1. La technique de PA par contraintes utilisée est la suivante :

1. Les points de départ de la PA sont les documents les mieux classés obtenus par une recherche probabiliste.
2. Initialement, les liens de type *"le voisin le plus proche"* et *"citation"* sont utilisés pour la PA ; ces liens représentent les relations les plus fortes entre les documents.
3. Dans les cycles suivants de PA seuls les liens de type *"le voisin le plus proche"* sont utilisés ; la relation *"citation"* est intéressante seulement avec les premiers documents.
4. Les poids des relations sont utilisés dans l'évaluation des niveaux d'activation des nœuds ; ils sont spécifiés comme des valeurs de *"crédibilité"* associées aux règles d'inférence représentant l'existence d'une relation entre deux nœuds.
5. Les documents utilisés comme une partie d'un chemin d'activation ne sont pas utilisés une prochaine fois s'ils sont réactivés.

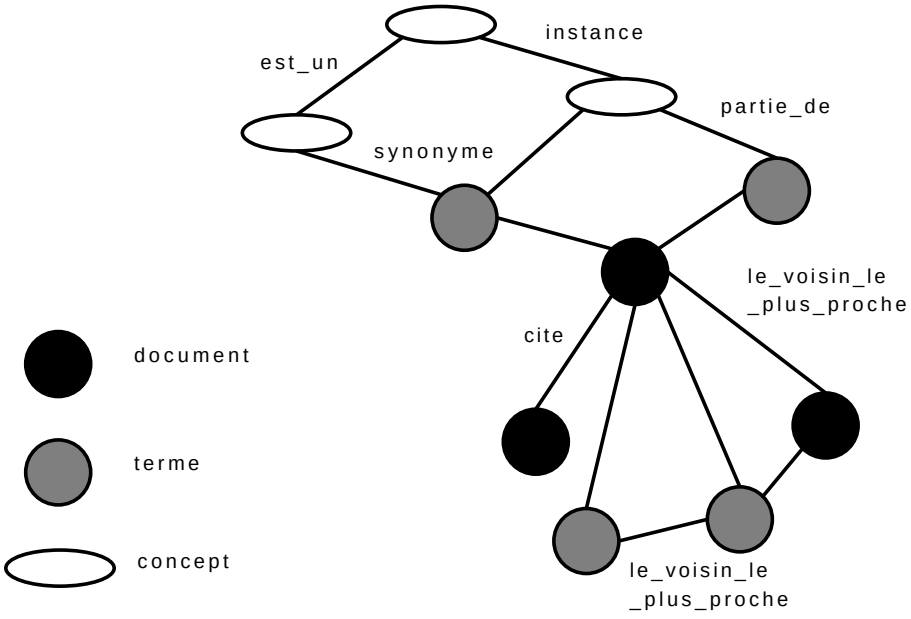


FIGURE 6.6 – La représentation du réseau sémantique dans I^{3}R

Les auteurs ont implémenté un paradigme de recherche appelé *"sources multiples de preuves"* en utilisant ces contraintes sur le modèle de PA de base. Ce paradigme se fonde sur l'hypothèse qu'un document est probablement pertinent si sa pertinence est obtenue à partir de différents indices ou preuves.

Les expériences ont montré la possibilité d'améliorer les performances d'un SRI générique en se basant uniquement sur l'utilisation des relations *"le voisin le plus proche"* et *"citation"*.

Cependant, l'ampleur de l'amélioration varie d'une collection de documents à une autre, montrant ainsi la difficulté de l'utilisation de la PA dans un SRI opérationnel. Néanmoins, ce travail peut être considéré comme l'une des meilleures tentatives de combiner un modèle de PA par contraintes avec des techniques probabilistes de RI.

6.4.4 L'expérience de l'Université Cornell

G. Salton et *C. Buckley* de l'Université Cornell ont décrit dans un article très intéressant une comparaison évaluative de quelques modèles de PA avec le modèle vectoriel [Salton & Buckley, 1988]. En particulier, l'efficacité d'un modèle de PA est évaluée et comparée à celle d'un modèle associatif linéaire basé sur le calcul vectoriel dans un espace vectoriel. Le modèle de PA utilisé dans cette évaluation est une version améliorée du modèle de PA simple, permettant la couverture de seulement deux liens en partant des nœuds initialement activés. La structure du réseau utilisé par ce modèle est aussi assez simple comme le montre l'exemple illustré par la figure 6.7. Sur ce type de réseau simple, les auteurs ont utilisé un modèle de PA simple amélioré par des facteurs de normalisation sur la fonction d'activation. Les poids des liens sont déterminés en utilisant la fréquence des termes qui mesure la fréquence d'occurrences d'un terme dans un document ou dans une requête. Ce modèle de PA est comparé à un modèle vectoriel qui bénéficie de plusieurs années d'expérimentation, et où les facteurs de normalisation ont été minutieusement établis afin d'obtenir les meilleures performances. En plus, le modèle vectoriel utilise la mesure IDF qui mesure l'importance relative d'un terme dans une collection de documents.

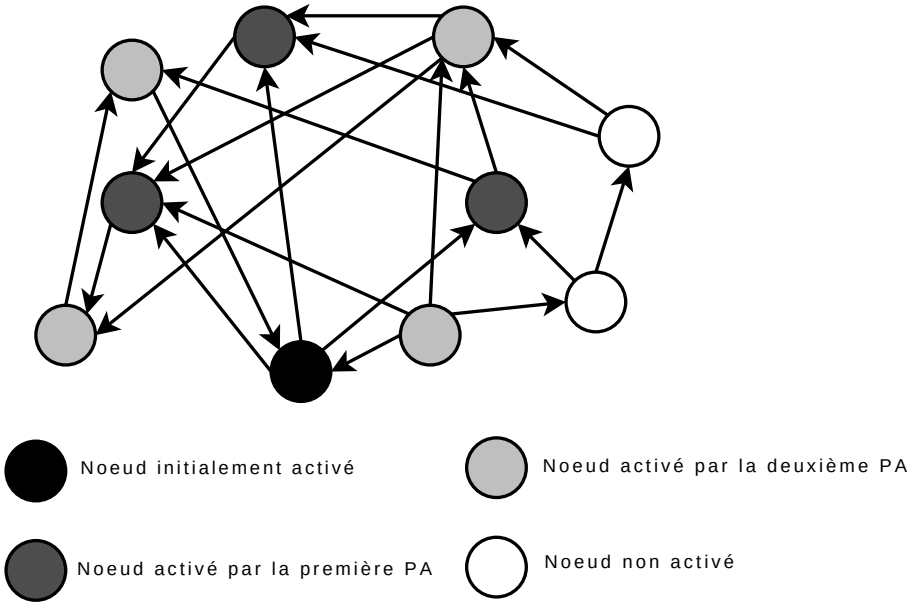


FIGURE 6.7 – Le réseau utilisé par Salton et Buckley pour les évaluations

Plusieurs collections tests et différents schémas de calcul de poids ont été utilisés pour l'évaluation. Cependant, malgré le grand effort fourni pour réaliser ce travail d'évaluation, on doit noter que le modèle de PA utilisé dans ce travail est un modèle très réduit de la PA et qu'il n'est pas surprenant que l'évaluation ait été plus favorable au modèle vectoriel.

6.4.5 La propagation d'activation et le connexionnisme

Le système *AIRS* (Associated Information Retrieval System) de *H. Kimoto* et *T. Iwadera* est un système de RI qui introduit les techniques de PA dans un thésaurus dynamique. Le concept principal de *AIRS* qui le distingue des autres systèmes est qu'il détermine le besoin de l'utilisateur à partir d'un échantillon de documents pertinents choisis par ce dernier pour produire *"les termes de l'information"*. Ces derniers sont utilisés pour construire un thésaurus dynamique qui génère, au temps d'exécution de la recherche, les mots-clés associés.

Le thésaurus dynamique est représenté comme un réseau où les nœuds représentent les termes et les liens les relations sémantiques entre les termes. Le réseau est obtenu en combinant les informations provenant d'un thésaurus statique et des termes de l'information. Les termes de l'information sont obtenus en classant les termes selon leurs fréquences et leurs positions dans un ensemble de documents pertinents fournis par l'utilisateur. En particulier, les co-occurrences des termes sont utilisées pour construire de nouveaux liens entre les nœuds dans le thésaurus dynamique. La structure du réseau est assez particulière et elle diffère de celles des réseaux connexionnistes classiques. En fait, les liens ne possèdent pas un poids mais ce sont les nœuds qui en ont un. Ces poids sont obtenus à partir des termes de l'information.

La recherche des documents pertinents pour une requête donnée est réalisée par un processus similaire à celui de la propagation d'activation par contraintes. L'ensemble des termes utilisé dans une requête est élargi en ajoutant des termes associés selon leurs poids et leurs liens. Des contraintes ont été établies afin de limiter la distance de la propagation d'activation sur le réseau et seuls les termes avec des poids supérieurs à un seuil prédéfini sont utilisés comme des termes de recherche associés.

Plusieurs expériences ont été réalisées, mais avec une petite collection de documents ; elles ont montré que la performance de ce modèle est meilleure que celle obtenue par un thésaurus statique. Cependant, il existe quelques inconvénients à leur approche. Le premier inconvénient est conceptuel : les auteurs prétendent utiliser une approche connexionniste. Cependant, une approche connexionniste basée sur l'utilisation d'un réseau de neurones serait assez différente. Ce qui distingue un réseau de neurones de la PA est principalement la présence d'une fonction d'activation non-linéaire et, en particulier, la présence d'une phase d'apprentissage qui est utilisée pour modifier les poids des liens ; ainsi la propagation d'activation sur le réseau reflétera un certain pattern désiré. Dans le système *AIRS*, il n'existe aucune fonction d'activation pour les nœuds et il n'y a pas de phase d'apprentissage. Si cette approche doit être considérée comme une approche connexionniste, elle devient une approche très pauvre dans ce domaine. Plusieurs recherches se sont intéressées à l'application des réseaux de neurones en RI [Dominich, 2002]. L'approche de *Kimoto* et *Iwadera* est plutôt une approche de PA que de connexionnisme. Un autre point est que ce modèle est assez similaire aux modèles statistiques classiques, et par conséquent des performances similaires aux autres approches (comme celle de *l'Université de Cornell*) sont attendues pour ce modèle. Finalement, la typologie complexe des relations entre termes fournies par le thésaurus "classique" n'est pas du tout utilisée. Tous les types des relations sont représentés en utilisant un seul type de lien, ressemblant ainsi plus à un réseau associatif qu'à un thésaurus.

6.5 Conclusion

Les applications présentées dans ce chapitre peuvent être considérées comme les principales applications de l'approche de PA en RI. Ce domaine de recherche s'est avéré très intéressant pour la RI et capable de fournir de bons résultats. Avec la multiplication des ressources sémantiques

propres aux domaines et les techniques automatiques de leur construction, la construction de réseaux sémantiques est devenue plus facile et la PA peut devenir une solution très intéressante pour répondre aux problèmes posés par la RI.

Chapitre 7

Application de l'approche de propagation d'activation pour la reformulation des requêtes

7.1 Introduction

La quantité d'information médicale disponible au grand public a considérablement augmenté ces dernières années. Pourtant, le processus de recherche d'information pour les usagers de santé peut être particulièrement complexe car ils ignorent souvent le terme médical à utiliser ou comment les concepts médicaux sont inter-reliés.

Les études récentes ont constaté que le grand public est de plus en plus à la recherche d'information médicale en utilisant des bases de données bibliographiques (comme Medline) et des sites Web spécialisés sur Internet. Le résultat de cette demande croissante a été "l'explosion" des informations médicales destinées aux usagers de santé [Miller et al., 2000].

Même si les usagers de santé sont à la recherche d'information médicale, ils ne sont pas nécessairement capables de trouver ce qu'ils cherchent. Les requêtes entrées par les usagers de santé aux systèmes de recherche d'information médicale sont souvent infructueuses ou insatisfaisantes [Tolle & Chen, 2000]. En plus des problèmes traditionnels que les utilisateurs rencontrent quand ils essaient de trouver des documents pertinents, les usagers de santé affrontent une difficulté supplémentaire liée au fait qu'ils ne sont pas des spécialistes, alors que la médecine est un domaine complexe qui nécessite des connaissances avancées.

Les interfaces visuelles qui utilisent les relations sémantiques pour explorer le contexte qui entoure la requête d'un usager de santé pourrait aider l'utilisateur dans la construction d'une requête qui mène à une recherche fructueuse [Slaughter, 2002]. Dans ce chapitre, nous explorons l'utilisation de l'ontologie comme structure de réseau sémantique pour l'expansion et la reformulation des requêtes des usagers de santé dans le domaine du cancer du sein. L'approche de propagation d'activation a été adoptée pour inférer de nouveaux concepts à partir de l'ensemble initial de concepts qui se trouve dans la requête d'un utilisateur. L'approche est inspirée des travaux de *Berger* et sert avant tout à ouvrir des voies de recherche dans cette direction [Berger et al., 2003]. L'originalité de notre travail réside dans l'utilisation d'une ontologie avec des mesures d'association basées sur des calculs sur des corpus de textes d'usagers de santé. Ce travail a des retombées sur la construction des interfaces utilisateurs qui améliorent la capacité des usagers de santé à définir leur besoin d'information.

7.2 Structure du réseau sémantique

La première étape de ce travail consiste à convertir l'ontologie construite dans les étapes précédentes en un réseau sémantique exploitable par la méthode de propagation d'activation. Comme nous avons expliqué dans le chapitre précédent, la méthode de propagation d'activation dans un réseau sémantique s'appuie sur deux éléments principaux : les nœuds qui désignent les concepts et les arcs qui désignent les relations.

7.2.1 Les concepts de l'ontologie

Les concepts de l'ontologie représentent les nœuds dans le réseau sémantique utilisé pour la propagation d'activation. Comme les noms des concepts utilisés dans Protégé sont uniques, ils sont utilisés comme étiquettes pour les nœuds dans le réseau sémantique. Les nœuds possèdent également un identifiant numérique pour faciliter l'accès et les calculs par la suite.

7.2.2 Les relations de l'ontologie

Les relations dans le réseau sémantique possèdent une étiquette et un poids numérique qui dénote la "force" de la relation entre deux concepts. Dans notre ontologie, deux concepts peuvent être reliés par plusieurs relations. Dans le réseau sémantique, les relations sont fusionnées en une seule et deviennent un arc entre deux nœuds qui a comme étiquette *"Rel_ id1_ id2"*, où *id1* représente l'identifiant du premier nœud et *id2* représente l'identifiant du deuxième nœud.

Dans le réseau sémantique, nous distinguons trois types de liens :

Les liens générateurs : inspirés des travaux de Shoval [Shoval, 1981], ces liens relient des concepts du domaine aux concepts "généraux". Par "général" on désigne des concepts qui ne font pas partie directement du domaine concerné par l'application. L'objectif de ces liens et de lier un concept du domaine aux différents concepts généraux qui le génèrent. Par exemple, le concept *Alopécie* (chute de cheveux) et les concepts *Cheveux* et *Chute*. Par ce choix, nous essayons de simuler le comportement des usagers de santé qui se basent beaucoup sur la description des concepts et non pas sur leur définition. Ainsi, nous pouvons identifier les concepts du domaine par l'ensemble des concepts généraux qui les décrivent dans la requête. Le calcul des poids de ces liens se fait pour l'instant manuellement. Si un concept i est relié à n concepts générateurs, le poids de chaque lien reliant le concept i à un concept générateur est égal à $\frac{1}{n}$.

Les liens du domaine : les poids de ces liens sont obtenus automatiquement par une mesure d'association calculée sur la base des collocations entre concepts. Les mesures d'association permettent de quantifier l'information partagée par des couples de termes ou de concepts et de repérer les groupes de mots qui apparaissent ensemble plus fréquemment que ne le voudrait le hasard. Le détail de cette méthode est donnée dans la section 7.2.2.

Les liens 0 : la dernière cagtégorie concerne les liens qui sont restés sans poids attribué après les deux processus. Plusieurs liens sont dans ce cas car il n'existait pas dans le corpus une co-occurrence entre les concepts qu'ils relient. Nous n'avons traité manuellement qu'une partie de ces liens. Nous avons examiné les co-occurrences avec les concepts voisins pour attribuer des poids à ces liens. Par exemple, le concept *Chimiothérapie* est lié au concept *Effet_ Indésirable_ Chimiothérapie* par la relation *Entraine*. Ce lien n'a pas d'occurrence tel qu'il

est dans le corpus et possède par conséquent un poids 0. Dans ce cas nous examinons, la co-occurrence entre le concept *Chimiothérapie* et les différents concepts subsumés par le concept *Effet_ Indésirable_ Chimiothérapie*. Par la suite, le maximum entre ces poids est attribué au lien reliant le concept *Chimiothérapie* au concept *Effet_ Indésirable_ Chimiothérapie*. Les liens reliant le concept *Effet_ Indésirable_ Chimiothérapie* aux concepts désignant les effets de la chimiothérapie se voient attribuer le poids entre le concept *Chimiothérapie* et le concept qui désigne l'effet en question. La figure 7.1 montre un exemple de ce processus de calcul.

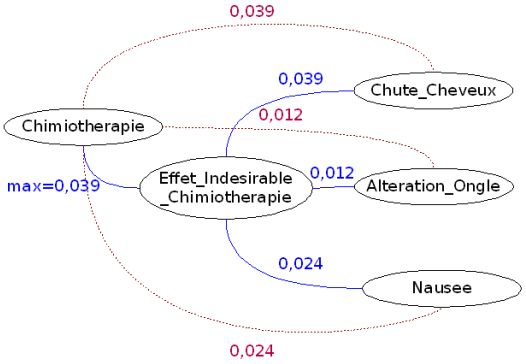


FIGURE 7.1 – Calcul du poids des liens 0

Les mesures d'association

Il existe beaucoup de formules différentes pour le calcul des mesures d'association : S. Evert en détaille plus d'une vingtaine sur son site `www.collocations.de`. Leur calcul se base généralement sur des tables de contingence semblables à la Table 7.1. Cette table de contingence contient les effectifs observés O pour les couples de mots apparaissant dans un contexte donné (co-occurrence directe, phrase, etc.). Les effectifs sont mesurés pour les couples de mots qu'il est possible de former à partir de deux mots x et y et l'ensemble des autres mots du corpus. L'effectif observé pour le couple de mots xy est noté O_{11}, celui du couple $\neg xy$ est O_{21}, etc. La taille du contexte utilisée pour leur calcul est variable, même si pour l'acquisition de termes on ne prend généralement en compte que la co-occurrence directe (mots adjacents).

	$Y = y$	$Y \neq y$
$X = x$	O_{11} $f(x, y)$	O_{12} $f(x, \neg y)$ $f_1(x) - f(x, y)$
$X \neq x$	O_{21} $f(\neg x, y)$ $f_2(y) - f(x, y)$	O_{22} $f(\neg x, \neg y)$ $N - f_1(x) - f_2(y) + f(x, y)$

TABLE 7.1 – Table de contingence pour deux éléments x et y. N correspond au nombre de tokens, $f_1(x)$ au nombre d'occurrences de x en première position dans le couple et $f_2(y)$ au nombre d'occurrences de y en deuxième position dans le couple.

L'idée sous-jacente à ces travaux est que le sens d'un terme est lié à ses contextes d'utilisation. Ainsi, les termes sont considérés comme sémantiquement proches s'ils apparaissent dans

des contextes similaires. Ces méthodes rejoignent ainsi les théories qui mettent l'accent sur l'importance de l'usage pour la sémantique lexicale, telles que celles du linguiste *Firth* ("You shall know a word by the company it keeps") et du philosophe *Wittgenstein* ("Meaning is use").

Différents niveaux de relations de co-occurrence peuvent être considérés :

- **Co-occurrences de 1er ordre** ou co-occurrences directes : deux mots sont considérés comme proches s'ils apparaissent dans le même contexte, c'est-à-dire s'ils sont directement co-occurrents (relation syntagmatique).
- **Co-occurrences de 2nd ordre** ou co-occurrences indirectes : deux mots sont similaires s'ils apparaissent dans des contextes similaires (relation paradigmatique). Ainsi deux mots M_1 et M_2 pourront être considérés comme sémantiquement proches s'ils partagent des co-occurrents M_i et ce même s'ils n'apparaissent jamais dans le même contexte [Denhière & Lemaire, 2003, Martinez, 2000, Rapp, 2003].

Pour notre travail, nous avons utilisé une mesure de co-occurrence de 1er ordre : *la fréquence jointe normalisée* [Phillips, 1989]. Le corpus des usagers de santé est utilisé pour le calcul de ces mesures. Ainsi, notre hypothèse de travail suppose qu'il existe une forte relation entre deux concepts s'ils apparaissent dans le même message. L'avantage de ce corpus est qu'il est constitué de messages d'usagers de santé qui sont relativement courts ; il définit ainsi naturellement le contexte pour la recherche des co-occurrences (un message). La co-occurrence entre deux concepts est calculée sur la base des co-occurrences entre les termes qui désignent chaque concept. La formule utilisée pour le calcul de ces mesures est la suivante :

$$\text{Fréquence jointe normalisée} = \frac{f_{ij}}{f_i + f_j - f_{ij}}$$

où :

- f_{ij} est le nombre de fois où les deux concepts apparaissent ensemble dans le même message,
- f_i est la fréquence du premier concept dans le corpus,
- f_j est la fréquence du deuxième concept dans le corpus.

La méthode d'identification des concepts au sein des messages des utilisateurs est décrite dans la section 7.3.

7.2.3 Conversion de l'ontologie

Le module de propagation d'activation a été implémenté en Java en utilisant la librairie *SpreadingActivation* développée par *S. Reed* [1]. L'ontologie a été convertie pour qu'elle puisse être utilisée par la librairie Java. Le réseau sémantique est stocké directement dans des classes Java qui contiennent les descriptions des nœuds et des liens. L'API *OWL API* a été utilisée pour parser l'ontologie et la réécrire au format utilisable par la librairie *SpreadingActivation*.

Le but du travail étant de tester les techniques de propagation d'activation pour la reformulation des requêtes en utilisant l'ontologie, nous avons conscience que le choix technique de stocker le réseau sémantique issu de l'ontologie dans des classes Java n'est pas le plus adapté pour qu'il soit implémenté dans un SRI "réel". La structure du réseau sémantique doit être en principe sauvegardée dans un document à part pour faciliter sa réutilisation et sa mise à jour.

1. `http://www.texai.org`

7.3 Identification des concepts de l'ontologie

Dans cette section, nous décrivons la technique adoptée pour l'identification des concepts dans les messages du corpus et dans les requêtes des usagers. L'approche adoptée est celle développée dans le travail de [Diallo, 2006]. Cependant, nous n'avons pas une phase de désambiguïsation en cas de termes polysémiques dans notre processus. Les travaux qui se sont intéressés à la désambiguïsation des termes concernent des travaux sur toute la langue et non pas dans des domaines restreints. La limitation du domaine de l'application réduit considérablement les cas de polysémie. De plus, nous estimons qu'un message offre peu de contextes pour désambiguïser efficacement un terme. Par conséquent, dans le cas des termes polysémiques, tous les concepts qui sont désignés par ce terme sont considérés et leurs occurrences se voient incrémentées. En ce qui concerne la phase de reformulation des requêtes, le processus de désambiguïsation est implicitement implémenté par l'approche.

Le processus procède en deux étapes : la première étape consiste en l'indexation basée sur les mots simples de la source textuelle à traiter (le message) et la seconde consiste en l'indexation conceptuelle à partir de l'ontologie. Le module qui effectue l'indexation basée sur les termes simples et sur les concepts ainsi que le module de recherche utilisent l'API Lucene de la fondation Apache [1].

7.3.1 Indexation basée sur les mots

Le processus d'indexation à base de mots simples est détaillé dans la figure 7.2. Ce processus suit l'approche classique d'indexation d'une collection de documents suivant le modèle vectoriel. En entrée du processus nous avons les documents à traiter. Ces documents passent par un processus de prétraitement (2) dont le but est d'effectuer un ensemble d'opérations de préparation des documents (tokenisation, purge des mots vides, désuffixation). Nous utilisons une liste de mots vides fournie avec l'API Lucene. L'algorithme de désuffixation utilisé est une adapatation de l'algorithme de Porter pour l'anglais. Après l'étape (2) nous obtenons une liste de mots constituant le vocabulaire de la source textuelle. C'est ce vocabulaire qui servira à représenter les vecteurs des documents. Le calcul des fréquences et des positions des mots dans chaque document est effectué à l'étape (3). Nous obtenons enfin la représentation des documents comme montrée sur la figure 7.2. La position des mots est importante car elle permet par la suite de repérer les termes dénotant les concepts dans les documents lors de la phase d'indexation conceptuelle.

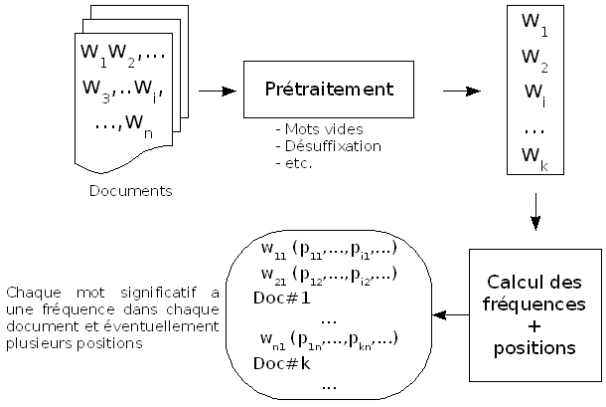


FIGURE 7.2 – Caractérisation de documents basée mots

1. lucene.apache.org

7.3.2 Indexation conceptuelle

L'indexation conceptuelle est réalisée en projetant sur les documents l'ontologie utilisée pour la caractérisation sémantique. L'ontologie est projetée sur la représentation des documents obtenue à l'étape précédente. Le détail de l'approche d'indexation conceptuelle, qui utilise le modèle d'ontologies décrit à la section suivante, est fourni par la figure 7.3.

Modèle de l'ontologie pour l'indexation conceptuelle

L'ontologie à utiliser pour l'indexation conceptuelle est représentée suivant un modèle défini de la façon suivante. Une ontologie O est constituée de l'ensemble C, Sub, lex, $RelLex$ où :

1. C est l'ensemble de concepts de l'ontologie,
2. Sub est la relation de subsumption entre concepts,
3. Lex est le lexique associé aux concepts de l'ontologie,
4. $RelLex$ est une fonction qui associe un élément de Lex à un concept de l'ontologie.

Chaque concept de l'ontologie est représenté par un ou plusieurs termes. Un exemple d'un extrait de la définition d'un concept issu de l'ontologie du cancer du sein développée est donné par le listing 7.1.

```
<owl:Class rdf:about="#Examen_Par_Mammographie">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom rdf:resource="#Sein"/>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="A_Comme_Position"/>
        </owl:onProperty>
      </owl:Restriction>
    </rdfs:subClassOf>
    <owl:disjointWith rdf:resource="#Examen_Par_Radiographie_Du_Poumon"/>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#Examen_Par_Radiographie"/>
    </rdfs:subClassOf>
    <rdfs:label xml:lang="fr">Examen du sein par radiographie</rdfs:label>
    <rdfs:label xml:lang="fr">Examen par mammographie</rdfs:label>
    <rdfs:label xml:lang="fr">Examen par mamo</rdfs:label>
    ...
    <rdfs:label xml:lang="fr">Examen par mammo</rdfs:label>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom>
          <owl:Class rdf:ID="Mammogramme"/>
        </owl:someValuesFrom>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="Produit"/>
        </owl:onProperty>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:label xml:lang="fr">Examen du sein par radio</rdfs:label>
  </owl:Class>
```

Listing 7.1 – Extrait de la définition d'un concept de l'ontologie

Projeter une ontologie sur un document

L'ontologie est projetée sur la représentation de la dite source afin d'identifier les concepts de l'ontologie. On procède par itération sur chaque document représenté. Les concepts eux-mêmes ne sont pas présents explicitement dans les documents, mais à travers les termes qui les dénotent. Il faut à chaque fois identifier le nombre de termes d'un concept présents dans un document donné. Tout d'abord, chaque terme subit la même opération de prétraitement que dans l'étape précédente. Ensuite une recherche dans la représentation du document est effectuée afin de vérifier si la suite de mots du terme de l'ontologie y est présente. Si le résultat est positif, le nombre d'occurrences du concept pour le document est incrémenté et la position du concept est notée. A la fin du processus, la fréquence de chaque concept pour chaque document est calculée afin d'obtenir une représentation de la source textuelle basée sur les concepts.

Il arrive que deux concepts soient désignés par deux termes qui se recouvrent. Cette situation peut se présenter lorsque l'un des concepts spécialise l'autre (exemple du cas du concept *Chimiothérapie* désigné par le terme "chimiothérapie" et du concept *Chimiothérapie_ Adjuvante* dont un des termes représentants est "chimiothérapie adjuvante"). L'algorithme peut détecter la présence de ces deux concepts à la même position, même si en réalité il s'agit uniquement du concept *Chimiothérapie_ Adjuvante*. Les positions des concepts sont alors utilisées pour ne garder que les concepts les plus spécifiques de la hiérarchie.

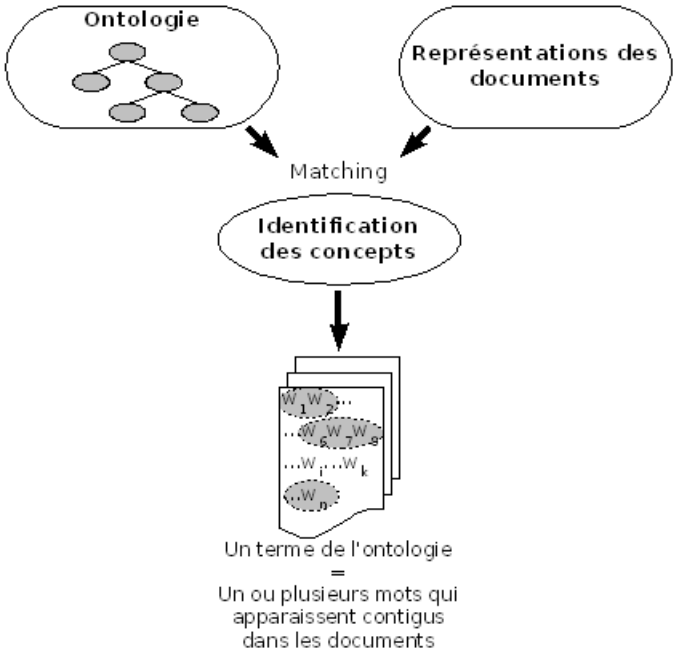


FIGURE 7.3 – projection de l'ontologie sur un ensemble de documents

7.4 La propagation d'activation pour l'extension des requêtes

Comme nous l'avons expliqué, la propagation d'activation consiste à propager une certaine quantité d'énergie entre les nœuds pendant un certain nombre de pas. Initialement, seuls les nœuds qui représentent les concepts de la question sont activés, les autres nœuds reçoivent une activation égale à 0. Ensuite, chaque nœud activé envoie son activation et la propage à travers

le réseau.

Pendant chaque cycle de la propagation d'activation les nœuds reçoivent l'activation de tous les nœuds auxquels ils sont reliés, sans tenir compte du type du lien. Des variables d'état liées aux nœuds et aux liens permettent de contrôler la direction de la propagation pour qu'elle ne se propage pas plusieurs fois sur le même chemin. Finalement, la propagation est arrêtée quand une des contraintes de contrôle de l'activation est vérifiée. Une fois la propagation d'activation terminée et selon notre hypothèse de recherche, les nœuds les plus activés seront les plus représentatifs de la requête de l'usager de santé. La figure 7.4 représente les différentes étapes de ce processus.

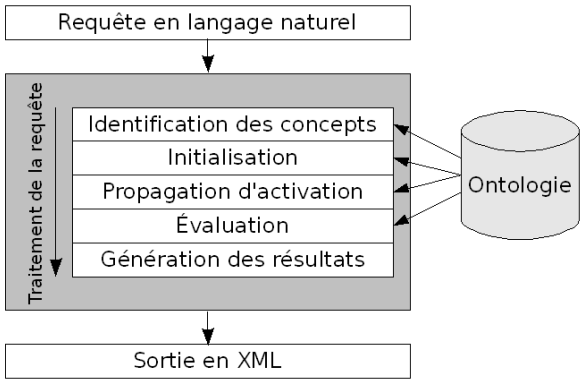


FIGURE 7.4 – Architecture de l'outil de reformulation des requêtes

7.4.1 Activation initiale

Pour les nœuds qui représentent des concepts de la requête, l'activation initiale est calculée à partir de la formule suivante :

$$a_i = \frac{f_i}{N} \tag{7.1}$$

Où :

– f_i est la fréquence du concept lié au nœud i dans la requête.

– N est le nombre total de concepts dans la requête.

Cette formule nous permet de définir l'activation initiale des nœuds à partir de leurs apparitions dans la requête, en limitant les valeurs à l'intervalle $[0, 1]$. Pour ceux qui ne représentent pas des concepts dans la requête, l'activation initiale est égale à 0.

7.4.2 Méthode de propagation d'activation

Comme nous l'avons expliqué, l'approche choisie dans ce travail consiste à propager une activation commençant des nœuds sources à travers les liens munis de poids le long du réseau. On appelle pulsation le processus consistant à propager une activation d'un nœud aux nœuds adjacents. Le processus est divisé en deux étapes : premièrement, une ou plusieurs pulsations sont déclenchées et deuxièmement, un contrôle d'arrêt détermine si le processus doit continuer ou s'arrêter. La propagation d'activation travaille selon la formule suivante :

$$I_j(p) = \sum_{i}^{k} (O_i(p-1).w_{ij}) \tag{7.2}$$

Chaque nœud j détermine l'entrée totale I_j à la pulsation p des nœuds qui lui sont reliés. Donc, la sortie $O_i(p-1)$ du nœud i à la pulsation précédente $p-1$ est multiplié par le poids w_{ij} du lien connectant le nœud i au nœud j, et la somme totale correspondant aux k nœuds connectés est calculée.

La valeur de sortie d'un nœud doit être ensuite déterminée. Dans la majorité des cas, on ne fait pas de distinction entre la valeur d'entrée et le niveau d'activation des nœuds. Avant de propager l'activation aux nœuds adjacents, une fonction calcule la sortie par rapport au niveau d'activation du nœud : $O_i = f(I_i)$. Différentes fonctions peuvent être utilisées pour déterminer la valeur de sortie du nœud, par exemple la fonction sigmoïde, ou la fonction d'activation linéaire. La fonction la plus souvent utilisée est la fonction seuil, qui détermine si un nœud doit être activé ou pas. Si le niveau d'activation dépasse le seuil, l'état du nœud est actif et sa valeur de sortie est propagée aux nœuds adjacents. Dans notre travail, la fonction seuil a été adoptée.

Quand le processus itératif de propagation d'activation est déclenché, une itération correspond à une pulsation. En partant de l'ensemble des concepts sources déterminés à la phase d'initialisation (cf. Section 7.4.1), l'activation est propagée aux nœuds adjacents selon la formule suivante [Berger et al., 2003] :

$$O_i(p) = \begin{cases} 0 & \text{si } I_i(p) < \tau, \\ \dfrac{F_i}{p+1} \cdot I_i(p) & \text{sinon, avec } F_i = \left(1 - \dfrac{C_i}{C_T}\right) \end{cases} \tag{7.3}$$

La sortie $O_i(p)$, envoyée du nœud i à la pulsation p, est calculée comme la fraction de F_i, qui limite la propagation selon le degré de connectivité du nœud i (contrainte sémantique cf. section 6.3.2), et $p+1$, qui exprime la diminution de la relation sémantique selon la distance du nœud i des sources d'activation (contrainte de distance cf. Section 6.3.2). En plus, F_i est calculé en divisant le nombre de concepts C_i directement connectés au nœud i par le nombre total des nœuds C_T qui composent le réseau. La valeur τ Représente le seuil pour l'activation.

Ensuite, tous les nœuds qui viennent d'être activés sont utilisés à la prochaine itération comme sources d'activation et le processus continue jusqu'à ce qu'il ait atteint un nombre maximum d'itérations. Quand le processus est terminé le système examine tous les nœuds et les classe par rapport à leur activation.

7.5 Évaluation de l'approche

7.5.1 Collection de test : Requêtes des usagers de santé

L'une des principales difficultés de notre travail a été l'absence d'un corpus de référence de requêtes d'usagers de santé dans le domaine du cancer du sein. Par conséquent, il a fallu construire manuellement un corpus de questions d'usagers de santé et identifier les concepts en relation avec ces questions. Pour cela, nous avons à nouveau utilisé les forums du cancer du sein pour trouver ce type de ressources. Huit questions qui concernent le cancer du sein ont été identifiées sur le forum de La Ligue Contre le Cancer et treize autres sur le forum de Doctissimo [1]. Ces questions n'ont pas été utilisées dans la phase de conception de l'ontologie. Nous avons ensuite analysé les questions et les réponses fournies dans les forums pour extraire les concepts en relation avec la question de l'usager. Ces concepts peuvent se trouver explicitement dans la question ou dans les réponses fournies. Dans le choix des questions, nous avons sélectionné celles qui contiennent au moins une réponse fournie par un professionnel de santé et où l'utilisateur semblait être satisfait

1. `http://forum.doctissimo.fr/`

par les réponses fournies. On a considéré que le dernier point était vérifié si le fil de discussion a été clôturé par l'usager qui l'a déclenché et si le dernier message contenait des remerciements et pas de nouvelle question. L'ensemble des 21 questions a été utilisé pour tester le système de propagation d'activation développé et pour la phase d'évaluation.

7.5.2 Expérimentations et résultats

Les mesures d'évaluation utilisées viennent du domaine de la recherche d'information (cf. Section 3.10.1). **Le rappel** mesure la capacité de la méthode à identifier tous les concepts en rapport avec la requête de l'utilisateur (cf. Section 7.5.1). Il se calcule en divisant le nombre total de concepts correctement identifiés par le nombre de concepts dans la requête de référence.

$$\text{Rappel} = \frac{\mid C \cap C_{ref} \mid}{\mid C_{ref} \mid}$$

Avec :

- $\mid C \mid$= nombre total de concepts identifiés,
- $\mid C_{ref} \mid$= nombre de concepts dans la requête de référence,
- $\mid C \cap C_{ref} \mid$= nombre de concepts correctement identifiés.

La précision mesure la capacité de la méthode à identifier des concepts corrects. Elle se calcule en divisant le nombre total de concepts correctement identifiés par le nombre total de concepts identifiés.

$$\text{Précision} = \frac{\mid C \cap C_{ref} \mid}{\mid C \mid}$$

On cherche ainsi à obtenir la plus grande précision et le plus grand rappel possibles.
Le listing 7.2 montre un exemple d'un message d'un usager de santé utilisé pour la phase d'expérimentation :

```
Apres avoir parcourue le forum qui est tres reconfortant me lance moi aussi
    dans la discution. J'ai passe hier une mamo qui a duree dans le temps :
    1ere mamo+ 2eme mamo+ eco+ mamo verdict = sein gauche ACR5 en raison d'
    un volumineux foyer de microcalcifications irréguliers donce. Un
    prelevement par macro biopsie est souhaitable. Je me pose plein de
    questions pour la suite. Pouvez-vous m'eclaire sur ce fameux ACR5 et le
    prelevement par macro. Il faut dire que depuis je filp pas mal. Merci d'
    avance pour vos reponse.
```

Listing 7.2 – Exemple d'une question d'un usager de santé

Le tableau 7.2 montre la liste des concepts identifiés dans ce message par la méthode d'indexation :

Concept	Nombre d'occurrences
Mammographie	4
Echographie	1
Diagnostic	1
Sein	1
ACR5	2
Micro-calcification	1
Macro-biopsie	1

TABLE 7.2 – Liste des concepts identifiés dans le message

Après étude du message et des différentes réponses fournies par les internautes, nous avons établi cette liste de concepts qui représentent les concepts les plus représentatifs du message de la patiente (cf. Tableau7.3) :

Classement	Concept
1	Classification ACR
2	ACR5
3	Macro-biopsie
4	Micro-calcification
5	Cancer du sein
6	Stades du cancer du sein
7	Diagnostic du cancer du sein

TABLE 7.3 – Liste des concepts les plus représentatifs du message de la patiente

La figure 7.5 montre une partie des résultats obtenus après application de la propagation d'activation, pour une valeur du seuil d'activation égale à 0.05, pour l'ensemble initial de concepts identifiés par la méthode d'indexation utilisée :

```
mammographie                    0.4771895
acr5                            0.3828946
diagnostic                      0.3808305
sein                            0.3602902
diagnostic_du_cancer_du_sein    0.3557931
examen_par_mammographie         0.3464278
classification_acr              0.3227496
echographie                       0.30703
macro-biopsie                   0.2953729
examen_par_echographie          0.2854907
micro-calcification             0.2643684
anesthesie_locale               0.1309442
cancer_du_sein                  0.1300047
medecin                         0.1176294
```

FIGURE 7.5 – Trace de l'exécution de l'algorithme de propagation d'activation

Ces résultats présentent un taux de rappel de $85,7\%$ pour un taux de précision de $42,8\%$. Sur cet exemple, la méthode de propagation d'activation a retrouvé la plupart des concepts pertinents (bon rappel). Cependant, le faible taux de précision est dû à plusieurs raisons. La collection de tests n'est pas très adaptée à l'évaluation d'une tâche de RI. Les tests sont faits sur des messages et non sur des requêtes. La méthode d'indexation identifie des concepts qui ne font pas partie de la question de l'usager mais font partie du message. Dans cet exemple, c'est le cas des concepts *Mammographie* et *Echographie*. Par contre, ces deux concepts sont utilisés dans l'ensemble initial des concepts activateurs, ce qui mène à du bruit dans la liste des résultats. Les relations d'hyperonymie et celles qui ne sont pas utiles pour la requête d'utilisateur sont également sources de bruit. Dans l'exemple, le concept *Médecin* est inféré à cause de la relation *Pratiqué_par*, mais qui n'est pas utile dans ce cas là.

La figure 7.6 montre l'ensemble des résultats obtenus sur la collection de tests utilisée, pour une valeur du seuil d'activation égale à $0,05$. Cette valeur du seuil d'activation fournit

les meilleurs résultats parmi ceux testés.

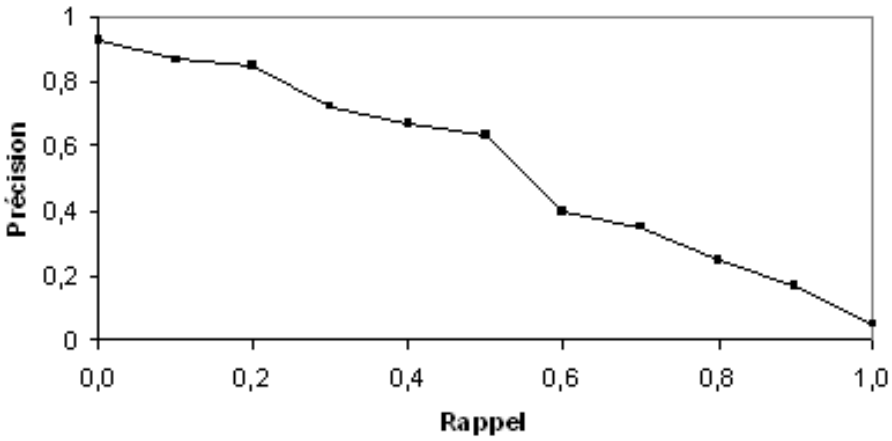


FIGURE 7.6 – Résultats des tests sur le corpus de référence

Les résultats obtenus montrent un niveau de rappel satisfaisant avec un niveau de précision qui diminue rapidement. En étudiant de plus près le mécanisme de propagation d'activation, nous avons constaté que les types de liens ont une influence primordiale sur la qualité de l'expansion. Les valeurs médiocres de précision sont pour la plupart dues aux relations de généralisation. La relation d'*hyperonymie* conduit vite à "l'éloignement" du sens initial voulu. Néanmoins, la méthode est efficace pour inférer les concepts les plus représentatifs de la question de l'usager de santé en haut de la liste des réponses.

7.6 Conclusion et discussion

Nous avons présenté dans ce chapitre une approche de reformulation automatique de requêtes par expansion. De façon générale, l'approche que nous avons présentée s'inscrit dans le cadre de l'application "partielle" des ontologies à la recherche d'information. Elle consiste à projeter une requête sur une ontologie, identifier les nœuds (concepts) de l'ontologie qui représentent au mieux le contenu de la requête, en utilisant différentes relations sémantiques. Nous avons proposé une approche utilisant la propagation d'activation dans un réseau sémantique. Nous avons évalué notre approche sur une collection que nous avons créée nous-mêmes pour les besoins de l'application. Rappelons que nous n'évaluons pas le fait de trouver des documents pertinents par rapport à une requête mais le fait d'identifier les concepts qui représentent le mieux une requête. Ces concepts vont à leur tour être soumis à un SRI pour retourner des documents pertinents. Les concepts "bruits" affectent fortement les résultats de recherche et ne doivent pas figurer dans le haut de la liste des concepts trouvés. Notre méthode présente des résultats satisfaisants pour le haut de la liste des réponses mais ils se dégradent rapidement pour des taux de rappel supérieurs à 50%. L'examen du mécanisme de propagation d'activation a révélé que la relation d'hyperonymie est la principale cause de la dégradation des résultats. Parmi les perspectives de ce travail figure la prise en compte de la sémantique des relations dans le calcul des poids entre les nœuds. En effet, notre calcul de relations, basé sur la co-occurrence des concepts dans le corpus, ne tient pas compte de la sémantique des relations entre les concepts. Cependant, la relation d'hyperonymie est souvent implicitement présente dans le corpus par la seule présence dans le message d'occurrences reliées par cette relation, sans qu'il y ait une

intention de l'utilisateur. Par exemple, l'utilisateur emploie le terme *cancer du sein* pour désigner le concept *Cancer_Sein* au début de son message, mais par la suite il va juste employer le terme *cancer* pour désigner le même concept (*Cancer_Sein*) parce qu'il sait que le contexte le définit ainsi. Or, dans notre application, ce phénomène va être considéré comme l'occurrence d'une relation d'hyperonymie. Par conséquent, les relations d'hyperonymie se voient attribuer des poids forts dans le réseau sémantique. Considérer la sémantique de la relation dans le processus de calcul peut réduire un tel phénomène.

Une autre perspective de notre travail est de tester d'autres méthodes d'identification de concepts. Les expériences menées par *Diallo*, dans son travail [Diallo, 2006], ont montré que l'utilisation de la désuffixation introduit du bruit dans l'identification des concepts. Il faudra donc réfléchir à des méthodes plus appropriées au langage utilisé dans les forums, qui se rapproche plus du langage parlé.

Conclusion

La terminologie utilisée par les spécialistes constitue parfois une barrière significative pour le grand public qui cherche des informations sur des thématiques liées à la santé ou qui essaye de comprendre des concepts techniques. Les usagers de santé ont souvent besoin de comprendre le "jargon", les termes utilisés parmi les spécialistes, pour mieux gérer leur situation. La recherche que nous avons menée a exploré principalement deux questions générales :

- Quelles sont les méthodes utiles pour l'identification, la création et l'analyse des termes utilisés par le grand public dans un corpus de textes d'un domaine particulier ?
- Quelles sont les caractéristiques des termes utilisés par les usagers de santé et les médiateurs de santé pour communiquer sur des problèmes d'ordre médical ?

S'intéresser à la première question apporte des informations aux études qui s'intéressent aux langages utilisés par les non-spécialistes d'un domaine donné. La deuxième question améliore la compréhension des termes utilisés par le grand public dans le domaine médical, et plus particulièrement le cancer du sein. Ce type de connaissances va faciliter le développement des outils et des ressources pour aider les usagers de santé à accéder à l'information médicale. Comme résultat sur le long terme, les usagers de santé vont être mieux préparés pour participer activement à leurs parcours de soin.

Des méthodes quantitatives et qualitatives ont été utilisées pour identifier et caractériser les termes utilisés par deux groupes : les médiateurs et les usagers de santé, pour communiquer sur le cancer du sein. L'étude a analysé l'usage des termes par les usagers et les médiateurs de santé et les a comparés les uns aux autres. L'analyse quantitative a fourni une perspective générale sur la terminologie, facilitant ainsi la définition de critères pour des analyses plus approfondies (e.g., analyse de la longueur des termes, analyse de la variabilité expressive des concepts). Les analyses ont été conduites sur plusieurs niveaux : termes, concepts, termes-concepts et relations. Plusieurs caractéristiques propres à la terminologie des usagers de santé et à celle des médiateurs de santé ont été observées et discutées.

Contributions

Notre travail est composé de deux parties principales :

Construction d'une ontologie du cancer du sein

Même s'il existe plusieurs méthodologies en terminographie ou en développement des terminologies, ces méthodes ne sont pas directement exploitables pour identifier les termes utilisés par les usagers de santé. Contrairement aux spécialistes du domaine, qui forment des communautés de discours bien spécifiques avec un fond commun d'expériences et de connaissances techniques,

les usagers de santé forment des groupes larges et diversifiés possédant des niveaux différents de connaissance du domaine. De plus, l'objectif des sous-langages, un point d'intérêt traditionnel de la terminographie, est le transfert précis et exact de l'information technique. Les non-spécialistes utilisent le langage général pour une grande variété d'objectifs. Par conséquent, bien que les méthodologies traditionnelles de terminographie ont servi comme point de départ, les procédures ont évolué pour s'adapter aux variations considérables des termes utilisés par les usagers de santé.

La méthodologie générale s'appuie sur les travaux de *C. Roche* et *B. Bachimont* et les principes de la sémantique différentielle à prendre en compte dans les étapes de construction des ressources onto-terminologiques. Dans ce cadre, nous avons travaillé à la réalisation des objectifs suivants :

- Production de ressources termino-ontologiques : nous avons construit deux corpus, le corpus des médiateurs de santé et le corpus des usagers de santé, et une ontologie dans le domaine du cancer du sein, comptant 1287 concepts.
- Développement de méthodes et d'outils pour l'acquisition de ces ressources : nous avons construit un parseur pour la collecte et le nettoyage du corpus des usagers de santé issu de deux forums de discussion sur le cancer du sein. Nous avons développé un outil pour l'extraction des n-grammes des corpus. Nous avons également développé un outil de recherche pour le regroupement des n-grammes dans des groupes conceptuels.
- Développement de méthodes pour l'analyse quantitative et qualitative de ces ressources : nous avons analysé l'ontologie construite sur plusieurs niveaux : termes, concepts, termes-concepts et relations.

L'analyse de l'ontologie construite et des résultats du mapping vers UMLS et CHV a mené à plusieurs conclusions :

- La principale différence entre la terminologie utilisée par les usagers de santé et celle des médiateurs ou professionnels de santé réside principalement au niveau des termes. Seuls 0,6% des concepts sont propres aux usagers de santé et 2% aux médiateurs de santé. Par contre, 10% des termes sont propres aux usagers de santé et 6,5% aux médiateurs de santé.
- Une autre différence majeure entre les deux terminologies réside au niveau des relations entre concepts. Les usagers de santé connectent les concepts entre eux d'une manière différente de celle des professionnels de santé. Nous avons modélisé cette différence en utilisant la relation *Relation_X* dans notre ontologie.
- Les termes utilisés par les usagers de santé sont fortement liés à leurs contextes. Ils sont souvent constitués de mots issus du langage général. Donc, sans des indices contextuels, le sens du terme est difficile à trouver. La terminologie technique est construite pour être précise : les termes dans des langues de spécialité désignent des concepts spécifiques. Cependant, si les termes techniques ne sont pas connus, les mots du langage général sont utilisés. L'utilisation des abréviations et des acronymes accentue encore plus le problème. En fait, les termes utilisés par les usagers de santé sont un "hybride" entre les mots du langage général et ceux de la langue de spécialité, ce qui conduit à une perte de précision, même pour les termes professionnels.
- Les formalismes de représentation des connaissances actuels n'offrent pas assez d'expressivité pour pouvoir exprimer une connaissance supplémentaire sur le niveau de "technicité" des termes.

Reformulation des requêtes des usagers de santé

Cette partie de notre thèse consiste à tester l'ontologie construite dans la première partie dans une application de reformulation des requêtes des usagers de santé en utilisant l'ap-

proche de propagation d'activation. L'approche adoptée est inspirée des travaux de H. Berger [Berger et al., 2003] avec une adaptation au niveau du calcul des poids des relations entre les concepts (les nœuds). Au niveau technique, nous nous sommes basés sur la librairie de S. Reed pour réaliser le module de propagation d'activation. Pour atteindre notre objectif, nous avons réalisé plusieurs tâches :

- Conversion de l'ontologie au format utilisé par la librairie Java : nous avons écrit un parseur qui utilise l'API d'OWL pour convertir l'ontologie au format approprié et pour fusionner les relations multiples entre deux concepts en un lien unique entre deux nœuds.
- Identification des concepts : nous avons développé un outil basé sur l'API Lucene de la fondation Apache pour l'identification des concepts dans les messages et dans les requêtes des usagers de santé. Pour le moment seule la forme des mots après désuffixation est prise en compte.
- Calcul des poids de liens entre les nœuds du réseau sémantique issu de l'ontologie : nous nous sommes basés sur le calcul de la valeur de la fréquence jointe normalisée entre les concepts dans le corpus des usagers de santé. Pour les relations génératrices le poids a été calculé sur la base du nombre de liens sortants. Pour les autres liens qui sont restés à 0, nous avons étudié les relations transitives pour leur attribuer manuellement un poids.
- Création d'un corpus de questions d'usagers de santé : l'absence d'un corpus de référence pour tester l'outil de reformulation des requêtes nous a conduits à construire manuellement un corpus de questions d'usagers de santé issu de deux forums de discussion sur le cancer du sein et d'identifier les concepts en rapport avec leurs questions. Nous avons obtenu un ensemble de 21 questions, ce qui est un nombre très inférieur aux corpus de référence existants. Ni les ressources disponibles ni le temps ne permettent de construire un plus grand corpus. Ce travail est essentiellement destiné à diriger les perspectives d'amélioration et d'implémentation réelle.
- Réalisation de l'outil de propagation d'activation : l'outil a été testé sur la base des requêtes du corpus de référence construit. Les résultats ont montré que la méthode est efficace pour inférer les concepts les plus représentatifs d'une requête d'usager de santé mais le taux de précision diminue rapidement à des niveaux de rappel supérieurs à 50%.

Perspectives

Les perspectives de notre travail sont multiples et concernent aussi bien l'amélioration des méthodes proposées que leur utilisation.

Nous avons vu que les méthodes traditionnelles de terminographie ne peuvent pas être appliquées directement sur les corpus des usagers. La méthode d'extraction des n-grammes produit encore beaucoup de bruit car la production de formes différentes dans ce type de corpus est très forte. Nous envisageons de réfléchir sur des méthodes plus adaptées pour l'aide à la construction de ressources termino-ontologiques à partir de corpus d'usagers dans des domaines de spécialités. Nous voulons également inclure les usagers de santé dans les étapes de construction et d'évaluation du système.

Un autre point de réflexion concerne le formalisme de représentation des connaissances qui permettra d'annoter les termes selon leur usage ou leur cible : professionnel de santé ou grand public.

Finalement, une question cruciale dans ce travail reste la gestion du cycle de vie de cette ontologie ? Comme nous l'avons expliqué, ce type de ressource est une cible mouvante sur plusieurs facteurs et peut évoluer à tout moment. Par conséquent, il faut des méthodes et des environne-

ments pour gérer ces changements. Il faut savoir gérer la mise à jour de concepts, de termes et de relations et surtout repérer quand apporter ces changements.

En ce qui concerne l'application de l'ontologie pour l'aide à la recherche d'information, les voies d'amélioration sont multiples :

- Tester d'autres algorithmes de PA : changement du seuil, changement de la fonction d'activation, changements des contraintes d'activation, ...
- Prendre en compte les opérateurs logiques dans une requête : et, ou, sauf, ...
- Prendre en compte la sémantique des relations dans l'attribution des poids entre les liens.
- Tester d'autres méthodes d'identification de concepts dans les corpus et les requêtes.
- Créer un plus grand corpus de référence pour l'évaluation du système.

A plus long terme, notre objectif est de créer un système interactif d'aide à la formulation des requêtes. En se basant sur les concepts présents dans l'ontologie, plusieurs heuristiques peuvent être appliquées à l'interface de l'utilisateur pour l'assister dans la formulation de sa requête. Par exemple, le concept *Sein* peut être relié à plusieurs autres concepts : (1) l'organe et sa fonction ; (2) les maladies ou les symptômes liés au sein ; (3) les traitemens de ces maladies ; (4) les examens médicaux liés au sein ; (5) des spécialistes en oncologie ; ou (6) d'autres concepts. Les utilisateurs peuvent sélectionner une ou plusieurs entrées à partir d'une liste de concepts, liés au *Sein*, groupés au sein d'une classification sémantique. Un avantage de ce type d'interaction est qu'il offre à l'utilisateur un ensemble de concepts qu'il n'a peut être pas considérés au début et qu'il pourrait trouver utiles pour sa recherche. La navigation dans les concepts du domaine peut même apporter des connaissances supplémentaires sur le domaine, surtout si l'on fournit à chaque fois des définitions adaptées aux usagers.

Travail à faire

Alors que les Étas-Unis ont déjà amorcé deux initiatives collaboratives et Open Source concernant le CHV (Consumer Health Vocabulary) et le LHI (Lay Health Informatics), les efforts de la France et de l'Europe restent minces dans ces domaines. Pourtant, les acteurs de ce type d'initiatives sont présents et il suffit de mettre l'effort et les moyens nécessaires pour les réunir. Dans le premier chapitre de cette thèse, nous avons cité les prinicipaux acteurs qui fournissent l'information médicale destinée aux usagers de santé : CISMeF Patient, HON Patient/Particulier, Vidal de la famille. A l'initiative de notre laboratoire, TIMC, nous avons participé à l'élaboration de deux projets TecSan dans le cadre des appels à projets ANR 2007 et 2008 qui rassemblent ces acteurs et d'autres partenaires pour créer une initiative similaire à celle des États-Unis en France. Le projet a été classé deux fois en liste complémentaire, mais sans financement. Ceci témoigne de l'intérêt naissant de la communauté des scientifiques pour ce type de services. Nous comptons continuer dans cette voie et allons proposer un autre projet dans le cadre de l'appel à projet : "Web innovant" [1]. Ce projet aura pour objectif de rassembler les partenaires adéquats pour former un vrai groupe de réflexion autour des nouveaux services du Web adaptés aux usagers de santé.

1. `http://www.telecom.gouv.fr/rubriques-menu/soutiens-financements/programmes-nationaux/volet-numerique-du-plan-relance/services-innovants-du-web/396.html`

Annexe A

Liste des relations de l'ontologie

0 **Associé_A**
Inverse : Associé_A
Définition : Possède une relation significative avec.

.1 **Topologiquement_Relié_A**
Inverse : Topologiquement_Relié_A
Définition : Désigne la manière dont les constituants ou les parties sont inter-reliés ou organisés.

.1.1 **Partie_De**
- Inverse : A_Comme_Partie
  Définition : Compose, avec un ou plusieurs d'autres unités physiques, une entité plus large.

.1.1.1 **Constitué_De**
- Inverse : Constitue
  Définition : Est structurellement composé de (en partie ou en entier) d'une collection d'éléments. e.g., Tissu/Cellule.

.1.1.2 **Contient**
- Inverse : Contenu_Dans
  Définition : Contient ou il est le récipient d'un fluide ou d'autres substances.

.1.1.3 **Ingrédient_De**
- Inverse : A_Comme_Ingrédient
  Définition : Un constituant d'une préparation ou se trouve naturellement dans une substance. e.g., caféine/café.

.1.1.4 **Composant_De**
- Inverse : A_Comme_Composant
  Définition : Une relation structurelle et fonctionnelle claire entre un objet entier et ses parties. e.g., moteur/voiture.

.1.2 **Connecté_A**
- Inverse : Connecté_A
  Directement attaché à une autre unité physique comme les tendons sont connectés aux muscles.

.1.2.1 **Branche_De**

- Inverse : A_Comme_Branche
  Définition : Surgit de la division de. Par exemple : arborisation des artères.

.1.2.2 **Interconnecté_Avec**

- Inverse : Interconnecté
  Définition : Sert à lier ou joindre ensemble deux ou plusieurs unités physiques.

.1.2.2 **Tributaire_De**

- Inverse : A_Comme_Tributaire
  Définition : Fusionne avec. Par exemple, la convergence des veines.

.1.3 **Position_De**

- Inverse : A_Comme_Position
  Définition : La position, le site, ou la région d'une entité ou le lieu d'un processus.

.1.3.1 **Adjacent_A**

- Inverse : Adjacent_A
  Définition : voisin ou proche de.

.1.3.2 **Entoure**

- Inverse : Entouré_Par
  Définition : Etablit les frontières de, ou définit les limites d'une autre structure physique.

.1.3.3 **Traverse**

- Inverse : Traversé_Par
  Définition : Franchit ou s'étend le long d'une autre structure physique ou une zone.

.2 **Fonctionnellement_Relié_A**

Inverse : Fonctionnellement_Relié_A
Définition : Sont reliés par l'accomplissement d'une certaine fonction ou activité.

.2.1 **Affecte**

- Inverse : Affecté_Par
  Définition : Produit un effet direct sur.

.2.1.1 **Absorbe**

- Inverse : Absorbé_Par
  Définition : Prend et assimile, s'imprègne et retient en soi.

.2.1.2 **Retarde**

- Inverse : Retardé_Par
  Met en retard ou diffère.

.2.1.3 **Complique**

- Inverse : Compliqué_Par
  Définition : Rend plus complexe, plus sévère ou résulte à des effets secondaires.

.2.1.4 **Altère**

- Inverse : Altéré_Par
  Définition : Modifie en mal, provoque un changement, une détérioration de l'état ou de la valeur de quelque chose.

.2.1.5 **Facilite**

- Inverse : Facilité_Par
  Définition : Rend facile ou simplifie.

.2.1.6 **Augmente**

- Inverse : Augmenté_Par
  Définition : Rend plus grand, plus considérable, accroître.

.2.1.7 **Diminue**

- Inverse : Diminué_Par
  Définition : Rend plus petit, moins important, réduit.

.2.1.8 **Interagit_Avec**

- Inverse : Interagit_Avec
  Définition : Agit, fonctionne ou opère ensemble avec.

.2.1.9 **Dirige**

- Inverse : Dirigé_Par
  Définition : Conduire, mener, Orienter ou guider dans une direction.

.2.1.10 **Prévient**

- Inverse : Prévenu_Par
  Définition : Arrête, empêche par prévention ou élimine une action ou une condition.

.2.1.11 **Traite**

- Inverse : Traité_Par
  Définition : Appliquer un remède dans le but de soigner ou améliorer une condition.

.2.2 **Entraîne**

- Inverse : Entraîné_Par
  Définition : Avoir pour conséquence, pour effet.

.2.2.1 **Cause**

- Inverse : Causé_Par
  Définition : Provoquer un effet sur un processus. Cela veut dire qu'un agent/instrument causal a conduit à l'effet. L'effet ici n'est pas un produit.

.2.2.2 **Produit**

- Inverse : Produit_Par
  Définition : Un agent causal ou un processus a conduit au produit (un objet).

.2.2.3 **Résulte_A**

- Inverse : Résultat_de
  Définition : Un processus emmène à un état particulier ou à un autre processus.

.2.3 **Exécute**

- Inverse : Exécuté_Par
  Définition : Un agent exécute, accomplit ou réalise une activité.

.2.3.1 **Réalise**

- Inverse : Réalisé_par
  Définition : Un agent exécute ou réalise une fonction.

.2.3.2 **Manifeste**

- Inverse : Manifesté_par
  Définition : Un agent montre un certain acte. Ceci n'est pas une propriété de l'agent mais un comportement qui est exécuté.

.2.3.3 **Pratique**

- Inverse : Pratiqué_par
  Définition : Un agent exécute une action habituellement ou régulièrement.

.2.4 **Se_Produit_Dans**

- Inverse : A_Comme_Occurrence
  Définition : A lieu ou se manifeste dans une population donnée.

.2.5 **Processus_De**

- Inverse : A_Comme_Processus
  Définition : Une fonction ou un changement dans l'action ou de l'état de quelque chose.

.2.6 **Utilise**

- Inverse : Utilisé_Par
  Définition : Se sert de ou l'emploie dans la réalisation d'une activité.

.3 **Temporellement_Relié_A**

Inverse : Temporellement_Relié_A
Définition : Reliés dans le temps.

.3.1 **Se_Produit_En_Même_Temps_que**

- Inverse : Se_Produit_En_Même_Temps_que
  Définition : Survient en même temps que, avec ou conjointement.

.3.2 **Précède**

- Inverse : Suit
  Définition : Survient plus tôt dans le temps.

.3.3 **Age_de**

- Inverse : A_Pour_Age
  Définition : Reliés par la durée d'existence.

.3.4 **Fréquence_Cyclique_De**

- Inverse : A_Fréquence_Cyclique
  Définition : Le nombre de fois un phénomène survient dans des cycles.

.3.5 **Retarde***

- Inverse : Retardé_Par

Définition : Mettre en retard ou différer.

.3.6 **Durée_De**

– Inverse : A_Pour_Durée
Définition : Lié au temps que dure un phénomène, une situation, un objet.

.3.7 **Temps_Occurrence**

– Inverse : A_Temps_Occurrence
Définition : Lié à un instant ou une intervalle du temps d'occurrence.

.4 **Conceptuellement_Relié_A**

– Inverse : Conceptuellement_Relié_A
Définition : Reliés par un concept, une notion ou une idée.

.4.1 **Analyse**

– Inverse : Analysé_Par
Définition : Étudie ou examine en utilisant des méthodes quantitatives ou qualitatives.

.4.1.1 **Evalue_Effet_De**

– Inverse : Effet_Evalué_Par
Définition : Analyse les influences ou les conséquences d'une fonction ou d'une action.

.4.1.2 **Diagnostique**

– Inverse : Diagnostiqué_Par
Définition : Une personne distingue ou identifie la nature ou les caractéristiques de.

.4.1.3 **Mesure**

– Inverse : Mesuré_Par
Définition : Établit ou marque les dimensions ou la quantité.

.4.1.4 **Évaluation_De**

– Inverse : A_Pour_Évaluation
Définition : Le jugement de la valeur ou du degré d'un attribut ou d'un processus.

.4.1.5 **Degré_De**

– Inverse : A_Pour_Degré
Définition : L'intensité relative d'un processus ou l'intensité relative ou le montant de qualité d'un attribut.

.4.1.6 **Mesure_De**

– Inverse : A_Pour_Mesure
Définition : Une valeur résultant d'une mesure qui établit ou marque les dimensions ou la quantité.

.4.1.7 **Comparé_A**

– Inverse : Comparé_A
Définition : Une relation comparative entre deux concepts. C'est une relation générale où la relation elle-même peut être plus spécifique. Par exemple : le concept A *est similaire au* concept B.

.4.2 **Propriété_De**

- Inverse : A_Pour_Propriété
  Définition : Les caractéristiques de, ou la qualité de.

.4.3 **Nécessite**

- Inverse : Nécessite_Par
  Définition : Ce qu'il lui est nécessaire ou essentiel.

.4.4 **Dérivé_De**

- Inverse : A_Pour_Dérivé
  Définition : En chimie, une substance structurellement reliée à une autre ou qui peut être obtenue à partir d'une autre.

.4.5 **Forme_Développée_De**

- Inverse : A_Pour_Forme_Développée
  Définition : Un stage antérieur dans le cycle de vie d'un objet.

.4.6 **Méthode_De**

- Inverse : A_Pour_Méthode
  Définition : La manière ou la séquence d'évènements dans la réalisation d'une action ou d'une procédure.

.4.7 **Problématique_Dans**

- Inverse : A_Pour_Problématique
  Définition : Est un problème ou un point de discussion, d'étude, de débat ou de conflit.

Autres relations

Relation_X
Inverse : Relation_X
Définition : Possède une relation significative avec.

Annexe B

Extrait de l'ontologie du cancer du sein sous Protégé

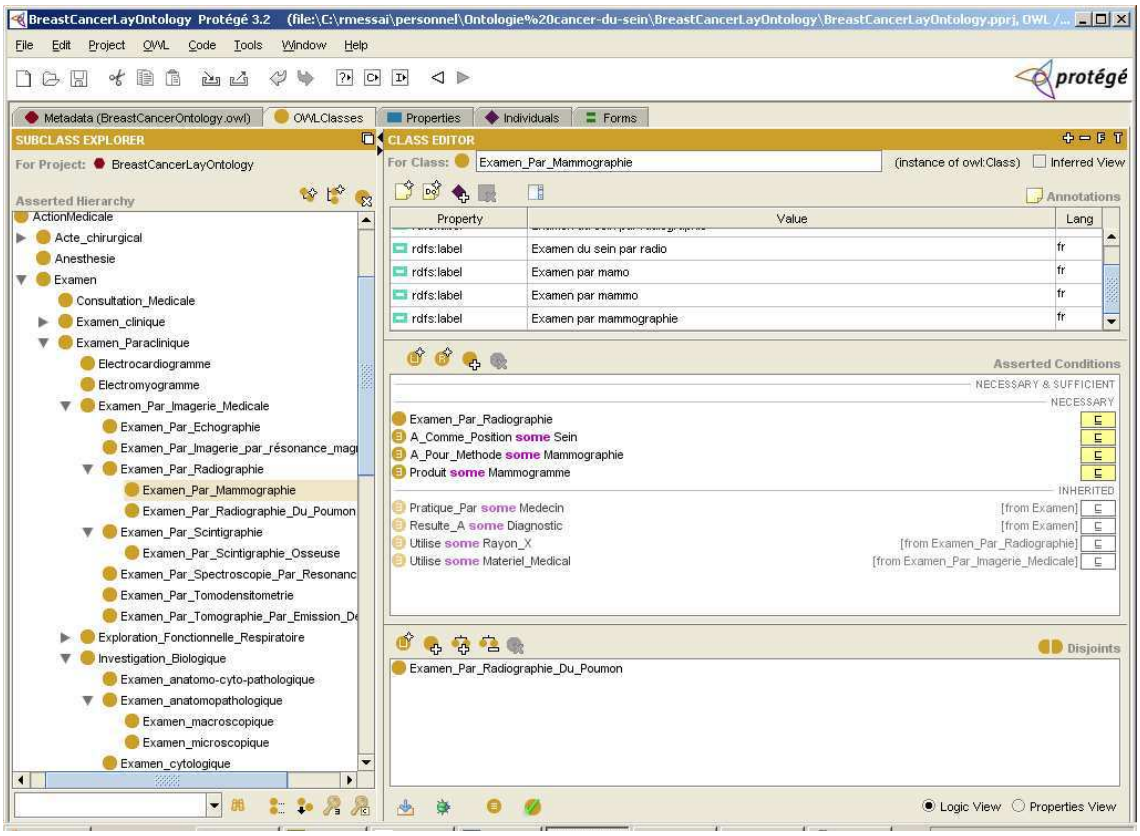


Figure B.1 – Extrait de l'ontologie du cancer du sein vue avec Protégé 3.2

Bibliographie

[Arpirez et al., 2001] Arpirez, J. C., Corcho, O., Fernandez-Lopez, M., & Gomez-Perez, A. (2001). Webode : a scalable ontological engineering workbench. In *First International Conference on Knowledge Capture (KCAP01)* (pp. 6–13). Victoria : ACM Press.

[Aussenac-Gilles et al., 2003] Aussenac-Gilles, N., Biébow, B., & Szulman, S. (2003). D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. In *Conférence TIA* Strasbourg.

[Baader et al., 2003] Baader, F., Horrocks, I., & Sattler, U. (2003). Description logics as ontology languages for the semantic web. In *Festschrift in honor of Jörg Siekmann, Lecture Notes in Artificial Intelligence* (pp. 228–248). : Springer-Verlag.

[Baader & Nutt, 2002] Baader, F. & Nutt, W. (2002). *the Description Logic Handbook*, chapter Basic Description Logics, (pp. 47–100). Cambridge University Press.

[Bachimont, 2000] Bachimont, B. (2000). *Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances*, chapter Ingénierie des connaissances, chapitre 19, (pp. 305–323). Paris : L'Harmattan.

[Bachimont et al., 2002] Bachimont, B., Isaac, A., & Troncy, R. (2002). Semantic commitment for designing ontologies : A proposal. In *13th International Conference on Knowledge Engineering and Knowledge Management*, volume Lecture Notes in Artificial Intelligence (pp. 114–121). : Springer.

[Bader & Theofanos, 2003] Bader, L. J. & Theofanos, F. M. (2003). Searching for cancer information on the internet : Analyzing natural language search queries. *J Med Internet Res*, 5(4), e31.

[Baeza-Yates & Ribeiro-Neto, 1999] Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc.

[Balas & Boren, 2000] Balas, E. A. & Boren, S. A. (2000). *Yearbook of medical informatics : Patient-centered systems*, chapter Managing clinical knowledge for health care improvement, (pp. 65–70). Stuttgart, Germany : Schattauer.

[Baneyx, 2007] Baneyx, A. (2007). *Construire Une Ontologie De La Pneumologie : Aspects Théoriques, Modèles Et Expérimentations*. PhD thesis, Université Paris 6.

[Baroni & Bernardini, 2004] Baroni, M. & Bernardini, S. (2004). Bootcat : Bootstrapping corpora and terms from the web. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)* (pp. 1313–1316). Lisbon, Portugal.

[Baroni & Ueyama, 2006] Baroni, M. & Ueyama, M. (2006). Building general and special purpose corpora by web crawling. In *Proceedings of the 13th NIJL International Symposium, Language Corpora : Their Compilation and Application* (pp. 31–40).

[Bataillon et al., 2006] Bataillon, R., Hascoet, J.-Y., Leneel, H., Caron, B., Samzun, J.-L., & Pencole, D. (2006). Vers une consultation médicale de prévention. *Santé publique*, 18, 5–6.

[Baziz, 2005] Baziz, M. (2005). *Indexation Conceptuelle Guidee Par Ontologie Pour La Recherche D'Information*. PhD thesis, Universite Paul Sabatier.

[Bechhofer et al., 2001] Bechhofer, S., Horrocks, I., Goble, C., & Stevens, R. (2001). Oiled : a reason-able ontology editor for the semantic web. In *Joint German/Austrian conference on Artificial Intelligence (KI01)*, volume 2174 of *Lecture Notes in Artificial Intelligence* (pp. 396–408). Berlin : Springer.

[Belkin & Croft, 1992] Belkin, N. J. & Croft, W. B. (1992). Information filtering and information retrieval : two sides of the same coin ? *Commun. ACM*, 35(12), 29–38.

[Benjamins & Fensel, 1998] Benjamins, R. & Fensel, D. (1998). The ontological engineering initiative-ka. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS 98* (pp. 287–301). Trento, Italy.

[Berger et al., 2003] Berger, H., Dittenbach, M., & Merkl, D. (2003). Activation on the move : Querying tourism information via spreading activation. In *DEXA* (pp. 474–483).

[Berners-Lee et al., 2000] Berners-Lee, T., Hendler, J., & Lassila, O. (2000). Semantic web. *Scientific America*, 1(1), 68–88.

[Bernhard, 2003] Bernhard, D. (2003). Ontology building based on text corpora. Master's thesis, Institut National Polytechnique de Grenoble.

[Bernhard, 2006] Bernhard, D. (2006). *Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales*. PhD thesis, Université Joseph Fourier.

[Bernier & Heumann, 1957] Bernier, C. L. & Heumann, K. F. (1957). Correlative indexes, iii, semantic relations among semantemes - the technical thesaurus. *American Documentation*, 8, 211–220.

[Blazquez et al., 1998] Blazquez, M., Fernandez-Lopez, M., Garcia-Pinar, J. M., & Gomez-Perez, A. (1998). Building ontologies at the knowledge level using the ontology design environment. In M. M. E. B.R. Gaines (Ed.), *11th International Workshop on Knowledge Acquisition, Modeling and Management (KAW98)* Banff.

[Blois, 1984] Blois, M. S. (1984). Information and medicine. Berkeley, CA : University of California Press.

[Bonnevay & Lamure, 2003] Bonnevay, S. & Lamure, M. (2003). Bases de connaissances anatomo-fonctionnelle : application au cerveau et au cœur. *Santé et Systémique*, 7(3-4), 47–75.

[Bordogna & Pasi, 2000] Bordogna, G. & Pasi, G. (2000). Flexible querying of web documents. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems(FQAS)*.

[Borst, 1997] Borst, W. N. (1997). *Construction of Engineering Ontologies*. PhD thesis, University of Tweenty, Enschede, NL-Centre for Telematica and Information Technology.

[Boughanem et al., 2005] Boughanem, M., Loiseau, Y., & Prade, H. (2005). Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In *AMR05, 3rd International Workshop on Adaptive Multimedia Retrieval* Glasgow, UK : Springer.

[Bourigault & Aussenac-Gilles, 2003] Bourigault, D. & Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. In *TALN* Batz-sur-Mer.

[Bowker, 1996] Bowker, L. (1996). Towards a corpus-based approach to terminography. *Terminology*, 3(1), 27–52.

[Brachman & Schmolze, 1985] Brachman, R. J. & Schmolze, J. G. (1985). An overview of the kl-one knowledge representation system. *Cognitive Science*, 9(2), 171–216.

[Bush, 1945] Bush, V. (1945). As we may think. *Atlantic Monthly*, 176, 101–108.

[Cabré, 1999] Cabré, M. T. (1999). *Terminology : Theory, methods, and applications*, volume 1. Amsterdam : John Benjamins Publishing Company.

[Callan et al., 1992] Callan, J. P., Croft, W. B., & Harding, S. M. (1992). The inquery retrieval system. In *In Proceedings of DEXA* (pp. 78–83).

[Cassileth et al., 1980] Cassileth, B. R., Zupkis, R. V., Sutton-Smith, K., & March, V. (1980). Information and participation preferences among cancer patients. *Ann Intern Med*, 92(6), 832–836.

[Chan & Woodruff, 1997] Chan, A. & Woodruff, R. (1997). Communicating with patients with advanced cancer. *J Palliat Care*, 13(3), 29–33.

[Chapman et al., 2003] Chapman, K., Abraham, C., Jenkins, V., & Fallowfield, L. (2003). Lay understanding of terms used in cancer consultations. *Psycho-Oncology*, 12, 557–566.

[Charlet, 2002] Charlet, J. (2002). L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Mémoire d'Habilitation à diriger des recherches, Université Pierre et Marie Curie.

[Church & Mercer, 1993] Church, K. W. & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1), 1–24.

[Cleverdon, 1967] Cleverdon, C. W. (1967). The cranfield tests on index language devices. In *Aslib Proceedings*, volume 19 (pp. 173–192).

[Cohen & Kjeldsen, 1987] Cohen, P. & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation on sematic networks. *Information Processing& Management*, 23(4), 255–268.

[Condamines, 2003] Condamines, A. (2003). Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques. Habilitation à diriger des recherches, CNRS & Université de Toulouse-Le Mirail.

[Corby et al., 2006] Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., & Gandon, F. (2006). Searching the semantic web : Approximate query processing based on ontologies. *IEEE Intelligent Systems*, 21(1), 20–27.

[Corcho et al., 2003] Corcho, O., Fernandez-Lopez, M., & Gomez-Perez, A. (2003). Methodologies, tools and languages for building ontologies : where is their meeting point ? *Data Knowl. Eng.*, 46(1), 41–64.

[Corcho et al., 2002] Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A., & Vicente, O. (2002). Webode : an integrated workbench for ontology representation, reasoning and exchange. In A. Gomez-Perez & V. R. B. (Eds.) (Eds.), *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, volume 2473 of *Lecture Notes in Artificial Intelligence* (pp. 138–153). Berlin : Springer.

[Crestani, 1997] Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6), 453–482.

[Croft et al., 1989] Croft, W., Lucia, T., Crigean, J., & Willet, P. (1989). Retrieving documents by plausible inference : an experimental study. *Information Processing & Management*, 25(6), 599–614.

[Crowell et al., 2005] Crowell, J., Zeng, Q., & Tse, T. (2005). A web application to support consumer health vocabulary development. In *AMIA 2005 Symposium Proceedings* (pp. 932).

[Darmoni et al., 2002] Darmoni, S., Thirion, B., Platel, S., Douyère, M., Mourouga, P., , & Leroy, J.-P. (2002). Cismef-patient : a french counterpart to medlineplus. *J Med Libr Assoc*, 90(2), 248–253.

[Decker et al., 1999] Decker, S., Erdmann, M., Fensel, D., & Studer, R. (1999). Ontobroker : Ontology based access to distributed and semistructured information. In *Semantic Issues in Multimedia Systems (DS8)* (pp. 351–369). Boston : Kluwer Academic Publisher.

[Deerwester et al., 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), 391–407.

[Denhière & Lemaire, 2003] Denhière, G. & Lemaire, B. (2003). Modélisation des effets contextuels par l'analyse de la sémantique latente. In d. In Bastien, J. (Ed.), *Actes des Deuxièmes Journées d'étude en Psychologie Ergonomique (EPIQUE 2003), Roquencourt : INRIA*.

[Diallo, 2006] Diallo, G. (2006). *Une Architecture à Base d'Ontologies pour la Gestion Unifiée des Données Structurées et non Structurées*. PhD thesis, Université Joseph Fourier.

[Domingue, 1998] Domingue, J. (1998). Tadzebao and webonto : Discussing, browsing and editing ontologies on the web. In *Proc. 11th Knowledge Acquisition Workshop (KAW98)* Banff.

[Dominich, 2002] Dominich, S. (2002). Application, analysis and evaluation of neural networks-based interaction information retrieval. In *Information Processing and Management*.

[Drouin, 2004] Drouin, P. (2004). Spécificités lexicales et acquisition de la terminologie. In *In Actes des 7e Journées internationales d'analyse statistique des données textuelles (JADT-2004)* (pp. 345–352). Louvain-la-Neuve, Belgique.

[Dwivedi et al., 2003] Dwivedi, A., Bali, R. K., Naguib, R. N. G., & NassaI, N. S. (2003). Clinical knowledge management for healthcare. In *Proceedings of the 4^{th} Annual IEEE Conf on Information Technology Applications in Biomedicine* UK.

[Eysenbach, 2003] Eysenbach, G. (2003). The impact of the internet on cancer outcomes. *CA Cancer J Clin*, 53, 356–371.

[Eysenbach & Till, 2001] Eysenbach, G. & Till, E. J. (2001). Ethical issues in qualitative research on internet communities. *British Medical Journal*, 323, 1103–1105.

[Farquhar et al., 1996] Farquhar, A., Fikes, R., & Rice, J. (1996). The ontolingua server : A tool for collaborative ontology construction. In *Proc. 10th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW96)*, volume 44.1-44.19 Banff.

[Fellbaum et al., 2006] Fellbaum, C., Hahn, U., & Smith, B. (2006). Towards new information resources for public health-from wordnet to medicalwordnet. *J Biomed Inform*, 39(3), 321–332.

[Fensel et al., 1999] Fensel, D., Angele, J., Decker, S., Erdmann, M., Schnurr, H., Staab, S., Studer, R., & Witt, A. (1999). On2broker : Semantic-based access to information sources at the www. *WebNet*, 1, 366–371.

[Fernandez et al., 1997] Fernandez, M., Gomez-Perez, A., & Juristo, N. (1997). Methontology : From ontological art towards ontological engineering. In *Symposium on Ontological Engineering of AAAI* Stanford, USA.

[Fernandez-Lopez et al., 1999] Fernandez-Lopez, M., Gomez-Perez, A., Pazos-Sierra, A., & Pazos-Sierra, J. (1999). Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems and their applications*, January/February, 37–46.

[Fogel et al., 2002] Fogel, J., Albert, M. S., Schnabel, F., Ditkoff, A. B., & Neugut, I. A. (2002). Use of the internet by women with breast cancer. *J Med Internet Res*, 4(2), e9.

[Fox & Rainie, 2000] Fox, S. & Rainie, L. (2000). *The online health care revolution : How the Web helps Americans take better care of themselves.* Technical report, Washington, D.C. : The Pew Internet & American Life Project.

[Frydel, 2006] Frydel, Y. (2006). *Internet au quotidien : un Français sur quatre.* Technical report, INSEE Première.

[Furnas et al., 1998] Furnas, G., Deerwester, S., Dumais, S., Landauer, T., Harshman, R., Streeter, L., & Lochbaum, K. (1998). Information retrieval using a singular value decomposition model of latent semantic structure. In *The 11th International Conference on Research and Development in Information Retrieval* (pp. 465–480). Grenoble, France : ACM Press.

[Furnas et al., 1987] Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 964–971.

[Gemoets et al., 2004] Gemoets, D., Rosemblat, G., Tse, T., & Logan, R. (2004). Assessing readability of consumer health information : an exploratory study. In *Medinfo*, volume 11 (pp. 869–873).

[Ghani et al., 2001] Ghani, R., Jones, R., & Mladenic, D. (2001). Mining the web to create minority language corpora. In *CIKM01 : Proceedings of the tenth international conference on Information and knowledge management* (pp. 279–286). New York, NY, USA : ACM Press.

[Gonzalo et al., 1998] Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL 98, Workshop on Usage of WordNet for NLP* (pp. 38–44).

[Gruber, 1991] Gruber, T. (1991). The role of common ontology in achieving sharable, reusable knowledge bases. In *Proc. Second International Conference on Principles of Knowledge Representation and Reasoning* (pp. 601–602). Cambridge, MA.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specification. knowledge acquisition. *Special issue : Current issues in knowledge modeling*, 5(2), 199–220.

[Gruninger & Fox, 1994] Gruninger, M. & Fox, M. S. (1994). The design and evaluation of ontologies for enterprise engineering. In *Workshop on Implemented Ontologies, European Workshop on Artificial Intelligence* Amsterdam, NL.

[Guarino et al., 1995] Guarino, N., Carrara, M., & Giaretta, P. (1995). *Towards Very Large Knowledge Bases, Knowledge Building and Knowledge Sharing*, chapter Ontologies and knowledge bases : towards a terminological clarification, (pp. 25–32). IOS Press, Amsterdam.

[Guarino et al., 1999] Guarino, N., C.Masolo, & Vetere, G. (1999). Ontoseek : using large linguistic ontologies for accessing on-line yellow pages and product catalogs. *IEEE Intelligent Systems*, 14(3), 70–80.

[Haas, 1997] Haas, S. W. (1997). Disciplinary variation in automatic sublanguage term identification. *J. Am. Soc. Inf. Sci.*, 48(1), 67–79.

[Haas & Hert, 2000] Haas, S. W. & Hert, C. A. (2000). Terminology development and organization in multi-community environments : the case of statistical information. In *Proceedings of the American Society for Information Science SIG/CR Classification Workshop* (pp. 51–72). Chicago.

[Haas & Hert, 2002] Haas, S. W. & Hert, C. A. (2002). Finding information at the u.s. bureau of labor statistics : overcoming the barriers of scope, concept, and language mismatch. *Terminology*, 8(1), 31–56.

[Haendel et al., 2008] Haendel, M. A., Neuhaus, F., Osumi-Sutherland, D., Mabee, P. M., Jr., J. L. M., Mungall, C. J., & Smith, B. (2008). *CARO : the Common Anatomy Reference Ontology*, chapter The Foundational Model of Anatomy Ontology, (pp. 311–333). Springer, New York.

[Heflin et al., 2003] Heflin, J., Hendler, J., & Luke, S. (2003). Shoe : A blueprint for the semantic web. In *In Fensel, D., Hendler, J. A., Lieberman, H. and Wahlster, W., éditeurs : Spinning the Semantic Web : Bringing the World Wide Web to Its Full Potential [outcome of a Dagstuhl seminar]* (pp. 29–63). : MIT Press.

[Hoffmann, 1979] Hoffmann, L. (1979). Towards a theory of lsp : Elements in a methodology of lsp analysis. *Fachsprache*, 1, 14–32.

[Horrocks & van Harmelen, 2001] Horrocks, I. & van Harmelen, F. (2001). *Reference Description of the DAML+OIL (March 2001)*. Technical report, Ontology Markup Language.

[Jones & Needham, 1968] Jones, K. S. & Needham, R. (1968). Automatic term classification and retrieval. *Information Processing and Management*, 4, 91–100.

[Jorgensen & Gotzsche, 2004] Jorgensen, K. J. & Gotzsche, P. C. (2004). Presentation on websites of possible benefits and harms from screening for breast cancer : cross sectional study. *BMJ*, 328, 148–155.

[Kabel et al., 2004] Kabel, S., Hoog, R. D., Wielinga, B., & Anjewierden, A. (2004). The added value of task and ontology-based markup for information retrieval. *Journal of the American Society for Information Science and Technology*, 55(4), 348–362.

[Kandel et al., 2004] Kandel, O., Duhot, D., Very, G., Lemasson, J.-F., & Boisnault, P. (2004). Existe-t-il une typologie des actes effectués en médecine générale ? *La Revue Du Praticien - Médecine Générale*, 18, 781–784.

[Kendall et al., 2002] Kendall, E., Dutra, M., & McGuinness, D. (2002). Towards a commercial ontology development environment. In *First International Semantic Web Conference (ISWC02)*, volume 2342 of *Lecture Notes in Computer Science* Berlin : Springer.

[Khan, 2000] Khan, L. R. (2000). *Ontology-based Information Selection*. PhD thesis, Faculty Of The Graduate School. University Of Southern California.

[Kilgarriff & Grefenstette, 2003] Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3), 333–347.

[Kim, 2005] Kim, H. (2005). Ontoweb : Implementing an ontology based web retrieval system. *JASIST*, 56(11), 1167–1176.

[Kiryakov et al., 2003] Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., & Goranov, M. (2003). Semantic annotation, indexing, and retrieval. In *2nd International Semantic Web Conference (ISWC2003)* Florida, USA.

[Kjeldsen & Cohen, 1987] Kjeldsen, R. & Cohen, P. (1987). The evolution and performance of the grant system. *IEEE Expert*, Summer, 73–79.

[Kogut et al., 2002] Kogut, P., Cranefield, S., Hart, L., Dutra, M., Baclawski, K., Kokar, M., & Smith, J. (2002). Uml for ontology development. *The Knowledge Engineering Review*, 17(1), 61–64.

[Lame, 2002] Lame, G. (2002). *Construction d'ontologies à partir de textes. Une ontologie du droit dédiée à la recherche d'information sur le Web.* PhD thesis, Ecole des Mines.

[Lassila & Swick, 1999] Lassila, O. & Swick, R. R. (1999). *Resource description framework (rdf) model and syntax specification.* Technical report, World Wide Web Consortium W3C Recommendation 22 February 1999 / W3C /.

[Lebart & Salem, 1994] Lebart, L. & Salem, A. (1994). *Statistique textuelle.* Dunod, Paris.

[Lee et al., 1993] Lee, J. H., Kim, W. Y., Kim, M. H., & Lee, Y. J. (1993). On the evaluation of boolean operators in the extended boolean retrieval framework. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 291–297). Pittsburgh, PA, USA.

[Lenz, 1984] Lenz, E. (1984). Information seeking : a component of client decisions and health behavior. *Advances in Nursing Science*, 6(3), 59–72.

[Leontis et al., 2006] Leontis, N. B., Altman, R. B., Berman, H. M., Brenner, S. E., Brown, J. W., Engelke, D. R., Harvey, S. C., Holbrook, S. R., Jossinet, F., Lewis, S. E., Major, F., Mathews, D. H., Richardson, J. S., Williamson, J. R., & Westhof, E. (2006). The rna ontology consortium : an open invitation to the rna community. *RNA*, 12, 533–541.

[Lerner, 2000] Lerner, E. . (2000). Medical communication : Do our patients understand ? *The American Journal of Emergency Medicine*, 18(7), 764–766.

[L'Homme, 2004] L'Homme, M.-C. (2004). *La terminologie : principes et techniques.* Paramètres.

[Liu & Curran, 2006] Liu, V. & Curran, J. R. (2006). Web text corpus for natural language processing. In *Proceeding of EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 233–240). Trento, Italy.

[Léon & Millon, 2005] Léon, S. & Millon, C. (2005). Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du web. In *In Actes de RECITAL 2005*, volume 1 (pp. 595–604).

[Lopez, 1999] Lopez, M. F. (1999). Overview of methodologies for building ontologies. In V. R. Benjamins, B. Chandrasekaran, A. Gomez-Perez, N. Guarino, & M. U. (Eds) (Eds.), *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)* Stockholm, Sweden.

[Luhn, 1957] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, October, 309–317.

[Maron & Kuhns, 1960] Maron, M. E. & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 216–244.

[Marshall, 2000] Marshall, P. D. (2000). Bridging the terminology gap between health care professionals and patients with the consumer health terminology (cht). In *AMIA Symp. 2000* (pp. 1082).

[Martinet, 2004] Martinet, J. (2004). *Un modèle vectoriel relationnel de recherche d'information adapté aux images.* PhD thesis, Université Joseph Fourier, Grenoble.

[Martinez, 2000] Martinez, W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. In *In Actes de JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles.*

[McBride, 2004] McBride, B. (2004). The resource description framework (rdf) and its vocabulary description language rdfs. In *In Handbook on Ontologies* (pp. 51–66).

[McCray et al., 1999] McCray, A. T., Loane, R. F., Browne, A. C., & Bangalore, A. K. (1999). Terminology issues in user access to web-based medical information. In *Proc AMIA Symp* (pp. 107–111).

[McCray & Tse, 2003] McCray, A. T. & Tse, T. (2003). Understanding search failures in consumer health information systems. In *AMIA Annu Symp Proc* (pp. 430–434).

[McGuinness, 1998] McGuinness, D. L. (1998). Ontological issues for knowledge-enhanced search. In *Proceedings of the First International Conference on Formal Ontology in Information Systems* (pp. 302–316). Trento, Italy.

[McGuinness & van Harmelen, 2004] McGuinness, D. L. & van Harmelen, F. (2004). *OWL web ontology language overview*. Technical report, W3c recommendation, World Wide Web Consortium.

[Meric et al., 2002] Meric, F., Bernstam, E. V., Mirza, N. Q., Hunt, K. K., Ames, F. C., Ross, M. I., Kuerer, H. M., Pollock, R. E., Musen, M. A., & Singletary, S. E. (2002). Breast cancer on the world wide web : cross sectional survey of quality of information and popularity of websites. *BMJ*, 324, 577–581.

[Miller et al., 2000] Miller, N., Lacroix, E., & Backus, J. (2000). Medline plus : Building and maintaining the national library of medicine's consumer health web service. *Bulletin of the Medical Library Association*, 88(1), 11–14.

[Moigno et al., 2002] Moigno, S. L., Charlet, J., Bourigault, D., Degoulet, P., & Jaulent, M. (2002). Terminology extraction from text to build an ontology in surgical intensive care. In *AMIA Symp* (pp. 430–435).

[Naets, 2005] Naets, H. (2005). La déclaration universelle des droits de l'homme : 329 langues pour la constitution automatique de corpus et de lexique. In *Actes de TALN 2005*, volume 2 (pp. 261–268).

[Nardi & Brachman, 2002] Nardi, D. & Brachman, R. J. (2002). *the Description Logic Handbook*, chapter An Introduction to Description Logics, (pp. 5–44). Cambridge University Press.

[NCI, 2006] NCI (2006). *Integrating Key Biomedical Terminologies : NCI Metathesaurus*. Technical report, National Cancer Institute.

[Neches et al., 1991] Neches, R., Fikes, R. E., Finin, T., Gruber, T. R., Senator, T., & Swartout, T. (1991). Enabling technology for knowledge sharing. ai magazine 12(3), 36-56. *AI Magazine*, 12(3), 36–56.

[Neelamegham & Jain, 1999] Neelamegham, R. & Jain, D. (1999). Consumer choice process for experience goods : An econometric model and analysis. *Journal of Marketing Research*, 36(3), 373–386.

[Nielsen-Bohlman, 2004] Nielsen-Bohlman, L. (2004). *Health literacy : a prescription to end confusion*. Washington DC : National Academies Press.

[Patrick et al., 2001] Patrick, T. B., Monga, H. K., Sievert, M. C., Hall, J. H., & Longo, D. R. (2001). Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *J Med Internet Res*, 3(3), e24.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligente systems : networks of plausible inference*. Morgan Kaufmann Publishers Inc.

[Phillips, 1989] Phillips, M. (1989). *Lexical structure of text. Discourse analysis monograph, no. 12.* English Language Research : University of Birmingham.

[Picht & Draskau, 1985] Picht, H. & Draskau, J. (1985). *Terminology : An introduction.* Surrey, UK : University of Surrey Press.

[Pifalo et al., 1997] Pifalo, V., Hollander, S., Henderson, C., DeSalvo, P., & Gill, G. P. (1997). The impact of consumer health information provided by libraries : The delaware experience. *Bulletin of the Medical Library Association*, 85(1), 16–22.

[Pincock, 2004] Pincock, S. (2004). Initiative to exchange cancer research information is launched. *BMJ*, 328, 728.

[Plovnick & Zeng, 2004] Plovnick, R. M. & Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology : A pilot study. *Journal of Medical Internet Research JMIR*, 6(3) :e27, 1–10.

[Popov et al., 2003] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). Kim : A semantic annotation platform. In *2nd International Semantic Web Conference (ISWC2003)* Florida, USA.

[Preece, 1981] Preece, S. (1981). *A spreading activation model for information retrieval.* PhD thesis, Univesity of Illinois, Urbana Champaign, USA.

[Péry-Woodley, 1995] Péry-Woodley, M. P. (1995). Quels corpus pour quels traitements automatiques ? *T.A.L*, 36(1-2), 213–232.

[Quillian, 1968] Quillian, R. (1968). *Semantic Information Processing*, chapter Semantic memory, (pp. 216–270). The MIT Press, Cambridge, MA, USA.

[Rapp, 2003] Rapp, R. (2003). Discovering the meaning of an ambiguous word by searching for sense descriptors with complementary context patterns. In *In Actes des Cinquièmes Rencontres Terminologie et Intelligence Artificielle* (pp. 145–155).

[Rastier et al., 1994] Rastier, F., Cavazza, M., & Abeille, A. (1994). Sémantique pour l'analyse - de la linguistique à l'informatique. Enseignement de Paris, France : Masson.

[Rector & Nowlan, 1994] Rector, A. L. & Nowlan, W. A. (1994). The galen project. *Comput Methods Programs Biomed*, 45(1-2), 75–78.

[Renahy, 2007] Renahy, E. (2007). *L'utilisation d'Internet pour la recherche d'information en santé.* Technical report, SIRS Infos. Paris : Inserm.

[Renahy et al., 2007] Renahy, E., Parizot, I., Lesieur, S., & Chauvin, P. (2007). *WHIST : Enquête web sur les habitudes de recherche d'informations liées à la santé sur Internet.* Technical report, Enquête INSERM.

[Richardson & Smeaton, 1995] Richardson, R. & Smeaton, A. F. (1995). *Using WordNet in a knowledge-based approach to information retrieval.* Technical report, Dublin, Ireland.

[Rijsbergen et al., 1981] Rijsbergen, C. J., Harper, D., & Porter, M. F. (1981). The selection of good search terms. *Information Processing and Management*, 17, 77–91.

[Rijsbergen, 1979] Rijsbergen, J. V. (1979). *Information retrieval.* Dept. of Computer Science, University of Glasgow.

[Robertson, 1977] Robertson, S. E. (1977). The probabilistic ranking principle in ir. *Journal of Documentation*, 33, 294–304.

[Robertson, 1981] Robertson, S. E. (1981). The methodology of information retrieval expirement. *Information retrival expirement*, 1, 9–31.

[Robertson et al., 1995] Robertson, S. E., Walker, S., & Hancock-Beaulieu, M. (1995). Large test collection experiments on an operational, interactive system : Okapi at trec. *Inf. Process. Manage.*, 31(3), 345–360.

[Roche, 2003] Roche, C. (2003). The differentia principle as a cornerstone of ontology. In *Knowledge Management and Philosophy Workshop in WM 2003 Conference* Luzern.

[Roche, 2005] Roche, C. (2005). Terminologie et ontologie. *Larousse*, 157, 1–11.

[Roche, 2007] Roche, C. (2007). Le terme et le concept : fondements d'une ontoterminologie. In *Actes de la Conférence TOTh 2007 : Terminologie et Ontologie : Théories et Applications* (pp. 1–22). Annecy.

[Rose et al., 2001] Rose, J. S., Fisch, B. J., Hogan, W. R., Levy, B., Marshall, P., Thomas, D. R., & Kirkley, D. (2001). Common medical terminology comes of age, part two : current code and terminology sets–strengths and weaknesses. *Journal of Healthcare Information Management*, 15(3), 319–330.

[Rosemblat et al., 2006] Rosemblat, G., Logan, R., Tse, T., & Graham, L. (2006). Text features and readability : Expert evaluation of consumer health text. In *Medical Internet. MEDNET.*

[Rosse & Mejino, 2007] Rosse, C. & Mejino, J. (2007). *The Foundational Model of Anatomy Ontology*, chapter Anatomy Ontologies for Bioinformatics : Principles and Practice, (pp. 59–117). Springer.

[Rosse & Mejino, 2003] Rosse, C. & Mejino, J. L. V. (2003). A reference ontology for biomedical informatics : the foundational model of anatomy. *Journal of Biomedical Informatics*, 36, 478–500.

[Rothschild, 1998] Rothschild, S. K. (1998). Cross-cultural issues in primary care medicine. *Disease-A-Month*, 44(7), 293–319.

[Rousselot, 2004] Rousselot, F. (2004). L'outil de traitement de corpus likes. In *In Actes de TALN.*

[Rousselot & Frath, 2000] Rousselot, F. & Frath, P. (2000). Terminologie et intelligence artificielle.

[Roussey et al., 2001] Roussey, C., Calabretto, S., & Pinon, J. (2001). A multilingual information system based on knowledge representation. In *ADBIS 2001 : advances in databases and information systems* (pp. 98–111). Vilnius.

[Rubin et al., 2006] Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., Sim, I., Chute, C. G., Solbrig, H., Storey, M.-A., Smith, B., Day-Richter, J., Noy, N. F., & Musen, M. A. (2006). National center for biomedical ontology : advancing biomedicine through structured organization of scientific knowledge. *OMICS*, 10, 185–198.

[Rubin, 2003] Rubin, R. (2003). Doctor-patient language gap isn't healthy. USA today.

[Rudd & Glanz, 1990] Rudd, J. & Glanz, K. (1990). *How Individuals Use Information for Health Action : Consumer Information Processing*, chapter Health Behavior and Health Education : Theory, Research, and Practice, (pp. 115–139). Jossey-Bass Publishers : San Francisco.

[Rumelhart & Norman, 1983] Rumelhart, D. & Norman, D. (1983). *Representation in memory.* Technical report, Department of Psychology and Institute of Cognitive Science, UCSD La Jolla, USA.

[Salton, 1968] Salton, G. (1968). Automatic information organization and retrieval. Mc Graw Hill, New York.

[Salton, 1971] Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

[Salton & Buckley, 1988] Salton, G. & Buckley, C. (1988). On the use of spreading activation methods in automatic information. In *SIGIR88 : Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 147–160). New York, NY, USA : ACM.

[Salton et al., 1983a] Salton, G., Buckley, C., & Yu, C. (1983a). An evaluation of term dependence models in information retrieval. *LNCS*, 146, 151–173.

[Salton et al., 1983b] Salton, G., Fox, E., & Wu, H. (1983b). Extended boolean information retrieval. *Communications of the ACM*, 26(11), 1022–1036.

[Salton & McGill, 1983] Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York : McGraw-Hill.

[Salton & McGill, 1986] Salton, G. & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.

[Satterlund et al., 2003] Satterlund, J. M., McCaul, D. K., & Sandgren, K. A. (2003). Information gathering over time by breast cancer patients. *J Med Internet Res*, 5(3), e15.

[Segall & Roberts, 1980] Segall, A. & Roberts, L. W. (1980). A comparative analysis of physician estimates and levels of medical knowledge among patients. *Sociology of Health and Illness*, 2(3), 317–334.

[Shoval, 1981] Shoval, P. (1981). Expert/consultation system for a retrieval data-base with semantic network of concepts. In C. J. Crouch (Ed.), *Theoretical Issues in Information Retrieval, Proceedings of the Fourth International Conference on Information Storage and Retrieval* (pp. 145–149). Oakland, California, USA : ACM.

[Singhal, 2001] Singhal, A. (2001). Modern information retrieval : A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), 35–42.

[Slaughter, 2002] Slaughter, L. (2002). *Semantic relationships in health consumer questions and physicians' answers : A basis for representing medical knowledge and for concept exploration interfaces*. PhD thesis, Faculty of the Graduate School of the University of Maryland.

[Slaughter et al., 2006] Slaughter, L. A., Soergel, D., & Rindflesch, T. C. (2006). Semantic representation of consumer questions and physician answers. *Int J Med Inform*, 75(7), 513–529.

[Smith, 2003] Smith, B. (2003). *Ontology*, chapter Blackwell Guide to the Philosophy of Computing and Information, (pp. 155–166). Oxford : Blackwell.

[Smith et al., 2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceuster, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., the OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., & Richard H Scheuermann14, Nigam Shah15, P. L. W. . S. L. (2007). The obo foundry : coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25, 1251–1255.

[Soergel et al., 2004] Soergel, D., Tse, T., & Slaughter, L. (2004). Helping healthcare consumers understand : an "interpretive layer" for finding and making sense of medical information. *Medinfo*, 11(Pt 2), 931–935.

[Sowa, 1984] Sowa, J. F. (1984). *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley.

[Spadaro, 2003] Spadaro, R. (2003). *European Union citizens and sources of information about health*. Technical report, EUROBAROMETER 58.0.

[Sparck Jones & Jackson, 1970] Sparck Jones, K. & Jackson, D. (1970). The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and Retrieval*, 5, 175–201.

[Spees, 1991] Spees, C. (1991). Knowledge of medical terminology among clients and families. *IMAGE : J Nurs Sch*, 23, 225–229.

[Stavri, 2001] Stavri, P. Z. (2001). Personal health information-seeking : A qualitative review of the literature. In *MedInfo*.

[Studer et al., 1998] Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering : principles and methods. *Data and Knowledge Engineering*, 25, 161–197.

[Sure et al., 2002a] Sure, Y., Angele, J., & Staab, S. (2002a). Ontoedit : Guiding ontology development by methodology and inferencing. In *University of California* (pp. 29–31). : Springer.

[Sure et al., 2002b] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002b). Ontoedit : collaborative ontology engineering for the semantic web. In *First International Semantic Web Conference (ISWC02)*, volume 2342 of *Lecture Notes in Computer Science* (pp. 221–235). Berlin : Springer.

[Swartout et al., 1997] Swartout, B., Ramesh, P., Knight, K., & Russ, T. (1997). Toward distributed use of large-scale ontologies. In *AAAI Symposium on Ontological Engineering* Stanford.

[Szulman et al., 1999] Szulman, S., Biébow, B., & Aussenac-Gilles, N. (1999). Vers un environnement intégré pour la structuration de terminologies : Terminae. Présentation Terminae.

[Thompson, 1984] Thompson, J. (1984). *Compliance*, chapter The experience of illness, (pp. 109–131). London, UK : Tavistock Publications.

[Tolle & Chen, 2000] Tolle, K. & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *JASIS*, 51(4), 352–370.

[Troncy & Isaac, 2002] Troncy, R. & Isaac, A. (2002). Doe : une mise en œuvre d'une méthode de structuration différentielle pour les ontologies. In *Actes des 13ème journées francophones sur Ingénierie des Connaissances (IC)* (pp. 63–74).

[Tse, 2003] Tse, T. (2003). *Identifying and Characterizing a Consumer Medical Vocabulary.* PhD thesis, College of Information Studies, University of Maryland, College Park.

[Tse & Soergel, 2003] Tse, T. & Soergel, D. (2003). Exploring medical expressions used by consumers and the media : An emerging view of consumer health vocabularies. In *AMIA Annu Symp Proc* (pp. 674–678).

[Turney, 2001] Turney, P. (2001). Answering subcognitive turing test questions : a reply to french. *J. Experimental and Theoretical Artificial Intelligence*, 13, 409–419.

[Uschold & King, 1995] Uschold, M. & King, M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing.*

[Vaufreydaz, 2002] Vaufreydaz, D. (2002). *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue.* PhD thesis, Université Joseph Fourier - Grenoble 1.

[Vickery, 1960] Vickery, B. C. (1960). Thesaurus - a new word in documentation. *Journal of Documentation*, 16(4), 181–189.

[Vivaldi et al., 2001] Vivaldi, J., Marquez, L., & Rodriguez, H. (2001). Improving term extraction by system combination using boosting. In *In Proceedings of ECML* (pp. 515–526).

[Voorhees, 1993] Voorhees, E. M. (1993). Using wordnet to disambiguate word sense for text retrieval. In *ACMSIGIR 93* (pp. 171–180). Pittsburg, PA, USA.

[WHO, 1995] WHO (1995). Icd-10. Vol. 2, Second Edition.

[Wiggers et al., 1990] Wiggers, J. H., o. Donovan, K., Redman, S., & Sanson-Fisher, R. W. (1990). Cancer patient satisfaction with care. *Cancer*, 66(3), 610–616.

[Yarowsky, 1992] Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *In Proceedings of COLING-92* (pp. 454–460).

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.

[Zeng et al., 2007] Zeng, Q., Kim, H., Goryachev, S., Keselman, A., Slaughter, L., & Smith, C. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. *Stud Health Technol Inform*, 129, 1117–1121.

[Zeng et al., 2001] Zeng, Q., Kogan, S., Ash, N., & Greenes, R. A. (2001). Patient and clinician vocabulary : how different are they ? *Medinfo*, 10(Pt 1), 399–403.

[Zeng et al., 2002] Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., & Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods Inf Med*, 41(4), 289–298.

[Zeng & Tse, 2004] Zeng, Q. T. & Tse, T. (2004). *Open Source and Collaborative Development of Consumer Health Vocabulary.* Technical report, Lister Hill National Center for Biomedical Communications. NLM.

[Zeng & Tse, 2006] Zeng, Q. T. & Tse, T. (2006). Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*, 13, 24–29.

Résumé

Internet est devenu une source importante d'information médicale pour les patients et leurs proches : recherche d'informations sur leurs maladies et les dernières recherches cliniques, ainsi que pour y constituer des communautés "numériques" de dialogue et de partage. Cependant, accès à Internet ne signifie pas nécessairement accès à l'information. Le manque de familiarité avec le langage médical constitue un problème majeur pour les usagers de santé dans l'accès à l'information et son interprétation. Le travail de cette thèse s'inscrit dans la problématique d'étude et de caractérisation de la terminologie des usagers de santé pour pouvoir proposer des services adaptés à leur langage et à leur niveau de connaissances. Sa production principale est une ontologie dans le domaine du cancer du sein orientée vers les usagers de santé. Cette ontologie est construite à partir d'un ensemble de corpus de textes représentant deux catégories : les médiateurs de santé et les usagers de santé. Les éléments de cette ontologie ont été analysés en utilisant des méthodes quantitatives et qualitatives sur plusieurs niveaux : termes, concepts et relations.

L'ontologie produite a constitué le noyau d'une application de reformulation de requêtes d'usagers de santé en utilisant l'approche de propagation d'activation dans un réseau sémantique. Les concepts de l'ontologie représentent les nœuds dans le réseau sémantique et les liens entre ces nœuds ont des poids, calculés soit automatiquement sur la base des co-occurrences des concepts dans un corpus de textes soit, manuellement selon le type des liens ; ces poids reflètent la "force" de la relation entre les nœuds.

Mots-clés: Terminologie des usagers de santé, ontologies, reformulation des requêtes, réseaux sémantiques, propagation d'activation.

Abstract

The Internet has become an important source of medical information for patients and their family members : search for information about their diseases and recent clinical research, building numeric communities for exchange and sharing of information and of personal experience. However, access to the Internet does not mean access to information. The lack of familiarity with the medical language is a major problem for health consumers in information access and understanding. The aim of this thesis is to analyse and characterize the terms used by non-professionals during their discourse on medical topics in order to propose services adapted to their language and to their level of knowledge. Its main production is a health consumer ontology in the breast cancer field, which is based on two types of text corpora : health mediators and health consumers. The elements of this ontology have been analysed on several levels : terms, concepts and relations.

The resulting ontology has been the core of a health consumers query reformulation application using spreading activation techniques through a semantic network. The concepts of the ontology represent the nodes in the semantic network. The links between the nodes have weights calculated, either automatically based on the co-occurrence of concepts in a corpus of texts or manually depending on the links type, reflecting the "strength" of the relation between the nodes.

Keywords: Consumer health terminology, ontologies, query reformulation, semantic networks, spreading activation.