Utilisation de ressources externes dans un modèle Bayésien de Recherche d'Information. Application à la recherche d'information multilingue avec UMLS.

Thi Hoang Diem Le

UNIVERSITE JOSEPH FOURIER
Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique
***

THESE
Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITE JOSEPH FOURIER - GRENOBLE I
Discipline : Informatique

Prépentée et soutenue publiquement le 31 mai 2009 par
Thi Hoang Diem LE

TITRE
Utilisation de ressources externes dans un modèle Bayésien de Recherche d'Information. Application à la recherche d'information médicale multilingue avec UMLS.

Composition du jury : :
M. Jean Caelen : Président du jury
Mme Anne Boyer : Rapporteur
M. Pierre Zweigenbaum : Rapporteur
M. Mohan Boughanem : Examinateur
Mme Catherine Berrut : Directeur de thèse
M. Jean-Pierre Chevallet : Co-directeur de thèse
Mme Dong Thi Bich Thuy : Co-directeur de thèse

Thèse préparée au sein de l'équipe MRIM du laboratoire LIG

Remerciements

C'est un très grand plaisir pour moi d'exprimer ma reconnaissance à ceux qui m'ont aidé tout au long de mes études doctorales.

Je tiens à remercier tout d'abord mes encadrants de thèse, sans qui je n'aurais pas pas pu mener à bien la thèse.

Je veux sincèrement remercier M. Jean Pierre Chevallet, enseignant chercheur à l'Université Pierre Mendès France, pour avoir été patient avec moi au cours de ce long travail et pour tout le temps qu'il a consacré à mes réunions de thèse en France ainsi qu'à Singapour. Je suis très reconnaissante d'avoir eu ses encouragements dans les moments difficiles, ceux qui m'ont beaucoup aidé.

Je tiens à remercier Mme Catherine Berrut, Professeur de l'Université Joseph Fourier, pour sa patience et ses remarques constructives, qui m'ont aidé à bien avancer le manuscrit de thèse.

J'adresse aussi mes remerciements à mon co-encadrant de thèse au Vietnam, Mme Dong Thi Bich Thuy, Professeur de l'Université Nationale du Vietnam à Ho Chi Minh Ville, qui m'a accueilli et supporté depuis mes études à l'université au Vietnam jusqu'en France.

Je voudrais également remercier M. Jean Caelen d'avoir accepté de présider le jury ; Mme Anne Boyer et M. Pierre Zweigenbaum d'avoir rapporté ma thèse, M. Mohand Boughanem d'avoir accepté d'être examinateur, malgré leur emploi du temps très chargé.

J'adresse mes remerciements aussi aux membres de l'équipe MRIM, où j'ai fait mon stage de DEA et puis un an de ma thèse. Merci à M. Bernard Cassagne de m'avoir aidé à régler les problèmes techniques, mes collègues Bao Quoc, Xuan Hung, Cong Phap, An Te, Trong Ton, Caroline, Loic, Said, Mbarek, Leila, Ali,... pour les pauses café agréables et leur encouragements. Je remercie les membres de l'équipe du labo IPAL et I2R à

Singapour qui ont collaboré avec moi dans les travaux.

Je n'oublie pas mes autres amis en France, au Vietnam et à Singapour qui m'ont également soutenu dans la vie étudiante avec tous types de difficultés mais aussi de la joie.

Je tiens à remercier du fond du coeur ma famille, mes grands parents, mes parents, mes beaux parents, mes soeurs qui me supportent toujours dans la vie ainsi que dans mes études.

Pour finir, je remercie mon mari, Nicolas, qui me supporte avec sa confiance, son amour et son soutien sans condition, toujours avec un grand plaisir. Je suis énormément reconnaissante pour tout cela.

Maintenant ma vie tourne une nouvelle page, je garderai bien en memoire cette page de mon parcours universitaire, avec tous les bons souvenirs, pour continuer vers une nouvelle aventure ...

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

Acronyme	Description	Page
RI	Recherche d'Information	12
RIM	Recherche d'Information Multilingue	12
SRI	Système de Recherche d'Information	12
RSV	Valeur de pertinence (Relevant Status Value)	21
VSM	Modèle vectoriel (Vector Space Model)	21
MAP	Mean Average Precision	24
UMLS	Unified Medical Language System	37
GAO	Graphe Acyclique Orienté	24
RB	Réseau Bayésien	45
RIRBRE	Modèle de la Recherche d'Information basé sur le Réseau Bayésien étendue avec une Ressource Externe	68
CLEF	Cross-Language Evaluation Forum	87
DFR	Divergence From Randomness	85

Résumé

Dans les systèmes de recherche d'information, une indexation à base de termes et une correspondance à base d'intersection introduisent le problème de la disparité à cause des variations linguistiques.

Avec l'objectif de résoudre ce problème, notre travail de thèse se positionne dans l'utilisation des ressources externes dans la recherche d'information. Ces ressources offrent non seulement les concepts pour une indexation plus précise et indépendante de langue, mais aussi une base de relations sémantiques entre ces concepts. Nous étudions en premier une indexation par concepts extraits à partir d'une ressource externe. Nous proposons ensuite de prendre en compte ces relations sémantiques entre les concepts dans la correspondance par un modèle de recherche d'information basé sur un réseau Bayésien des concepts et leurs relations sémantiques. Ainsi, nous étudions les extensions de l'indexation conceptuelle à des documents et requête structurés et multi-médias. Les fonctions de reclassement et de combinaison ont été proposées afin d'améliorer la performance de la recherche dans ces contextes.

La validation des propositions est effectuée par des expérimentations dans la recherche d'information multilingue médicale, avec l'utilisation du méta thésaurus UMLS comme ressource externe.

Mots clés : Recherche d'information multilingue, indexation conceptuelle, ressource externe, réseau Bayésien, UMLS.

Chapitre 1

Introduction générale

Sommaire

1.1 La recherche d'information

La Recherche d'information (RI) est un domaine des technologies de l'information qui consiste à rechercher dans une grande masse d'informations, les documents qui satisfont les besoins d'un utilisateur. Ce dernier exprime son besoin sous forme d'une requête normalement en langue naturelle. Dans ce domaine, l'adéquation entre les informations recherchées et le besoin de l'utilisateur correspond, dans le système de RI (SRI), à un calcul de correspondance entre documents et requêtes.

La recherche d'information multilingue (RIM) a pour objectif de satisfaire un besoin sous forme d'une requête dans une langue autre que la langue des documents.

Le processus de RI (monolingue ou multilingue) à partir d'une requête donnée se déroule en deux phases principales :

- La phase d'*indexation* des documents, qui a pour objectif de représenter le contenu des documents sous une forme manipulable par l'ordinateur.
- La phase d'*interrogation* qui interprète la requête et utilise une fonction de correspondance, afin de trouver les documents pertinents. Ces derniers sont jugés les plus similaires avec les informations demandées par l'utilisateur.

Nous étudions par la suite les défis dans la RI.

1.2 Des variations linguistiques aux défis dans la RI

1.2.1 Les variations linguistiques

Le langage naturel est le moyen usuel pour exprimer des informations. Pourtant, le langage naturel est très ambigu. De plus, il existe des variations à différents niveaux : morphologique, lexicale, syntaxique. Cela représente un vrai défi pour la RI lorsque l'on veut représenter le contenu des textes.

- **La variation morphologique**

  La variation morphologique est un phénomène linguistique qui concerne la structure des mots. Il y a deux types de morphologie : " inflectionelle" et "dérivationelle". Le premier type de morphologie décrit les changements dans la structure interne entre les mots du terme sans effet sur leur catégorie grammaticale et un faible effet sur leur sens, par exemple : la forme plurielle, possessive, comparative, participe passé, etc. Le deuxième type de morphologie peut influer ou non sur les catégories grammaticales des mots.

  Les problèmes concernant ce type de variation sont normalement résolus par des techniques de lemmatisation et d'expansion de requête afin d'augmenter le taux de

rappel. Dans l'expansion de requête, les mots ajoutés à la requête sont les variants morphologiques des mots retrouvés dans la requête. La lemmatisation a pour but de normaliser les différents variants à un lemme. Cette technique peut être soit statistique, soit linguistique : elle prend en compte le contexte grammatical des mots. L'efficacité des outils de lemmatisation dépend aussi de la complexité de chaque langue. En conclusion, cette technique est considérée comme efficace pour augmenter la performance d'un SRI.

- **La variation lexicale**

  Ce type de variation concerne normalement le cas de synonymie. La synonymie est la "Relation sémantique entre des mots ou des expressions dont les sens sont identiques ou très proches"[1], par exemple : "mort", "décès". Cependant, il faut noter que la synonymie absolue, dont les deux mots synonymes sont interchangeables dans tous les contextes, est très rare. A cause de la synonymie, les documents et les requêtes ne décrivent pas toujours le même sens par le même mot. Les méthodes d'expansion des requêtes par des termes synonymes avec les termes des requêtes ont été proposées en consultant une base lexicale contenant des termes synonymes comme Wordnet. Pourtant, C. de Loupy a fait une évaluation du taux de synonymie dans un texte et il fait les constatations intéressantes suivantes :

  "Même si les taux de polysémie et de synonymie théoriques sont importants, le taux réel d'utilisation des différents sens d'un terme et des différents mots possibles pour exprimer un concept donné est relativement faible" [23].

  Ce phénomène concerne normalement les concepts fréquents qui apportent le moins d'information, ou les moins discriminants. Ce travail conclut que les phénomènes de synonymie dans un texte ne provoquent pas de problème pour un SRI. Malgré cela, une expansion de requête semble quand même intéressante dans le cas où les termes de la requête sont moins pertinents que leurs synonymes proposés à ajouter.

- **La variation syntaxique**

  La variation syntaxique concerne le changement dans la structure des composants d'un syntagme, par exemple : "coupe du monde", "coupe mondiale". Comme les syntagmes sont plus expressifs dans la présentation du contenu d'un texte, des travaux sur l'indexation avec des termes syntagmes nominaux ont été proposés et montrent une augmentation de la performance de la RI [7]. Cependant, l'exploitation des autres types de syntagmes, par exemple les syntagmes verbaux, et leur identification reste une question dont la réponse n'est pas claire.

- **Les relations sémantiques**

[1] http ://www.granddictionnaire.com

Les relations sémantiques entre termes sont importantes dans la compréhension du texte. Leur classification est très variée selon différents points de vue et a beaucoup de réponse jusqu'à ce jour. Parmi celles-ci, la classification de Faradanne [15] est assez complète et générale pour tous les domaines. Pourtant, l'extraction et le typage de manière claire de toutes ces relations sémantiques dans un SRI sont un défi à cause de l'ambiguïté de la langue naturelle.

1.2.2 Défis

L'augmentation de la performance d'un SRI reste un défi permanent pour les chercheurs dans ce domaine. Le volume des grandes masses de documents et l'ambiguïté linguistique de la langue naturelle sont des obstacles que doivent passer les SRI pour augmenter leur performance. Les constatations suivantes ouvrent différentes directions de recherche :

- Sur la description ou l'indexation des documents et de la requête : L'indexation basée sur les termes n'est pas assez précise pour décrire le contenu d'un document ou d'une requête à cause de leur ambiguïté.
- Sur la correspondance : Un document pertinent ne partage pas toujours les mêmes termes avec la requête. De ce fait, les documents pertinents qui ne contiennent pas les mêmes termes avec la requête ne sont pas trouvés. C'est le problème de la disparité de termes (mistmatch). Par exemple : la différence dans le style d'écriture des différents types de documents (scientifique, médical, social, etc.) et la faible taille des requêtes de l'utilisateur sont des obstacles que la RI doit passer. De plus, les relations sémantiques entre les termes ne sont pas encore bien traitées dans la RI. Par exemple la relation "Is-a" ou "Synonym" entre un terme d'indexation du document et celui de la requête dans la figure 1.1. L'exploitation de différents types de relations entre les termes ou concepts doit permettre d'augmenter la performance de la recherche.

1.3 Contribution

Dans l'objectif de résoudre les défis mentionnés, nous proposons un modèle de RIM avec une indexation plus précise ainsi qu'une correspondance ayant la capacité de prendre en compte les relations sémantiques.

Pour une indexation plus précise et indépendante de la langue, nous proposons une indexation conceptuelle à base d'une ressource externe. Un concept est généralement similaire à la notion de catégorie [33]. Un concept peut être également défini comme

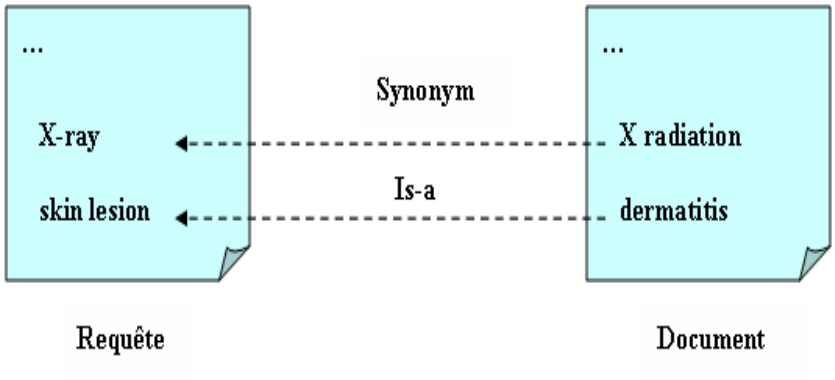


FIG. 1.1 – Exemple des relations sémantiques entre terme d'indexation du document et de la requête

une notion abstraite et compréhensible par l'humain, mais indépendante des supports matériels, des langues et des représentations informationnelles. C'est pour cette raison que l'indexation conceptuelle a la capacité de résoudre le problème des variations des termes, qu'elle soit morphologique, lexicale ou syntaxique. De plus, la conceptualisation unifie les termes dans les langues différentes en une forme unique, c'est à dire qu'elle fait tomber la barrière de la langue. De ce fait, une indexation conceptuelle pour la RI multilingue sera effectuée de manière identique que la RI monolingue. Les concepts sont extraits à partir d'une ressource externe. Une ressource externe, dans le contexte de la RI, est constituée d'informations sur la description des termes ou concepts, leur classification et leur structuration ainsi que leurs relations.

Pour résoudre le problème de disparité, étant donné une présentation conceptuelle des documents et de la requête, nous nous posons la question :
"*Quel modèle de RI est approprié pour prendre en compte des relations sémantiques dans la correspondance* ?"

Le modèle basé sur le réseau Bayésien est le modèle approprié par sa capacité à modéliser explicitement des liens entre concepts (via la forme graphique) et de prendre en compte le poids de ces liens dans la fonction de correspondance (via le procédure d'inférence probabiliste). Nous proposons donc un modèle à base de réseau Bayésien des concepts et des relations sémantiques. Les relations sémantiques entre concepts sont extraites à partir d'une ressource externe. Elles jouent le rôle de pont entre concepts sémantiquement liés afin de résoudre la disparité. De plus, nous intégrons la mesure de similarité entre concepts sémantiquement liés dans la fonction de correspondance via le processus d'inférence de probabilité sur le réseau Bayésien. Cette intégration n'a pas besoin d'ajouter de nouveaux concepts, donc elle ne change pas la distribution d'origine des

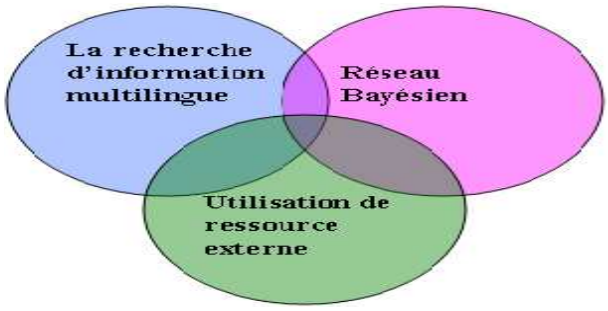


FIG. 1.2 – Les domaines concernés dans le cadre de la thèse

concepts dans les documents ou la requête comme la méthode d'expansion des documents ou des requêtes.

Le travail de thèse s'articule dans le cadre de la RI monolingue et multilingue, autour de l'utilisation d'une ressource externe, et du modèle basé le sur un réseau Bayésien. La figure 1.2 illustre le cadre de cette thèse.

Dans un second temps, à partir d'une indexation à base de concepts, nous étudions une extension à des documents et requête structurés et multi-médias. La structuration des documents et de la requête sous forme des dimensions du domaine permet de reclasser les documents selon les dimensions qu'ils partagent avec la requête. Ce reclassement renforce la correspondance entre document-requête et est donc supposé améliorer la performance de recherche. De plus, les concepts peuvent représenter non seulement le contenu des textes, mais aussi des images. C'est pour cette raison que dans la RI sur les documents multi-médias, les images et les textes peuvent être indexés par un ensemble de concepts commun. Nous suggérons une fusion de la recherche à base du contenu textuel avec la recherche à base du contenu d'images. Cette fusion permet aussi une meilleure correspondance document-requête.

Afin d'appliquer concrètement ces travaux et de valider les résultats, le développement et les évaluations des propositions dans le domaine médical ont été effectués.

1.4 Structure de la thèse

La thèse est organisée en 8 chapitres :

Le chapitre 1 est une introduction générale du domaine de la RI et les défis concernant notre contribution. Le développement se décompose en trois grandes parties qui se répartissent de la façon suivante.

Tout d'abord la partie 1 qui correspond à l'état de l'art, comprenant les trois chapitres du numéro deux au numéro quatre :

- Le chapitre 2 concerne un rappel sur les modèles de RI et la RI multilingue.
- Dans le chapitre 3, nous nous intéressons aux ressources externes et à leur application dans la RI générale ainsi que la dans RI dédiée à un domaine spécifique tel que le domaine médical. Nous y introduisons aussi UMLS et ses applications en RI.
- Le chapitre 4 est un état de l'art de l'approche Bayésienne dans la RI.

La partie 2 est la contribution, se divise en trois chapitre :

- Le chapitre 5 se rapporte à la proposition du modèle basé sur le réseau Bayésien des concepts et de leurs relations sémantiques. Ce chapitre constitue la contribution principale de la thèse.
- Le chapitre 6 se concentre sur la validation du modèle Bayésien proposé. Elle concerne l'application et l'évaluation des propositions dans le domaine médical avec l'utilisation d'UMLS. Il s'agit d'une application de la RI sur des images médicales avec la collection de CLEF images médicales. Nous validons le modèle proposé et l'intérêt de la prise en compte des relations sémantiques.
- Le chapitre 7 aborde l'extension à des documents et requête structurés et multimédias. Cette partie de la contribution, aussi validée sur la collection ImageCLEFMed, ouvre des perspectives intéressantes pour les travaux futurs.

Enfin, le chapitre 8 conclut notre travail de thèse, synthétise les différents aspects de la contribution et décrit les nouvelles perspectives.

Première partie

ETAT DE L'ART

Dans cette partie de l'état de l'art, le chapitre 2 est un panorama général des modèles de la RI monolingue et multilingue, ainsi que des approches d'amélioration de la performance de recherche. Le chapitre 3 présente la notion de ressource externe et l'état de l'art de son utilisation dans la RI. Nous introduisons aussi la ressource UMLS et ses applications pour la RI dans le domaine médical. Ce chapitre met en valeur l'intérêt des ressources externes dans la RI. Le chapitre 4 est l'état de l'art des modèles de la RI à base de réseau Bayésien, sur lequel s'appuie notre modèle proposé.

Chapitre 2

La recherche d'information monolingue et multilingue

Sommaire

2.1 Les modèles de RI

Il existe trois familles de modèles de RI principales dans l'état de l'art : modèle booléen, modèle vectoriel et modèle probabiliste.

- Le *Modèle booléen* est basé sur la théorie logique et des ensembles. Un document d est représenté par un ensemble de termes ; une requête q est représentée par une expression logique de termes. La similarité entre document et requête est la validité de l'implication $d \Rightarrow q$. Les termes ne sont alors pas pondérés, et les documents retrouvés ne sont pas triés.
- Dans le *modèle vectoriel* (VSM), documents et requêtes sont représentés dans l'espace de dimension N, avec N étant le nombre total des termes d'indexation. Un document est représenté par un vecteur des poids des termes correspondants :

  $$D = \{w_{d_1}, w_{d_2}, ..., w_{d_N}\}$$

  et une requête est représentée de manière similaire par :

  $$Q = \{w_{q_1}, w_{q_2}, ..., w_{q_N}\}$$

  où w_{d_i} et w_{q_i} sont les poids du terme t_i dans le document D et la requête Q.
  Avec cette représentation des documents et de la requête, il existe plusieurs méthodes de calcul de la similarité document-requête, ou le RSV (Relevant status value). Dans le système SMART [73], le *cosinus* a été utilisé :

  $$RSV(D,Q) = \frac{\sum_{i=1}^{N} w_{q_i}.w_{d_i}}{(\sqrt{\sum {w_{q_i}}^2}).(\sqrt{\sum {w_{d_i}}^2})}$$

- Le principe de la recherche du *modèle probabiliste* est basé sur la probabilité de pertinence d'un document par rapport à une requête[70]. Soit R la pertinence, le modèle détermine les probabilités $P(R = 1|D)$ et $P(R = 0|D)$. Ces probabilités signifient la probabilité d'obtenir une information pertinente (ou non-pertinente) si on retrouve le document D. Une autre type de modèle probabiliste, modèle de langue, est proposée par Ponte [59]. Les modèles de langue déterminent la probabilité $P(Q|D)$ - la probabilité que la requête Q puisse être générée à partir du document D [47]. Les modèles basés sur le réseau Bayésien, ou réseau d'inférence, s'intègrent aussi dans ce modèle probabiliste. L'état de l'art de ces modèles est présenté dans le chapitre 3.

2.2 La RI multilingue

Avec la grande masse de données disponibles aujourd'hui, le nombre de documents disponibles dans les différentes langues augmente. Dans le domaine de la RI, un utilisateur qui soumet une requête dans sa langue préférée, peut aussi avoir besoin de trouver des documents dans d'autres langues. La raison est qu'il exprime son besoin d'information bien dans une langue mais qu'il veut aussi chercher des documents dans d'autres langues. La recherche d'information multilingue (RIM) a pour objectif de satisfaire ce besoin. Par définition, la RIM vise à faire tomber la barrière de la langue afin de permettre aux utilisateurs de trouver les documents dans des langues différentes de la langue de la requête.

Afin d'atteindre ce but, en général, la solution adoptée pour un système de RIM est de traduire la requête et les documents dans une même langue, quelle que soit leur langue d'origine. Après cette transformation, la RIM devient une RI monolingue. Il y a trois approches possibles pour cette traduction :

- la traduction de la requête dans la langue des documents ;
- la traduction des documents dans la langue de la requête ;
- la traduction des documents et de la requête dans une langue intermédiaire ou un langage pivot. Ce langage pivot peut être une langue naturelle populaire (comme l'anglais) ou une langue artificielle, par exemple UNL (Universal Networking Language)[1].
- l'utilisation de concepts pour représenter les sens des termes issus de langues différentes.

La traduction de la requête est moins coûteuse à mettre en œuvre que la traduction des documents. En effet, le nombre de termes d'une requête est normalement beaucoup plus faible que dans un document. Pourtant, la traduction a souvent besoin du contexte des termes et les requêtes courtes ne contiennent pas suffisamment d'informations contextuelles par rapport aux documents. Les techniques de traduction de requêtes ou de documents comprennent :

- l'utilisation d'un dictionnaire bilingue ;
- l'utilisation d'un système de traduction automatique ;
- l'utilisation de textes parallèles

L'approche de l'utilisation de dictionnaires bilingues consiste à traduire simplement les termes [58]. Les limites de cette méthode concernent la dépendance à la couverture du dictionnaire, ainsi que l'ambiguïté de la traduction. En effet, dans le cas où un terme peut avoir plusieurs sens, il faut alors décider du sens correct pour la traduction.

[1] www.undl.org

Utiliser un système de traduction automatique rend forcément dépendant de la qualité de traduction du logiciel utilisé [54]. De plus, il est difficile de trouver un bon logiciel de traduction automatique quelles que soient les langues.

Dans le cas de l'utilisation de textes parallèles, il s'agit de remplacer un texte dans une langue source par un texte aligné dans la langue cible. Il ne s'agit pas à proprement parler de traduction, mais plutôt du remplacement entre une paire de textes parallèles, ou alignés. La construction d'une base de textes parallèle est effectuée sur les corpus parallèles. Une unité d'alignement peut être un document, un paragraphe, une phrase, un syntagme ou un mot.

Les performances de la RIM sont normalement inférieures à celles de la RI monolingue. Nie [53] a utilisé un modèle de traduction probabiliste à partir de corpus parallèles pour la RIM. Dans les expérimentations sur TREC6 et TREC7, cette méthode obtient un meilleur résultat qu'une méthode basée sur un dictionnaire bilingue, et elle est comparable avec une méthode basée sur une système de traduction automatique avec une précision allant jusqu'à 94% de la RI monolingue.

2.3 L'évaluation de la performance de recherche

Après la construction d'un SRI, l'évaluation de la performance de recherche concerne le niveau de satisfaction de l'utilisateur pour les informations qu'il obtient et la qualité en temps de réponse. De cette manière, nous sommes en mesure de savoir quelles approches sont efficaces. L'évaluation d'un SRI joue donc un rôle très important. La qualité du résultat et la vitesse sont deux critères principaux à évaluer.

2.3.1 La pertinence

La qualité des résultats, c'est à dire la pertinence des documents, est jugée par le système et ensuite par l'utilisateur. Après la procédure de recherche, le SRI donne son résultat sous la forme d'une liste de documents retrouvés par le système : c'est la pertinence système. A son tour, l'utilisateur va juger la qualité de cette liste (pertinence utilisateur). Afin de ne pas introduire d'ambiguïté par la suite, nous utilisons le terme «pertinence» pour parler de la pertinence pour l'utilisateur. La pertinence d'un système est mesurée via :

- Le taux de rappel : c'est la capacité du système à donner tous les documents pertinents. C'est la proportion des documents pertinents retrouvés parmi tous les documents pertinents dans la collection des documents.

- Le taux de précision : c'est la capacité du système à retrouver seulement les documents pertinents. C'est la proportion des documents pertinents parmi les documents retrouvés.
- La MAP (Mean Average Precision) est une mesure plus récente. Cette mesure, qui est standard dans la communauté de TREC (Text Retrieval Conference)[2], donne une évaluation de la qualité de la recherche par les niveaux de rappel. Pour un besoin d'information, la valeur de précision moyenne est la moyenne des précisions obtenues pour l'ensemble des k premiers documents après que chaque document pertinent a été retrouvé. Ces valeurs sont ensuite utilisées pour calculer la moyenne pour un ensemble de requêtes. Concrètement, pour chaque requête $q_j \in Q$, si l'ensemble des documents pertinents de q_j est $d_1, ...d_{mj}$, et R_{jk} est le rang des documents dans le résultat, compté à partir du premier document jusqu'au document d_{mj}, on a [49] :

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{mj} \sum_{k=1}^{mj} Precision(R_{jk}) \tag{2.1}$$

2.3.2 La vitesse

Le temps d'attente pour avoir le résultat est aussi un critère d'évaluation de la satisfaction de l'utilisateur par rapport au système. La vitesse du système est jugée non seulement via le temps de réponse après l'entrée d'une requête mais aussi via le temps pendant les interactions avec l'utilisateur pour la reformulation et le raffinement de la requête.

2.4 Amélioration de la performance de la RI

Avec les contributions des chercheurs dans le domaine de la RI, les SRI évoluent. Le problème n'est plus seulement de trouver les documents, mais d'améliorer les résultats afin de mieux satisfaire l'utilisateur. Comme le fonctionnement d'un SRI se décompose en deux phases, les études sur l'augmentation de la performance d'un SRI portent donc également sur ces deux phases :

- Sur la phase d'indexation : cela concerne l'étude de l'amélioration de la qualité des termes d'indexation et du modèle d'indexation. Par exemple, il existe des recherches sur l'indexation relationnelle ainsi que sur les termes d'indexation composés.
- Sur la phase d'interrogation : cela concerne l'étude sur l'extension de la requête et sur la fonction de correspondance.

[2] `http://trec.nist.gov/`

Le but final de ces travaux est d'augmenter la performance du système, autrement dit d'élever la précision et le rappel dans une proportion significative.

2.5 Conclusion

Ce chapitre est un rappel des modèles de RI et de RIM, ainsi que de leur performance. Le but général de l'évolution des SRI se centre sur l'amélioration de la performance de recherche. L'utilisation des ressources externes est l'une des approches pour cette évolution. Nous dressons dans le chapitre suivant un état de l'art sur l'utilisation des ressources externes dans la RI.

Chapitre 3

La Recherche d'Information basée sur une ressource externe

Sommaire

3.1 Introduction

Grâce à une grande disponibilité des ressources externes (générales ou spécifiques), les systèmes à base de connaissances externes se développent. Dans le domaine de la RI, les ressources externes, i.e. les ressources hors de la collection des documents et de la requête, jouent un rôle important. Ces ressources offrent des connaissances sur la corrélation entre des mots et des termes, donc une meilleure compréhension du texte. La compréhension du texte évolue alors du niveau linguistique vers le niveau sémantique. De ce fait, dans la RI, la correspondance entre un besoin d'information et des documents peut évoluer du niveau de correspondance entre sacs de mots vers la correspondance sémantique. Nous consacrons ce chapitre à un état de l'art, d'une part sur les approches de l'utilisation des ressources externes dans la RI en général, et d'autre part sur une ressource externe du domaine médical : UMLS, et les travaux qui l'utilisent dans la RI médicale.

3.2 Les ressources externes

Il existe une grande variété de ressources externes. Certains types de ressources sont populaires, nous reprenons leur définition issue du grand dictionnaire[1] dans le domaine des technologies de l'information :

- *Vocabulaire contrôlé* : «Dans un domaine préalablement défini (d'ordre scientifique, technique, professionnel ou autre, et en général pour une langue donnée), le choix de termes sélectionnés, classés et indexés en vue de faciliter l'indexation, le stockage et la recherche des publications traitant des concepts apparentés à ces termes.»
- *Taxonomie* : «Construction d'un plan de classification de concepts utilisant des classes disjointes de concepts agrégés.»
- *Thésaurus* : «Vocabulaire contrôlé et dynamique de termes ayant entre eux des relations sémantiques et génériques, et qui s'applique à un domaine particulier de la connaissance.»
- *Ontologie* : «Ensemble d'informations dans lequel sont définis les concepts utilisés dans un langage donné et qui décrit les relations logiques qu'ils entretiennent entre eux».

Le vocabulaire contrôlé ne contient que des termes d'indexation pour faciliter l'accès et l'indexation des documents, alors que la taxonomie classifie, organise les termes et possiblement ajoute des relations hiérarchiques entre groupes de termes. Le thésaurus possède en plus des relations sémantiques entre termes du type de la causalité, l'association,... L'ontologie, quant à elle, est plutôt la représentation formelle des informations qui

[1] http ://www.granddictionaire.com

sont définies sous la forme des concepts avec leurs relations.

Parmi les ressources externes, Wordnet [51] et EuroWordnet [55] sont les plus utilisés dans la RI générale, alors que UMLS (Unified Medial Language System) est utilisé pour la RI médicale. Par ailleurs, les méthodes pour enrichir des ressources sont aussi étudiées [35].

Nous introduisons par la suite plus de détails sur l'application de ces ressources dans la RI générale.

3.3 L'utilisation des ressources externes dans la RI

L'utilisation de ressources externes aux documents dans la RI a déjà été largement explorée. Les applications principales concernent :

- L'extraction de concepts et l'indexation conceptuelle : il s'agit d'identifier les concepts correspondant aux termes et de les utiliser dans l'indexation.
- L'expansion de requêtes et de documents consiste à ajouter les concepts sémantiquement liés aux concepts des requêtes ou des documents.
- Le calcul de similarité sémantique entre concepts via les informations sur les concepts et leurs relations dans les ressources externes.

3.3.1 Indexation conceptuelle

La notion de concept a différentes définitions selon différents points de vue. En pratique, un concept est identifié par des informations associées qui le décrivent. Ces informations sont souvent des textes et des termes, mais aussi des morceaux d'images, des définitions logiques ou des contraintes. Par exemple CYC [44] est un large ensemble de concepts sous un format lisible par une machine dans lequel les concepts sont décrits par des expressions logiques et un ensemble de termes connexes. ConceptNet [46] est plus informel, sans langage logique. Il capture des connaissances de bon sens (common sense knowledge) en insistant sur les relations sémantiques riches. Tandis que dans Wordnet [51], les noms, verbes et adjectifs en anglais sont organisés en ensembles de synonymes, ou "synsets". Dans UMLS, un concept regroupe différents termes qui proviennent de différentes sources.

L'identification de concepts dans un texte pour l'indexation conceptuelle nécessite de résoudre l'ambiguïté des termes [10], [5]. Dans un domaine spécifique comme le domaine médical, la désambiguïsation est légèrement moins compliquée que dans le domaine général parce qu'un concept dans le domaine général peut porter différents sens dans différents domaines.

L'indexation par termes a une limite : le problème de "term mismatch" (ou disparité) à cause des variations morphologiques, lexicales ou syntaxiques des termes. Comme les concepts portent sur le sens, ils sont donc plus abstraits que les termes et ils sont supposés mieux décrire le contenu des documents et par conséquent, meilleurs pour l'indexation.

L'indexation au niveau conceptuel par contre, fait face au problème de l'ambiguïté dans l'identification des concepts à partir de textes. Cette ambiguïté vient du fait qu'un terme peut possésder plusieurs sens. Ce problème dépend des ressources qui définissent les concepts et leurs sens. Un concept d'une ressource générale est normalement plus ambigu qu'un concept d'une base de connaissances d'un domaine spécifique. Parmi les ressources lexicales, Wordnet est la plus exploitée. Gonzalo [34] a commencé avec l'utilisation des synsets de Wordnet comme espace d'indexation au lieu de l'espace de termes dans le modèle d'espace vectoriel. L'indexation conceptuelle avec une désambiguïsation manuelle de la collection de test a amélioré le résultat de recherche (29%) par rapport à l'indexation par termes. Cependant, sans cette désambiguïsation manuelle des requêtes, le résultat obtenu est seulement identique avec l'indexation basée sur les termes.

Dans le travail de Baziz [10, 12, 9, 11], les documents sont représentés par des synsets de Wordnet. Il a proposé deux schémas de recherche :

- Premier schéma : Le document est représenté comme un "noyau sémantique" extrait à partir des synsets de Wordnet. Dans ce travail, Baziz a fait l'hypothèse qu'un concept correspond à un sens. Afin de choisir le meilleur sens possible des termes à désambiguïser, il a utilisé différentes mesures de similarité sémantiques pour calculer la similarité d'un concept par rapport aux autres concepts dans les documents. L'hypothèse de cette désambiguïsation est que parmi tous les concepts candidats d'un terme, le plus vraisemblable est celui qui a le plus de liens avec les autres concepts dans le même document. Un document ou une requête peut donc être représenté par un réseau ou un "noyau sémantique" des concepts liés. Ces concepts sont pondérés par une variation de $tf.idf$ en prenant en compte des similarités sémantiques les uns par rapport aux autres. Le modèle vectoriel est ensuite utilisé. L'expérimentation avec cette méthode a été réalisée dans le cadre de la tâche anglais GIRT de la campagne CLEF2004. Elle est classée en deuxième sur 3 groupes de participants, avec une précision moyenne de 0.3855.
  Nous reprenons un exemple de Baziz, la Figure 3.1 représente un réseau sémantique possible contenant les nœuds $\{C_2^1, C_7^2, C_1^3, C_1^4, C_4^5, C_2^m\}$ résultant de la combinaison du 2ème sens du premier terme T_1, du 7ème sens de T_2,.., du 2ème sens de T_m. Les liens entre les concepts ou nœuds représentent les valeurs de proximité sémantique entre les nœuds. Elles sont calculées en utilisant des mesures de similarité.

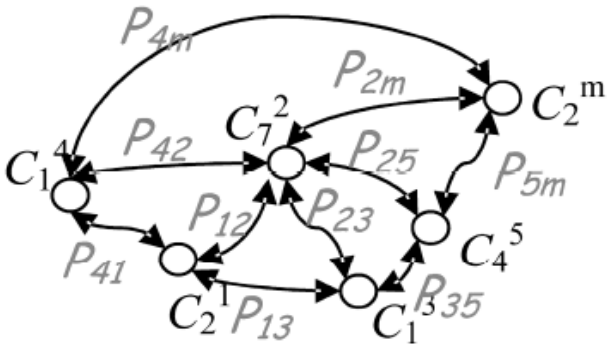


FIG. 3.1 – Exemple de réseau sémantique construit à partir d'une configuration de concepts candidats [8].

– Dans le deuxième schéma, documents et requêtes sont représentés par des sous-arbres de synsets de Wordnet. Les liens entre synsets dans ces arbres sont les relations de type Is-a. Un sous-arbre minimum est construit pour représenter à la fois documents et requêtes. Basés sur ce sous arbre minimum, les sous arbres des documents et des requêtes sont élargis avec de nouveaux concepts. La similarité document-requête est proposée comme le degré d'appartenance de la requête au document. Cette dernière est calculée par une méthode d'appariement flou. Ce schéma donne des résultats comparables avec le premier schéma. Pourtant cette méthode se limite aux cas où les requêtes peuvent être représentées par un sous arbre de concepts.
Baziz a illustré ce schéma par la Figure 3.2.

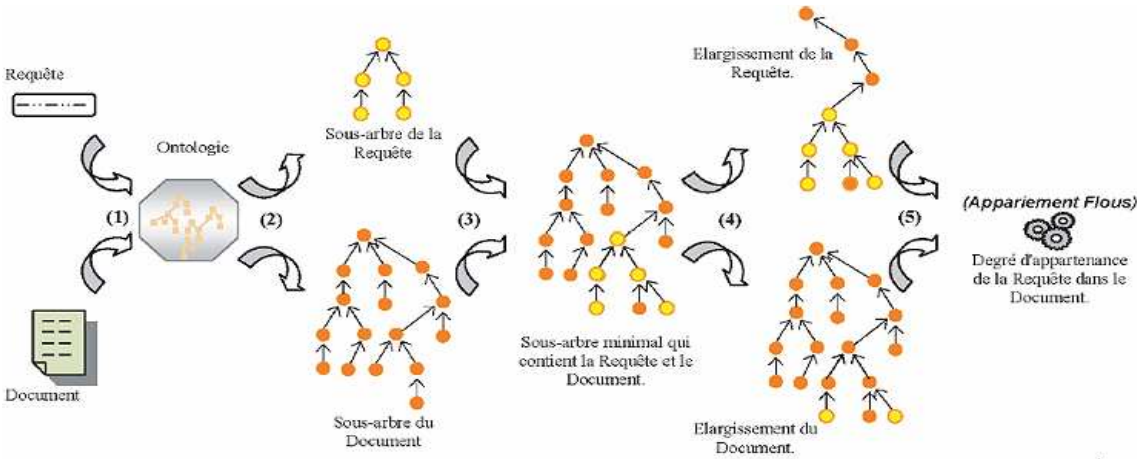


FIG. 3.2 – Schéma général de l'approche de représentation basée sur les sous arbres [8].

3.3.2 L'extension de requête ou de documents

La deuxième application de ressource externe en RI a pour objectif de résoudre le problème de "term mismatch", c'est-à-dire que documents et requête ne partagent pas le même ensemble des termes d'indexation. La solution est l'extension de l'ensemble des termes d'indexation des documents ou de la requête.

L'extension de la requête est définie comme un traitement pour élargir le champ de la recherche d'une requête donnée en ajoutant des termes similaires avec ceux de la requête. L'extension des documents est similaire à la requête. L'objectif est d'obtenir une meilleure performance de recherche. La figure 3.3 décrit un exemple de cette extension.

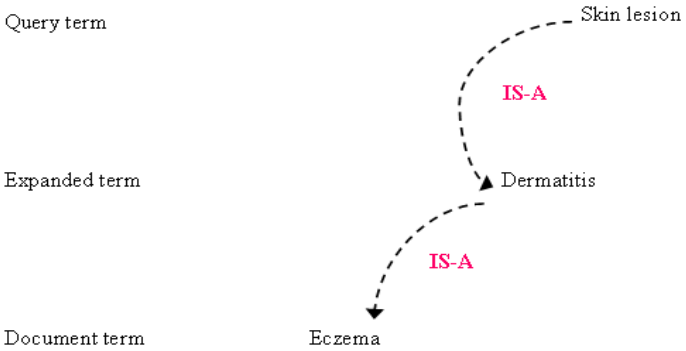


FIG. 3.3 – Exemple de l'extension

Les problèmes traités dans cette approche sont liés principalement à deux aspects :

- *Quels termes doit-on utiliser pour étendre la requête ?*
  En ajoutant des termes à la requête, le taux de rappel peut être augmenté. Cependant, si les termes ajoutés ne sont pas vraiment nécessaires pour la recherche, le taux de précision peut diminuer.
- *Comment ces nouveaux termes doivent-ils être ajoutés dans la requête ?*
  Il existe deux manières d'ajouter de nouveaux mots à la requête :
  - De manière manuelle : par une interaction avec l'utilisateur pour l'aider à définir le plus exactement possible son besoin d'information.
  - De manière automatique : c'est le système qui reformule la requête sans intervention de l'utilisateur. Nos études portent sur ce type d'extension de requête et donc quand nous parlons d'extension, par défaut il s'agit de ce type d'extension.

L'extension de la requête est plus populaire que l'extension des documents dans un grand nombre de recherches. L'état de l'art de l'extension de requête peut se classer en trois groupes, en fonction du contexte et dans le but de trouver les termes similaires à ajouter au document ou à la requête [24] :

- *L'approche basée sur une collection (ou analyse globale)* : le contexte global du terme dans la collection (par exemple la distribution, co-occurrence des termes) est étudié pour trouver les termes similaires à ajouter [32]. Par ailleurs, Helen [57] a montré les limites de co-occurrence dans l'extension de requête. Elle conclut que la performance du SRI n'augmente pas de manière très significative avec cette méthode pour les raisons suivantes :
  - La similarité entre deux termes est maximale quand ces deux termes ont des fréquences documentaires comparables. Alors, si X est un terme de la requête, les termes ajoutés (ou les termes jugés les plus similaires) auront une fréquence documentaire comparable.
  - Quand les termes de la requête et ses voisins les plus proches ont une fréquence documentaire comparable, ils ont une même capacité de discrimination documentaire.
  - Les termes de forte fréquence documentaire sont de faibles discriminants pour distinguer un document pertinent des non-pertinents. Ces termes ajoutés à la requête par n'importe quelle méthode d'expansion seront de mauvais discriminateurs.
- *L'approche basée sur la requête (ou analyse locale)* : Le contexte du terme dans ce cas est réduit à un sous ensemble des documents dans les documents retournés, il vient du (pseudo) retour de pertinence de l'utilisateur ou du profil d'utilisateur ou des requêtes dans le passé [41]. Cette approche est donc orientée requête et dépend de la performance du système ainsi que du jugement de l'utilisateur. Billerbeck [14] a montré des améliorations de cette méthode dans la RI sur le web, avec une augmentation de 26% à 29% par rapport à la non-expansion de requête. Cui [24] analyse un profil utilisateur, qui contient les informations des requêtes et des documents retrouvés, ainsi que des documents vus (cliqués) dans les réponses. Il considère que les documents cliqués par l'utilisateur sont les documents pertinents pour la requête correspondante. De ce fait, les termes de ces documents et de la requête correspondante ont des corrélations. L'expérimentation de cette méthode a montré une amélioration significative de la performance pour tous les types de requêtes (longues ou courtes). La limite de cette approche basée sur le retour de pertinence de l'utilisateur réside principalement dans la difficulté de récupérer les retours de pertinence, à propos de la qualité ainsi que de la quantité des jugements.
- *L'approche basée sur des ressources externes* : c'est une exploration des ressources externes comme Wordnet [84], [48]. Cette méthode trouve des termes à ajouter grâce aux liens sémantiques explicites disponibles dans ces ressources.

Les méthodes d'extension de documents sont similaires à l'extension de requête. L'extension de documents est utilisée principalement dans la recherche de discours, par exemple par Singhal [75]. Dans la RI, l'extension de documents a été étudiée plus récemment par Billerbeck [13] et Wang [85]. Ces travaux ont montré l'efficacité de l'extension de documents.

3.3.3 Les mesures de similarité sémantique ou distance sémantique

L'estimation de la ressemblance du sens entre termes ou concepts est un domaine difficile du TALN (Traitement automatique de la langue naturelle) et de la RI, par exemple la ressemblance entre termes ou concepts qui sont liés par la relation d'*hyponymie/hyperonymie* (ou *Is-A*), et de *meronymie/holonymie* (ou *Partie-Entier*). La notion de *similarité sémantique* (ou son inverse : *distance sémantique*) est utilisée pour exprimer la ressemblance du sens de ces termes ou concepts. Certaines mesures de similarité sémantique, ou distance sémantique, ont été proposées en utilisant les ressources disponibles, notamment la taxonomie dans Wordnet. Ces mesures peuvent être divisées en 4 groupes détaillés ci-dessous.

Méthode basée sur le comptage de lien (Edge counting)

Cette méthode considère la position où se trouvent les concepts sur la taxonomie. L'hypothèse est que plus il y a de liens entre deux concepts et plus proches ils sont, c'est-à-dire plus ils sont similaires.

- La mesure de similarité de Leacock et Chorodow [43] est basée sur le plus court cheminement (basé sur "IS-A") entre deux "synsets" ou concepts c_1, c_2 dans la taxonomie de Wordnet :

$$sim(c_1, c_2) = -log\frac{minLen(c_1, c_2)}{2D} \tag{3.1}$$

  $minLen(c_1, c_2)$ est la longueur du plus court chemin entre c_1, c_2 trouvé sur la taxonomie ; D est la profondeur maximale de la taxonomie. Par exemple dans la figure 3.4, supposons que la hiérarchie a la profondeur $D = 5$, la similarité sémantique est calculée comme suit :

$$sim(\text{"Argent","Crédit"}) = -log(2/(2 * 5)) = 0.7$$

- Rada [61] a utilisé la notion de *distance sémantique* de la manière suivante : deux concepts sont d'autant plus similaires que la valeur de la *distance sémantique* entre

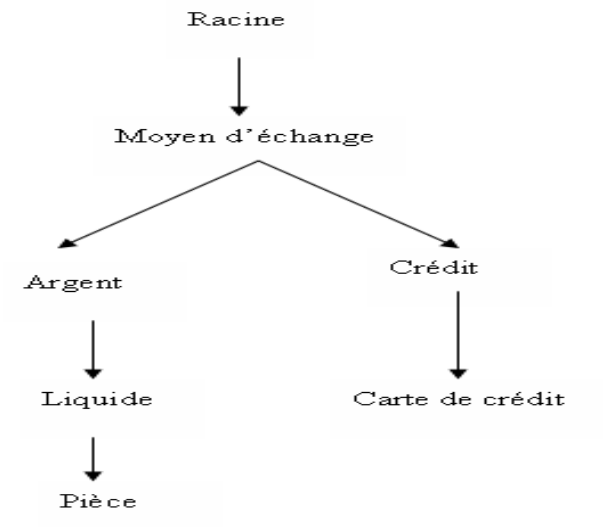


FIG. 3.4 – Exemple une taxonomie des concepts pour le calcul de la similarité sémantique

eux est faible. Cette mesure est calculée en se basant sur le chemin entre deux concepts c_1, c_2 dans la taxonomie :

$$DS(c_1, c_2) = 2L - n \tag{3.2}$$

où L est la taille de la taxonomie et n longueur minimale du chemin entre c_1, c_2. Un chemin d'origine c_1 et d'extrémité c_2 est défini par une suite finie d'arcs consécutifs, reliant c_1 à c_2. La longueur d'un chemin est la somme des arcs qui constituent le chemin.

La *distance sémantique* entre deux documents est la distance moyenne entre tous les couples de concepts u, v de deux ensembles de concepts E_1, E_2 de ces documents :

$$DS(D_1, D_2) = \begin{cases} 0 & \text{si } E_1 = E_2 \\ \frac{1}{|E_1||E_2|} \sum_{u \in E_1} \sum_{v \in E_2} DS(u, v) & \text{sinon} \end{cases} \tag{3.3}$$

L'expérience a montré que cette méthode a une corrélation de 50% par rapport aux jugements humains.

– Hirst et St-Onge [36], quant à eux, supposent que "deux concepts sont sémantiquement proches si leur synsets dans Wordnet sont liés par un court chemin qui ne change pas souvent de direction". Les directions sont décrites par : vers le haut, vers le bas et horizontale. Sa mesure de similarité sémantique considère tous les types de relations dans Wordnet :

$$sim(c_1, c_2) = C - len - k \times d \tag{3.4}$$

où *len* est la longueur du chemin de c_1 à c_2 ; d est le nombre de changements de direction sur le chemin ; C et k sont des constantes.

Méthode basée sur le contenu informationel du nœud

- Resnik [64] a défini la notion de classe comme l'ensemble des synsets qui contiennent un terme w dans la taxonomie de Wordnet : $classes(w) = \{C|w \in words(C)\}$. Ainsi, $words(C)$ est l'ensemble des termes de la classe C. L'hypothèse est que plus une classe est fréquente dans le corpus, plus la classe contient d'information. Le *contenu informationel* d'une classe C est donc estimé par une fonction de probabilité d'occurrence $P(C)$ de la classe C dans le corpus :

$$CI(C) = -log(P(C)) \tag{3.5}$$

où :

$$P(C) = \frac{Freq(C)}{N} \tag{3.6}$$

et N est la taille totale d'échantillon de texte et :

$$Freq(C) = \sum_{w \in termes(C)} \frac{1}{classes(w)} \times Freq(w) \tag{3.7}$$

$Freq(w)$ est la fréquence d'occurrence du terme w dans la collection. La similarité entre deux concepts est considérée comme la similarité entre deux classes qui contiennent ces deux concepts. L'idée pour estimer la similarité entre deux concepts est que : la similarité entre deux concepts est liée avec l'information qu'ils partagent en commun, indiquée par le *plus spécifique concept* $psc(c_1, c_2)$ qui les subsume. Le *plus spécifique concept* est supposé être le concept qui a le contenu informationel le plus grand. La similarité entre deux classes ou deux concepts dans ces deux classes est proposée comme suit :

$$\begin{aligned} sim(C_1, C_2) &= maxCI(C_i) \\ &= -log(P(psc(c_1, c_2))) \end{aligned} \tag{3.8}$$

avec C_i l'ensemble des classes qui dominent C_1 et C_2.

- Jiang et Corath [40] proposent de prendre en compte aussi les contenus d'information dans la fonction de la similarité. La distance sémantique est calculée comme suit :

$$dis(c_1, c_2) = 2log(P(psc(c_1, c_2))) - (log(P(c_1)) + log(P(c_2))) \tag{3.9}$$

- De manière similaire, Lin [45] a proposé :

$$sim(c_1, c_2) = \frac{2 \times log(P(psc(c_1, c_2)))}{log(P(c_1)) + log(P(c_2))} \tag{3.10}$$

Méthode basée sur des critères (Feature based)

Cette méthode, proposée par Tversky [81], considère les critères (features) des termes dans la fonction de similarité. Ces critères, ou l'ensemble des descriptions, concernent diverses connaissances sur les termes, par exemple les attributs, les relations,... des termes dans une ontologie.

L'hypothèse est que : deux concepts sont d'autant plus similaires s'ils ont plus de critères communs et moins de critères non-communs :

$$sim(t_1, t_2) = \frac{|d_1 \cap d_2|}{|d_1 \cap d_2| + \alpha|d_1 \backslash d_2| + (\alpha - 1)|d_2 \backslash d_1|} \tag{3.11}$$

où d_1, d_2 sont les ensembles des descriptions de termes t_1, t_2 et α est défini comme l'importance des critères non communs. Cette mesure est dans l'intervalle $[0, 1]$. Elle augmente avec la similarité et baisse avec la différence des descriptions. Cette méthode nécessite donc une base de connaissances riche en descriptions de termes.

Méthode hybride

Dans cette méthode, les chemins liant deux termes ou concepts dans la taxonomie et leurs critères peuvent être pris en compte dans la fonction de similarité. Rodriguez [71] a proposé une fonction de similarité qui est la somme pondérée des similarités basées sur l'ensemble des synonymies, sur les liens sémantiques ainsi que sur les critères. Cette mesure peut être utilisée aussi pour des similarités basées sur une ontologie unique ou multiple.

Comparaison des mesures de similarité sémantique

Les méthodes de Leacock [43], Hirst [36], Resnik [64], Jiang [40] et Lin [45] ont été expérimentées avec Wordnet et comparées dans le travail de Budanitsky [18]. L'évaluation de ces méthodes est réalisée par corrélation par rapport aux jugements humains. La méthode de Jiang et la méthode de Leacock apparaissent comme étant les meilleures globalement. La table. 3.1 montre un exemple de ces résultats. Cependant, cette comparaison est limitée à quelques douzaines de couples de termes capables d'être jugés par l'humain. Seule la méthode de Jiang considère toutes les relations dans Wordnet, toutes les autres méthodes ne considèrent que les relations d'hyponymie.

En terme de l'application de la similarité sémantique dans la RI, les recherches ont été menées selon des buts différents : pour résoudre la désambiguïsation ou le problème de la disparité (ou "mismatch"). Pour résoudre le problème de la désambiguïsation, le travail de Baziz [10] a été mentionné dans la section 3.3.1. Pour résoudre le problème

TAB. 3.1 – Corrélation entre les mesures de similarité sémantique et les jugements humains par Miller et Charles (MC) [52] ou par Rubenstein et Goodenough (RG) [72]

Similarité sémantique	MC	RG
Hirst	.744	.786
Leacock	.816	**.838**
Resnik	.774	.779
Jiang	**.850**	.781
Lin	.829	.819

de la disparité, Rada [61] a appliqué sa méthode de calcul de similarité sémantique dans la fonction de correspondance. La similarité document-requête est la somme normalisée des similarités sémantiques de tous les couples de termes. Pourtant, les expérimentations sont limitées à une très petite collection avec l'évaluation de la performance manuelle. De manière similaire, Richardson [66], [67] a appliqué aussi ces mesures dans l'appariement direct document-requête. Il a comparé 3 méthodes : la mesure de similarité basée sur le contenu de l'information de Resnik, celle basée sur le comptage de liens et la pondération $tf.idf$ [76]. Son expérimentation est menée avec la collection TREC avec l'utilisation de Wordnet. Les résultats obtenus ont montré que les deux mesures de similarité sémantique donnent des résultats comparables et significativement moins bons que ceux obtenus par la méthode avec $tf.idf$.

3.4 UMLS

UMLS (Unified Medical Language System)[2] est un projet de la NLM (National Library of Medicine) depuis 1986. L'objectif est de surmonter les deux problèmes suivants :

- Il existe des manières variées pour décrire un concept dans les différentes ressources linguistiques, sous des formes différentes.
- Des informations utiles du domaine sont distribuées dans beaucoup de ressources et systèmes disparates.

Ces deux problèmes rendent les tâches d'accès, d'indexation, etc. des informations du domaine plus difficiles. Le projet d'UMLS est né en vue de résoudre ces problèmes. Avec cet objectif, UMLS est la fusion d'environ 140 sources de données terminologiques du domaine biomédical. Il contient également des outils linguistiques en vue de faciliter les tâches d'accès, de recherche, d'intégration, et d'agrégation des informations biomédicales

[2] `www.nlm.nih.gov/research/umls/`

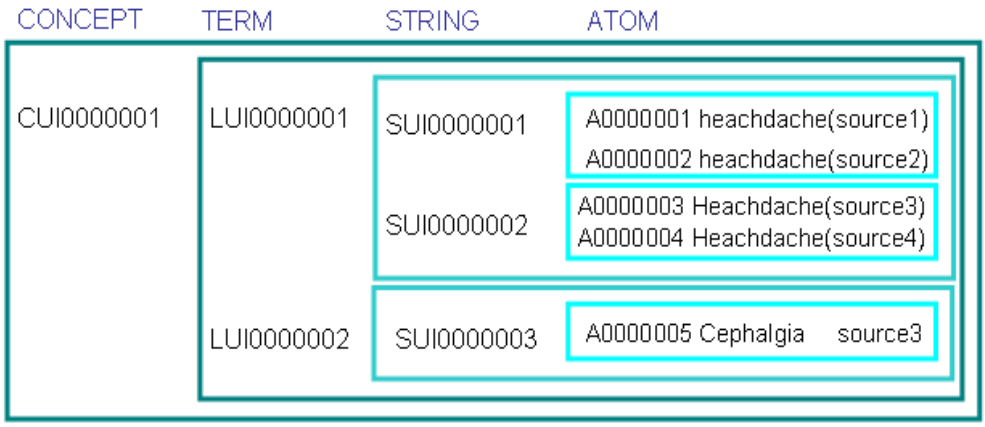


FIG. 3.5 – Exemple de la structuration d'un concept dans UMLS

et de santé. Il comprend trois composants principaux : le *Méta thesaurus*, le *Semantic Network*, et le *Specialist Lexicon*.

La structuration d'un concept dans le Méta thesaurus d'UMLS comprend quatre niveaux :

L'atome : c'est le plus petit élément dans la structure. Il représente les instances d'une chaîne de caractères venant de différentes sources ;

Les chaînes : représente les variations de forme d'une chaîne de caractères. C'est le groupement des atomes qui ont la même forme de chaîne de caractères ;

Le terme : représentent les variations de désignation d'un concept. Ce sont donc les termes des synonymes qui groupent un ensemble de chaînes ;

Le concept : représente le sens des termes. C'est le groupement des termes synonymes.

La figure 3.5 montre un exemple de cette structuration. La figure 3.6 montre la structure de Méta thésaurus et d'un réseau sémantique. Les détails sur chaque composant d'UMLS sont présentés dans l'annexe A.

3.5 Les applications d'UMLS dans la RI

Depuis sa construction, UMLS a été largement utilisé dans plusieurs tâches d'accès d'information, d'indexation, de catalogage, etc. dans le domaine biomédical. Pour la RI, les applications d'UMLS concernent :

- L'extraction de concepts pour l'indexation conceptuelle.
- La traduction de requête pour la RI multilingue.
- L'exploitation de relations sémantiques.

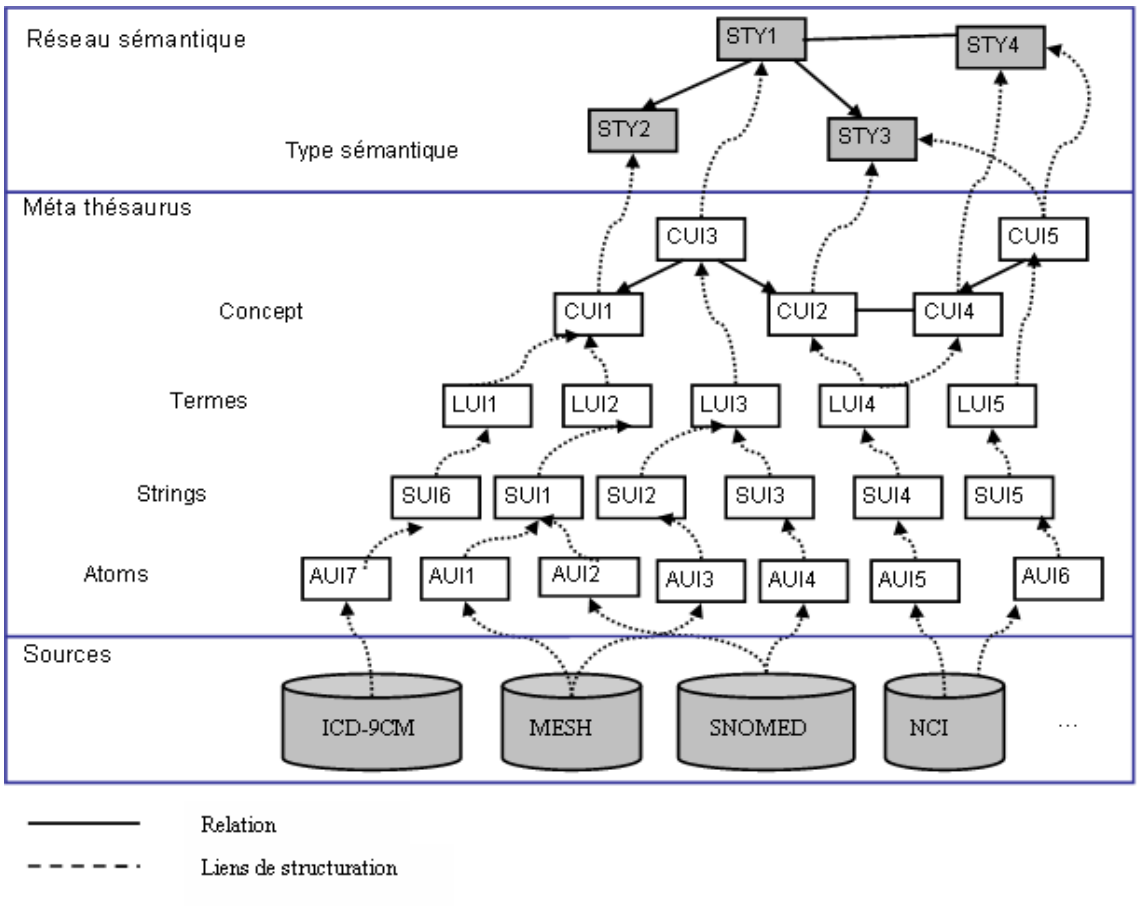


FIG. 3.6 – Structure du Métathésaurus et du réseau sémantique

– L'extension de requêtes.

L'indexation conceptuelle permet de normaliser ou d'unifier les variations de surface linguistique ou les différentes langues des termes qui ont le même sens. Aronson [5] a travaillé initialement sur l'identification des concepts d'UMLS à partir de texte pour la RI dans le domaine médical. Il a construit une requête hybride formée des termes de la requête et des concepts extraits de cette requête. Il a ainsi construit des vecteurs mixtes avec comme dimension des termes et des concepts. Les résultats ont montré une amélioration de 4% de MAP par rapport au vecteur des termes seuls. L'expérimentation a porté sur la collection de test d'UMLS qui contient des citations de Medline. Cette méthode semble utiliser des informations redondantes car l'index comprend à la fois des termes et des concepts qui correspondent à un même sens.

Une autre application d'UMLS concerne la RI multilingue. Eichman [27] a utilisé les formes textuelles multilingues d'un concept pour traduire la requête de l'anglais au français et à l'espagnol. Le système de SMART [73] a été utilisé pour comparer cette méthode avec la RI multilingue basée sur un dictionnaire. Les résultats ont montré que l'utilisation d'UMLS pour la traduction de requêtes est meilleure pour les requêtes en espagnol, mais moins favorable pour les requêtes en français.

Les relations sémantiques dans UMLS sont aussi des informations potentiellement utiles. Pour exploiter ces connaissances, Vintar [83] a annoté les relations sémantiques entre les concepts via les relations sémantiques entre les types sémantiques dans le Réseau Sémantique d'UMLS. Cette annotation atteint un taux de 17% de précision selon le jugement de l'expert. La poursuite de ce travail, [68], consiste à annoter les relations sémantiques entre les concepts qui sont co-occurrents dans les phrases. L'expérimentation a été menée avec des citations de Medline sur le système Rotondospider de Eurospider. Ce système, qui indexe à la fois les concepts et leurs relations sémantiques annotées, prouve que cette méthode améliore la performance par rapport à l'indexation avec concepts seuls.

La dernière application d'UMLS dans la littérature est dans l'expansion de requêtes. Aronson [4] a étendu une requête par les concepts correspondant au texte de la requête et les noms de ces concepts dans UMLS. Cette méthode a montré une amélioration de 9% à 15% de MAP par rapport à une requête non étendue. Ce résultat est comparable avec le retour de pertinence d'utilisateur de [77] qui utilise les termes dans MESH pour l'indexation.

3.6 Conclusion

Nous avons vu dans ce chapitre un état de l'art sur les SRI basés sur des bases de connaissances externes.

Nous avons abordé dans la première partie du chapitre des approches d'utilisation de base de connaissances dans les SRI généraux. Ces méthodes concernent : l'indexation conceptuelle, l'extension de requête ou des documents et l'utilisation des mesures de similarité sémantique. La plupart des travaux utilisent Wordnet comme ressource externe.

L'indexation conceptuelle a montré que des améliorations de la performance de recherche sont possibles, quelle que soit la méthode de désambiguïsation manuelle (Gonzalo [34]) ou automatique (Baziz [8]).

L'extension des requêtes ou des documents, à l'aide des ressources, a résolu quelque part le problème de "term mismatch" en élargissant l'ensemble des termes d'indexation par des termes similaires. Par contre, cette approche a aussi les limites suivantes :

- En ajoutant des nouveaux termes, la méthode d'expansion change la nature de la collection et de la requête, ainsi que la distribution d'origine des termes.
- Elle provoque le risque d'ajouter une masse de termes inutiles. Si c'est le cas, les termes ajoutés diminuent la performance de la recherche.
- La pondération des nouveaux concepts ajoutés est un problème qui n'est pas encore résolu de manière significativement efficace dans l'état de l'art.

Les mesures de similarité sémantique sont proposées en vue d'estimer la ressemblance sémantique des termes en utilisant notamment la taxonomie dans Wordnet. Ces mesures permettent d'identifier les termes les plus similaires sémantiquement avec un terme donné. Ces termes similaires sont ajoutés à la requête ou participent au calcul de la correspondance entre documents et requêtes [82],[61], etc. Ces applications ont montré des effets positifs.

Dans la dernière partie de ce chapitre, nous avons examiné des travaux récents sur l'utilisation d'UMLS dans la RI médicale. Ces travaux ont montré des améliorations effectives des SRI. Actuellement, l'exploitation des relations sémantiques, surtout de hiérarchie des concepts dans UMLS est très peu étudiée.

Nous constatons que l'exploitation des ressources linguistiques dans la RI est une approche intéressante. Spécialement, l'utilisation des relations sémantiques peut éventuellement résoudre le problème de disparité entre les termes de la requête et ceux des documents. La question posée est la suivante : Comment intégrer ces relations de manière explicite dans le modèle de RI pour améliorer la performance de recherche ? Nous nous sommes orientés vers un modèle à base de réseaux Bayésiens qui a la capacité de modéliser explicitement des liens entre concepts (via la forme graphique) et de prendre

en compte le poids de ces liens dans la fonction de correspondance (via la procédure d'inférence probabiliste). De plus, ce modèle permet de prendre en compte des relations entre concepts sans changer ou ajouter l'ensemble des concepts dans les documents et dans la requête. C'est une différence importante par rapport à la méthode d'extension de la requête. La théorie de l'approche Bayésienne ainsi que l'état de l'art des modèles basés sur réseau Bayésien dans la RI sont présentés dans le chapitre suivant.

Chapitre 4

La Recherche d'information basée sur un réseau Bayésien

Sommaire

4.1 Introduction

L'approche Bayésienne est connue comme une bonne solution aux problèmes contenant de l'incertitude. Le concept de chance et d'incertitude est présent dans la vie quotidienne (ex : incertitude sur la météo). Comme la RI est aussi un processus incertain, l'application de l'approche Bayésienne dans la RI a été étudié. Une des premières études est celle de Turtle et Croft, [80], [78]. Elle montre qu'un modèle de RI basé sur un réseau Bayésien est plus général et peut englober d'autres modèles comme le modèle probabiliste, booléen ainsi que la pondération tf-idf du modèle vectoriel. Ainsi, Ricardo [6] indique que *«le Réseau Bayésien fournit un formalisme clair pour combiner les différentes sources d'évidences (requêtes passées, cycle de rétroaction, formulation de requêtes) pour le calcul de la correspondance entre la requête et les documents»*.

Les notions et les théorèmes principaux de la théorie des probabilités nécessaires à la compréhension du réseau Bayésien sont présentés dans l'annexe B. Dans les sections suivantes nous abordons : la définition du réseau Bayésien et l'inférence des probabilités sur le réseau Bayésien dans la section 4.2 ; l'état de l'art des modèles à base de réseau Bayésien dans la RI dans la section 4.3 et enfin la section de la conclusion.

4.2 Le réseau Bayésien

4.2.1 Les notions de base

- Un **Graphe** $\mathcal{G}$ est un couple (S, R) où :
  - S est l'ensemble fini des nœuds (ou sommets).
  - R est un sous ensemble de $S \times S$. Les éléments dans R sont appelés les arcs(ou arêtes) du graphe. Dans le cas où $\mathcal{G}$ est un graphe orienté, un arc(n_i, n_j) est caractérisé par le nœud initial n_i et le nœud terminal n_j.
- Un **chemin** dans un graphe orienté $\mathcal{G}(S, R)$ est une suite finie d'arcs :

  $$che(n_1, n_k) = \{(n_1, n_2), ..., (n_{k-1}, n_k)\}$$

  tels que :

  $$\forall i = 1, ..., k-1 : (n_i, n_{i+1}) \in R$$

  La longueur du chemin est :

  $$len(che(n_1, n_k)) = k - 1$$

- Un graphe orienté est *acyclique* si :

  $$\nexists che(n_i, n_i), \forall n_i \in \mathcal{S}$$

4.2.2 Définition

Le **réseau Bayésien** (RB), aussi appelé *réseau de croyance, réseau graphique* ou *réseau causal*, est un *graphe acyclique orienté* GAO, défini par un triplet $(\mathcal{N}, \mathcal{A}, \mathcal{P})$ [38], [39] où :

- $\mathcal{N}$ est l'ensemble n_i des variables, représentant des événements associés aux nœuds dans le graphe. Chaque variable a un ensemble fini d'états mutuellement exclusifs.
- A est l'ensemble des arcs orientés représentant les *relations* de cause à effet ou de dépendance entre les nœuds qu'ils relient. Ces arcs ne doivent former aucun cycle. Par exemple :
  Avec $n_i, n_j \in \mathcal{N}$, $(n_i, n_j) \in \mathcal{A}$ est un arc initié de n_i et terminé à n_j. Il représente la relation cause-effet de n_i à n_j.
- $\mathcal{P}$ est la distribution des probabilités dans le RB. C'est l'ensemble des probabilités conditionnelles des nœuds sachant leurs parents dans le RB. Une probabilité conditionnelle P(B|A) exprime la force (strength) de l'influence de la relation de A à B. C'est la raison pour laquelle le RB peut décrire à la fois qualitativement des dépendances entre variables (via la forme du graphe) et quantitativement ces dépendances (via les probabilités conditionnelles).

La figure 4.1 est un exemple de RB.

Dans le RB, on définit la fonction *parent* d'un nœud comme suit :

$$\begin{aligned} pa: \quad \mathcal{N} &\rightarrow \mathcal{N}^* \\ x &\rightarrow \{x'_1, x'_2, ...\} \end{aligned} \tag{4.1}$$

où N^* représente l'ensemble de tous les ensembles dans N, et :

$$pa(x) = \{x' \in \mathcal{N} | (x', x) \in \mathcal{A}\}$$

$pa(x)$ est donc l'ensemble des nœuds liés avec x par des arcs orientés vers x.

Par exemple dans la figure 4.1 :

$$pa(b) = \{a\}$$

$$pa(c) = \{a\}$$

$$pa(d) = \{b, c\}$$

On définit aussi la fonction *enf* (enfant) comme suit :

$$\begin{aligned} enf: \quad \mathcal{N} &\rightarrow \mathcal{N}^* \\ y &\rightarrow \{y'_1, y'_2, ...\} \end{aligned} \tag{4.2}$$

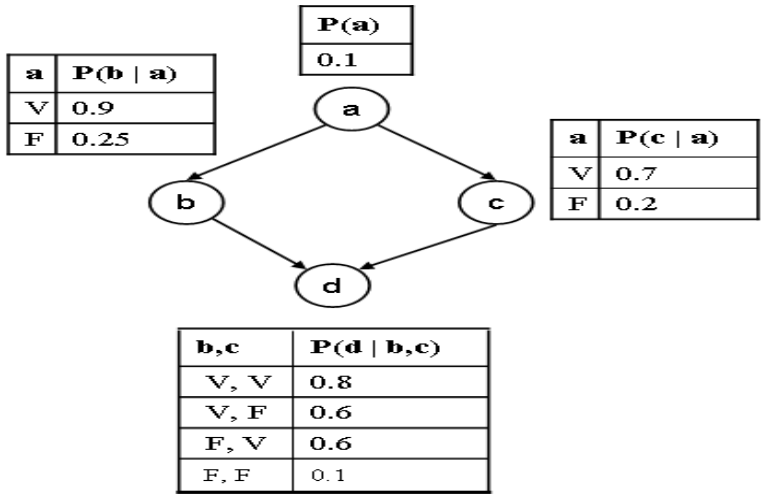


FIG. 4.1 – Exemple d'un réseau Bayésien

$$enf(x) = \{y' \in \mathcal{N} | (y, y') \in \mathcal{A}\}$$

$enf(y)$ est donc l'ensemble des nœuds liés avec y par des arcs d'origine de y.

Par exemple aussi dans la figure 4.1 :

$$enf(a) = \{b, c\}$$

$$enf(b) = enf(c) = \{d\}$$

Si $pa(n)$ est différent du vide, on peut évidement assurer que :

$$\forall n \in N : n \in enf(pa(n))$$

4.2.3 Règle de chaîne (chain rule) ou théorème de Bayes généralisé

Avec $U = (A_1, ..., A_n)$ un ensemble de variables d'un univers, la probabilité jointe est :

$$P(U) = P(A_1, ..., A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)...P(A_n|A_1...A_{n-1}) \quad (4.3)$$

S'il faut accéder aux tables de probabilité jointe sur toutes les variables dans l'univers pour calculer $P(U)$, $P(U)$ grandit exponentiellement avec le nombre des variables. Un réseau Bayésien sur U qui est une manière de stocker des informations dont $P(U)$ peut être calculé quand nécessaire, est une représentation plus compacte de $P(U)$ parce que chaque variable dans le réseau ne dépend que de ses parents. Avec sa représentation compacte de $P(U)$ et l'indépendance conditionnelle assurée, on a :

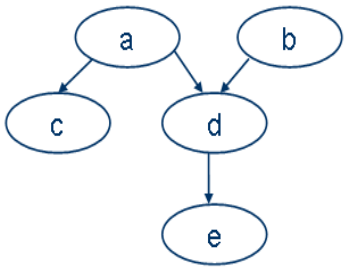


FIG. 4.2 – Exemple d'un RB pour le calcul de la probabilité jointe

$$P(A_i|A_1...A_{i-1}) = P(A_i|pa(A_i)) \tag{4.4}$$

où $pa(A_i)$ l'ensemble des parents de A_i. La distribution de probabilité jointe $P(U)$ dans 4.3 devient :

$$P(U) = \prod_i P(A_i|pa(A_i)) \tag{4.5}$$

Par exemple, dans la figure 4.2, la probabilité jointe est :

$$P(a,b,c,d,e) = P(a)P(b|a)P(c|a,b)P(d|a,b,c)P(e|a,b,c,d) \tag{4.6}$$

En assumant que l'indépendance conditionnelle s'applique sur le RB, cette probabilité jointe devient :

$$\begin{aligned} P(a,b,c,d,e) &= P(a)P(b|pa(b))P(c|pa(c))P(d|pa(d))P(e|pa(e)) \\ &= P(a)P(b|a)P(c|a)P(d|a,b)P(e|d) \end{aligned} \tag{4.7}$$

4.2.4 L'inférence des probabilités dans le réseau Bayésien

Le processus de mise à jour de la probabilité des variables lorsqu'une nouvelle évidence e arrive est vu comme un processus d'inférence probabiliste ou processus de révision des croyances. Ce processus est NP-difficile en général [38], [39], [56], à cause de la complexité du calcul sur le réseau quand le réseau a un grand nombre de variables, des relations. Dans les réseaux simples, une méthode directe pour la mise à jour des probabilités est l'utilisation des matrices d'association (Link matrix).

Pour chaque nœud, sa matrice d'association liste toutes les combinaisons possibles des états de ses parents. Dans le cas de valeurs binaires, cette matrice est de taille 2×2^n pour n parents d'un nœud A et spécifie la probabilité que A prenne la valeur $Vrai$ ou $Faux$ pour toutes les combinaisons de ses parents. Le processus d'inférence du réseau

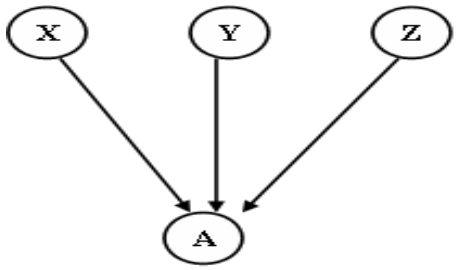


FIG. 4.3 – Exemple d'un nœud avec parents

Bayésien peut utiliser les possibilités fournies par des parents pour mettre des conditions sur la matrice d'association afin d'estimer la confiance sur A ou $P(A = vrai)$, noté par bel(A).

Nous prenons l'exemple d'un nœud A qui a trois parents X, Y, Z, et :

$$bel(X) = P(X = vrai) = x, \tag{4.8}$$

$$bel(Y) = P(Y = vrai) = y, \tag{4.9}$$

$$bel(Z) = P(Z = vrai) = z, \tag{4.10}$$

La matrice d'association dans ce cas spécifie $2^3 = 8$ combinaisons des parents :

$$\begin{bmatrix} \overline{XYZ} & \overline{XY}Z & \overline{X}Y\overline{Z} & X\overline{YZ} & XY\overline{Z} & X\overline{Y}Z & \overline{X}YZ & XYZ \\ \overline{XYZ} & \overline{XY}Z & \overline{X}Y\overline{Z} & X\overline{YZ} & XY\overline{Z} & X\overline{Y}Z & \overline{X}YZ & XYZ \end{bmatrix}$$

Dans cette matrice, la première ligne correspond au cas $A = faux$ et la deuxième correspond au cas $A = vrai$

$$\begin{aligned} bel(A) \quad = \quad & P(A|\overline{XYZ})(1-x)(1-y)(1-z) + P(A|\overline{XY}Z)(1-x)(1-y)z + \\ & P(A|\overline{X}Y\overline{Z})(1-x)y(1-z) + P(A|X\overline{YZ})(1-x)yz + \\ & P(A|XY\overline{Z})xy(1-z) + P(A|X\overline{Y}Z)x(1-y)z + \\ & P(A|\overline{X}YZ)(1-x)yz + P(A|XYZ)xyz \end{aligned} \tag{4.11}$$

Pour une variable A qui a n parents $\{t_1, ..., t_n\}$, sa probabilité est :

$$bel(A) \quad = \quad \sum_{i=1}^{2^n} P(A|\pi_i) \times \prod_{t_j \in \pi_i} P(t_j) \times \prod_{\neg t_k \in \pi_i} (1 - P(t_k)) \tag{4.12}$$

avec π_i est une configuration des parents de q parmi 2^n. Le calcul de la probabilité conditionnelle sur un grand nombre des parents avec la matrice d'association n'est pas pratique.

La raison est que le nombre des combinaisons des parents augmente exponentiellement avec le nombre de parents. Nous présentons par la suite des expressions canoniques déduites pour certaines combinaisons spéciales des parents. Ces expressions sont applicables dans la RI.

Les expressions pour des combinaisons spéciales

Dans la RI, Turtle et Croft [79], [80] ont proposé des expressions pour calculer la probabilité de manière efficace. Ces expressions sont déduites à partir de la marginalisation par-dessus les formes canoniques des matrices d'associations. Ces matrices sont proposées pour les opérateurs booléens ou la recherche probabiliste.

- *La OU-combinaison* : Dans le cas de la OU-combinaison, sachant que A est vrai si X ou Y ou Z est vrai et A est faux seulement si X et Y et Z sont faux, la matrice d'association devient :

$$L_{OU} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

En utilisant la formule de probabilité totale on a :

$$\begin{aligned} bel(A) = P(A = vrai) &= (1-x)(1-y)z + (1-x)y(1-z) + x(1-y)(1-z) + \\ &\quad xz(1-y) + xy(1-z) + yz(1-x) + xyz \end{aligned} \tag{4.13}$$

mais :

$$\begin{aligned} &(1-x)(1-y)(1-z) + (1-x)(1-y)z + (1-x)y(1-z) + x(1-y)(1-z) + \\ &\quad xz(1-y) + xy(1-z) + yz(1-x) + xyz = 1 \end{aligned} \tag{4.14}$$

donc :

$$bel(A) = P(A = vrai) = 1 - (1-x)(1-y)(1-z) \tag{4.15}$$

$$P(A = faux) = (1-x)(1-y)(1-z) \tag{4.16}$$

- *La ET-combinaison* : Dans le cas de la ET-combinaison, A est vrai si X et Y et Z sont vrai et faux dans les autres cas. La matrice d'association devient :

$$L_{ET} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

En utilisant la formule de probabilité totale on a aussi :

$$\begin{aligned} P(A = faux) &= (1-x)(1-y)(1-z) + (1-x)(1-y)z + (1-x)y(1-z) + \\ &\quad x(1-y)(1-z) + xz(1-y) + xy(1-z) + yz(1-x) \end{aligned} \tag{4.17}$$

On peut simplifier la formule :

$$bel(A) = P(A = vrai) = xyz \tag{4.18}$$

$$P(A = faux) = 1 - xyz \tag{4.19}$$

– *La NON-combinaison* :
De manière similaire avec la NON-combinaison, on a :

$$bel(A) = P(A = vrai) = 1 - x \tag{4.20}$$

– *La Max-combinaison*
Dans cette combinaison, on a simplement :

$$bel(A) = P(A = vrai) = max\{x, y, z\} \tag{4.21}$$

– *Matrice de la somme des poids (Weighted-Sum matrix)*
Dans ce cas, chaque nœud parent a un poids associé, de même pour le nœud enfant. La confiance sur A dépend des poids des parents qui sont vrais : plus son poids est grand, plus il a une grande influence sur cette confiance. Considérant que $w_x, w_y, w_z \geq 0$ sont les poids des parents X, Y, Z, que $0 \leq w_a \leq 1$ est le poids de l'enfant A et que $w_x + w_y + w_z = t$, la matrice de somme des poids a la forme suivante :

$L_W =$

$$\begin{bmatrix} 1 & 1-\frac{w_z w_a}{t} & 1-\frac{w_y w_a}{t} & 1-\frac{w_x w_a}{t} & 1-\frac{(w_y+w_z)w_a}{t} & 1-\frac{(w_x+w_z)w_a}{t} & 1-\frac{(w_x+w_y)w_a}{t} & 1-w_a \\ 0 & \frac{w_z w_a}{t} & \frac{w_y w_a}{t} & \frac{w_x w_a}{t} & \frac{(w_y+w_z)w_a}{t} & \frac{(w_x+w_z)w_a}{t} & \frac{(w_x+w_y)w_a}{t} & w_a \end{bmatrix}$$

qui produit :

$$\begin{aligned} bel(A) = P(A = vrai) &= \frac{w_z w_a}{t}(1-x)(1-y)z + \frac{w_y w_a}{t}(1-x)y(1-z) + \\ &\quad \frac{w_x w_a}{t}x(1-y)(1-z) + \frac{(w_y + w_z)w_a}{t}(1-x)yz + \\ &\quad \frac{(w_x + w_z)w_a}{t}x(1-y)z + frac(w_x + w_y)w_a txy(1-z) + w_a xyz \\ &= \frac{(w_x x + w_y y + w_z z)w_a}{t} \\ P(A = faux) &= 1 - \frac{(w_x x + w_y y + w_z z)w_a}{t} \end{aligned}$$

4.2.5 Bilan

Dans cette section, nous avons présenté la procédure d'inférence des probabilités sur le RB. Nous abordons l'algorithme de l'inférence sur le RB. Il s'agit de l'utilisation des matrices d'association pour calculer la probabilité postérieure, ou la croyance d'un nœud A, noté $bel(A)$, connaissant les probabilités de ses parents. Cependant, la complexité de ce calcul augmente exponentiellement avec le nombre de parents de A. En l'appliquant à la RI, Turtle et Croft ont proposé certaines expressions canoniques pour simplifier ce calcul pour quelques types de combinaisons des parents.

Le récapitulatif des expressions est le suivant :

Pour un nœud A qui a $pa(A) = \{t_1, t_2, ..., t_n\}$ et $bel(t_i) = p_i$:

$$bel_{not}(A) = 1 - p_1 \tag{4.22}$$

$$bel_{or}(A) = 1 - \prod_i (1 - p_i) \tag{4.23}$$

$$bel_{and}(A) = \prod_i p_i \tag{4.24}$$

$$bel_{max}(A) = max\{p_1, ...p_n\} \tag{4.25}$$

$$bel_{sum}(A) = \frac{\sum_i p_i}{n} \tag{4.26}$$

$$bel_{wsum}(A) = \frac{\sum_i w_i p_i}{\sum_i w_i} \text{ avec } w_i \text{ est le poids de } t_i \tag{4.27}$$

Ces expressions sont efficaces et éventuellement applicables dans notre modèle proposé dans le chapitre suivant. Nous allons dresser par la suite l'état de l'art des travaux précédents sur ce type de modèle.

4.3 Les modèles Bayésiens dans la littérature des systèmes de recherche d'information

Avec sa capacité à résoudre efficacement des problèmes concernant l'incertitude, le modèle Bayésien a été proposé comme une nouvelle fonction de correspondance dans la RI. Différentes formes de modèles Bayésiens pour la RI ont été proposées selon différents points de vue. Cela entraîne une grande variété de méthodes pour calculer la probabilité a posteriori. Ainsi, il existe donc différentes manières de combiner de multiples sources d'évidences ou de connaissances.

4.3.1 Description générale du modèle de RI basé sur le réseau Bayésien

Etant donné :

- la collection des documents $D = \{d_1, d_2, ...d_N\}$
- la requête q
- l'ensemble des vocabulaires d'indexation $\Gamma = \{t_1, ..., t_{N_\Gamma}\}$. L'indexation est définie par :

$$\begin{aligned} Indx : \quad & D \rightarrow \Gamma^* \\ & Indx(d) = \{t_{d_1}, ...t_{d_i}\} \end{aligned}$$

avec Γ^* l'ensemble des sous ensembles de Γ. Dans les SRI, la requête est représentée aussi par les termes d'indexation, la fonction d'indexation peut être étendue pour la requête :

$$Indx(q) = \{t_{q_1}, ...t_{q_j}\}$$

Le modèle de RI basé sur un RB classique peut être défini par (Ψ, Δ) avec :

- $\Psi(\mathcal{N}, \mathcal{A}, \mathcal{P})$ est le RB qui comprend :
  - $\mathcal{N}$: ensemble des nœuds qui représentent $D \cup \{q\} \cup \Gamma$. En fait, il s'agit d'une bijection à partir de l'ensemble $D \cup \{q\} \cup \Gamma$ vers les nœuds. Pour simplifier la notation, nous définissons $\mathcal{N} = D \cup \{q\} \cup \Gamma$
  - $\mathcal{A}$: ensemble des arcs :

$$\mathcal{A} = \{(d, t_i) | t_i \in Indx(d)\} \cup \{(q, t_j) | t_j \in Indx(q)\}$$

  - $\mathcal{P} = \{P(n_i | pa(n_i))\}, \forall n_i \in \mathcal{N}$: la distribution des probabilités des variables dans Ψ
- Δ est la fonction de correspondance entre les documents et la requête.

$$\begin{aligned} \Delta : \quad & D \times \{q\} \rightarrow [0, 1] \\ & \Delta(d, q) = P(q|d) \end{aligned}$$

La procédure de RI dans ce modèle est l'inférence de probabilité sur le RB : un document d observé provoque une propagation des probabilités sur le réseau, des parents à ses enfants et termine à q. Le but de l'inférence est de calculer $\Delta(q, d) = P(q|d)$ qui représente la fonction de correspondance.

Par exemple dans le modèle de Turtle et Croft [80], sachant qu'un document d dans la collection est observé, on a :

$$P(d) = 1 \tag{4.28}$$

et :

$$P(q) = \sum_{d=Vrai,Faux} P(q,d) \text{ (la marginalisation)} \quad (4.29)$$

$$= \sum_{d=Vrai,Faux} P(q|d) \times P(d) \text{ (règle fondamentale)} \quad (4.30)$$

$$= P(q|d) \times P(d) + P(q|\neg d) \times P(\neg d) \quad (4.31)$$

$$= P(q|d) \times P(d) + P(q|\neg d) \times (1 - P(d)) \quad (4.32)$$

$$= P(q|d) \text{ (parce que } P(d) = 1) \quad (4.33)$$

donc :

$$\Delta(d,q) = P(q|d) = P(q) = bel(q) \quad (4.34)$$

Avec ce schéma général du modèle de RI basé sur le RB, les travaux sur ce type de modèle dans la littérature se différencient dans :

- Le vocabulaire d'indexation : termes, syntagmes.
- La structure du RB (par exemple la direction des arcs, intégration des nouveaux arcs, ...)
- Le calcul des probabilités

Dans la section suivante, nous aborderons des travaux dans l'état de l'art de ce modèle.

4.3.2 Modèle de réseau d'inférence

Selon le point de vue de Turtle et Croft [79], [80], la RI est une inférence ou un processus de raisonnement dans lequel nous estimons la probabilité qu'un document, qui est vu comme une évidence, satisfasse le besoin d'information de l'utilisateur. Leur modèle proposé, appelé *réseau d'inférence*, est l'un des tout premiers modèles et a beaucoup attiré l'attention sur le modèle Bayésien pour la RI. Ils ont montré que ce modèle est une extension du modèle probabiliste et permet l'intégration des connaissances (requêtes anciennes, la rétroaction de pertinence, etc.) dans un cadre unique.

La figure 4.4 décrit la forme graphique du réseau d'inférence. Dans ce réseau, les nœuds comprennent des nœuds documents (d_i), des nœuds termes (t_i), des nœuds requêtes (q_i) et un nœud représentant le besoin d'information (I). Les requêtes sont combinées par des opérations logiques pour formuler un besoin d'information. Chaque variable du nœud terme du réseau correspond à l'événement où un terme d'indexation est associé à un document ou une requête. La variable du nœud document correspond à l'événement où un document est observé, de même pour la requête. La probabilité antérieure d'un document

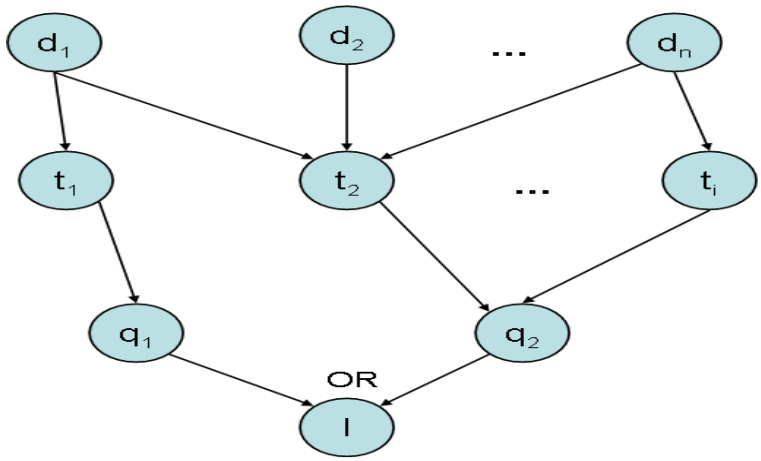


FIG. 4.4 – Réseau d'inférence

est la probabilité d'observer ce document dans la collection. Cette observation de document est la cause d'observation de tous ses termes associés. La valeur de pertinence entre document et requête est interprétée comme la probabilité postérieure de la requête (ou la croyance que le besoin d'information soit satisfait), sachant qu'un document est observé dans la collection. En pratique, étant donné les probabilités antérieures associées avec des documents et des probabilités conditionnelles de leurs sous nœuds, les probabilités postérieures des autres nœuds dans le réseau peuvent être alors calculées dans le processus d'inférence. Les matrices d'association sont utilisées pour ce calcul. Le processus d'inférence se termine au nœud requête lorsque sa valeur de pertinence est calculée.

Par ailleurs, ce modèle montre théoriquement qu'il a la capacité de simuler les schémas de pondération des autres modèles ($tf.idf$, modèle probabiliste, Booléen). Par exemple dans la figure 4.5, pour illustrer la pondération $tf.idf$, q étant une représentation de la requête, A, B, C étant les nœuds documents ; w_a, w_b, w_c étant les poids tf normalisés de A, B, C ; idf_q étant le poids idf normalisé de q ; et :

$$w_q = idf_q.(w_a + w_b + w_c) \tag{4.35}$$

Quand A est initialisé, en utilisant la somme matricielle des poids, la croyance sur q est :

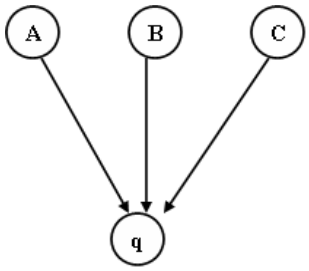


FIG. 4.5 – Exemple d'un réseau Bayésien dans la simulation du schéma de pondération tf.idf

$$\begin{aligned} bel(q) &= \frac{w_a w_q}{w_a + w_b + w_c} \\ &= \frac{tf_a.idf_q(w_a + w_b + w_c)}{w_a + w_b + w_c} \\ &= tf_a.idf_q \end{aligned}$$

qui est donc une forme de la pondération $tf.idf$.

D'ailleurs, ce modèle de Turtle et Croft a également la capacité d'intégrer différentes sources d'évidences. Ces sources d'évidences peuvent venir des retours de pertinence d'utilisateurs. Des expérimentations de ce modèle ont été menées dans le système InQuery [78], [19]. Des résultats comparables avec un modèle probabiliste ont été rapportés, ainsi qu'une augmentation des performances en combinant différentes formulations logiques de requêtes.

4.3.3 Modèle du réseau de croyance

D'après le point de vue de Ribeiro [65], un SRI est vu comme un système de correspondance de concepts, la RSV est interprétée comme la probabilité de correspondance entre une requête et un document dans un espace de concepts. Le modèle proposé, le réseau de croyance (Figure 4.6), est supposé plus général par sa capacité à subsumer le réseau d'inférence. Cette hypothèse est théoriquement prouvée by Baeza-Yates [6].

Dans le modèle de croyance de Bruza [16],[17], documents et requêtes sont pourtant vus comme des *objets d'information*. Ces objets sont caractérisés par un ensemble de descripteurs, appelés *expressions d'indexation*. Une expression d'indexation est formée de termes et de connecteurs, qui modélisent les relations entre ces termes. Elle peut être une phrase, un syntagme ou juste une combinaison des termes.

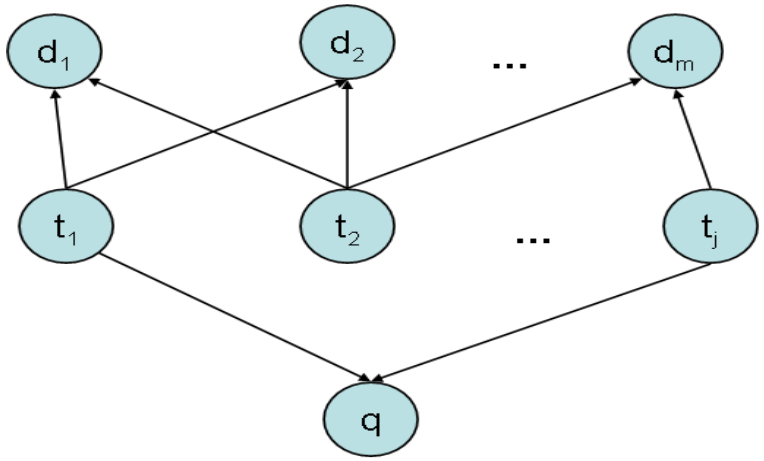


FIG. 4.6 – Réseau de croyance de Baeza

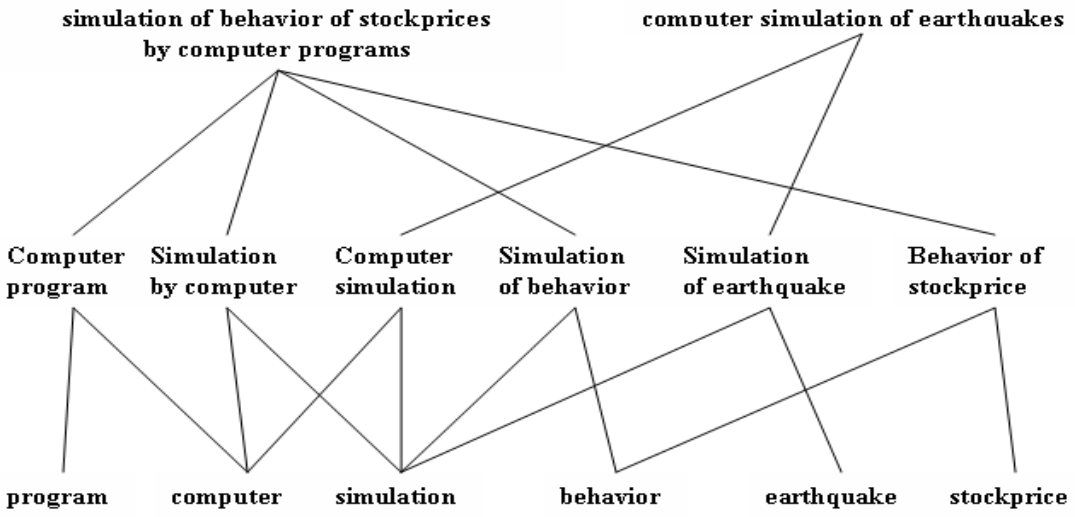


FIG. 4.7 – Exemple d'un réseau de Bruza

Dans le cadre de ce modèle, les connecteurs entre termes sont soit des prépositions prédéfinies, soit nuls (espace blanc, c'est-à-dire qu'il n'y a pas de mots pour les connecter). Par exemple : *people IN (need OF (information (retrieval)))* ; dans cet exemple, il y a ces connecteurs : «IN», «OF» et «nul» (entre «information» et «retrieval»). Le tableau 4.1 liste les connecteurs utilisés dans ce modèle et leurs probabilités prédéfinies. Une expression d'indexation est liée avec ses composants par la relation *sous-expression-de*. La figure 4.8 montre un exemple de la décomposition des expressions d'indexation en composants par des connecteurs.

Le principe pour juger que le document d est approprié à la demande d'information q est le suivant : d est approprié à q si q peut être prouvé à partir des expressions d'indexation de l'objet d'information d via des règles d'inférence.

Les règles d'inférence règlementent l'inférence de la probabilité vers les nœuds enfants en considérant le type de connecteurs et son contexte. Cette probabilité représente la probabilité que le nœud enfant soit impliqué par ses parents via différents types de connecteurs.

Le calcul de la probabilité postérieure des nœuds est simplifié par l'application du théorème de réduction de réseau. Ce théorème permet une estimation directe et simple des probabilités en réduisant le calcul de probabilité à certains niveaux des nœuds à partir des feuilles seulement. Il propose de calculer la valeur de pertinence entre document d et requête q comme suit :

$$P(q|d) = \prod_{y \in Bot(q) - Bot(d)} P(y| \wedge \pi(y))$$

avec $Bot(q) - Bot(d)$ qui dénote les ensembles des nœuds dans les deux plus bas niveaux de l'index expression, c'est-à-dire les feuilles et leurs parents, $\pi(y)$ dénote l'ensemble des nœuds parents du nœud y.

Voici un exemple repris dans le même travail [17] (cf. figure 4.7) :

Supposons que le document d contienne *«simulation of behavior of stockprices by computer programs»* et que la requête q soit *«computer simulation of earthquakes»*, nous avons les deux plus bas niveaux des index d'expression comme suit :

Bot(q)=comput ∘ sim, sim of *earthq, comput, sim, earthq*

Bot(d)=behav of *stpr, comput ∘ prog, sim* of *behav, sim by comput, sim, behav, stpr, comput, prog*

Supposons que les probabilités des termes soient : *comput* (0.25), *sim* (0.25), *behav* (0.125), *stpr* (0.125), *prog* (0.125), *earthq* (0.125) et celles des connecteurs : ∘(0.53), *of* (0.15). En appliquant le théorème de réduction de réseau :

Bot(q)-Bot (d) = comput ∘ sim, sim of *earthq, earthq*

Connecteur	Probabilité
$\circ$	0.5366
and	0.0492
as	0.0004
at	0.0348
between	0.0052
by	0.0061
for	0.0327
from	0.0039
in	0.0632
of	0.1529
on	0.0370
or	0.0026
over	0.0066
through	0.0035
to	0.0170
with	0.0248

TAB. 4.1 – Les connecteurs et leurs probabilités

La RSV est donc :

$$\begin{aligned} Pr(q|i) &= Pr(comput \circ sim | comput \wedge sim) Pr(\text{sim } of \text{ earthq} | sim \wedge earthq) Pr(earthq) \\ &= 0.53 \times 0.15 \times 0.125 \\ &= 0.0103 \end{aligned}$$

La méthode de décomposition de syntagme en termes d'indexation par des mots connecteurs n'est pas flexible et n'est pas indépendante des langues. Les termes décomposés ne sont pas nécessairement utiles à l'indexation. De plus, les probabilités des connecteurs sont prédéfinies heuristiquement. Ces points faibles affectent de façon importante les performances de la RI.

Pour mieux représenter des syntagmes dans un RB, Ho [60], [37] propose d'utiliser la structure *tête-modificateurs* des syntagmes. Basés sur cette structure, des syntagmes sont décomposés en plusieurs sous-syntagmes ou termes simples par des règles de décomposition. Les liens d'un syntagme vers ses composants sont les liens syntaxiques, ils portent donc plus de sens que les liens simples entre termes connectés par certaines prépositions. La figure 4.9 illustre un exemple de cette méthode de décomposition d'un syntagme.

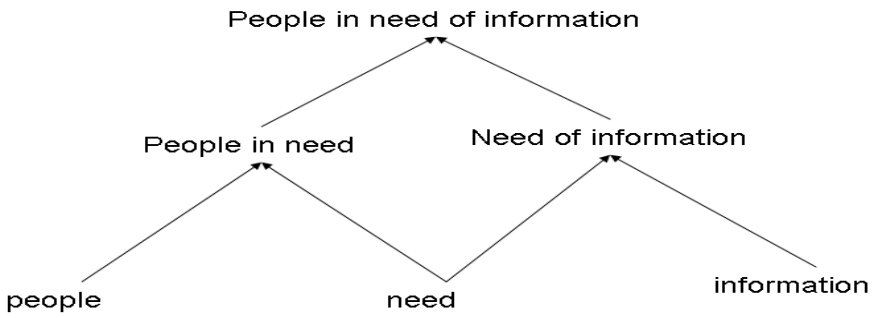


FIG. 4.8 – Exemple de la décomposition de syntagme de Bruza

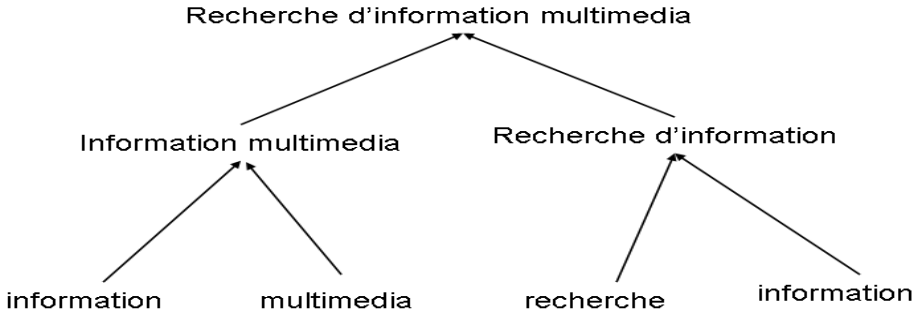


FIG. 4.9 – Exemple la décomposition de syntagme de Ho

Ce modèle apporte des augmentations de performances par rapport aux modèles VSM. Cependant, dans ce modèle il reste encore certains éléments assez heuristiques dans les formules d'estimation des probabilités. De plus, il ne prend pas en compte des liens sémantiques entre des termes ou syntagmes.

4.4 Conclusion

Dans ce chapitre, nous avons présenté l'approche Bayésienne et l'état de l'art des modèles Bayésien dans la RI.

Les modèles basés sur un RB combinent la théorie des probabilités avec la théorie des graphes. Ce modèle a non seulement la capacité de modéliser les variables et leurs dépendances, mais aussi la capacité de modéliser les évidences et leurs influences sur les variables via un processus d'inférence. Ce dernier permet de mettre à jour les probabilités des variables dans tout le réseau. Pour cette tâche, l'utilisation de la matrice d'association est une méthode simple appliquée pour des réseaux simples. Ce calcul peut être simplifié pour certains types de combinaison spéciale des parents, proposés par Turtle et Croft [79], [80].

Nous avons étudié aussi les modèles de RI basés sur un RB. Les modèles de Turtle et Croft, Berthier [65] et Baeza-Yates [6] permettent d'englober les modèles VSM, Booléen et probabiliste. Des améliorations de performance de recherche ont été apportées par combinaison logique de la requête. La prise en compte des liens entre terme d'indexation dans ce genre de modèle a été ensuite proposée par Bruza [16, 17] et Ho [60, 37], mais ces travaux utilisent seulement des liens syntaxiques.

Pour mieux capturer le contenu et la sémantique des textes dans le contexte de RI, nous proposons dans le chapitre suivant le modèle Bayésien pour la RI dans lequel documents et requêtes sont représentés par des concepts et leurs liens sémantiques.

Deuxième partie

CONTRIBUTION

La RI à base d'intersection des termes est limitée au niveau des index et de la correspondance. La raison est que les documents pertinents ne partagent pas toujours les mêmes termes que la requête. Notre contribution principale a pour objectif de surmonter ces limites.

En utilisant une ressource externe, l'indexation conceptuelle est supposée atteindre un niveau plus précis. Pour une meilleure correspondance, nous suggérons de prendre en compte des relations hiérarchiques entre concepts dans le modèle de RI à base de réseau Bayésien. Ces relations sont extraites à partir d'une ressource externe et jouent le rôle de pont entre les concepts du document et de la requête. Ces ponts sont donc supposés résoudre le problème de la disparité. Cette proposition est présentée dans le chapitre 5.

Dans le chapitre 6, les expérimentations pour la RI dans le domaine médical sont menées pour valider le modèle proposé. Pour cela, nous utilisons les collections ImageCLEFMed 2006, 2007 ainsi que la ressource externe UMLS.

A partir d'une indexation à base de concept, nous étudions dans le chapitre 7 l'extension à des documents et requête structurés et multi-médias. Pour des documents et requête structurés par les dimensions du domaine, nous proposons une fonction de reclassement des documents en se basant sur l'intersection des dimensions entre document-requête. Pour la recherche sur les documents multi-médias, une fusion de la recherche à base d'image et de texte est proposée. Ces propositions sont aussi validées sur les collections ImageCLEFMed 2006, 2007.

Chapitre 5

Proposition d'un modèle Bayésien à base de concepts et de relations sémantiques

Sommaire

5.1 Introduction

Dans ce chapitre, nous proposons un modèle Bayésien à base de concepts et de relations sémantiques. Ce modèle utilise une ressource externe.

Pour l'indexation, les index sont les concepts extraits à partir de la ressource externe. Dans le contexte d'indexation par concepts et où les documents contiennent des concepts qui sont plus spécifiques que ceux de la requête, nous supposons que la prise en compte des relations sémantiques entre les concepts des documents et de la requête dans la fonction de correspondance est nécessaire.

Par exemple, le document contient « dermatitis » alors que la requête contient « lésion de la peau ».

Des méthodes comme l'expansion de requêtes ont été proposées en vue de prendre en compte implicitement ces liens par l'ajout des concepts sémantiquement liés. Elles ont montré certaines améliorations, mais aussi des limites :

- En ajoutant de nouveaux concepts, la méthode d'expansion change la nature de la collection et de la requête, ainsi que la distribution d'origine des concepts.
- Elle provoque le risque d'ajouter une masse de concepts inutiles. C'est le cas où le choix de concepts ajoutés n'est pas le bon ou les concepts ajoutés sont trop éloignés de la demande d'information. Dans ce cas les concepts ajoutés diminuent la performance de la recherche.
- Le choix de la pondération des concepts ajoutés est un problème qui n'est pas encore résolu de manière claire dans l'état de l'art.

De notre point de vue, un modèle à base de réseau Bayésien convient très bien pour prendre en compte des liens sémantiques. Dans ce genre de modèle, les documents et la requête restent tels quels et les liens sémantiques sont modélisés explicitement dans un cadre unique. En effet, la forme graphique du réseau permet de modéliser facilement les relations entre les nœuds. En plus, le mécanisme d'inférence probabiliste permet de modéliser la nature incertaine de la RI, ainsi que les poids des liens sémantiques entre les index.

Notre modèle s'inspire du réseau d'inférence de Turtle et Croft. Les différences entre notre modèle et celui de Turtle et Croft ainsi que des autres modèles Bayésiens dans l'état de l'art résident dans :

- L'intégration de la ressource externe dans le modèle :
  - Les termes d'indexation sont des concepts au lieu de termes
  - L'intégration des relations hiérarchiques dans le réseau Bayésien. Ces relations sont extraites et déduites à partir d'une ressource externe.
- Les pondérations des liens et le calcul des probabilités.

Le but de ce modèle est d'une part de tendre vers une indexation plus précise et d'autre part de prendre en compte les relations hiérarchiques entre concepts de documents et ceux de la requête afin de résoudre le problème de la disparité.

Nous introduisons par la suite les définitions et notations nécessaires pour la description du modèle proposé dans la section 5.3. Nous montrons aussi dans la section 5.4 la réduction au modèle à base d'intersection (par exemple le modèle VSM) de notre modèle dans le cas où les relations sémantiques ne sont pas intégrées. Nous concluons ce chapitre dans la dernière section.

5.2 Les définitions et notations

Nous définissons d'abord les éléments dans notre modèle :

5.2.1 Ressource externe

Une ressource externe peut être définie simplement par $\Upsilon(\mathcal{T}, \mathcal{C}, \mathcal{R})$ avec :

- $\mathcal{T} = \{t_1, t_2, ...\}$ est l'ensemble des termes. Un terme est un mot ou un syntagme nominal qui a une signification particulière dans un domaine.
  Par exemple : le terme « radio » a les significations suivantes dans différents domaines :
  - *la photographie de l'intérieur du corps : radio des poumons* (dans le domaine médical)
  - *l'appareil avec lequel on écoute des émissions : écouter la radio* (dans le domaine technique)
- $\mathcal{C} = \{c_1, c_2, ...\}$ est l'ensemble des concepts.
  Par exemple :

  $\mathcal{C} = \{C0018681\}$(«headache», «head pain»,...), C0043309(«Roentgen Rays», «X-ray»),...

  Chaque concept est le représentant d'un ensemble de termes. Les termes associés à un concept doivent avoir le même sens.
  Nous définissons la fonction ζ (*nom de concept*), qui représente l'ensemble des termes associés à un concept :

$$\begin{aligned} \zeta : \ \mathcal{C} &\rightarrow \mathcal{T}^* \\ c &\rightarrow \{t_{c1}, t_{c2}, ...\} \end{aligned}$$

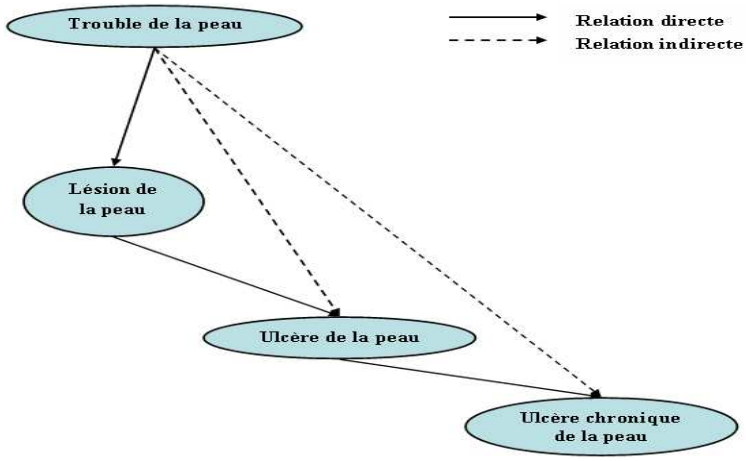


FIG. 5.1 – Exemple des relations directes ou indirectes entre les concepts

avec $\{t_{ci}\}$ l'ensemble des termes associés au concept c.

Par exemple :

ζ(C0043309)={«Roentgen Rays», «X-ray», «X Radiation», «Röntgenstraal»,...}

- Dans le cadre de notre modèle, nous définissons $\mathcal{R}$ comme une relation entre concepts. Cette relation est de type hiérarchique (hyponymie-hyperonymie et holonymie-méronymie). Le graphe de $\mathcal{R}$ est un sous ensemble de $\mathcal{C} \times \mathcal{C}$ qui satisfait :
  - La transitivité : $\forall c_i, c_j \in \mathcal{C}$, si $\exists c_k$ tel que $(c_i, c_k) \in \mathcal{R}, (c_k, c_j) \in \mathcal{R}$ alors $(c_i, c_j) \in \mathcal{R}$.

    Une relation (c_i, c_j) est dite *indirecte* s'il existe un chemin $che(c_i, c_j)$ entre c_i et c_j d'une longueur supérieure ou égale à deux. Une relation qui n'est pas directe est dite *directe*. La figure 5.1 illustre une exemple de relation directe et indirecte entre concepts.
  - La non réflexivité : $\forall c_i, c_j \in \mathcal{C}$, si $(c_j, c_i) \in \mathcal{R}$ alors $(c_i, c_j) \notin \mathcal{R}$. Le réseau des concepts et de leurs relations est donc acyclique.

5.2.2 La présentation conceptuelle des documents et de la requête

On reprend les notations dans la section 5.3.2 : $D = \{d_1, d_2, ...d_N\}$ est la collection de documents, N est le nombre total de documents dans la collection, et q est la requête.

La présentation conceptuelle d'un document ou d'une requête est constituée par les concepts d'une ressource externe les mieux appropriés aux syntagmes présents dans le document ou la requête.

Etant donné Θ l'ensemble de tous les termes de la collection et de la requête, le

vocabulaire des termes dans un document ou dans la requête est défini par :

$$\begin{aligned} V: \quad D \cup \{q\} &\rightarrow \Theta^* \\ d &\rightarrow \{t_{1d}, t_{2d}, ...\} \\ q &\rightarrow \{t_{1q}, t_{2q}, ...\} \end{aligned} \tag{5.1}$$

Avec la ressource externe donnée $\Upsilon(\mathcal{T}, \mathcal{C}, \mathcal{R})$, on définit la fonction d'indexation $Indx$ (ou la conceptualisation) par :

$$Indx: \quad \Theta^* \rightarrow \mathcal{C}^*$$

avec :

$$Indx(V(d)) = \{c_i \in \mathcal{C} | \exists t_j \in V(d) \text{ tel que } t_j \in \zeta(c_i)\} \tag{5.2}$$

La fonction d'indexation peut être étendue à la requête :

$$Indx(V(q)) = \{c_i \in \mathcal{C} | \exists t_j \in V(q) \text{ tel que } t_j \in \zeta(c_i)\} \tag{5.3}$$

On définit $\forall c_i \in Indx(V(d))$:

$$f(c_i, d) = \sum_{t_j} f(t_j, d), \forall t_j \in V(d) \text{ tel que } t_j \in \zeta(c_i) \tag{5.4}$$

où f est la fonction qui donne la fréquence (nombre d'occurrences) du concept ou terme dans le document ou la requête.

De manière similaire $\forall c_i \in Indx(V(q))$:

$$f(c_i, q) = \sum_{t_j} f(t_j, q), \forall t_j \in V(q) \tag{5.5}$$

Pour simplifier l'écriture, nous noterons par la suite $Indx(d)$, $Indx((q))$ à la place de $Indx(V(d))$, $Indx(V(q))$.

Par exemple, étant donné un document d :

$$d = \{t_1, t_2, t_3\}$$

avec :

$$f(t_1, d) = 2,\ f(t_2, d) = 1,\ f(t_3, d) = 1$$

et dans une ressource externe Υ :

$$t_1, t_2 \in \zeta(c_1), t_3 \in \zeta(c_2)$$

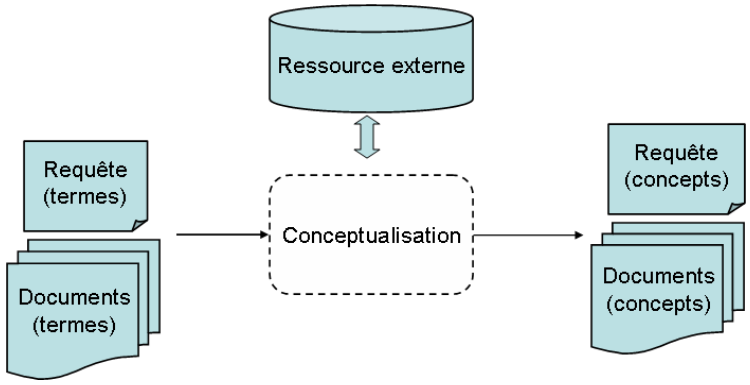


FIG. 5.2 – La conceptualisation des documents et de la requête en utilisant une ressource externe

La conceptualisation permet une présentation conceptuelle du document comme suit :

$$Indx(d) = \{c_1, c_2\}$$

avec :

$$f(c_1, d) = 3,\ f(c_2, d) = 1$$

Dans le cas où un concept représente des termes dans les langues différentes, cette indexation conceptuelle sera adaptée pour la RIM sur ces langues comme sur la RI monolingue parce qu'il n'y a plus la barrière de langues entre les concepts. La figure 5.2 illustre cette conceptualisation.

5.3 Modèle de RI basé sur le réseau Bayésien étendu avec une ressource externe (RIRBRE)

Avec une collection D de documents, une requête q et une ressource externe Υ donnée, nous définissons notre modèle de RI basé sur le réseau Bayésien par :

$$(\Upsilon, \Psi, \Delta)$$

où :

- Υ est la ressource externe intégrée dans le modèle.
- Ψ est le RB représentant les documents D, la requête q, les concepts ainsi que les relations entre D, q et les concepts.
- Δ est la fonction de correspondance entre un document d et la requête q.

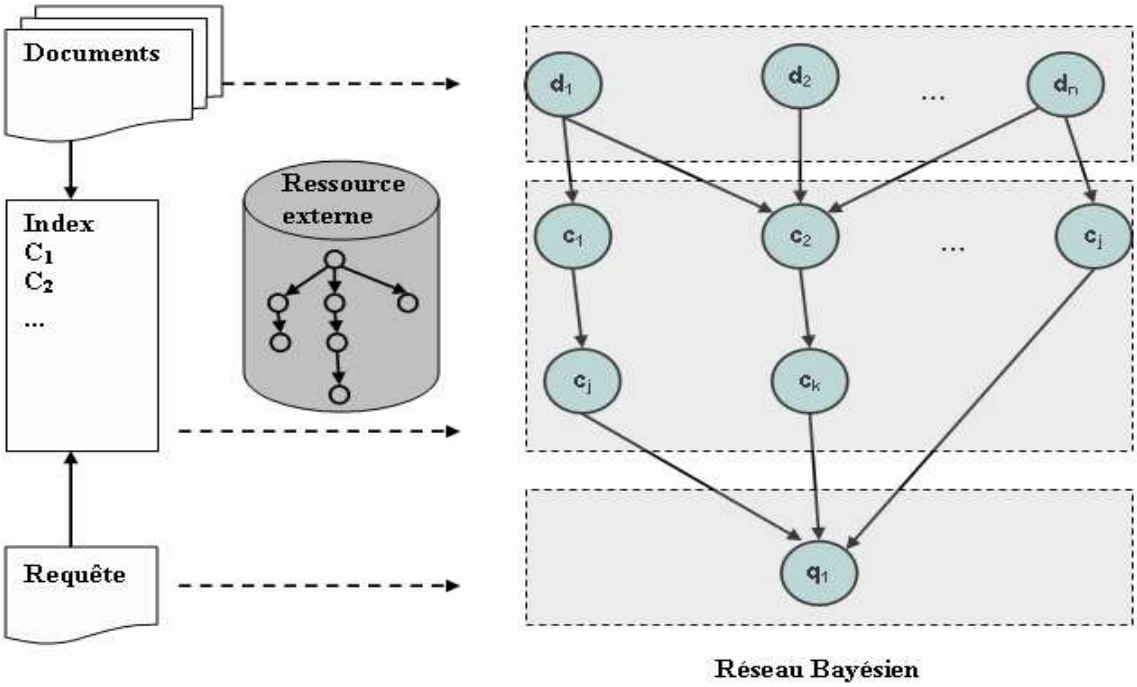


FIG. 5.3 – Modèle de la RI basé sur le réseau Bayésien étendue avec une ressource externe

La figure 5.3 illustre notre modèle. Dans cette figure, le réseau Bayésien est construit par les nœuds représentant les documents D, la requête q et les concepts de l'indexation. Les relations entre concepts sont extraites à partir de la ressource externe.

Nous présentons la description détaillée des éléments du modèle par la suite.

5.3.1 Le réseau Bayésien des documents et de la requête

Notre réseau RB est défini aussi par un triplet $\Psi(\mathcal{N}, \mathcal{A}, \mathcal{P})$.

Les nœuds $\mathcal{N}$

L'ensemble des documents D, de la requête q et leurs concepts Γ est modélisé par l'ensemble des variables. Chaque variable a deux états : observé et non observé. Dans le RB, ces variables sont représentées par les nœuds $\mathcal{N}$. Il s'agit d'une bijection $\mathcal{F}$: $D \cup \{q\} \cup \Gamma \rightarrow \mathcal{N}$. Par exemple une variable de concept c est notée par $\mathcal{F}(c)$. Pour simplifier la notation, nous remplaçons dans la suite $\mathcal{F}(c)$ par c.

Les nœuds dans notre RB sont donc définis comme suit :

$$\mathcal{N} = D \cup \{q\} \cup \{Indx(q)\} \cup \{Indx(d) | d \in D\}$$

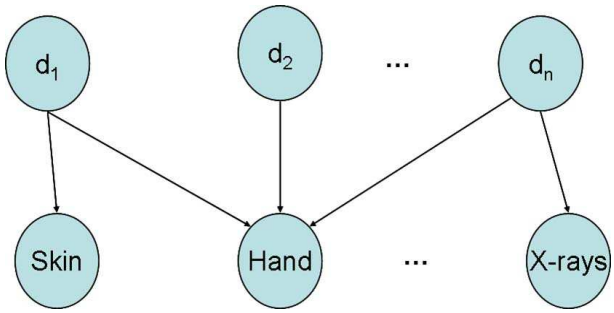


FIG. 5.4 – Les arcs entre les documents et les concepts

Les arcs $\mathcal{A}$

Les arcs dans notre RB sont composés de trois sous-ensembles correspondant à trois types d'arcs :

$$\mathcal{A} = \mathcal{A}_{\mathcal{CD}} \cup \mathcal{A}_{\mathcal{CQ}} \cup \mathcal{A}_{\mathcal{CC}}$$

Ces arcs ne doivent former aucun cycle.

- $\mathcal{A}_{\mathcal{CD}}$: Les arcs entre nœuds concepts et nœuds documents
  Ces arcs représentent les liens de l'indexation, i.e entre les documents et leurs index :

  $$\mathcal{A}_{\mathcal{CD}} = \{(d, c) | d \in D, c \in Indx(d)\}$$

  Ces arcs sont orientés des documents vers les concepts. Les liens entre un nœud document d et les concepts qu'il contient impliquent que ces concepts rapportent au document certaines informations, par définition de la fonction $Indx$.
  La figure 5.4 est un exemple de ces arcs.
- $\mathcal{A}_{\mathcal{CQ}}$: Les arcs entre les concepts et la requête
  De manière similaire, ces arcs modélisent les liens entre la requête et ses concepts :

  $$\mathcal{A}_{\mathcal{CQ}} = \{(c, q) | c \in Indx(q)\}$$

  Le nœud requête est lié à un nœud concept qu'il contient par un lien dirigé vers le nœud requête.
  L'exemple de ces arcs est illustré dans la figure 5.5.
- $\mathcal{A}_{\mathcal{CC}}$: Les arcs entre concepts
  Ces arcs sont extraits à partir de la ressource externe, donc $\mathcal{A}_{\mathcal{CC}} \subset \mathcal{R}$:

  $$\mathcal{A}_{\mathcal{CC}} = \{(c_i, c_j) | c_i, c_j \in Indx(q) \cup Indx(d), \forall d \in D \text{ et } (c_i, c_j) \in \mathcal{R}\}$$

  La figure 5.6 est un exemple de l'extraction des relations entre concepts à partir de la ressource externe pour construire le RB.

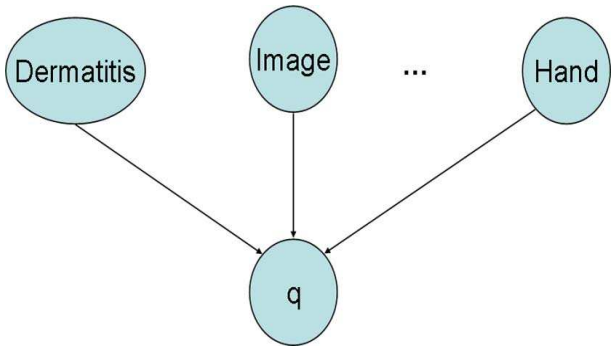


FIG. 5.5 – Les arcs entre la requête et les concepts

La vérification de cycle dans le RB

Pour que notre RB soit correct, il faut vérifier qu'il est acyclique. Voyons les types de nœuds dans le RB :

- les nœuds documents n'ont que des arcs divergents, i.e il n'y a que des arcs d'origine de ces nœuds. Il n'existe alors pas de chemin $che(d, d)$ dans notre RB, donc il n'y a pas de cycle.
- le nœud de la requête, quant à lui, n'a que des arcs convergents. Il n'y pas non plus de cycle sur ce nœud.
- pour les nœuds concepts, il n'y a pas non plus des cycles sur ces nœuds dans le RB par définition de $\mathcal{R}$

Notre RB est donc acyclique.

Les probabilités conditionnelles $\mathcal{P}$

Nous reprenons le principe du RB dans cette partie : $\mathcal{P}$ inclut les probabilités conditionnelles de tous les nœuds dans le RB sachant leurs parents :

$$\mathcal{P} = \{P(n|pa(n))|n \in \mathcal{N})\}$$

Nous définissons ces probabilités dans la section 5.3.3 du processus d'inférence des probabilités.

5.3.2 La fonction de correspondance

Le dernier élément dans notre modèle à base de RB, le Δ, est la fonction de correspondance entre les documents et la requête. Nous reprenons le principe du modèle de RI basé sur un RB classique :

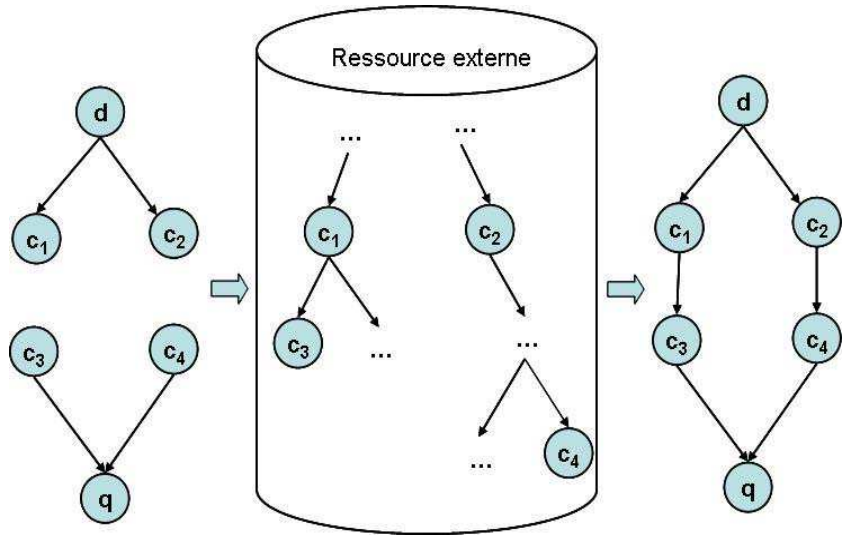


FIG. 5.6 – Exemple de l'extraction des relations à partir de la ressource externe pour construire le réseau bayésien

$$\Delta : \quad D \times q \quad \rightarrow [0,1]$$

Nous reprenons aussi dans le modèle de Turtle et Croft [80] (cf. section) :

$$\Delta(d,q) = P(q|d) = bel(q)$$

où $bel(q)$ est la croyance (belief) que q soit satisfait. La procédure de RI dans ce modèle est le processus d'inférence de probabilité sur le RB : un document d observé provoque une inférence des probabilités sur le réseau, des parents à ses enfants et termine à q. Le but de l'inférence est de calculer $bel(q)$, ou la croyance que q soit observée sachant d.

Nous présentons par la suite le processus d'inférence de probabilité pour calculer $bel(q)$.

5.3.3 Le processus d'inférence des probabilités ou de la recherche d'information dans le modèle proposé

Le processus de recherche d'information dans notre modèle est un processus d'inférence des probabilités sur le RB étant donné un document d observé. Il s'agit de la mise à jour des probabilités des nœuds dans le RB sachant que d est observé. Le but est de calculer la probabilité que la requête soit observée.

L'inférence des probabilités dans les RB est NP-difficile. Avec un document d observé, nous étudions d'abord la capacité d'éliminer certains nœuds ou arcs afin de simplifier le calcul des probabilités. Cette élimination est juste l'ignorance de ces nœuds ou arcs dans la procédure d'inférence de probabilité dans le contexte où d est observé. Ces nœuds et arcs doivent être conservés tels quels dans le RB original Ψ pour l'observation des autre documents.

Nous définissons cette élimination par la fonction $\mathcal{E} : \mathcal{E}(d, \Psi(\mathcal{N}, \mathcal{A}, \mathcal{P})) = \Psi'(\mathcal{N}', \mathcal{A}', \mathcal{P}')$. A chaque fois qu'un document d est observé, notre procédure d'inférence de probabilité sera exécutée sur Ψ'.

Les éliminations proposées concernent :

- $\{d_j | d_j \in D \wedge d_j \neq d\}$: tous les documents autres que d sont éliminés.
- $\{c_k | c_k \in \mathcal{N} \wedge c_k \notin (Indx(d) \cup Indx(q))\}$: tous les nœuds concepts qui n'appartiennent ni à d, ni à q.
- Comme le but du modèle proposé est de prendre en compte des relations hiérarchiques entre les concepts du document d et les concepts de la requête q, et pas les relations entre les concepts du document d ou entre les concepts de la requête, nous pouvons donc éliminer :
  - $\{(c_i, c_j) | c_i, c_j \in Indx(d)\}$: les relations entre concepts de d, par exemple dans la figure 5.7

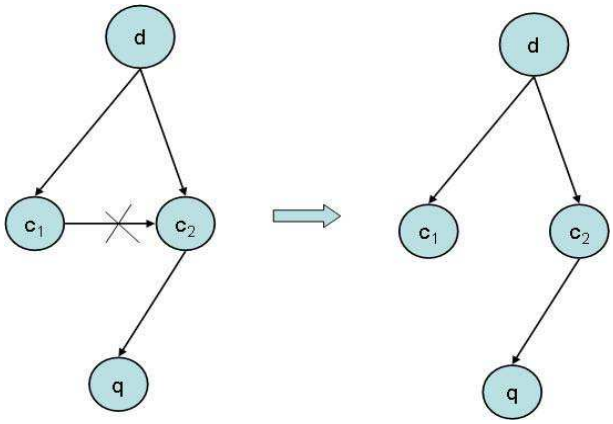


FIG. 5.7 – Exemple d'une réduction des relations entre concepts du document

  - $\{(c_i, c_j) | c_i, c_j \in Indx(q)\}$: les relations entre concepts de la requête, l'exemple est illustré par la figure 5.8

La figure 5.9 illustre l'exemple d'un réseau et de l'élimitation les nœuds pour un document observé.

Avec cette élimination pour chaque document d observé, dans le RB réduit Ψ', les

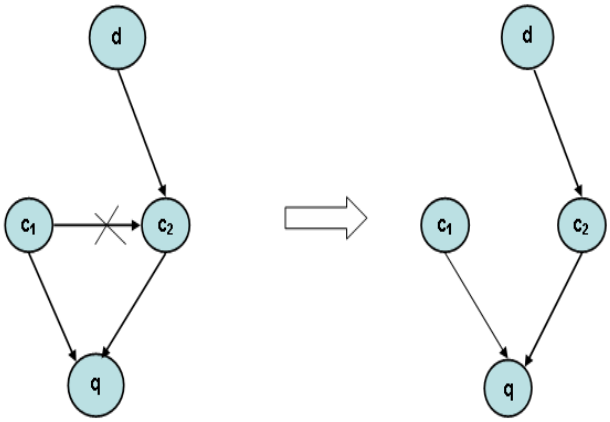


FIG. 5.8 – Exemple d'une réduction des relations entre concepts de la requête

nœuds dont il faut calculer les probabilités dans la procédure de l'inférence sont :

$$\mathcal{N}' = d \cup \{q\} \cup Indx(d) \cup Indx(q)$$

et les arcs :

$$\mathcal{A}' = \{\mathcal{A}_{\mathcal{CD}} \cup \mathcal{A}_{\mathcal{CQ}} \cup \{(c_i, c_j) \in \mathcal{A}_{\mathcal{CC}} | c_i \in Indx(d) \wedge c_j \in Indx(q)\}\}$$

Algorithme d'inférence des probabilités

Sur Ψ', nous pouvons adapter la méthode d'inférence des probabilités de Turtle et Croft : les probabilités ou la croyance des nœuds pourront être calculées en utilisant des matrices d'association (cf. section 4.2.4).

Le processus d'inférence comprend 3 étapes :

- l'initialisation de la probabilité antérieure quand un document d est observé ;
- l'inférence des probabilités sur les nœuds de concepts ;
- l'estimation de $bel(q)$ ou $P(q = Vrai)$. Le rang des documents pour chaque requête pourra s'établir sur ces valeurs.

Nous détaillons chaque étape par la suite.

Initialisation de la probabilité antérieure

Étant donné un document d observé, différents schémas d'initialisation de sa probabilité antérieure peuvent être effectués selon les différentes interprétations de cet événement.

$P(d)$ représente la probabilité que d soit observé. Quand l'événement « d est observé » est certain, on peut définir :

$$P(d) = 1 \tag{5.6}$$

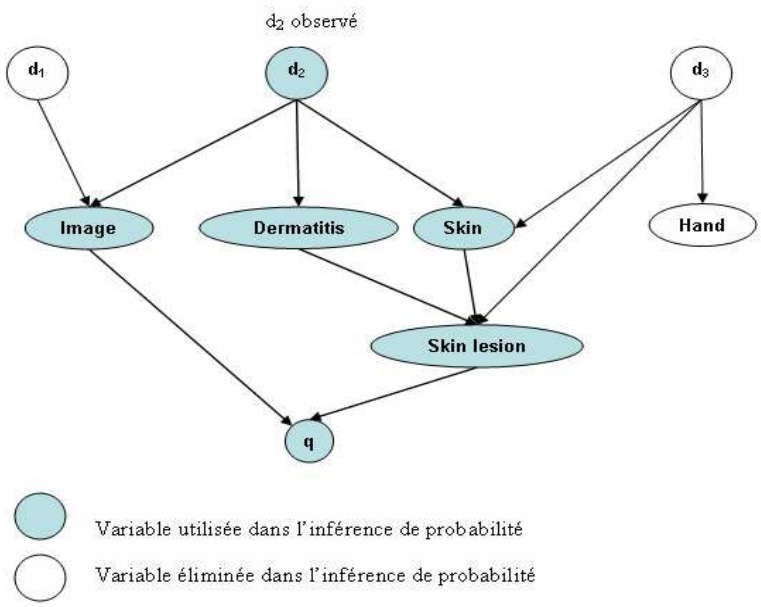


FIG. 5.9 – Exemple du réseau et de l'élimitation des nœuds quand d_2 est observé.

L'inférence des probabilités sur des concepts

$\forall c_i \in \mathcal{N}$, sa probabilité sera calculée en se basant sur ses parents. Dans notre RB, il y a deux cas :

- $c_i \in Indx(d)$, c'est le cas où d est parent de c_i.
- $c_i \notin Indx(d)$ dans ce cas d n'est pas parent de c_i. Dans Ψ', c_i doit être parmi des concepts de la requête, i.e $c_i \in Indx(q)$.

Le calcul de la probabilité de c_i ou $bel(c_i)$ dans chaque cas est décrit ci-dessous.

- Si $c_i \in Indx(d)$
  Soit par exemple les nœuds « corps » et « maladie de la peau » dans la figure 5.10, on définit :

$$bel(c_i) = P(c_i|d) \times P(d) = P(c_i|d) \tag{5.7}$$

  La probabilité conditionnelle $P(c_i|d)$ représente la probabilité qu'on observe le concept c_i sachant le document d observé. Dans le contexte de la RI, cela correspond à la pondération de c_i par rapport à d. Les différents schémas de pondération des index, par exemple ceux à base de $tf.idf$ normalisé , peuvent être utilisés pour ce modèle. On définit :

$$bel(c_i) = P(c_i|d) \quad = \quad w(c_i, d) \tag{5.8}$$

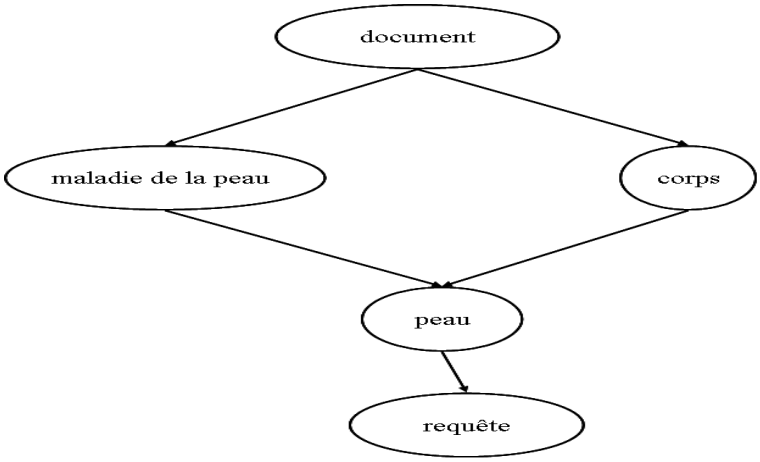


FIG. 5.10 – Exemple où plusieurs concepts d'un document sont liés avec un concept de la requête.

où $w(c_i, d)$ est une fonction de la pondération du c_i dans d.
Cette fonction peut être aussi appliquée pour la pondération d'un concept dans la requête. Elle est définie par :

$$\begin{aligned} w: \quad D \cup \{q\} \times \mathcal{N} &\rightarrow [0,1] \\ (d,c) &\rightarrow x \\ (q,c) &\rightarrow y \end{aligned}$$

– Si $c_i \notin Indx(d)$, i.e $c_i \in Indx(q)$
C'est le cas où c_i est un concept de la requête. Il y a deux possibilités :
- Si $\nexists c_j \in Indx(d)$ tel que $(c_j, c_i) \in \mathcal{A}'$, i.e il n'existe pas de concept dans d qui ait un arc vers c_i. Dans ce cas, c_i n'est pas observé, on a donc :

$$bel(c_i) = 0$$

- Si $\exists c_j \in Indx(d)$ tel que $(c_j, c_i) \in \mathcal{A}'$, c'est le cas où il existe au moins un concept dans d qui a un arc vers c_i. Par exemple dans la figure 5.10, avec $c_i =$ «*peau*».
Un tel concept peut avoir plusieurs concepts parents. Il faut donc choisir une combinaison raisonnable des probabilités des parents. Dans le cas où un concept de la requête c_i est lié à plusieurs concepts c_j d'un document d, la probabilité de c_i est calculée en fonction de la combinaison des probabilités de ses parents c_j. Il faut donc choisir le type de combinaison des parents qui correspond bien au contexte de la RI.

Dans notre réseau, toutes les relations entre concepts sont de même type hiérarchique, les concepts parents de c_i sont supposés plus ou moins sémantiquement proches ensembles. S'il existe un c_j qui est important (cette importance est représentée par sa probabilité, ou son poids dans le document) pour le contenu du document d, c_i doit être aussi important. Par contre, si tous les concepts parents ont faible importance (par exemple les concepts qui ne sont pas discriminants dans la collection), c_i doit être aussi de faible importance. Ce n'est pas le nombre de parents qui compte dans ce cas. L'importance de c_q dépend donc de l'importance maximum d'un concept parmi ses concepts parents, et pas de l'ensemble des parents.

C'est pour cette raison que nous choisissions a Max-Combinaison (cf. section 4.2.4) des parents dans le calcul de la probabilité de c_i. Cette combinaison prend en compte seulement le concept qui a la probabilité la plus importante. Nous pensons aussi que les autres combinaisons qui prennent en compte la probabilité de tous les parents correspondent moins à ce contexte.

En plus, nous proposons de prendre en compte le poids de la relation entre c_i, c_j, ou la dépendance entre eux $P(c_i|c_j)$. Ce poids est défini par sim, qui est une fonction pour mesurer la similarité entre deux concepts. Nous définissons cette fonction comme suit :

$$\begin{aligned} sim: \quad & \mathcal{R} \rightarrow [0,1] \\ & (c_i, c_j) \rightarrow x \end{aligned}$$

Nous définissons :

$$bel(c_i) = bel(c_h) \times sim(c_h, c_i) \text{ avec } c_h = argmax_{c_k \in pa(c_i) \cap Indx(d)}(bel(c_k)) \quad (5.9)$$

Nous résumons donc le calcul de la probabilité des concepts comme suit :

$$bel(c_i) = \begin{cases} w(c_i, d) \text{ si } c_i \in Indx(d) \\ bel(c_h) \times sim(c_h, c_i), \text{ avec } c_h = argmax_{c_k \in pa(c_i) \cap Indx(d)}(bel(c_k))), \\ \qquad \text{si } c_i \in Indx(q) \text{ et } \exists c_j \in Indx(d) \text{ telque } (c_j, c_i) \in \mathcal{A}' \\ 0 \text{ sinon} \end{cases} \quad (5.10)$$

où $sim(c_i, c_k)$ est l'estimation de la similarité entre deux concepts.

L'estimation de bel(q) : La probabilité postérieure ou la croyance de la requête

La similarité document-requête dans ce modèle est la probabilité postérieure de la requête sachant un document observé. Cela peut être interprété comme la croyance que

le besoin d'information soit satisfait lorsque le processus d'inférence est effectué pour un document fourni.

Le principe du calcul est le suivant : plus les concepts de la requête sont observés sachant que le document d est observé, plus le besoin d'information de la requête est satisfait par rapport à d. De plus, les poids des concepts dans la requête sont aussi importants à prendre en compte dans le calcul de la similarité, parce qu'un concept qui a un poids important doit avoir une forte influence sur la requête qui le contient. La raison est que le poids d'un concept dans la requête représente son importance ou sa contribution dans le contenu de la requête. Autrement dit, la « quantité » (nombre) ainsi que la « qualité » (poids) des concepts de la requête doivent être pris en compte dans le calcul de la probabilité de la requête. C'est pour cette raison que la matrice de la somme des poids (cf. section 4.2.4) semble une solution raisonnable. La correspondance est donc calculée comme suit :

$$\Delta(d,q) = P(q|d) = bel_{wsum}(q) \quad = \quad \frac{\sum_{c_i \in Indx(q)} w(c_i,q) \times bel(c_i)}{\sum_{c_j \in Indx(q)} w(c_j,q)} \tag{5.11}$$

où $w(c_i,q)$ est le poids du concept c_i dans la requête q. Ce poids peut être calculé en utilisant différents schémas de pondération des concepts dans la requête.

À la fin de cette étape, la similarité entre tous les couples de document-requête est établie.

5.3.4 Un exemple de procédure d'inférence des probabilités

Voyons un réseau RB dans la figure 5.11. Chaque document parmi d_1, d_2, d_3 observé provoque une élimination sur le réseau origine pour avoir les réseaux simplifiés correspondants. La procédure d'inférence des probabilités est effectuée sur ces réseaux simplifiés.

Supposons que les poids des concepts dans les documents et la requête soient :

$$w(Image, d_1) = 0.1$$

$$w(Image, d_2) = 0.1$$

$$w(Dermatitis, d_2) = 0.4$$

$$w(Skin, d_2) = 0.2$$

$$w(Skin, d_3) = 0.1$$

$$w(Image, q) = 0.2$$

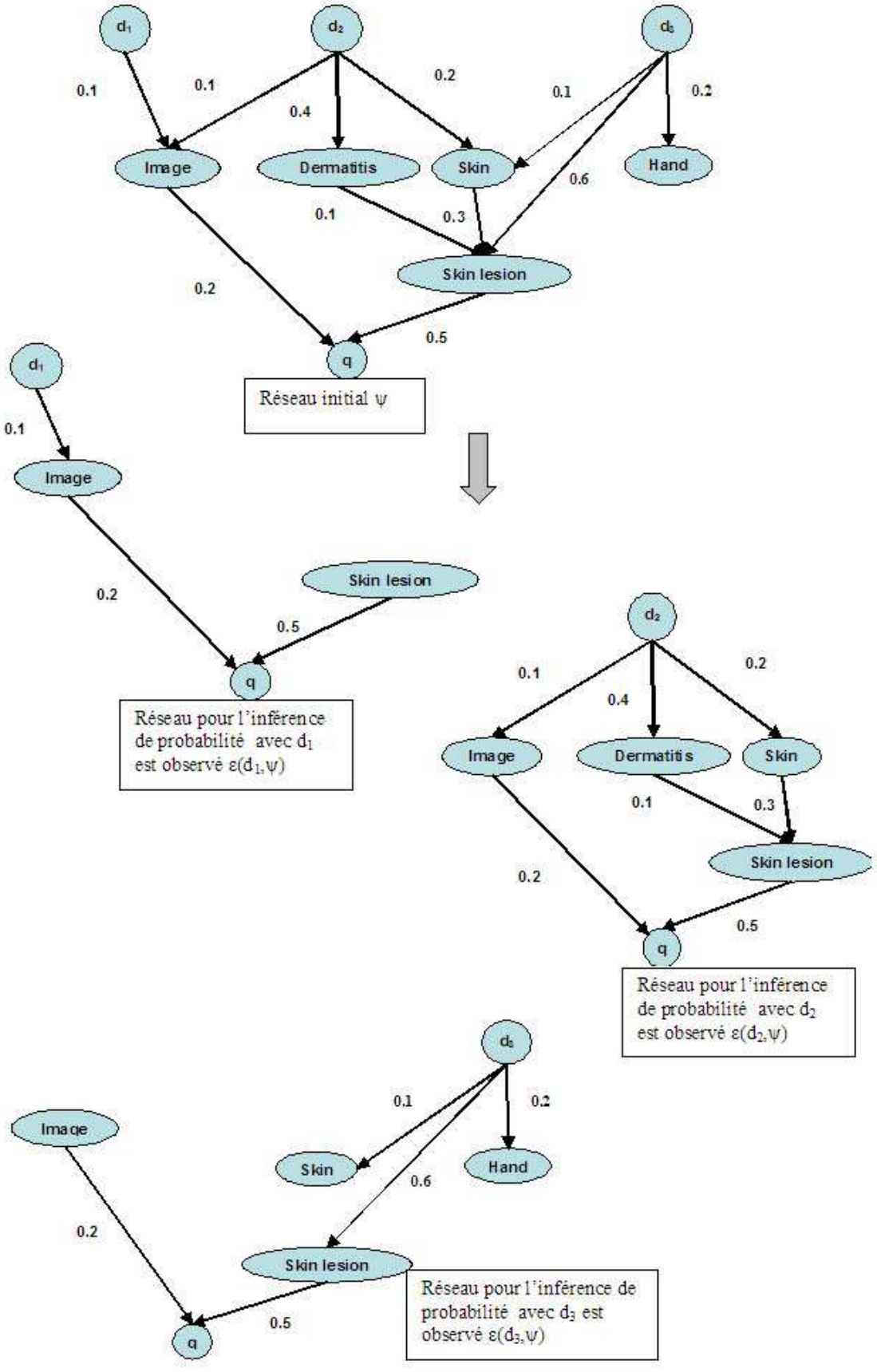


FIG. 5.11 – Exemple du réseau pour la procédure de l'inférence des probabilités

$$w(Skinlesion, d_3) = 0.6$$

$$w(Skinlesion, q) = 0.5$$

$$w(Hand, d_3) = 0.2$$

et que les valeurs de la similarité entre concepts soient :

$$sim(Skinlesion, Dermatitis) = 0.1$$

$$sim(Skinlesion, Skin) = 0.3$$

- Si d_1 est observé :

$$bel(d_1) = 1$$

On a :

$$bel(Image) = w(Image, d_1) = 0.1$$

$$bel(Skinlesion) = 0$$

La similarité document-requête :

$$\begin{aligned} P(q|d_1) &= \frac{\sum_{c_i \in Indx(q)} w(c_i, q) \times bel(c_i)}{\sum_{c_j \in Indx(q)} w(c_j, q)} \\ &= \frac{w(Image, q) \times bel(Image) + w(Skinlesion, q) \times bel(Skinlesion)}{w(Image, q) + w(Skinlesion, q)} \\ &= \frac{0.2 \times 0.1 + 0}{0.2 + 0.5} \\ &= 0.0286 \end{aligned}$$

- Si d_2 est observé :

$$bel(d_2) = 1$$

Les probabilités des concepts du document sont :

$$bel(Image) = w(Image, d_2) = 0.1$$

$$bel(Dermatitis) = w(Dermatitis, d_2) = 0.4$$

$$bel(Skin) = w(Skin, d_2) = 0.2$$

Inférence de la probabilité sur le nœud de concept de la requête :

$$Argmax_{c_k = Dermatitis, Skin}(bel(c_k)) = Dematitis$$

$$\begin{aligned} bel(Skinlesion) &= bel(Dermatitis) \times Sim(Skinlesion, Dermatitis) \\ &= 0.4 \times 0.1 \\ &= 0.04 \end{aligned}$$

La similarité document-requête :

$$\begin{aligned} P(q|d_2) &= \frac{\sum_{c_i \in Indx(q)} w(c_i, q) \times bel(c_i)}{\sum_{c_j \in Indx(q)} w(c_j, q)} \\ &= \frac{w(Image, q) \times bel(Image) + w(Skinlesion, q) \times bel(Skinlesion)}{w(Image, q) + w(Skinlesion, q)} \\ &= \frac{0.2 \times 0.1 + 0.5 \times 0.04}{0.2 + 0.5} \\ &= 0.0571 \end{aligned}$$

- Si d_3 est observé :

$$bel(d_3) = 1$$

On a :

$$bel(Skin) = w(Skin, d_3) = 0.1$$

$$bel(Skinlesion) = w(Skin, d_3) = 0.6$$

$$bel(hand) = w(Hand, d_3) = 0.2$$

$$bel(Image) = 0$$

La similarité document-requête :

$$\begin{aligned} P(q|d_3) &= \frac{\sum_{c_i \in Indx(q)} w(c_i, q) \times bel(c_i)}{\sum_{c_j \in Indx(q)} w(c_j, q)} \\ &= \frac{w(Skinlesion, q) \times bel(Skinlesion) + w(Image, q) \times bel(Image)}{w(Image, q) + w(Skinlesion, q)} \\ &= \frac{0.5 \times 0.6 + 0}{0.2 + 0.5} \\ &= 0.429 \end{aligned}$$

Le rang des documents dans l'ordre décroissant de la pertinence est donc :

d_3

d_2

d_1

5.4 La réduction du modèle proposé au modèle à base d'intersection

Nous allons dans cette section de la thèse montrer que notre modèle a la capacité de simuler les modèles à base d'intersection, par exemple le modèle VSM, dans le cas où il

n'y a pas de relations sémantiques entre les concepts. A partir de ce point, nous pourrons montrer les apports de la prise en compte des liens sémantiques par la comparaison de la performance selon le cas où ces liens existent ou pas dans le réseau.

Dans le cas où il n'y a pas de relations sémantiques entre les concepts des documents et ceux de la requête, c_i dans la formule (5.11) est donc soit un concept que le document d et la requête q ont en commun, soit il n'appartient pas au document.

On a donc $\forall c_i$:

$$bel(c_i) = \begin{cases} P(c_i|d) = w(c_i, d) \text{ si } d \in pa(c_i) \\ 0 \text{ sinon} \end{cases}$$

De plus, dans la formule (5.11), la valeur de $\sum_{c_j \in Indx(q)} w(c_j, q)$ est constante pour tout $bel(q)$. Dans le contexte de la RI, on peut alors éliminer ce paramètre dans la fonction de correspondance. La formule (5.11) devient :

$$\begin{aligned} \Delta(d, q) &= \sum_{c_i \in Indx(q)} w(c_i, q) \times bel(c_i) \\ &= \sum_{c_i \in (Indx(q) \cap Indx(d))} w(c_i, q) \times w(c_i, d) \end{aligned} \tag{5.12}$$

Cette formule est donc comme dans les modèles à base d'intersection. Par exemple, dans le cas du modèle VSM : Si on définit $w(c_i, q)$ tel que :

$$w(c_i, q) = \frac{w'(c_i, q)}{\sqrt{\sum_{j \in Indx(q)} w'(c_j, q)^2}} \tag{5.13}$$

et :

$$w(c_i, d) = \frac{w'(c_i, d)}{\sqrt{\sum_{h \in Indx(d)} w'(c_h, d)^2}} \tag{5.14}$$

où w' est une autre fonction de la pondération des concepts dans la requête ou dans les documents.

L'équation (5.12) devient :

$$\begin{aligned} \Delta(d, q) &= \sum_{c_i \in (Indx(q) \cap Indx(d))} \left(\frac{w'(c_i, q)}{\sqrt{\sum_{k \in Indx(q)} w'(c_k, q)^2}} \times \frac{w'(c_i, d)}{\sqrt{\sum_{h \in Indx(d)} w'(c_h, d)^2}} \right) \\ &= \frac{\sum_{c_i \in (Indx(q) \cap Indx(d))} w'(c_i, q) \times w'(c_i, d)}{\sqrt{\sum_{k \in Indx(q)} w'(c_k, q)^2} \times \sqrt{\sum_{h \in Indx(d)} w'(c_h, d)^2}} \end{aligned} \tag{5.15}$$

Cette formule est le cosinus du modèle VSM. Nous utilisons cette simulation dans les expérimentations pour évaluer notre modèle avec et sans les relations sémantiques pour voir les comportements du modèle.

5.5 Conclusion

Nous avons proposé dans ce chapitre un modèle de RI à base de réseau Bayésien étendu avec une ressource externe. Nous résumons le modèle dans le tableau 5.1.

Dans ce modèle, que nous appelons RIRBRE, les documents et la requête sont décrits sous forme de réseau de concepts. Les relations hiérarchiques entre concepts, qui sont extraites à partir d'une base de connaissances externe, sont aussi intégrées dans ce modèle par des arcs entre les nœuds concepts. De ce fait, ce modèle permet l'intégration des relations sémantiques entre concepts explicitement.

Le processus de RI est interprété par une inférence de probabilités sur le réseau. Ce processus permet de prendre en compte « naturellement » la similarité ou la dépendance entre concepts dans la fonction de correspondance. Cette similarité représentée par le poids de l'arc dans le RB, peut être calculée par une mesure de similarité sémantique.

De plus, le modèle proposé a la capacité de simuler les modèles à base d'intersection, en particulier le modèle VSM. C'est le cas où les relations entre concepts ne sont pas intégrées. Cela permet de comparer les deux modèles dans un cadre unique et l'effet de l'intégration des relations entre concepts.

Ce modèle résout le problème de la disparité dans l'indexation conceptuelle. Pour chaque couple document-requête, le problème de la disparité entre eux est résolu par une solution au niveau local par la prise en compte des relations entre concepts du document et de la requête. Cette solution n'influence pas la correspondance entre d'autres documents et la requête. C'est un point positif par rapport à la méthode de l'expansion de la requête qui résout ce problème par une solution qui change le besoin d'information et a donc un effet global. Le cadre général du modèle permet en plus d'englober d'autres modèles.

Dans le chapitre suivant nous abordons l'application de ce modèle dans le domaine médical et sa validation expérimentale.

TAB. 5.1 – Récapitulatif du modèle proposé

La collection : $D = \{d_1, d_2, ...d_N\}$ La requête : q
Le modèle Bayésien pour la RI à base de ressource externe : (Υ, Ψ, Δ) Υ : la ressource externe intégrée dans le modèle. Ψ : RB représentant les documents D, la requête q, les concepts ainsi que les relations entre D, q et les concepts. Δ : la fonction de correspondance entre un document d et la requête q.
Ressource externe : $\Upsilon(\mathcal{T}, \mathcal{C}, \mathcal{R})$ $\mathcal{T}$: l'ensemble des syntagmes nominaux. $\mathcal{C}$: l'ensemble des concepts. $\mathcal{R}$: l'ensemble des relations entre les couples de concepts.
Le réseau Bayesien : $\Psi(\mathcal{N}, \mathcal{A}, \mathcal{P})$ $\mathcal{N} = D \cup \{q\} \cup \{Indx(q)\} \cup \{Indx(d) \mid d \in D\}$ $\mathcal{A} = \mathcal{A}_{\mathcal{CD}} \cup \mathcal{A}_{\mathcal{CQ}} \cup \mathcal{A}_{\mathcal{CC}}$ Les arcs entre nœuds concepts et nœuds documents : $\mathcal{A}_{\mathcal{CD}} = \{(d, c) \mid d \in D, c \in Indx(d))\}$ Les arcs entre les concepts et la requête : $\mathcal{A}_{\mathcal{CQ}} = \{(c, q) \mid c \in Indx(q)))\}$ Les arcs entre concepts : $\mathcal{A}_{\mathcal{CQ}} = \{(c_i, c_j) \mid c_i, c_j \in Indx(q) \cup Indx(d), \forall d \in D))\}$ Les probabilités conditionnelles : $\mathcal{P} = \{P(n \mid pa(n)) \mid n \in \mathcal{N})\}$
La fonction de correspondance : $\Delta(d, q) = bel(q)$
Le calcul de la fonction de correspondance, pour chaque document d dans la collection par rapport à la requête, est effectué via le processus d'inférence de la probabilité sur le réseau réduit : $\Psi'(\mathcal{N}', \mathcal{A}', \mathcal{P}') = \mathcal{E}(d, \Psi(\mathcal{N}, \mathcal{A}, \mathcal{P}))$ $\mathcal{E}$ est une fonction d'élimination ou réduction du réseau original selon le document d observé.

Chapitre 6

Validation du modèle proposé : application d'UMLS dans la RI médicale multilingue

Sommaire

6.1 Introduction

Dans ce chapitre, nous réalisons des expérimentations dans un contexte concret de RI pour valider nos propositions. Il s'agit d'une application de la RI multilingue dans le domaine médical. Pour cela les corpus de documents issus de la collection-test de ImageCLEFMed 2006, 2007 sont utilisés. La ressource externe qu'on utilise est l'UMLS (présenté en section. 3.4).

Nous décrivons les évaluations effectuées dans le tableau 6.1. Les expérimentations

TAB. 6.1 – Description des évaluations effectuées

Evaluation	Index		Modèle					ImageCLEFMed	
	Terme	Concept	FREQ	DFR	BM25	VSM	RIRBRE	2006	2007
Indexation par terme (A)	X		X	X	X	X		X	
Indexation par concept (B)		X	X	X	X	X		X	X
Modèle RIRBRE (C)		X					X	X	X

suivantes sont menées :

- L'indexation conceptuelle versus indexation par terme (A versus B) dans la section 6.3 : une comparaison de l'indexation par concept et de l'indexation par terme dans les modèles de RI classiques (FREQ, VSM [73], BM25 [69], DFR [2]) est effectuée pour la RI et la RIM . Nous montrons que l'indexation par concept donne les meilleurs résultats globaux avec le modèle VSM. Ce résultat est ensuite utilisé pour le comparer avec le modèle RIRBRE.
- Le modèle à base de réseau Bayésien en utilisant une ressource externe (RIRBRE) pour la RIM avec et sans relations sémantiques entre concepts dans la section 6.4 : nous validons l'intérêt de la prise en compte des relations sémantiques du modèle proposé par l'amélioration de la performance par rapport au cas où il n'y a pas de relations. Par conséquent, cette comparaison permet B (VSM) versus C.

Nous présentons dans la section suivante le contexte d'expérimentation.

6.2 Contexte d'expérimentation

Cette section introduit le contexte d'expérimentation. Il concerne la collection, les outils d'extraction de concepts et les systèmes de RI existant que nous utilisons pour faire nos expérimentations.

"Show me images of a hand x-ray."
"Zeige mir Röntgenbilder einer Hand."
"Montre-moi des radiographies de la main."

FIG. 6.1 – Exemple d'une requête dans ImageCLEFMed2006

6.2.1 Collection ImageCLEFMed

ImageCLEFMed est une partie de CLEF (Cross-Language Evaluation Forum), qui est la campagne d'évaluation annuelle pour la recherche d'information multilingue depuis l'année 2000. ImageCLEFMed concerne la tâche de recherche d'images médicales à partir de documents hétérogènes et multilingues qui incluent à la fois des textes et des images.

L'ensemble des données (dataset) d'ImageCLEFMed 2006 contient Casimage, MIR, PEIR et PathoPIC (environ 50000 images au total) [30]. Deux nouveaux jeux de données, myPACS et CORI, sont ajoutés pour ImageCLEFMed2007. La collection inclut des images et leurs annotations au format XML. Ces annotations contiennent des textes en majorité en anglais, mais aussi en français et en allemand. Les détails de la description de la collection ainsi qu'un exemple d'une annotation peuvent être trouvés dans l'annexe C.

Il y a 25 requêtes dans ImageCLEFMed 2006 et 30 requêtes dans ImageCLEFMed 2007. Chaque requête comprend aussi du texte en trois langues et des images. La figure 6.1 montre un exemple d'une requête dans ImageCLEFMed 2006.

6.2.2 Extraction de termes

L'extraction de termes à partir de texte est illustrée dans la figure 6.2. Les documents et la requête sont analysés par l'analyseur morphosyntaxique TreeTagger [1] [74]. A partir de la sortie de TreeTagger, après suppression des mots vides, les termes sont sélectionnés ensuite pour représenter le contenu des documents et de la requête.

[1] http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

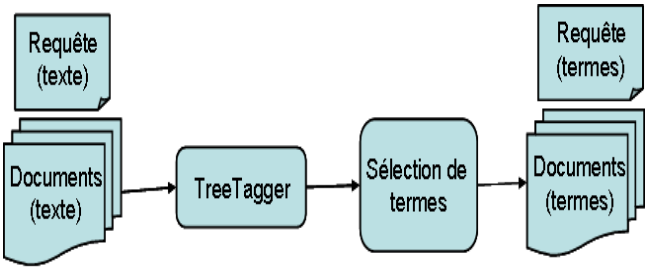


FIG. 6.2 – Le schéma de l'extraction des termes

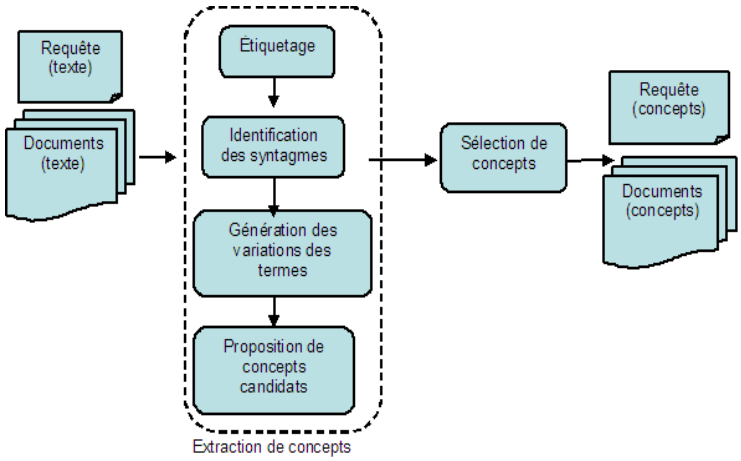


FIG. 6.3 – Le schéma de l'identification des concepts

6.2.3 Identification de concepts

Pour une indexation conceptuelle, il faut d'abord identifier des concepts à partir des termes. Pour cela, il faut extraire les concepts à partir de texte puis sélectionner les meilleurs. La figure 6.3 montre ce schéma. Il existe 2 outils d'extraction de concepts à partir de texte : Metamap [1], [3] et XIotaMap [63]. Avec la disponibilité de ces outils, nous utilisons Metamap pour extraire des concepts dans les textes en Anglais et XIotamap pour les textes en Français et Allemand.

Metamap

Pour extraire des concepts dans les textes anglais, UMLS nous offre l'outil Metamap [1], [3]. Les détails de la méthode de Metamap sont présentés dans l'Annexe D. Pour chaque syntagme, Metamap sort la liste des concepts candidats. Ces concepts correspondent à tous les composants dans la structure du syntagme. Les syntagmes sont structurés en tête-modificateurs. Chaque candidat est accompagné de valeurs qui mesurent sa

similarité avec le syntagme d'entrée. Metamap propose aussi le meilleur concept : celui qui est le plus similaire avec le syntagme. Les étapes nécessaires d'extraction des concepts de Metamap sont :

- **L'étiquetage** : pour associer les POS (Part Of Speech) à chaque mot.
- **L'identification des syntagmes** : pour identifier des syntagmes qui sont éventuellement associés avec un concept dans les ressources. Dans Metamap, cette étape est effectuée en utilisant le parseur SPECIALIST qui est basé sur le SPECIALIST lexicon dans UMLS.
- **La génération des variations** : générer toutes les variations possibles des syntagmes. Cette tâche est basée sur les variations lexicales, des abréviations, les positions des composants du syntagme, etc. Il est nécessaire de mesurer les distances entre le syntagme original et ses variations.
- **La proposition de concepts candidats** : en utilisant les sources terminologiques.
- **L'évaluation** de l'exactitude des concepts candidats.

XIotamap

Comme Metamap n'analyse que des textes en anglais, Maisonnasse [63] a construit un outil similaire pour analyser des textes en français et en allemand, appelé XIotaMap. XIotamap ne traite pas l'ambiguïté structurelle : il extrait tous les concepts qui correspondent aux variations des termes dans le texte.

Sélection de concepts

Avec la liste des concepts sortis par Metamap et XIotamap, nous avons deux possibilités pour sélectionner des concepts à partir de cette sortie :

- Ne prendre que le meilleur concept global qui correspond le mieux avec le terme. Cependant, ce concept peut être trop précis, par exemple le concept *«right middle lobe»*. Cela risque de réduire le rappel de la recherche. C'est pour cette raison que nous utilisons une autre solution d'extraction de concepts :
- Dans le cas d'un syntagme, nous prenons le meilleur concept global et en plus, les concepts qui correspondent à chaque composant de syntagme, i.e les concepts qui correspondent aux termes qui «composent» le syntagme, ce syntagme peut lui-même également correspondre à un concept. Il s'agit aussi donc de tous les concepts qui «composent» un concept. Donc, à partir de *«right middle lobe»*, nous identifions ces concepts : C0225757 («right middle lobe»), C0796494 («lobe»), C0205090 («right»), C0549183 («middle»).

Cette méthode s'approche de la méthode de Mao & Chu [50] dans le but d'identifier l'information précise (le concept correspond au plus long syntagme) mais aussi dans le but de garder les informations des termes qui composent le syntagme (ce sont des mots racinéis dans le travail de Mao & Chu, et les concepts correspondants dans le nôtre).

Cette méthode est une manière d'agrandir l'ensemble des concepts pour éviter le problème de la première méthode. Comme il n'y a aucune désambiguïsation du sens, il est difficile d'agrandir l'ensemble des concepts sans ajouter du bruit dans le résultat final. Cette solution semble la plus équilibrée.

Il faut noter aussi que ces deux outils d'extraction de concepts ne traitent que des syntagmes, et pas de verbes. Quelques informations risquent d'être perdues, par exemple le concept *«chest infection»* ne sera pas trouvé à partir de *«the chest is infected»*.

Nous verrons dans la section suivante l'utilisation des concepts extraits pour l'indexation conceptuelle dans la RI.

6.2.4 Le système de RI X-IOTA

Le système X-IOTA [20],[29] est une intégration flexible des différents modèles classiques de RI, y compris :

- FREQ (Fréquence et Intersection) : C'est la pondération et correspondance de base.
- VSM (Vector Space Model) : C'est le modèle classique de Salton [73] avec la pondération $tf.idf$ et mesure de correspondance cosinus.
- BM25 : C'est un modèle probabiliste proposé par Robertson [69].
- DFR (Divergence From Randomness) : c'est aussi un modèle probabiliste d'Amati [2].

Nous utilisons ces modèles ou pondérations classiques dans l'expérimentation de la RI à base de termes et de concepts par la suite.

6.3 Comparaison de la RI à base de concepts et de termes

Dans cette section, nous étudions les différences de performance de recherche entre des modèles de RI à base de termes et celui à base de concepts dans les modèles classiques. L'objectif est de valider l'intérêt de l'indexation conceptuelle ainsi que de trouver le modèle qui donne la meilleure performance avec les concepts. Ce modèle sera ensuite utilisé pour le comparer avec le modèle RIRBRE proposé.

Comme la collection est de nature multilingue, nous expérimentons à la fois la RI sur chaque langue et la RIM sur trois langues. La méthode pour la RIM est présentée par la

TAB. 6.2 – Exemple de fusion des termes

Document/requête	Anglais	Français	Allemand	Fusion-Union
q	t1	t2	t3	t1, t2, t3
d1	t1	t2		t1, t2
d2	t1			t1

suite.

6.3.1 La méthode pour la RIM

Dans la collection ImageCLEFMed, la distribution des textes des documents n'est pas équilibrée : il y a des documents ayant des textes dans 3 langues différentes (anglais, français et allemand), alors que d'autres ont des textes dans 2 langues différentes ou une seule langue.

- **Pour la RIM à base de termes :**
  Nous étudions les méthodes pour effectuer la recherche sur la collection multilingue :
  - *Fusion des index* (union) : Fusionner les termes d'un document ou de la requête dans différentes langues dans un seul index. La fusion ici est simplement une union des ensembles des termes. L'interrogation est effectuée comme pour la RI monolingue. Cependant, comme la distribution de texte des documents dans la collection n'est pas équilibrée, les documents qui possèdent plus de texte dans des langues différentes ont possiblement le plus grand nombre des termes en commun avec la requête. Par exemple dans le tableau 6.2, le document d1 qui a le texte en deux langues a 2 termes en commun avec la requête q (t1,t2), alors que le document d2 n'a qu'un seul terme commun (t1). Une pondération normale des termes n'est pas adaptée dans ce cas.
  - *Décomposer un document multilingue en plusieurs sous-documents*, chaque sous-document correspond au texte dans une langue : Les sous-documents sont interrogés et enfin un seul sous-document avec la valeur de RSV (Relevant status value) maximale sera retenu.
  - *Combinaison des résultats* : Dans la collection de test ImageCLEFMed, comme chaque requête est fournie en trois langues, le problème de la traduction ne se pose pas car nous pouvons effectuer une recherche monolingue séparément sur les trois langues. Dans les expérimentations pour la campagne CLEF, rien n'a été clairement indiqué quand au choix des langues des requêtes. Il est vrai, qu'en réalité, un système de recherche d'information multilingue ne dispose pas des

requêtes dans chaque langue. Utiliser les trois langues à la fois n'est donc pas réaliste, sauf si l'on fait l'hypothèse que l'utilisateur a posé sa requête dans une des langues, et un traducteur automatique l'a traduite dans les deux autres langues. De plus, la collection n'est pas totalement parallèle, la distribution des textes dans les trois langues n'est pas équilibrée : on n'a pas la totalité du corpus pour chaque langue. En l'absence d'indexation inter-langue comme nous l'avons fait avec l'indexation conceptuelle, nous avons choisi de considérer les langues séparément, en faisant l'hypothèse (non réaliste) de la traduction parfaite de la requête.

Les résultats de recherche monolingue dans chaque langue sont donc combinés pour le résultat final de la recherche sur cette collection multilingue. La figure 6.4 illustre cette méthode : l'interrogation en utilisant X-IOTA est effectué séparément sur : Requête Anglais-Document Anglais ; Requête Français-Document Français ; Requête Allemand-Document Allemand. Les trois listes des documents retrouvés dans chaque langue sont ensuite combinées par une fusion : la valeur de RSV finale d'un document par rapport à une requête est le maximum des valeurs de RSV de ce document dans chaque langue. Cette méthode donne le meilleur résultat dans notre expérimentation. C'est pour cette raison que nous utilisons cette méthode dans les expérimentations suivantes pour la RIM.

- **Pour la RIM à base de concepts :**

  L'indexation conceptuelle est supposée faire tomber la barrière de la langue, i.e les termes synonymes dans les langues différentes doivent correspondre à un concept identique. Pourtant ceci n'est pas assuré en pratique à cause de l'incomplet des ressources externes, des outils d'extraction de concepts,...Nous rencontrons donc le même problème que pour l'indexation par terme à cause de la distribution inéquilibrée des textes. C'est pour cette raison que nous effectuons aussi la méthode de fusion des résultats pour la RIM à base de concepts.

L'interrogation des documents et de la requête est effectuée avec X-IOTA. Le schéma de combinaison des résultats pour la RIM à base de concepts est comme à base de termes, ce qui est illustré dans la figure 6.4.

Les résultats de cette expérimentation sont présentés dans la section suivante.

6.3.2 Résultats

Nos expérimentations ont été effectuées en utilisant X-IOTA [20],[29]. Le tableau 6.5 montre les résultats d'indexation conceptuelle avec les modèles FREQ, VSM, DFR, BM25 sur chaque langue (Anglais, Français, Allemand) et la combinaison des trois langues

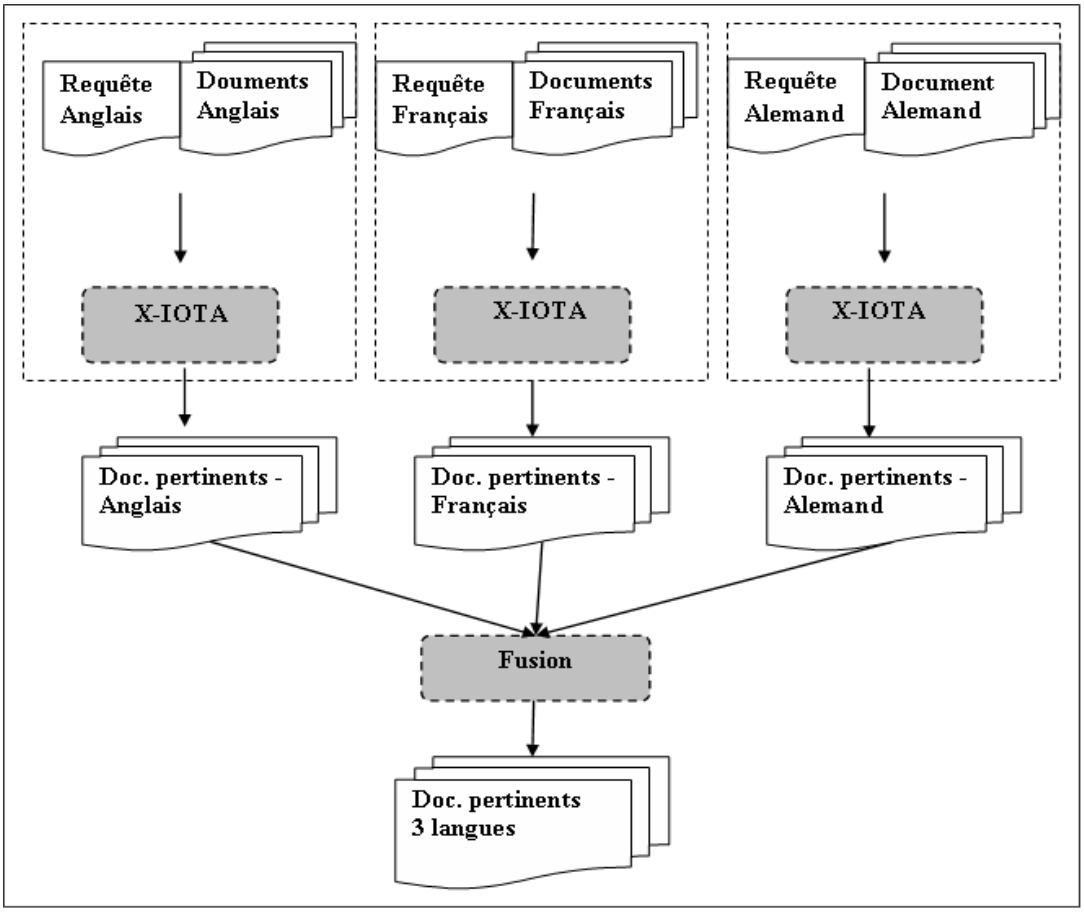


FIG. 6.4 – Le schéma de fusion des résultats pour la recherche d'information multilingue

TAB. 6.3 – Statistique des termes par langue dans UMLS version en 2005

	Nombre de termes dans UMLS	Pourcentage des termes/ total des termes dans UMLS
Anglais	4.297.431	63.45%
Français	156.404	2.31%
Allemand	168.186	2.48%

pour la RIM (noté par «3 Langues»). Le tableau 6.6 est la comparaison de l'indexation par terme (noté par «T») et par concept (noté par «C») pour chaque langue ou une combinaison de toutes les langues avec différents modèles. Les résultats sont évalués par MAP en utilisant le programme trec_eval [2], programme d'évaluation proposé par les campagnes d'évaluation de recherche d'information. La figure 6.6 décrit aussi ce résultat. Nous constatons que pour l'indexation conceptuelle, le MAP s'est beaucoup amélioré par rapport à l'indexation par terme dans toute l'interrogation sur les textes en Anglais, mais mois favorable sur deux autres langues. Le meilleur résultat de MAP pour la RIM est obtenu avec l'indexation conceptuelle avec le modèle VSM (0.204 de MAP). Les modèles DFR et BM25 donnent des résultats comparables avec le modèle VSM sur l'Anglais, mais beaucoup moins élevés sur 3 langues.

Nous expliquons par la suite les résultats obtenues dans les tableaux 6.3, 6.4 :

- La méthode à base de fréquence ne marche pas bien dans nos expérimentations :
  Ce résultat n'est pas une surprise, la fréquence (tf) seule est connue pour ne pas être une bonne pondération dans la littérature. La prise en compte des autres éléments, idf par exemple, donne effectivement de meilleurs résultats.
- La recherche avec concepts est moins bonne qu'avec des termes sur les langues Français et Allemand :
  Cela est à cause du problème de l'identification de concepts à partir de texte. L'identification de concepts à partir de texte en Français et Allemand n'est pas suffisamment efficace. La raison principale est que le nombre des termes associés aux concepts en Français et en Allemand dans la ressource UMLS est faible par rapport à l'Anglais, comme illustré dans le tableau 6.3.
- La recherche sur la partie du corpus limitée à l'Anglais marche le mieux, si on la compare avec les deux autres langues (Français, Allemand) :
  La raison de cette différence est que la collection n'est que partiellement parallèle : la distribution des textes par langue n'est pas équilibrée. Le nombre de textes en Anglais (40708) est beaucoup plus important qu'en Français (1899) et en Allemand

[2] http ://trec.nist.gov/data/reljudge_eng.html

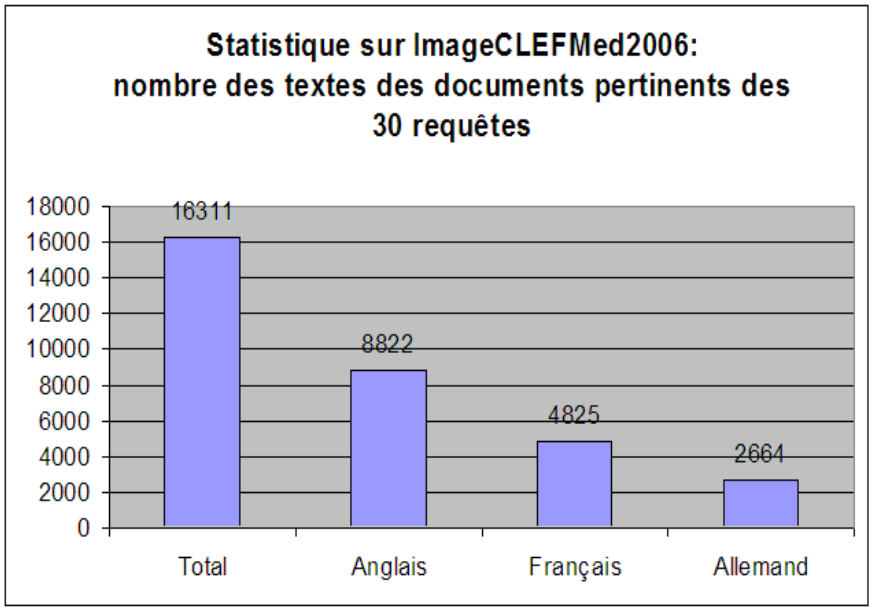


FIG. 6.5 – Statistique des documents pertinents par langue

(7805) (cf. Annexe C). Parmi les documents pertinents pour les 30 requêtes données, le nombre de document en Anglais (8822) est aussi plus grand qu'en Français (4825) et en Allemand (2664), comme illustré dans la figure 6.5. La capacité à trouver des documents pertinents à partir des textes en Français et en Allemand est donc aussi plus limitée. Par ailleurs, l'identification de concepts à partir de texte en Français et Allemand n'est pas suffisamment efficace. De ce fait, la capacité à trouver les documents pertinents à partir des textes en Anglais est aussi plus élevée.

– La combinaison de trois langues marche beaucoup mieux avec VSM que les autres modèles :

  Cela a rapport avec la fusion des résultats de l'interrogation séparée sur chaque langue. Pour un document multilingue (il y en a 7805 dans ImageCLEFMed2006) ; la méthode utilisée pour cette fusion consiste seulement à ne retenir que la valeur de RSV maximum calculée sur des index associés à chaque langue du document. Concrètement, la méthode de fusion consiste à produire trois listes des documents retrouvés, une par langue, puis à fusionner ces trois listes, à filtrer les documents identiques en ne prenant que le maximum de RSV, finalement à re-trier selon les RSV.

  Pour les modèles DFR et BM25, la distribution des valeurs de RSV de la recherche

TAB. 6.4 – Les intervalleles des RSV dans chaque langue des documents retrouvés pour la première requête

	Min RSV	Max RSV	RSV Moyen
Anglais	8.82	68.21	15.98
Français	80.95	112.19	85.76
Allemand	7.61	26.95	10.19

sur chaque langue est très différente. La raison est que la fonction de correspondance de DFR tient compte de la taille de la collection et que la fonction de correspondance de BM25 tient compte de la taille de chaque document. Cependant, ces éléments sont très variables sur chaque langue. Cela va donc directement modifier dans l'absolu la valeur de RSV par langue. Ces différences d'échelles dans les valeurs de RSV ne sont normalement pas un problème car ce qui compte ne n'est pas la valeur elle-même mais simplement l'ordre des document retrouvés. Si plusieurs requêtes sont faites, donc s'il y a fusion des listes, la valeur de RSV, ou plus particulièrement l'échelle de valeurs de RSV devient importante. C'est un problème de fusion bien connu qui est traité par ailleurs par exemple pour la recherche d'images. Par exemple avec modèle DFR : les intervalles des RSV dans chaque langue des documents retrouvés pour la première requête sont très larges (cf. le tableau 6.4). La méthode de fusion par maximum privilégie donc de manière artificielle les documents en Français tout simplement parce que leur RSV est en moyenne plus élevée (84.76) même s'ils sont moins bien classés dans le résultat de la recherche sur l'Anglais (15.98). Pourtant, la performance de la recherche sur le Français est nettement moins bonne qu'en Anglais, comme cela est expliqué précédement. C'est pour cette raison que notre méthode de fusion en l'appliquant sur les modèles DFR ne donne pas de bons résultats. Le modèle BM25 a un problème similaire à DFR.

Pour le modèle VSM, cette fusion marche parce que la valeur de RSV de ce modèle ne tient pas compte ni de la taille de la collection ni de la taille des documents. Dans chaque langue, pour la premierè requête, la valeur moyenne des RSV est de 0.08 pour l'Anglais, 0.04 pour le Français et 0.05 pour l'Allemand.

Pour toutes ces raisons, seuls les résultats sur la langue Anglaise sont véritablement significatifs pour la comparaison de l'indexation conceptuelle et de l'indexation par terme. Ces résultats, comme illustrés dans le tableau 6.4, montrent alors l'avantage de l'indexation conceptuelle, notamment avec le modèle VSM. Nous utilisons le meilleur résultat obtenu par le modèle VSM sur 3 langues comme le baseline pour comparer avec le modèle RIRBRE par la suite. Des exemples de cette expérimentation peuvent être trouvés

TAB. 6.5 – Comparaison (sur valeur de MAP) sur ImageCLEFMed2006 de différents modèles utilisant une indexation conceptuelle

	FREQ	VSM	DFR	BM25
Anglais	0.049	0.209	**0.210**	0.208
Français	0.003	**0.070**	0.047	0.043
Allemand	0.007	**0.016**	0.014	0.011
3 Langues	0.004	**0.204**	0.097	0.098

TAB. 6.6 – Comparaison de différents modèles utilisant une indexation par concept (C) et par terme (T) sur ImageCLEFMed2006

	FREQ		VSM		DFR		BM25	
	T	C	T	C	T	C	T	C
Anglais	0.028	**0.049**	0.166	**0.209**	0.053	**0.210**	0.157	**0.208**
Français	0.033	0.003	0.064	**0.070**	0.061	0.047	0.050	0.043
Allemand	0.010	0.007	0.017	0.016	0.018	0.014	0.013	0.011
3 Langues	0.021	0.004	0.176	**0.204**	0.06	**0.097**	0.167	0.098

dans l'Annexe E.

6.4 Application du modèle RIRBRE proposé pour la RIM

Dans cette section nous évaluons le modèle RIRBRE proposé dans le chapitre 5. Il s'agit de la RIM sur 3 langues. La méthode de RIM est identique à celle introduite dans la section 6.3.1. La figure 6.7 décrit notre schéma de recherche avec le modèle RIRBRE. Etant donné les documents et la requête représentés par les concepts (la conceptualisation est présentée dans la section 6.2.3), le processus de RI est effectué via les deux étapes suivantes :

- La construction du RB Ψ.
- Processus d'inférence sur le RB pour calculer la fonction de correspondance.

6.4.1 La construction du RB Ψ

Etant donné la collection D, la requête q et la ressource externe UMLS, nous construisons le RB Ψ comme suit :

- Pour chaque document d dans la collection, on crée le nœud du document, les nœuds des concepts dans d si ces concepts n'existent pas encore dans le réseau, ainsi que

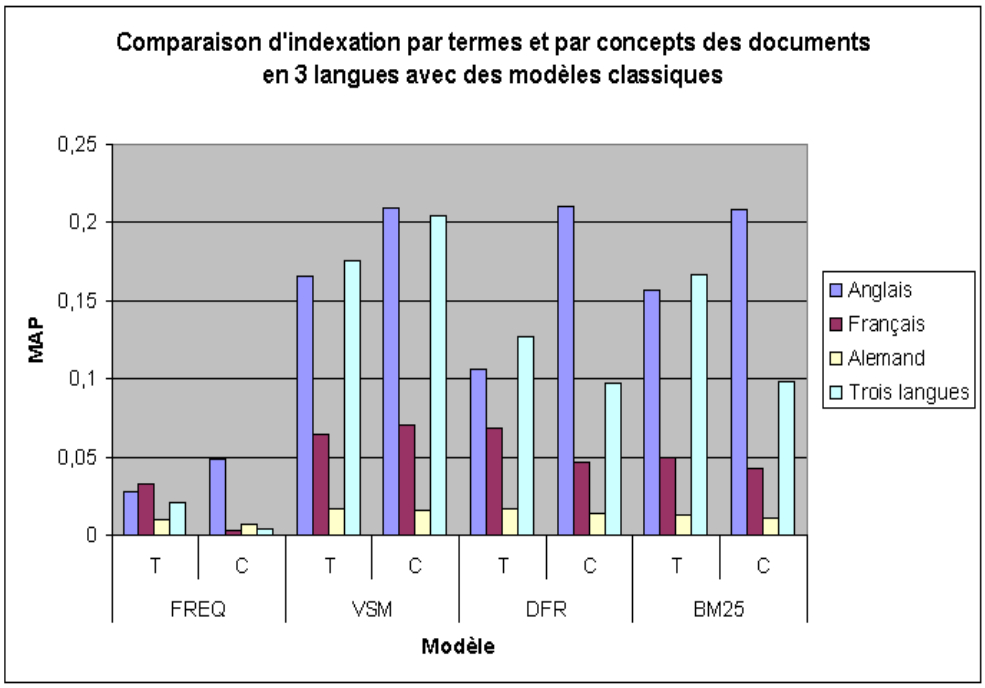


FIG. 6.6 – Comparaison d'indexation par terme et par concept des documents en 3 langues avec des modèles classiques

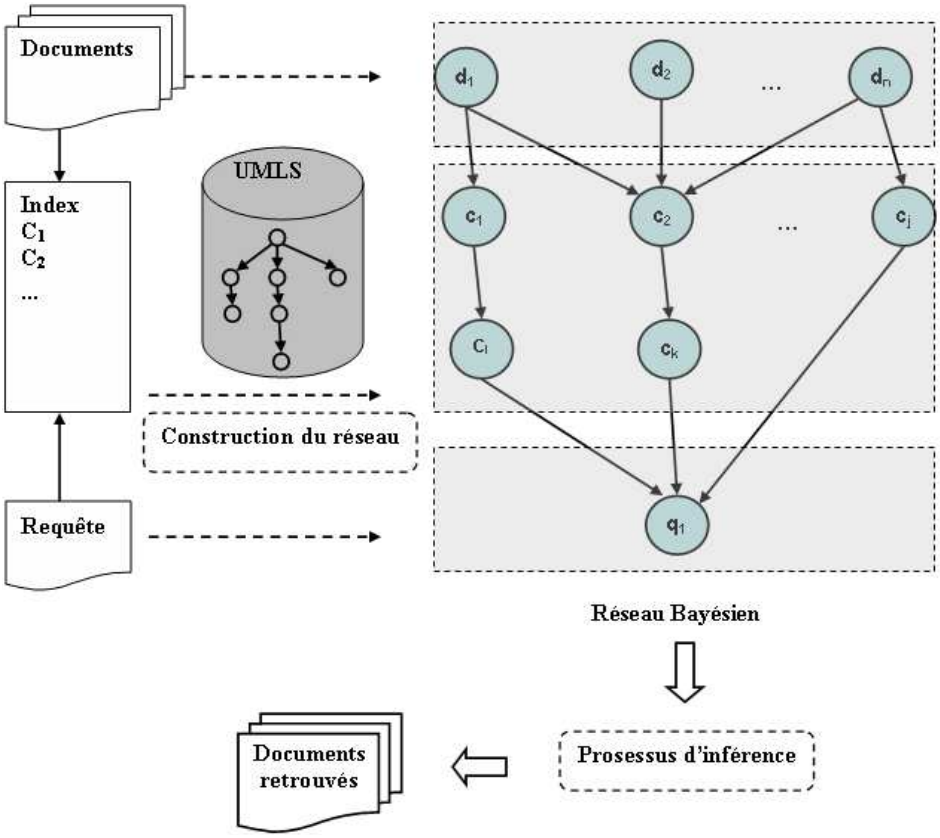


FIG. 6.7 – Schéma de la recherche avec le modèle RIRBRE

des liens entre le nœud document vers les nœuds concepts qu'il contient. La requête est traitée de manière similaire, mais les liens entre le nœud requête et les nœuds concepts qu'elle contient sont orientés vers la requête.

- En consultant la base de données d'UMLS, si on trouve une relation directe ou si on peut déduire une relation indirecte entre deux concepts c_i, c_j à partir des relations directes dans UMLS, on ajoute alors un arc entre ces deux concepts dans notre RB. Le choix des types de relations sémantiques à intégrer dans le modèle est une question importante car ces relations influent fortement dans le processus d'inférence. En effet, l'idée de l'inférence dans notre modèle signifie : la croyance qu'un document d déduise une requête q, il s'agit de la valeur de pertinence de d par rapport à q. De ce fait, les liens choisis qui participent à cette inférence respectent aussi cette idée. Nous ne considérons que les relations (c_1, c_2) qui satisfont la proposition suivante : si un document ou une requête parle de c_1, il parle probablement aussi de c_2. Les types de relations nous intéressant sont :
  - **la synonymie** est la *«relation sémantique entre des mots ou des expressions dont les sens sont identiques ou très proches»* [3].
    Par exemple : *épouse-conjointe, écran-terminal.*

  - **l'hyponymie-hyperonymie** (ou générale-spécifique) : l'hyponymie est «la relation entre un terme général et ses exemples (instances)» [31]. L'hyperonymie est l'inverse de l'hyponymie.
    Par exemple : *Pays-Vietnam, recherche d'information-recherche d'information multilingue.*
  - **l'holonymie-méronymie** (ou partie-tout) : l'holonymie est *«la relation hiérarchique existant entre deux concepts ou deux signes linguistiques, dans laquelle le premier est une partie d'un tout que constitue le second.»* [4]. La méronymie est l'inverse de l'holonymie.
    Par exemple : *doigt-main.*

  Dans le contexte conceptuel de notre modèle, un concept est théoriquement un regroupement des termes synonymes. Il n'y a donc pas de relation de synonymie entre concepts. Nous modélisons donc les relations du type hiérarchique : l'hyponymie-hyperonymie et l'holonymie-méronymie dans notre modèle.
  L'extraction de ces relations est effectuée sur les fichiers de données relationnelles de UMLS.
  Par exemple dans le fichier des relations hiérarchiques MRHIER.RRF d'UMLS :

[3] http ://www.granddictionnaire.com
[4] http ://www.granddictionnaire.com

«C0000039|A6326244|4|A6358980|NDFRT||A6916787.A6361870.A6331156.A6328886.A6356446.A6358980|||»
avec *«A6916787.A6361870.A6331156.A6328886.A6356446.A6358980»* l'ensemble des atomes {«A6916787»,«A6361870»,«A6331156»,«A6328886»,«A6356446»,«A6358980»} des concepts qui sont sur le chemin de la racine vers «C0000039» dans une hiérarchie d'une ressource composante d'UMLS. Ces concepts se relient donc avec «C0000039» par des relations hiérarchiques directes ou indirectes. En consultant le fichier des informations sur les concepts MRCONSO.RRF, on peut connaitre les concepts correspondants à ces atomes, par exemple :
C1371271|ENG|P|L4865588|PF|S5546833|N|A6916787||||SRC|RHT|V-NDFRT||
qui indique que «A6916787» est un atome du concepts «C1371271».
On peut donc extraire la relation hiérarchique entre les concepts «C0000039» et «C1371271».
Pour tous les couples de concepts c_i, c_j qui sont liés par une relation directe (c_i, c_j) ou par une relation indirecte, (i.e il existe un chemin $che(c_i, c_j)$) dans l'UMLS, nous mettons simplement un arc (c_i, c_j) dans le RB.

6.4.2 Processus d'inférence sur le RB pour calculer la fonction de correspondance

Comme proposé dans la section 5.3, la procédure de RI dans RIRBRE est le processus d'inférence des probabilités. Chaque document d dans la collection D doit être observé une fois dans la procédure de RI. Pour un document d observé, le processus d'inférence des probabilités est effectué sur le réseau réduit $\mathcal{E}(d, \Psi(\mathcal{N}, \mathcal{A}, \mathcal{P})) = \Psi'(\mathcal{N}', \mathcal{A}', \mathcal{P}')$. Ce processus comprend trois étapes :

L'initialisation de la probabilité

Si d est observé, nous remettons les propabilités des variables dans le réseau comme suit :

$$P(d) = 1$$

L'inférence des probabilités sur les concepts

Comme nous l'avons présenté dans la section 5.3.3, la probabilité d'un concept c_i dans le RB est calculé comme suit :

$$bel(c_i) = \begin{cases} w(c_i, d) \text{ si } c_i \in Indx(d) \\ bel(c_h) \times sim(c_h, c_i), \text{ avec } c_h = argmax_{c_k \in pa(c_i) \cap Indx(d)}(bel(c_k))), \\ \qquad \text{si } c_i \in Indx(q) \text{ et } \exists c_j \in Indx(d) \text{ telque } (c_j, c_i) \in \mathcal{A}' \\ 0 \text{ sinon} \end{cases} \tag{6.1}$$

$sim(c_i, c_k)$ est l'estimation de la similarité entre c_i, c_k. Elle peut être définie par une valeur empirique α dans intervalle $[0, 1]$ ou par une mesure de similarité sémantique.

Pour le deuxième schéma, nous utilisons la mesure de similarité de Leacock (cf. section 3.3.3) pour la fonction de sim par son efficacité. Pour tous $c_i \neq c_j$:

$$sim(c_i, c_j) = -log \frac{minLen(c_i, c_j)}{2 * L} \tag{6.2}$$

avec L est la profondeur de la hiérarchie des concepts d'UMLS. C'est la longueur maximale des chemins entre la racine et les feuilles dans la hiérarchie ; $minLen(c_i, c_j)$ représente la longueur minimale des chemins entre c_i, c_j dans la ressource externe. Dans notre cas, c_i, c_j se trouvant sur la hiérarchie dans UMLS, $minLen(c_i, c_j)$ sera la longueur du chemin unique entre eux. Cependant, nous avons besoin de normaliser cette mesure pour que la valeur de la similarité soit dans l'intervalle $[0, 1]$. Nous définissons donc la normalisation par la valeur maximum possible (quand $minLen(c_i, c_j) = 1$) :

$$sim(c_i, c_j) = \frac{-log \frac{minLen(c_i, c_j)}{2*L}}{-log \frac{1}{2*L}} \tag{6.3}$$

Le calcul de la pertinence

La pertinence est calculée par la fonction de correspondance comme suit :

$$\begin{aligned} \Delta(d, q) = P(q|d) = bel_{wsum}(q) &= \frac{\sum_{c_i \in q} w(c_i, q) \times P(c_i)}{\sum_{c_j \in q} w(c_j, q)} \\ &= \frac{\sum_{c_i \in q} w(c_i, q) \times bel(c_i)}{\sum_{c_j \in q} w(c_j, q)} \end{aligned} \tag{6.4}$$

Sachant que l'indexation conceptuelle donne le meilleur résulat avec le modèle VSM (cf. section précédente), nous voulons utiliser notre modèle pour simuler le modèle VSM. Il s'agit du cas où il n'y a pas de relations sémantiques entre concepts. Cela est le «baseline» dans nos expérimentations pour voir le comportement quand les relations sémantiques

entre concepts seront intégrées. Un baseline est le résultat d'une évaluation, utilisé comme une référence pour comparer avec des autres évaluations. Comme mentionné dans la section 5.4, pour cette simulation, nous définissons :

$$w(c_i, q) = \frac{w'(c_i, q)}{\sqrt{\sum_{j \in q} w'(c_j, q)^2}} \tag{6.5}$$

et :

$$bel(c_i) = w(c_i, d) = \frac{w'(c_i, d)}{\sqrt{\sum_{h \in Indx(d)} w'(c_h, d)^2}} \tag{6.6}$$

w' est la pondération $tf.idf$ des concepts dans la requête ou dans les documents.

6.4.3 L'algorithme de l'inférence des probabilités

L'algorithme de l'inférence des probabilités est comme suit :
Pour chaque document d dans la collection D :
- Initialiser $P(d) = 1$
- Remettre la probabilité de tous les autres nœuds dans le RB à zéro.
- Pour chaque concept c_i de d :
  - Calculer $P(c_i)$
  - Pour chaque nœud enfant de c_i :
    Si c'est un concept c_j de la requête :
    Calculer sa probabilité postérieure $P(c_j)$.

  - Fin Pour
- Fin Pour
- Calculer la probabilité postérieure de la requête $P(q)$ ou la $RSV(d, q)$.

Fin Pour

6.4.4 Les évaluations du modèle RIRBRE

Notre modèle basé sur un réseau Bayésien est validé par des expérimentations sur la collection CLEF image médicale 2006, 2007.

Nous avons évalué notre modèle avec deux cas : avec ou sans relations sémantiques entre les concepts dans le RB. Ces expérimentations nous permettent de valider les apports de ces relations dans la performance de la recherche. Les descriptions des évaluations sont comme suit :

– Le run RB1 : Dans cette évaluation du modèle, il n'y a pas de relations sémantiques entre concepts dans notre RB. Il est aussi le baseline de l'expérimentation sur ce modèle. Le RB1 correspond au cas où dans le RB $\Psi(\mathcal{N}, \mathcal{A}, \mathcal{P})$:

$$\mathcal{A} = A_{CQ} \cup A_{CD}$$

– Les runs RB2 à RB5 : sont les évaluations du modèle avec les relations sémantiques intégrés. Dans ce cas :

$$\mathcal{A} = A_{CQ} \cup A_{CD} \cup A_{CC}$$

Les relations intégrées sont les relations hiérarchiques (l'hyponymie-hyperonymie et l'holonymie-méronymie), qui correspondent aux relations nommées PAR-CHD (parent-enfant) et BR-RN (general-restreint) dans UMLS. La relation Is-a est incluse dans le type PAR-CHD. Nous intégrons aussi des liens indirects déduits à partir de la hiérarchie d'UMLS.

Dans les runs RB2 à RB4, la similarité entre deux concepts α est prédéfinie $(0.1, 0.2, 0.3)$. Ces valeurs permettent d'observer l'effet de la prise en compte de poids des relations dans la correspondance. Alors que dans le run RB5, cette similarité est mesurée par la méthode de Leacock qui cacule plus précisement sur le context des relations.

Les résultats de cette expérimentation sont présentés par la suite.

6.4.5 Résultats

Les résultats de l'expérimentation sur ImageCLEFMed 2007 sont présentés dans le tableau 6.7.

Dans les expérimentations, nous constatons toujours une amélioration de MAP du modèle Bayésien avec les relations hiérarchiques par rapport au baseline (RB1). Le run RB5 qui utilise la mesure de similarité sémantique de Leacock donne le meilleur résultat global. Cela valide donc l'utilisation de la mesure de similarité sémantique entre concepts de Leacock.

Ces résultats correspondent à nos attentes, l'intégration des relations sémantiques dans le modèle basé sur un réseau Bayésien augmente plus ou moins la performance de recherche. Cette amélioration est assez stable. De plus, la précision sur les 5 premiers documents retrouvés (Précision@5) n'est pas dégradée. Cela est un point positif du modèle. Dans la méthode d'expansion de la requête, l'ajout des concepts est risqué pour la performance de la recherche parce que le risque d'ajouter des concepts qui dégradent la performance et la précision est grand. Dans notre modèle, nous gardons les concepts tel quels dans la collection d'origine, et n'ajoutons que les liens entre eux. L'amélioration de la performance dans notre modèle dépend aussi de la possibilité de trouver des relations

TAB. 6.7 – Résultats sur ImageCLEFmed 2007 avec le modèle RIRBRE

Nom du test	MAP	Précision@5
RB1 (Baseline)	0.229	0.333
RB2	0.231	0.333
RB3	0.232	0.333
RB4	0.230	0.333
RB5	**0.234**	0.333

TAB. 6.8 – Les meilleurs résultats sur ImageCLEFmed 2006 et 2007 avec le modèle RIRBRE

Nom du test	MAP (ImageCMEFMed2006)	MAP (ImageCMEFMed2007)
RB1 (Baseline)	0.204	0.229
RB5	**0.206**	**0.234**

dans les concepts de la requête et concepts dans les documents. Dans le cas des requêtes très courtes comme notre collection de test, la possibilité de trouver des relations est relativement plus faible que dans des longues requêtes. Par conséquent, l'amélioration de la performance est donc assez faible.

Les figures 6.8 et 6.9 illustrent les courbes de rappel-précision des runs. Sur ces figures, on peut constater l'amélioration de la précision sur tous les points de rappel. Les meilleurs résultats obtenus globalement sur les deux collections ImageCLEFMed 2006 et ImageCLEFMed 2007 sont présentés dans le tableau 6.8. Nous obtenons toujours les meilleurs résultats avec RB5 dans ces deux collections.

Nous avons étudié les résultats de plus près : requête par requête ; et examiné les documents mieux et moins bien classés. Pour cela, nous avons examiné les valeurs de MAP de la recherche avec les runs RB1, RB2, RB3, RB4, RB5 de chaque requête sur 3 langues dans ImageCLEFMed 2007 (tableau 6.9 et 6.10).

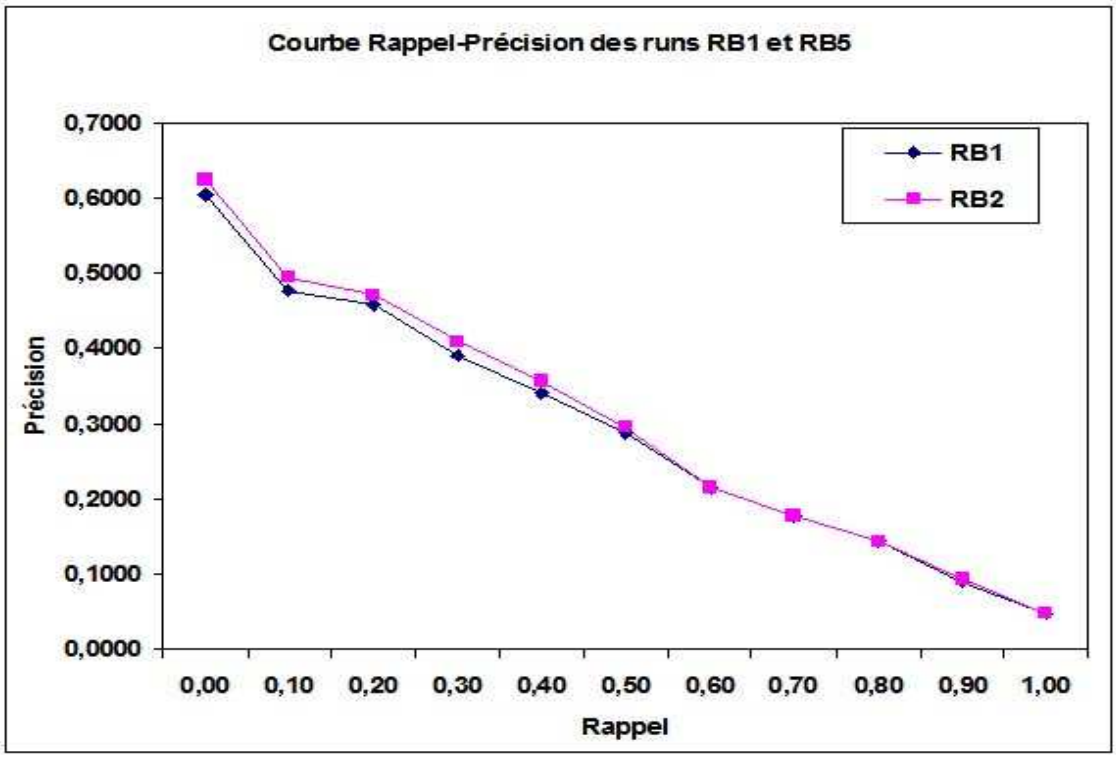
Voyons l'exemple de la requête :

Show me microscopic pathologies[C0030664] of cases[C0868928] with chronic myelogenous leukemia[C0023470].

Les termes soulignés sont suivis par les concepts correspondants.

Parmi les premiers documents réellements pertinents, on peut trouver ce document :

BLOOD-RES : Blood[C0005767] : Acute Myelogenous Leukemia[C0023467] : Micro blood[C0005768] film[C0086296] myeloblasts and some more mature forms



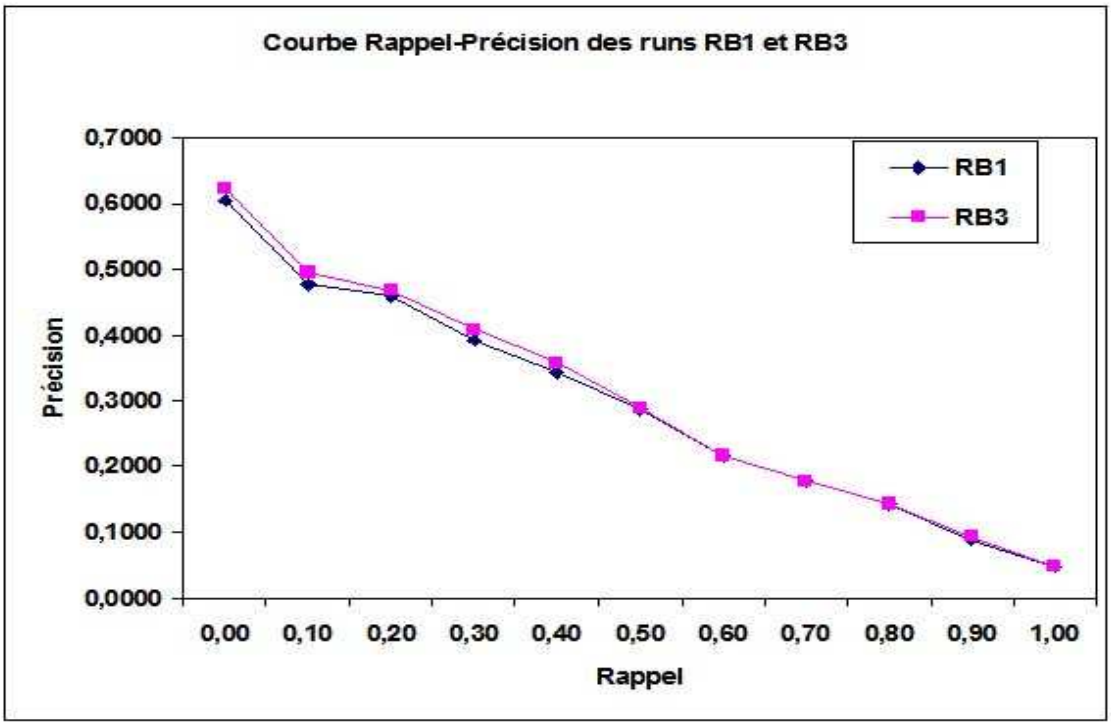


FIG. 6.8 – Courbe de rappel-précision des runs RB2, RB3 par rapport à RB1

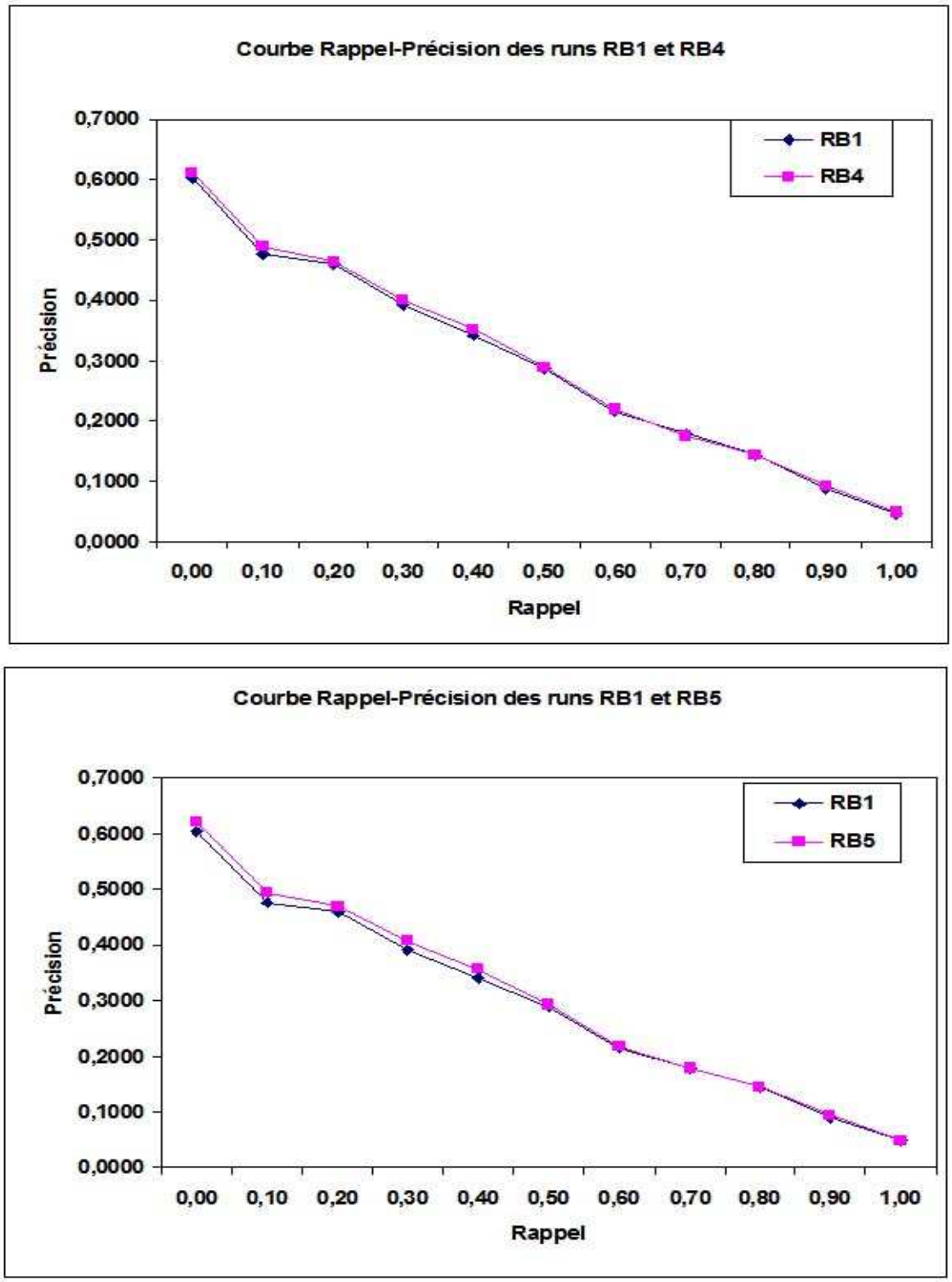


FIG. 6.9 – Courbe de rappel-précision des runs RB4, RB5 par rapport à RB1

TAB. 6.9 – Statistique sur les résultats

	RB1	RB2	RB3	RB4	RB5
Nombre de requêtes améliorées	Baseline	10	9	8	11
Nombre de requêtes baissées	Baseline	2	7	8	5
Nombre de documents pertinents retrouvés / total documents pertinents	1503/2654	1503/2654	1509/2654	1485/2654	1511/2654

On constate que ce document ne partage pas les mêmes concepts avec la requête ; avec le modèle à base d'intersection (comme modèle VSM), ce document n'est pas retrouvé. Cependant, dans notre modèle, ce document est retrouvé parce qu'il y a la relation du type PAR-CHD entre
«chronic myelogenous leukemia[C0023470]» et « Acute Myelogenous Leukemia[C0023467]».

Des résultats et exemples plus détaillés peuvent êtres trouvés dans l'Annexe E.

6.5 Conclusion

Dans ce chapitre, nous avons effectué des expérimentations de RI multilingue dans le domaine médical. La collection est ImageCLEFMed et la ressource utilisée est le méta thésaurus UMLS.

L'indexation conceptuelle pour la RI monolingue ou multilingue avec les modèles classiques a montré des améliorations par rapport à l'indexation à base de termes dans la plupart des test. Le modèle VSM a montré le meilleur résultat global avec l'indexation conceptuelle. Ces expérimentations montre qu'une indexation conceptuelle nécessite en premier une bonne extraction de concept. Dans nos expérimentations, cette tâche est de bonne qualité que sur les textes en Anglais uniquement parce que la ressource externe n'a pas assez de termes en Français et en Allemand. Pour cette raison, concernant la comparaison de l'indexation conceptuelle et de l'indexation par terme, seuls les résultats sur la langue Anglaise sont véritablement significatifs. Ces résultats, comme illustrés dans le tableau 6.4, montrent alors l'avantage de l'indexation conceptuelle.

Pour le modèle proposé à base de RB des concepts et des relations sémantiques, l'expérimentation montre que la prise en compte des relations sémantiques dans le cadre du réseau Bayésien améliore plus ou moins la performance de recherche. Dans le contexte des requêtes courtes, les relations retrouvées sont limitées, l'amélioration de la performance est aussi légère. L'amélioration est supposée plus significative quand il y a plus de relations retrouvées.

TAB. 6.10 – Les résultats requête par requête

Requête	RB1(VSM)	RB2	RB3	RB4	RB5
1	0	0	0	0	0
2	0,0639	0,0637	0,0606	0,0446	0,063
3	0,3675	0,3917	0,4025	0,3981	0,3983
4	0,0623	0,0625	0,0629	0,0633	0,0629
5	0,0238	0,0238	0,0238	0,0238	0,0238
6	0,0798	0,0806	0,0807	0,0804	0,0807
7	0,0011	0,0011	0,001	0,0009	0,001
8	1	1	1	1	1
9	0,0002	0,0002	0,0002	0,0002	0,0002
10	0,0169	0,0169	0,0169	0,0169	0,0169
11	0,0938	0,0938	0,0938	0,0938	0,0938
12	0,514	0,514	0,5138	0,5078	0,5138
13	0	0	0	0	0
14	0,1242	0,1244	0,1235	0,1108	0,1246
15	0,0578	0,0578	0,0576	0,0556	0,0576
16	0,1043	0,1043	0,1043	0,1043	0,1043
17	0,2829	0,283	0,2887	0,2816	0,2887
18	0,088	0,0886	0,0907	0,0887	0,0886
19	0,0217	0,0217	0,0217	0,0207	0,0217
20	0,2607	0,2607	0,2607	0,2607	0,2607
21	0,1142	0,1142	0,114	0,1138	0,114
22	0,747	0,747	0,7451	0,7451	0,7451
23	0,1838	0,1888	0,1898	0,1977	0,1899
24	0,3042	0,2988	0,2963	0,2921	0,2965
25	0,4411	0,4411	0,4411	0,4411	0,4411
26	0,2511	0,2511	0,2511	0,2511	0,2511
27	0,6655	0,6655	0,6655	0,6655	0,6655
28	0	0,0024	0,0056	0,0128	0,0056
29	0,4922	0,4998	0,505	0,5086	0,5047
30	0,5232	0,5254	0,5279	0,5294	0,5278
Moyenne	0,2295	0,2308	0,2315	0,2303	0,2314

En conclusion, l'utilisation d'UMLS comme ressource externe a donc montré d'une part les intérêts dans l'indexation conceptuelle pour la RI monolingue et multilingue avec des modèles classiques, surtout dans le modèle VSM. D'autre part, l'intégration des relations sémantiques d'UMLS dans le cadre d'un modèle de RI basé sur un réseau Bayésien a validé son apport dans l'amélioration de la performance de recherche.

Dans le chapitre suivant, nous introduisons des extensions de l'indexation conceptuelle à des documents et requête structurés et multi-médias.

Chapitre 7

Extension à des documents et des requêtes structurés et multi-médias

7.1 Introduction

Dans ce chapitre, nous abordons des extensions de l'indexation conceptuelle, d'une part pour des requêtes et des documents structurés et d'autre part, pour des requêtes et des documents multi-médias. Dans le domaine médical, nous montrons l'intérêt d'une indexation conceptuelle et l'usage d'une ressource externe de grande taille. Nous validons cette extension par des expérimentations sur les collections de la campagne d'évaluation ImageCLEFMed 2006 et 2007. Nous détaillons ces extensions dans les sections suivantes.

7.2 Extension à des documents et à des requêtes structurés

Nous avons constaté que les requêtes précises dans un domaine technique (ex : médical) possèdent une structure : il s'agit en fait des différents «éléments» qui composent une requête. Par exemple la requête «radiographie d'une fracture du fémur», fait clairement référence à la *modalité* de l'image recherchée (radiographie), à une *anatomie* précise (l'os du fémur) et à une *pathologie* (une fracture). Ces différents éléments de requête font référence à des parties très distinctes d'une ontologie qui, dans notre exemple, est celle de la médecine. Cela semble correspondre aux premiers niveaux de l'arbre d'une ontologie de la médecine. Des études similaires sur les catégories hiérarchiques dans une terminologie ont été faites [25], [26]. En examinant UMLS, même si ce n'est pas une véritable ontologie, nous avons retrouvé un découpage de tous les concepts en grandes catégories. Elles correspondent aux éléments qui structurent les requêtes. Nous avons fait ce constat

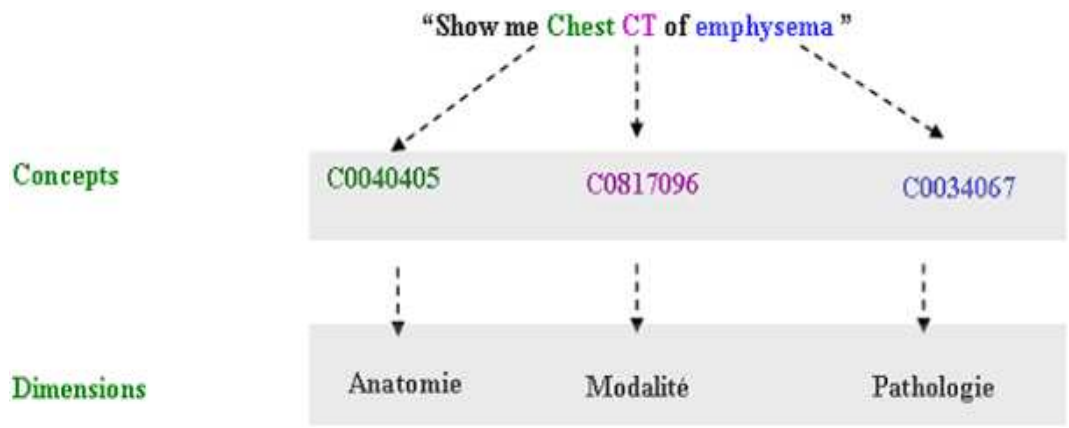


FIG. 7.1 – Exemple d'une requête et sa structuration par des dimensions

sur toutes les requêtes de la collection CLEF médicale (2006 et 2007). Nous faisons donc l'hypothèse que les requêtes précises dans un domaine technique sont structurées en référence à la hiérarchie d'une ontologie de ce domaine. Cette hypothèse a également été faite dans le travail de Radhouani [28], [22], [62]. Nous appelons ces éléments de structuration des requêtes *les dimensions* de la requête.

De manière pratique, nous avons décidé d'utiliser les groupes sémantiques d'UMLS (cf. annexe A) pour exprimer ces dimensions. Nous nous sommes intéressés à cette notion de dimension car nous avons constaté une nette amélioration des résultats dès lors que notre système tient compte de ces dimensions dans les réponses. Concrètement, nous avons mis en place après la première étape de l'indexation conceptuelle, un reclassement des réponses suivant les dimensions de la requête. Ainsi, les documents pertinents ne doivent pas partager seulement des concepts, mais aussi les dimensions de la requête.

Par exemple nous avons résolu la requête *«Show me x-ray images with fractures of femur»* de la manière suivante. Un document pertinent pour cette requête doit contenir :

- une dimension **anatomie**. Cela concerne la structure ou les organes du corps. Par exemple : le fémur, la tête, le rein,...
- une dimension **pathologie**. Cela concerne des maladies et des effets qu'elles provoquent. Par exemple : fracture, lésion, cancer,...
- une dimension **modalité**. Cela concerne des modes d'examen sur les organes afin de diagnostiquer les maladies. Par exemple : Rayon-X.

Nous avons constaté que les documents qui répondent, même fortement, à une partie seulement des dimensions ne sont pas pertinents.

La figure 7.1 illustre un exemple de la structuration des dimensions d'une requête.

Notons que cette structuration n'est possible que si les concepts ont été correctement

identifiés et si la ressource utilisée possède une hiérarchie ou une structuration qui nous permette d'organiser l'ensemble des concepts en dimensions. Notons également que l'indexation se réalise sur deux niveaux d'abstraction : les concepts et les dimensions. Dans la suite, nous montrons comment les prendre en compte dans la fonction de correspondance.

7.2.1 Fonction de reclassement

Nous avons testé l'hypothèse suivante [63] : «*les documents pertinents doivent inclure au moins une des trois dimensions si elles existent dans la requête*».

La fonction de reclassement correspondante force une intersection non vide des dimensions de la requête avec celle des documents :

$$Inclusion(d,q) = \Delta(d,q) \times \beta$$

avec :

$$\beta = \begin{cases} 0 & \text{si } (DM(q) \neq \emptyset)) \text{ et } ((DM(d) \cap DM(q) = \emptyset) \\ 1 & \text{sinon} \end{cases} \tag{7.1}$$

$DM(d)$ est l'ensemble des dimensions qui sont présentes dans le document d, similairement $DM(q)$ est l'ensemble des dimensions qui sont présentes dans la requête q. Cette fonction supprime tous les documents qui n'ont pas au moins une dimension en commun avec la requête.

Nous avons expérimenté une autre fonction de reclassement des documents, appelée fonction d'Intersection. Cette fonction favorise l'importance de l'intersection des dimensions entre documents et requête. Elle est basée sur l'hypothèse suivante :

La valeur de pertinence est proportionelle au nombre de dimensions en commun entre le document et la requête.

Cette fonction est définie par :

$$Intersection(d,q) = \Delta(d,q) \times \eta$$

avec :

$$\eta = \begin{cases} |DM(d) \cap DM(q)| & \text{si } (DM(q) \neq \emptyset) \\ 1 & \text{sinon} \end{cases} \tag{7.2}$$

Ce classement est moins «brutal» que le précédent car il réorganise les résultats sans supprimer de documents.

TAB. 7.1 – Fonctions de reclassement appliquées aux modèles de pondération sur ImageCLEFMed 2006

Modèle	Baseline(3 Langues)	+F Inclusion		+F Intersection	
	MAP	MAP	Amélioration	MAP	Amélioration
FREQ	0.004	0.033	+725%	0.017	+325%
VSM	0.204	0.239	+17%	**0.264**	+29%
DFR	0.097	0.135	+39%	0.177	+82%
BM25	0.098	0.140	+43%	0.180	+84%
RIRBRE(RB5)	0.206	0.226	+10%	**0.270**	+31%

7.2.2 Validation

Nous avons comparé ces deux fonctions de reclassement en les appliquant à une indexation conceptuelle avec des pondérations classiques ainsi que sur notre modèle RIRBRE.

Pour les pondérations classiques, nous avons évalué les résultats de l'indexation conceptuelle seule (Baseline), puis avec application d'une fonction de reclassement (cf table 7.1).

L'application de la fonction de reclassement produit une amélioration significative dans le MAP pour tous les modèles de pondération. La colonne baseline dans le tableau 7.1 correspond à une indexation conceptuelle sans fonction de reclassement. La fonction d'Intersection donne une meilleure amélioration de MAP que la fonction d'Inclusion. Cela est probablement dû au fait qu'elle prend en compte la taille de l'intersection. Pour les pondérations classiques, le meilleur résultat global a été obtenu avec le modèle vectoriel (VSM). Cette indexation a donné le meilleur résultat officiel de la campagne de ImageCLEFMed2006. Nous avons donc montré qu'une indexation conceptuelle devance toutes les autres méthodes non conceptuelles, c'est à dire utilisant des termes comme index.

Nous avons obtenu pour le modèle RIRBRE, une amélioration supplémentaire. Nous avons recommencé les expérimentations sur la collection ImageCLEFMed 2007. Les résultats sont présentés dans le tableau 7.2. On peut constater que le reclassement par Intersection appliqué au modèle RIRBRE, améliore les résultats dans tous les runs. L'usage de la pondération basée sur le réseau Bayésien donne systématiquement un meilleur résultat que la pondération du modèle vectoriel (VSM). Le meilleur résultat est obtenu avec notre modèle et la similarité sémantique de Leacock (RB2, RB3, RB5).

TAB. 7.2 – Fonction de reclassement par Intersection avec ImageCLEFMed 2007

Run	Baseline	+F Intersection	
	MAP	MAP	Amélioration
RB1=VSM	0.229	0.268	17%
RB2	0.230	**0.278**	21%
RB3	0.231	**0.278**	20%
RB4	0.230	0.275	19%
RB5	**0.234**	**0.278**	19%

7.3 Extension à des documents et à une requête multi-médias

La collection de documents ImageCLEFmed est en fait une collection multi-média. En effet, les documents sont des compte-rendus médicaux associé à des images. Nous avons travaillé dans une équipe qui a analysé les images et produit une indexation. Nous avons expérimenté une combinaison de notre indexation conceptuelle des textes avec l'indexation des images. Avec la ressource UMLS, nous avons la capacité d'indexer les images et les textes par un espace commun de concepts. Nous avons donc étudié la fusion des résultats de recherche sur chaque modalité (texte ou image) [42], [21]. Nous avons expérimenté une fusion à base de combinaison linéaire normalisée. La valeur de pertinence finale (RSV) entre une requête multi-média $q = (q_I, q_T)$ (q_I : images, q_T : texte) et un document $d = (d_I, d_T)$ est calculée comme suit :

$$RSV(q,d) = \frac{\epsilon RSV_I(q_I, d_I)}{\max\limits_{z \in \mathcal{D}_I} RSV_I(d_I, z)} + \frac{(1-\epsilon) RSV_T(q_T, d_T)}{\max\limits_{z \in \mathcal{D}_T} RSV_T(q_T, z)}$$

où RSV_I est le maximum de la similarité visuelle entre toutes les images de q_I et toutes les images de d_I, RSV_T est la RSV textuelle. $\mathcal{D}_I$ dénote la base d'images, et $\mathcal{D}_T$ dénote la base de texte. Le facteur ϵ permet de contrôler la fusion entre la similarité textuelle et visuelle.

La figure 7.2 illustre le schéma général de cette fusion pour la RI multi-modale texte-image sur la collection ImageCLEFMed 2006. Une base associant des images et des concepts d'UMLS a été construite manuellement pour réaliser l'indexation conceptuelle des images après un apprentisage sur cette base. Notre équipe a ainsi obtenu un sytsème de RI pour les images. Nous avons fusioné le résultat de ce système avec notre système d'indexation textuelle.

L'indexation conceptuelle des images a été réalisée par C. Lacoste [42].

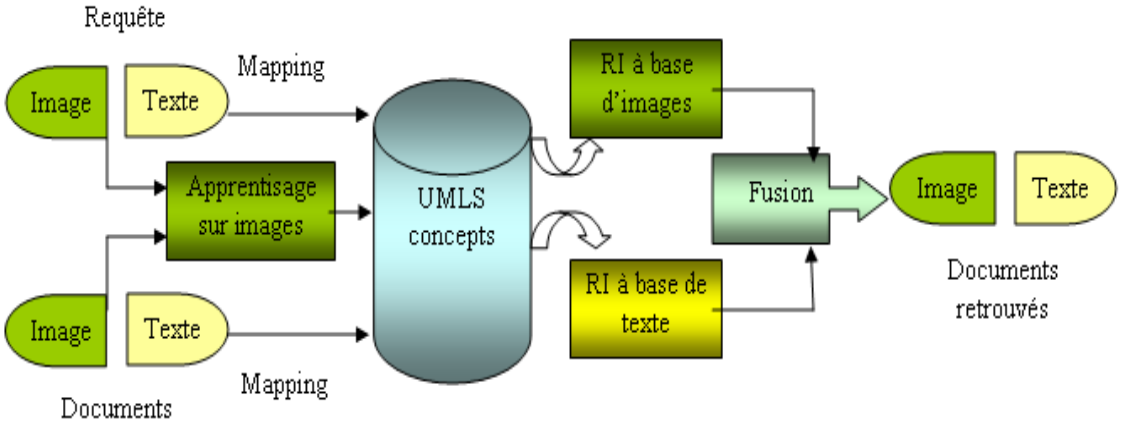


FIG. 7.2 – Schéma général de la recherche d'information multi-modale texte-image

TAB. 7.3 – Résultats de la combinaison de l'indexation Texte-Image pour la RI multi-modalité

Indexation		ϵ dans la combinaison de l'indexation Texte-Image pour la RI multi-média										
Texte	Image	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
VSM		0.264	0.297	0.306	0.317	0.322	0.326	**0.328**	0.318	0.293	0.249	0.064
BM25		0.180	0.206	0.216	0.228	0.238	**0.243**	0.241	0.233	0.215	0.182	0.064
DFR		0.177	0.220	0.227	0.229	0.234	**0.241**	**0.241**	0.237	0.217	0.184	0.064
FREQ		0.017	0.038	0.039	0.040	0.042	0.046	0.052	0.056	0.066	**0.073**	0.064

Le tableau 7.3 montre les résultats de la fonction de fusion de l'indexation sur les deux modalités. Nous avons testé la fusion avec différents modèles de pondération de concepts. La valeur $\epsilon = 0$ correspond à l'indexation purement textuelle alors que la valeur $\epsilon = 1$ correspond à une indexation purement visuelle. Les valeurs intermédiaires représentent différents taux de fusion entre l'image et le texte. Les résultats sur les textes uniquement sont bien meilleurs que l'indexation uniquement à base d'images.

La figure 7.3 illustre les courbes des résultats de cette fusion. Ces courbes montrent que toute combinaision du texte et de l'image donne un meilleur résultat qu'une indexation mono-média. Ce résultat montre donc l'intérêt d'une fusion pour la recherche multi-modalité par rapport à celle mono-modalité.

7.4 Conclusion

Dans cette partie du travail, nous avons présenté une extension à une requête et des documents structurés et multi-médias.

Dans le contexte des requêtes structurées par des dimensions du domaine médical,

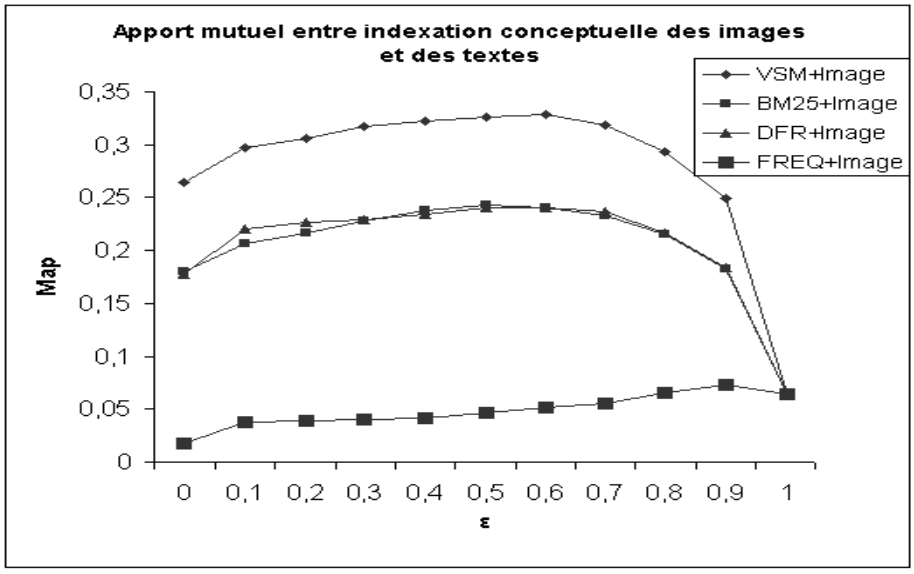


FIG. 7.3 – Apports mutuels entre indexation conceptuelle des images et des textes dans la fusion

nous avons proposé une fonction de reclassement sur la valeur de pertinence en utilisant les connaissances de ces dimensions dans UMLS. Cette méthode améliore les résultats dans les modèles classiques (classé comme meilleur résultat pour la recherche à base de texte avec modèle VSM dans la campagne ImageCLEFMed2006) ainsi que dans le modèle RIRBRE. L'utilisation de la ressource externe UMLS nous permet donc de structurer et de prendre en compte ces dimensions afin d'améliorer la performance de recherche.

Pour la RI multi-modale (texte + image), les images et textes peuvent être indexés par des concepts issues de la ressource externe. Nous avons proposé une méthode de fusion des résultats de la recherche à base de texte et à base d'image afin d'obtenir un meilleur résultat que la recherche seulement basée sur le texte ou l'image. Ce résultat a été classé aussi en tête dans la campagne ImageCLEFMed2006 dans la catégorie recherche multi-modale.

Finalement, ces extensions montrent que les ressources externes permettent de structurer les données à plusieurs niveaux d'abstraction : concept, dimension. De plus, elles offrent une vision unifiée des données, que ce soit les images ou le texte. De ce fait, le contenu sémantique des documents est aussi mieux capturé et géré.

Troisième partie

Conclusions et perspectives

Chapitre 8

Conclusions et perspectives

Sommaire

8.1 Conclusion

Notre thèse s'intègre dans le contexte de la Recherche d'Information Multilingue (RIM), utilisant un modèle à base de réseau Bayésien sur des ressources externes. La RI à base d'intersection des termes atteint plusieurs limites. La première est la limite de l'indexation basée sur des termes dont l'index n'est pas assez précis. Cela est à cause des variations morphologiques, lexicales ou syntaxiques des termes dans la langue naturelle. La seconde concerne la disparité entre documents et requête, c'est-à-dire qu'ils ne partagent pas les mêmes termes, spécialement dans le cas où la requête contient des termes plus généraux que ceux des documents.

L'objectif général de la thèse est alors de résoudre ces problèmes et d'améliorer la performance de la recherche. La contribution principale de notre travail réside dans un modèle de RI qui est capable d'améliorer la qualité des index ainsi que la correspondance entre les documents et la requête.

Pour résoudre le premier problème, nous proposons d'utiliser des concepts au lieu de termes à l'aide d'une ressource externe. L'indexation conceptuelle a la capacité de résoudre le problème des variations morphologiques, lexicales ou syntaxiques des termes. De plus, la conceptualisation unifie possiblement les termes dans les langues différentes en une forme unique, elle fait tomber donc la barrière de la langue.

Pour résoudre le deuxième problème, nous proposons un modèle basé sur un réseau Bayésien pour prendre en compte les liens sémantiques entre les concepts de la requête et ceux des documents. Ce modèle est capable de trouver des documents qui ont des concepts sémantiquement liés avec les concepts de la requête. Nous introduisons l'intégration de la mesure de similarité sémantique entre les concepts sémantiquement liés dans ce modèle.

Pour valider nos propositions, nous avons expérimenté sur la collection CLEF images médicales 2006, 2007 pour la RIM dans le domaine médical. La ressource utilisée dans notre application est l'UMLS. Ces expérimentations mettent en avant l'intérêt de nos propositions. Nous constatons des résultats positifs en utilisant l'indexation conceptuelle par rapport à l'indexation par termes. De même, l'intégration des relations sémantiques dans un modèle de recherche d'information à base de réseau Bayésien apporte des améliorations.

Pour aller plus loin dans l'indexation conceptuelle, nous avons étudié deux extensions possibles dans le but d'améliorer la performance de la recherche. La première extension consiste à appliquer une fonction de reclassement sur les valeurs de pertinence en structurant chaque document et la requête à partir de dimensions du domaine médical. Une seconde extension consiste à utiliser la RI multi-modalités. Pour cela, nous proposons de fusionner les résultats de la recherche a partir de textes avec les résultats de la recherche

à partir du contenu des images. La mise en place de ces extensions est d'autant plus intéressante qu'elles obtiennent les meilleurs résultats de leur catégorie dans la campagne d'évaluation ImageCLEFMed 2006 avec le modèle VSM, et une nouvelle amélioration avec le modèle RIRBRE.

8.2 Perspectives

Notre travail de thèse ouvre des perspectives intéressantes à différents niveaux.

8.2.1 Court terme

Dans le but d'améliorer et de valider plus en détail nos travaux, il semble utile et intéressant d'orienter la suite de cette étude vers les perspectives suivantes.

Il faudrait étudier la prise en compte du score de confiance dans l'identification des concepts dans le modèle. Dans l'état actuel, tous les concepts identifiés sont considérés de manière identique. Un tel score permettrait de mieux identifier les concepts les plus pertinents et ainsi d'améliorer encore la performance de la recherche.

Une évaluation sur d'autres collections de test permettrait de compléter la validation du modèle. Comme les requêtes dans la collection CLEF Images médicales sont courtes, la capacité de retrouver les relations sémantiques entre concepts du document et ceux de la requête est assez limité. L'apport de la prise en compte de ces relations dans notre modèle de RI n'est pas encore très significatif. Il serait donc intéressant d'expérimenter sur d'autres collections de test.

8.2.2 Long terme

Nous proposons aussi les pistes de recherche suivantes qui permettront de généraliser encore plus ce modèle, d'étendre ses possibilités et d'augmenter ses performances.

Nous désirons intégrer l'étude sur le typage des relations sémantiques entre concepts et leurs similarités correspondantes. Dans notre modèle, le typage des relations sémantiques et leurs pondérations appropriées ne sont pas encore bien étudiés. Un tel modèle sera plus global et permettra de prendre en compte tous les types de relations sémantiques avec leurs pondérations correspondantes.

Nous envisageons aussi d'étendre le modèle avec d'autres schémas de pondération des concepts. Notre modèle a montré la capacité de simuler le modèle VSM. Nous pensons qu'il est intéressant d'étudier d'autres schémas de pondération et de calculer les probabilités afin de comparer et d'améliorer encore la performance de recherche.

Enfin, nous souhaitons étendre le modèle proposé pour un SRI multi-média, qui est capable d'indexer et d'interroger des données multi-médias dans un cadre unique.

Annexe A. UMLS

Le Meta thésaurus d'UMLS

Le Méta thesaurus est la partie la plus importante par sa taille et son contenu. Il contient plus de 5.5 millions de termes en 17 langues et 1.1 million de concepts. Il est maintenu par les spécialistes de NLM avec deux mises à jour par année. Le principe de sa construction est le groupement à différents niveaux d'abstraction des unités des textes venant des sources terminologiques.

Les relations dans le Méta thésaurus viennent des relations dans les sources originales et de l'ajout de nouvelles relations fait manuellement par des spécialistes. Ces relations dans le Méta thésaurus peuvent être divisées en 2 groupes :

- Les relations Intra-source : Ce sont des relations entre atomes de même source de vocabulaire. Ces relations existent donc aussi entre concepts créés à partir de ces atomes. Ces relations comprennent : relations hiérarchiques (relations de parents immédiats, d'enfants immédiats, et d'enfants de même parents) ou non-hiérarchiques (association, cause-effet,...) et relations statistiques (co-occurrence).
- Les relations Inter-source : Ce sont des relations entre concepts venant de différentes sources terminologiques. Il y a les relations de synonymie ou non-synonymie dans ce groupe.

Il y a 12 types de relation sémantique. Chaque relation peut être précisée par plusieurs noms de relation. La relation de synonymie entre termes est représentée implicitement dans la structure des concepts (un concept est le groupement des termes synonymes) ou explicitement via les relations entre concepts.

AQ : admis qualificatif (Allowed qualifier)

CHD : relation de parent-enfant dans une source vocabulaire de Méta thésaurus (has

child relationship in a Metathesaurus source vocabulary)

PAR : relation de enfant-parent dans une source vocabulaire de Méta thésaurus (has parent relationship in a Metathesaurus source vocabulary)

QB : peut être qualifié par (can be qualified by)

RB : relation d'hypéronymie (has a broader relationship)

RL : similaire ou ressemblant, synonymie (similar, "alike", synonym)

RN : relation d'hyponymie (has a narrower relationship)

RO : relation qui est différente de synonymie, RN, RB.

RQ : lié ou possible synonyme (related and possibly synonymous)

RU : lié, mais non spécifié (Related, unspecified)

SIB : enfants de même parents dans une source vocabulaire de Méta thésaurus (has sibling relationship in a Metathesaurus source vocabulary)

SY : synonymie vérifiée par une ressource (source asserted synonymy)

Le réseau sémantique d'UMLS

Afin de fournir une catégorisation cohérente de tous les concepts dans le Méta thésaurus, le *réseau sémantique (Semantic Network)* est ajouté à cette structure et construit par les experts de NLM. Ils ont défini 135 *types sémantiques* qui représentent une catégorisation de tous les concepts, et les relations sémantiques entre ces types sémantiques. Cette structure est un ajout dû à la fusion des thésaurus. Elle permet de "couvrir" cette fusion d'une classification hiérarchique [1].

Les types sémantiques sont eux même groupés en 15 *groupes sémantiques*. Par exemple, les types sémantiques T017(*Anatomical Structure*), T029(*Body Location or Region*), T023(*Body Part, Organ, or Organ Componen*) sont groupés par le groupe sémantique ANAT(Anatomy) comme suit :

ANAT|Anatomy|T017|Anatomical Structure

ANAT|Anatomy|T029|Body Location or Region

ANAT|Anatomy|T023|Body Part, Organ, or Organ Component

Les groupes sémantiques comprendent :

ACTI : Activities and Behaviors

ANAT : Anatomy

CHEM : Chemicals and Drugs

CONC : Concepts and Ideas

DEVI : Devices

[1] En réalité il s'agit d'un treillis avec 54 types de relations différents.

DISO : Disorders
GENE : Genes and Molecular Sequences
GEOG : Geographic Areas
LIVB : Living Beings
OBJC : Objects
OCCU : Occupations
ORGA : Organizations
PHEN : Phenomena
PHYS : Physiology
PROC : Procedures

Tous ces groupements ou catégorisations dans UMLS sont réalisés manuellement par des spécialistes de NLM.

Specialist lexicon d'UMLS

Le lexique spécialiste dans UMLS offre les informations lexicales pour des systèmes spécialisés en TALN (Traitement Naturel du Langage Naturel), par exemple Metamap pour extraire de concepts. C'est un lexique Anglais général qui inclut aussi beaucoup de termes dans le domaine biomédical. Chaque entrée d'un terme comprend des informations syntaxiques, morphologiques et orthographiques. Voici un exemple :

base=anaesthetic
spelling_variant=anesthetic
entry=E0008769
cat=noun
variants=reg

Base de données de Méta thésaurus UMLS

En pratique, la base de données du Méta thésaurus est un ensemble de fichiers relationnels.

- Définition de concepts, noms de Concept et leurs sources :
  MRCONSO.RRF
- Attributs
  MRSAT.RRF, MRDEF.RRF, MRSTY.RRF, MRHIST.RRF
- Relations
  MRREL.RRF, MRCOC.RRF, MRCXT.RRF, MRHIER.RRF, MRMAP.RRF, MRSMAP.RRF

- Données sur Metathesaurus
  MRFILES.RRF, MRCOLS.RRF, MRDOC.RRF, MRRANK.RRF, MRSAB.RRF, AMBIGLUI.RRF, AMBIGSUI.RRF, CHANGE/
  MERGEDCUI.RRF, CHANGE/MERGEDLUI.RRF, CHANGE/DELETEDCUI.RRF, CHANGE/DELETEDLUI.RRF, CHANGE/
  DELETEDSUI.RRF, MRCUI.RRF
- Index
  MRXW_BAQ.RRF, MRXW_DAN.RRF, MRXW_DUT.RRF, MRXW_ENG.RRF, MRXW_FIN.RRF, MRXW_FRE.RRF, MRXW_GER.RRF, MRXW_HEB.RRF, MRXW_HUN.RRF, MRXW_ITA.RRF, MRXW_NOR.RRF, MRXW_POR.RRF, MRXW_RUS.RRF, MRXW_SPA.RRF, MRXW_SWE.RRF, MRXNW_ENG.RRF, MRXNS_ENG.RRF

Exemple de définition de concepts dans MRCONSO :

C0016700|ENG|P|L0016700|VO|S1346628|Y|A1306532|||2871-3517|CSP|ET|2871-3517|freeze fracture|0|N|256|

C0016700|FRE|P|L3253434|PF|S3781115|Y|A7447611||M0008836|D005614| MSHFRE|MH|D005614 |Cryofracture|3|N||

Annexe B. La théorie des probabilités

La théorie mathématique de la probabilité a été proposée initialement par Blaise Pascal (1623-1662) et Pierre de Fermat (1601-1665). Elle se développe fortement au dix-septième siècle, avec de nombreuses applications dans les domaines de la science, la technologie et la gestion.

La théorie des probabilités peut être définie comme l'étude de l'influence de la connaissance sur la croyance ou la confiance. Les statistiques sur un fait passé peuvent être considérées comme un type de connaissance qui peut être conditionné et utilisé pour mettre à jour la croyance ou la confiance.

Les notions principales

Nous rappelons ici les principales notions du domaine des probabilités :

- Une **Expérience** est une procédure ou une opération qui produit un des résultats de l'ensemble des résultats possibles. Par exemple : jeter un dé.
- Un **Résultat ou état (Outcome)** est un résultat spécifique d'une expérience. Par exemple : après avoir jeté le dé, le numéro est 1.
- Un **Univers (Sample space)** Ω est l'ensemble de tous les résultats possibles qui peuvent être obtenus au cours d'une expérience. Par exemple dans l'expérience de prendre une carte, l'univers est l'ensemble des 52 résultats possibles.
- Un **Événement(Event)** est l'ensemble des résultats ou un sous-ensemble de l'univers. Par exemple l'événement la carte "roi" est prise dans l'expérience de prendre une carte.
- Des **événements mutuellement exclusifs (Événements ou disjoint)** sont des événements qui n'ont pas de résultat en commun. Par exemple : les événements

"personne X est un homme" et "personne X est une femme" sont mutuellement exclusifs.

Espace de probabilité

Un espace de probabilité ou espace probabilisé est un triplet (Ω, E, P) où :
- Ω est l'univers.
- E est l'ensemble des événements.
- $P : E \rightarrow [0, 1]$ est la probabilité des événements. On a : $P(\Omega) = 1$.

Variable aléatoire

Etant donné l'espace de probabilité (Ω, E, P), une variable aléatoire X est une fonction $X : \Omega \rightarrow Y$. Cette fonction associe aux éléments de Ω des valeurs numériques ou non numériques dans Y : $X\{\Omega\} = \{x_1, x_2, ...\}$ avec $x_i \in Y$.

Chaque x_i est un état de X, et $X = x_i$ est un événement.

Par exemple :

Loterie avec 100 billets. On tire un billet au hasard. C'est une expérience aléatoire. On peut prendre comme univers $\Omega = \{b1, b2, b3, b4, ..., b100\}$. Ω a 100 éléments. On ajoute les hypothèses suivantes : Sur les 100 billets, 1 billet gagne 1000$, 5 gagnent 200$ et 10 gagnent 100$. A chaque billet on peut associer son gain (nul ou non).

X est une variable aléatoire qui à chaque élément de l'univers, associe un nombre réel (gain). L'ensemble des valeurs prises par la variable aléatoire G est :

$$X(\Omega) = \{0; 100; 200; 1000\}$$

$X(\Omega)$ est l'ensemble des gains possibles. On note $(X = 100)$ l'événement "Gagner 100$", qui est l'ensemble des billets qui gagnent 100$, il y en a 10.

Les trois axiomes

Une **probabilité** $P(A)$ d'un événement A est un nombre dans l'intervalle unité $[0, 1]$, avec l'événement A étant un sous-ensemble de l'univers S. Les trois axiomes qui doivent être satisfaits sont :
- $P(A) = 1$ seulement si A est certain.
- $P(S) = 1$ et $P(\emptyset) = 0$

– Si les événements $A_i, i = 1, 2, ...$ sont mutuellement exclusifs, on a :

$$P(\vee_i A_i) = \sum_i P(A_i) \tag{8.1}$$

Probabilité conditionnelle

La **probabilité conditionnelle** est le concept principal dans les traitements des réseaux Bayésiens. On peut l'exprimer de la façon suivante : *sachant un événement B, la probabilité de A est x, noté $P(A/B)=x$.* Par exemple : sachant qu'un patient est fumeur, la probabilité qu'il ait un problème au poumon est...

Règle fondamentale du calcul des probabilités :

$$P(A, B) = P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A) \tag{8.2}$$

Indépendance et indépendance conditionnelle

Deux événements A, B sont **indépendants** ssi :

$$P(A, B) = P(A) \times P(B)$$

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

Deux événements A, B sont **indépendants conditionnellement** à C ssi :

$$P(A|B, C) = P(A|C)$$

$$P(B|A, C) = P(B|C)$$

$$P(A, B|C) = P(A|C)P(B|C)$$

Théorème des probabilités totales

– **La marginalisation** :
Étant données les variables A, B qui ont un ensemble d'états (state) mutuellement exclusifs $a_1..a_n, b_1..b_m$, la distribution de la probabilité de $P(A = a_i)$, notée $P(a_i)$ peut être calculée comme suit :

$$P(a_i) = \sum_j P(a_i, b_j) \tag{8.3}$$

TAB. 8.1 – Un exemple de table de la probabilité jointe $P(A, B)$

	b_1	b_2	b_3
a_1	0.16	0.12	0.12
a_2	0.24	0.28	0.08

Ce calcul est appelé la marginalisation de l'événement B qui est marginalisé à partir de $P(A, B)$ et $P(a_i)$ est le résultat. La notation générale est :

$$P(A) = \sum_B P(A, B) \tag{8.4}$$

Par exemple, étant donné la table de la probabilité jointe $P(A, B)$ (cf. table 8.1), on peut avoir :

$$\begin{aligned} P(a_1) &= \sum_j P(a_1, b_j) \\ &= P(a_1, b_1) + P(a_1, b_2) + P(a_1, b_3) \\ &= 0.16 + 0.12 + 0.12 \\ &= 0.4 \end{aligned}$$

et :

$$P(A) = (0.4, 0.6)$$

– **La probabilité totale** :
Soit un événement A résultant de plusieurs causes B_i. Sachant les probabilités a priori $P(B_i)$ et les probabilités conditionnelles de A pour chaque B_i, on peut calculer $P(A)$:

$$P(A) = \sum_i P(A|B_i)P(B_i) \tag{8.5}$$

Cette formule est déduite à partir des règles fondamentales (8.2) et de la marginalisation (8.4).

Le théorème de Bayes

Les formules précédentes permettent de calculer les probabilités dans l'ordre de cause à effet. Pourtant, dans certains cas, la question inverse peut se poser : quand un événement a s'est produit, quelle est la probabilité que ce soit la cause b qui l'ait produite ? A

partir des règles fondamentales du calcul de probabilité (8.2), la règle de Bayes peut être déduite :

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} \tag{8.6}$$

avec $P(a), P(b)$ qui sont les probabilités antérieures respectivement de a et de b ; $P(b|a)$ est la probabilité postérieure de b sachant a et $P(a|b)$ est la vraisemblance (likelihood) de b sachant a.

Annexe C. La collection ImageCLEFMed

La description

La table. 8.2 et la table. 8.3 décrivent le contenu d'ImageCLEFMed [2].

TAB. 8.2 – Les collections dans ImageCLEFMed 2007

Nom de collection	Type d'image	Type d'annotation
Casimage	Radiologie et pathologie	Description des cas cliniques
MIR	Médecine nucléaire	Description des cas cliniques
PEIR	Radiologie et pathologie	Metadata du base de données HEAL
PathoPIC	Pathologie	Descriptions des images
MyPACS	Radiologie	Description des cas cliniques
CORI	Images d'endoscopie	Description des cas cliniques

La figure 8.1 décrit la statistique sur la distribution des annotations des collections dans ImageCLEFMed2007

Exemple

Exemple d'une annotation dans la collection :

```
<CASIMAGE_CASE>
<ID>2403</ID>
<Description>
Radiographie de l'épaule gauche du 03.01.1994 : Petite interruption de la corticale
```

[2] http ://ir.ohsu.edu/image/2007protocol.html

TAB. 8.3 – Description des donées des collections dans ImageCLEFMed 2007

Collection	Cas	Images	Annotations	Annotations par langue			Taille
				Anglais	Français	Allemand	
Casimage	2076	8725	2076	177	1899	0	1.28 GB
MIR	407	1177	407	407	0	0	63.2 MB
PEIR	32319	32319	32319	32319	0	0	2.50 GB
PathoPIC	7805	7805	15610	7805	0	7805	879 MB
myPACS	3577	15140	3577	3577	0	0	390 MB
CORI	1496	1496	1496	1496	0	0	34 MB

```
à la jonction entre la surface articulaire de la tête humérale et le trochiter.
IRM du 12.01.1994 : Hétérogénéité de signal du trochiter surtout visible en T2
avec suppression de graisse (cf 3ième et 4ême images) traduisant une fracture
qui prend tout le trochiter. A noter également une réaction liquidienne entre
le trochiter et le muscle deltoïde. Par ailleurs il n'y a pas de rupture de la
coiffe des rotateurs ni de lésion du long chef du biceps. Pas de lésion des
bourrelets glénoïdiens.
</Description>
<Diagnosis> Fracture arrachement non déplacée du trochiter.</Diagnosis>
<CaseID>44110006</CaseID>
<ClinicalPresentation>
Patient de 29 ans. Status après chute en snowboard.
</ClinicalPresentation>
<Commentary>
Le remaniement est nettement plus important sur l'IRM que sur les radiographies
standard. Il s'agit en fait d'une fracture arrachement du trochiter qui est sous-
estimé par des radiographies standard. Cette fracture arrachement n'est pas déplacée.
</Commentary>
<Language>French</Language>
<Title>
Collection Kindynis - Ostéoarticulaire
</Title>
</CASIMAGE_CASE>
```

Les requêtes d'ImageCLEFMed

Chaque requête d'ImageCLEFMed est une phrase dont il y a plusieurs termes. Ces requêtes sont relativement courtes mais le besoin d'information peuvent être très précise ou très général. Tous les besoins d'information (sauf la requête 8 d'ImageCLEFMed2006) en général include d'un type d'analyse (XRay, MRI,...), d'une maladie (tuberculosis,

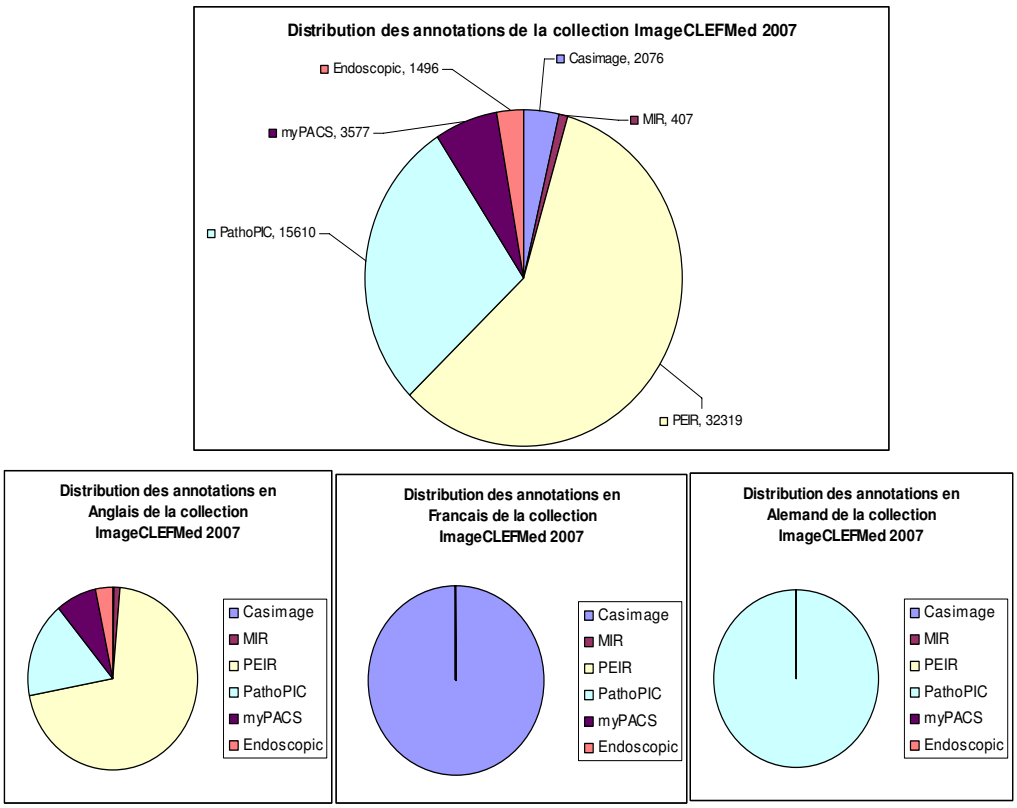


FIG. 8.1 – Statistique sur la distribution des annotations des collections dans ImageCLEFMed2007

lésion, ...), d'une partie du corps (tête, lung,...) ou d'une combinaison de ces trois éléments.

Les difficultés de ces requêtes concernent :

- Le cas où il y a les variations des termes dans la requête et dans les documents, i.e le terme la requête et les documents peuvent contenir des termes synonymes.
- Le cas où le besoin d'information est très général alors que les documents pertinents peuvent contenir des informations plus spécifiques.
- Le cas où la requête contient plusieurs éléments (type d'analyse, maladie, partie du corps) qui doivent être préférablement tous satisfaits dans les documents pertinents. Dans ce cas, non seulement l'intersection des termes d'indexation mais aussi de ces éléments doit être considérés.
-

Les requêtes d'ImageCLEFMed2006

1
Show me images of the oral cavity including teeth and gum tissue.
Zeige mir Bilder der Mundhöhle mit Zähnen und Zahnfleisch.
Montre-moi des images de la cavité buccale incluant des dents et du tissu des gencives.
2
Show me images of a frontal head MRI.
Zeige mir MR Frontalaufnahmen des Kopfes.
Montre-moi des images IRM frontal du crâne.
3
Show me images of a knee x-ray.
Zeige mir Röntgenbilder des Knies.
Montre-moi des radiographies du genou.
4
Show me x-ray images of a tibia with a fracture.
Zeige mir Röntgenbilder einer gebrochenen Tibia.
Montre-moi des radiographies du tibia avec fracture.
5
Show me x-ray images of a hip joint with prosthesis.
Zeige mir Röntgenbilder eines Hüftgelenks mit Prothese.
Montre-moi des radiographies d'articulation de la hanche avec une prothèse.
6
Show me images of a hand x-ray.
Zeige mir Röntgenbilder einer Hand.
Montre-moi des radiographies de la main.
7
Show me ultrasound images with a triangular result.
Zeige mir Ultraschallbilder mir dreieckigem Ergebnis.
Montre-moi des échographies de résultats triangulaires.
8
Show me images of PowerPoint slides.
Zeige mir Bilder von Powerpoint Folien.
Montre-moi des images de diapositives PowerPoint.
9
Show me images of an EEG or an ECG.
Zeige mir Bilder von EEG oder EKG.
Montre-moi des images d'un EEG ou d'un ECG.
10
Show me chest CT images with nodules.
Zeige mir CT Bilder der Lunge mit Knötchen.
Montre-moi des CTs du thorax avec nodules.
11
Show me ultrasound images with gallstones.
Zeige mir Ultraschallbilder mit Gallensteinen.

Montre-moi des échographies de calculs biliaires.
12
Show me a chest x-ray with tuberculosis.
Zeige mir Röntgenbilder der Lunge mit Tuberkulose.
Montre-moi des radiographies de la poitrine avec une tuberculose.
13
Show me CT images with a brain infarction.
Zeige mir CT Bilder eines Gehirnschlages.
Montre-moi des images CT avec un infarctus cérébral.
14
Show me MRI images of the brain with a blood clot.
Zeige mir MR Bilder des Gehirns mit Blutgerinnsel.
Montre-moi des images IRM du cerveau avec un caillot sanguin.
15
Show me x-ray images of vertebral osteophytes.
Zeige mir Röntgenbilder von vertebralen Osteophyten.
Montre-moi des radiographies d'ostéophytes vertébraux.
16
Show me ultrasound images of a foetus.
Zeige mir Ultraschallbilder eines Fötus.
Montre-moi des échographies d'un foetus.
17
Show me abdominal CT images of an aortic aneurysm.
Zeige mir CT Bilder des Abdomens mit einem Aneurismus der Aorta.
Montre-moi des CTs abdominaux d'un anévrisme aortique.
18
Show me blood smears that include polymorphonuclear neutrophils.
Zeige mir Blutabstriche mit polymophonuklearer Neutrophils.
Montre-moi des échantillons de sang incluant des neutrophiles polymorphonucléaires.
19
Show me images with multinucleated giant cells.
Zeige mir mikroskopische Bilder von Vasculitis.
Montre-moi des images microscopiques d'une vasculite.
20
Show photographs with lung tissue.
Zeige mir Fotos von Lungengewebe.
Montre des photographies de tissu pulmonaire.
21
Show me images of an infected wound.
Zeige mir Bilder einer infizierten Wunde.
Montre-moi des images d'une plaie infectée.
22
Show me photographs of tumours.
Zeige mir Fotos von Tumoren.
Montre-moi des photographies de tumeurs.

23
Show me CT or x-ray images showing the heart.
Zeige mir CT Bilder oder Röntgenbilder des Herzens.
Montre-moi des images CT ou des radiographies qui montrent le coeur.
24
Show me images of muscle cells.
Zeige mir Bilder von Muskelzellen.
Montre-moi des images de cellules musculaires.
25
Show me microscopic images of tissue from the cerebellum.
Zeige mir Mikroskopien von Kleinhirngewebe.
Montre-moi des images microscopiques de tissu du cervelet.
26
Show me x-ray images of bone cysts.
Zeige mir Röntgenbilder von Knochenzysten.
Montre-moi des radiographies de kystes d'os.
27
Show me images containing a Budd-Chiari malformation.
Zeige mir Bilder mit einer Budd-Chiari Verformung.
Montre-moi des images qui contiennent une malformation de Budd-Chiari.
28
Show me microscopic images showing parvovirus infection.
Zeige mir Mikroskopien mit einer Parvovirusinfektion.
Montre-moi des images microscopiques qui montrent une infection parvovirale.
29
Show me microscopic images of bacterial meningitis.
Zeige mir mikroskopische Bilder einer bakteriellen Hirnhautentzündung (Meningitis).
Montre-moi des images microscopiques de méningite bactérienne.
30
Show me images of findings with Alzheimer's Disease.
Zeige mir Bilder von Fällen mit einer Alzheimer Diagnose.
Montre-moi des images d'observations avec la maladie d'Alzeimer.

Les requêtes d'ImageCLEFMed2007

%Visually possible queries:
1
Cardiac MRI
MR-Bild Herz
IRM cardiaque
2
Mediastinal CT
Mediastinales CT
CT mediastinal
3

```
Photograph of dark brown skin lesion
Foto einer dunkelbraunen Hautläsion
Photo d'une lésion brune foncée de la peau
4
Xray hip fracture
Röntgenbild eines Hüftbruches
Radio d'une fracture de la hanche
5
Ultrasound with rectangular sensor
Ultraschallbild mit rechteckigem Sensor
Ultrason avec capteur rectangulaire
6
Leg of person
Bein einer Person
Jambe d'une personne
7
Xray dental implant or filling
Röntgenbild einer Zahnfüllung oder einer Zahnprothese
Radio d'un implant dentaire ou d'un plombage
8
Images acute otitis media
Bilder einer akuten Mittelohrentzündung
Image d'otite moyenne aigüe
9
Medial meniscus MRI
MR des medialen Meniskusses
IRM du ménisque interne
10
Gout images foot
Bilder eines Fu%es mit Gicht
Image d'un pied avec goutte
11
Glioblastoma CT
CT Glioblastom
CT d'un glioblastome
12
Gastrointestinal endoscopy with polyp
Magen-Darm-Endoskopie mit Polyp
Endoscopie grastroentestinale avec polype
13
Fetal MRI
MR-Bild Fetus
IRM d'un foetus
14
Mediastinum PET
```

PET Mittelfell
PET médiastinal
15
Lung xray tuberculosis
Röntgenbild Lunge Tuberkulose
Radio pulmonal de tuberculose
16
CT liver abscess
CT Leberabszess
CT d'un abcès du foie
17
Pathology non hodgkins lymphoma
Pathologiebild Non-Hodgkin-Lymphom
Image pathologique lymphome non-hodgkin
18
Photography of insect bite
Foto Insektenstich
Photo avec piqûre d'insecte
19
MRI or CT of colonoscopy
MR oder CT einer Koloskopie
IRM ou CT avec un colonoscopie
20
Stress fracture xray
Röntgenbild Ermüdungsbruch
Radio d'une fracture de stress
21
Pathology image with HIV
Pathologiebild mit HIV
Image pathologique avec VIH
22
Merkel cell carcinoma
Merkelzellkarzinom
Carcinome à cellules de Merkel
23
Bile duct cancer pathology image
Pathologiebild Gallengang Krebs
Image pathologique d'un cancer des voies biliaires
24
Gastrointestinal neoplasm
Magen-Darm-Blastom
Carcinome gastrointestinal
25
Tuberous sclerosis
Tuberöse Sklerose

Sclérose tubaire
26
Myocardial infarction pathology image
Pathologiebild Herzinfarkt
Image pathologique d'un infartus du myocarde
27
Mitral valve prolapse
Mitralklappenprolapssyndrom
Collapsus de la valve mitrale
28
Image of nursing
Bild Krankenpflege
Image de soins infirmiers
29
Pulmonary embolism all modalities
Lungenembolie alle Modalitäten
Embolie pulmonaire, toutes les formes
30
Microscopic giant cell
Mikroskopie Riesenzellen
Image microscopique de cellules géantes

Annexe D. Extraction de concepts avec XIotamap et Metamap

Méthode d'extraction de concepts de XIotamap

Les étapes d'extraction de concepts de cet outil sont :

- **L'étiquetage** avec TreeTagger [74].
- **La génération de variation** : Pour chaque mots, seulement les variantes majuscule-minuscule et stemming sont pris en compte.
- **La proposition de concepts candidats** : les concepts candidats sont extraits à partir des chaines des mots dans les textes.
- La désambiguïsation : Basée sur le contexte du domaine, ce qui nous amène a exclure certains thésaurus. Par exemple, on identifie *«x-ray»* comme *radiography* et non pas comme *«physical phenomenon»*.

Méthode d'extraction de concepts de Metamap

Ses étapes d'identification de concepts en pratique sont :

- **L'analyse morphologique** et l'identification des syntagmes nominaux.
- **La génération des variations** pour chaque syntagme nominal : basé sur acronymes, abréviations, synonymes, inflexion, épellation (spelling), et aussi des différentes combinaisons des composants du syntagme.
- **L'évaluation d'un score de confiance** pour chaque syntagme basée sur 4 critères. Une valeur normalisée entre 0 (la plus faible correspondance) et 1 (la plus forte correspondance) est calculée pour chaque critère. Le résultat d'évaluation est

Type de variation	Valeur de la distance
Spelling	0
Inflexion	1
Synonymie ou acronym/abbréviation	2
Dérivation	3

TAB. 8.4 – Distance de variants dans Metamap

enfin calculé : la couverture et la cohésion reçoivent 2 fois le poids de la centralité et de la variation.

- *La centralité* : cette valeur est égale à 1 si le syntagme générée comprend la tête du syntagme original et 0 dans les autre cas.
- *La variation* : cette valeur estime la différence entre la phrase générée et le syntagme original. La valeur de la variation totale est la somme des variations des mots composant la phrase. La dernière est une valeur prédéfinie selon la nature du type de variation. Ces valeurs sont listées dans la table 8.4.
- *La couverture* (coverage) permet d'évaluer la correspondance entre le syntagme candidat et le syntagme original. Un syntagme original ayant un concept correspondant exactement dans UMLS aura une couverture totale ; un syntagme original dont aucun de ses termes n'est lié a un concept présent dans UMLS à une couverture nulle. La couverture (coverage) pour évaluer combien le syntagme candidat et le syntagme original sont inclus dans la correspondance.
- *La cohésion* est similaire à la couverture sauf qu'elle insiste sur l'importance des composant connexe. Ce dernier est la séquence maximale continue des mots qui participent dans la correspondance.
- *La participation* (involvement) : cette valeur remplace la couverture dans le cas où l'ordre des mots est ignoré.

- **Le choix du meilleur concept** en fonction des 4 critères. Il n'y a pas d'étape de désambiguïsation.

Exemple de sortie de Metamap

Exemple de sortie de Metamap pour la requête :
"*Show me a chest x-ray with tuberculosis*" :

```
utterance('00000000.tx.16',"Show me a chest x-ray with tuberculosis.").
phrase('Show',[verb([lexmatch([show]),inputmatch([Show]), étiquette(verb))]).
phrase(me,[pron([lexmatch([me]),inputmatch([me]), étiquette(pron))]).
```

```
phrase('a chest x-ray',[det([lexmatch([a]), étiquette(det)])
                       ,mod([lexmatch([chest]), étiquette(noun)])
                       ,head([lexmatch([X-ray]), étiquette(noun)])]).
candidates([
  ev(-923,'C0202783','Chest x-ray',
    'Radiographic procedure on chest (procedure)',[chest,x,ray]
    ,[[1,1],[1,1],0],[2,2],[2,2],0],[3,3],[3,3],0]]),
  ev(-895,'C0856599','Breast X-ray',[breast,x,ray]
    ,[[1,1],[1,1],4],[2,2],[2,2],0],[3,3],[3,3],0]]),
  ev(-861,'C0034571','X-ray','roentgenographic',[x,ray]
    ,[[2,2],[1,1],0],[3,3],[2,2],0]]),
  ev(-861,'C0043299','X-ray','Diagnostic radiologic examination',[x,ray]
    ,[[2,2],[1,1],0],[3,3],[2,2],0]]),
  ev(-861,'C0043309','X-ray','Roentgen Rays',[x,ray]
    ,[[2,2],[1,1],0],[3,3],[2,2],0]]),
  ev(-861,'C1306645','X-ray','Plain film (procedure)',[x,ray]
    ,[[2,2],[1,1],0],[3,3],[2,2],0]]),
  ev(-844,'C0885876','X-rays','Homeopathic Preparations',[x,rays]
    ,[[2,2],[1,1],0],[3,3],[2,2],1]]),
  ev(-812,'C0086894','ray','Rajiformes'
    ,[ray],[[[3,3],[1,1],0]]),
  ev(-779,'C0851346','Rays','Radiation',[rays]
    ,[[3,3],[1,1],1]]),
  ev(-756,'C0078606','xanthosine','xanthosine',[xanthosine]
    ,[[2,2],[1,1],2]]),
  ev(-645,'C0817096','Chest','Chest',[chest]
    ,[[1,1],[1,1],0]]),
  ev(-590,'C0039992','Thoracic','Thorax',[thoracic]
    ,[[1,1],[1,1],2]]),
  ev(-590,'C0729233','Thoracic'
    ,'Dissecting aneurysm of the thoracic aorta',[thoracic]
    ,[[1,1],[1,1],2]]),
  ev(-562,'C0006141','Breast','Breast',[breast]
    ,[[1,1],[1,1],4]]),
  ev(-562,'C1268990','Breast','Entire breast',[breast]
    ,[[1,1],[1,1],4]]),
  ev(-545,'C0929301','Mammary','Mammary gland',[mammary]
    ,[[1,1],[1,1],6]]),
  ev(-530,'C0024659','Mammae','Mammary Glands, Animal',[mammae]
    ,[[1,1],[1,1],9]])]).
mappings([map(-923,[
  ev(-923,'C0202783','Chest x-ray'
  ,'Radiographic procedure on chest (procedure)',[chest,x,ray]
  ,[[1,1],[1,1],0],[2,2],[2,2],0],[3,3],[3,3],0]])])]).
phrase('with tuberculosis',[prep([lexmatch([with]),inputmatch([with]), étiquette(prep)])
```

```
  ,head([lexmatch([tuberculosis]),inputmatch([tuberculosis]), étiquette(noun)])
  ,punc([inputmatch([.]), étiquette(punctuation)])])
candidates([
  ev(-1000,'C0041296','Tuberculosis','Tuberculosis',[tuberculosis]
    ,[[1,1],[1,1],0]]),
  ev(-928,'C0443330','Tuberculous','Tuberculous (qualifier value)',[tuberculous]
    ,[[1,1],[1,1],3]]),
  ev(-900,'C1184740','Tubercle','Tubercle',[tubercle]
    ,[[1,1],[1,1],6]]),
  ev(-884,'C0332250','Tubercular','Tubercular (qualifier value)',[tubercular]
    ,[[1,1],[1,1],9]])]).
mappings([map(-1000,
  [ev(-1000,'C0041296','Tuberculosis','Tuberculosis',[tuberculosis]
    ,[[1,1],[1,1],0]])])]).
```

La requête est décomposée (dans l'étiquette `utterance`) pour identifier des syntagmes nominaux (étiquette `phrase`). Chaque syntagme nominal est analysé pour identifier tous les composants dans sa structure : les têtes et modificateurs sont identifiés. Pour chaque composant, les concepts candidats sont listés dans l'étiquette `candidates` (chaque `ev` est un candidat), avec des valeurs d'évaluation pour chaque concept.

Par exemple, dans `[[2,2],[1,1],0]` pour le 3ème candidat, les valeurs `[2,2],[1,1]` signifient que : le second terme `<<x>>` du terme `<<chest x-ray>>` dans la requête correspond avec le terme `<<X>>` du candidat `<<X-ray>>`, qui est un nom du concept `C0034571`. Le nombre qui suit est la distance de la correspondance calculée comme présenté précédent. Le meilleur choix de concepts global apparaît dans l'étiquette `mappings`. C'est le concept qui a un terme correspondant à la longueur maximum du syntagme d'origine.

Annexe E. Résultats et exemples des expérimentations

Dans cette annexe, nous montrons des exemples concrets dans les résultats obtenus sur la langue anglaise de la collection ImageCLEFMed2006. L'objectif est de montrer, par les exemples dans nos expérimentations, les avantages et limites de l'indexation par concept par rapport à celle par terme, et puis les avantages et limites de la prise en compte des relations dans le modèle Bayésien. La figure 8.2 et le tableau 8.5 illustrent les résultats (évalués par MAP) du modèle VSM de l'indexation par terme (VSM_Terme), par concept (VSM_Concept) et le modèle Bayésien (Bayes RB5) sur 30 requêtes d'ImageCLEFMed2006. Avec la recherche seulement sur les textes en Anglais de la collection, nous constatons des améliorations de :

- Indexation par concept par rapport à l'indexation par termes sur modèle VSM : 25% de MAP avec 23/30 requêtes améliorées.
- La prise en compte des relations entre concepts dans le modèle Bayésien par rapport au modèle VSM de l'indexation conceptuelle : 5.8% de MAP avec 20/30 requêtes améliorées.

A partir de ces résultats, nous extrayons ensuite quelques exemples pour montrer l'avantage et la limite de VSM_Concept par rapport à VSM_Terme , ainsi que l'avantage et limite de la prise en compte des relations dans le modèle Bayésien par rapport au VSM_Concept.

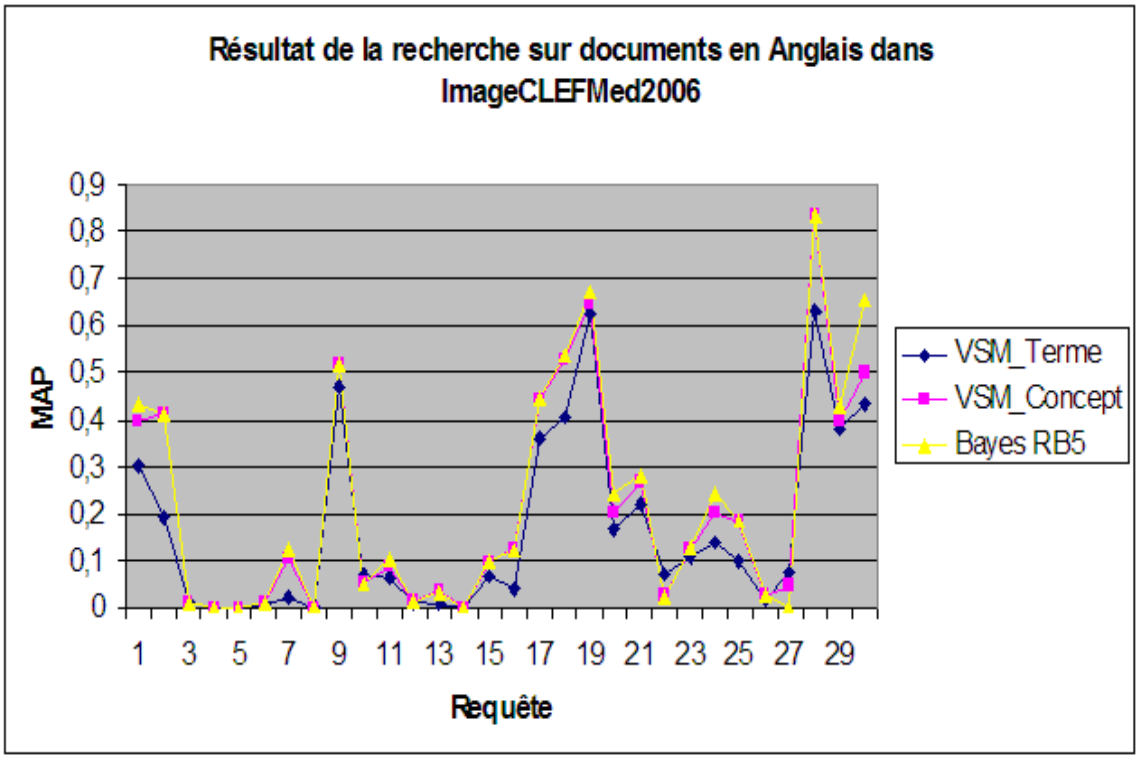


FIG. 8.2 – Les résultats (évalués par MAP) du modèle VSM de l'indexation par terme (VSM_Terme), par concept (VSM_Concept) et le modèle Bayésien (Bayes RB5) sur 30 requêtes d'ImageCLEFMed2006.

Avantage et limite de VSM_Concept par rapport à VSM_Terme

Avantage de l'indexation conceptuelle : Résolution de la synonymie

Requête 1 : Show me images of the oral cavity including teeth and gum tissue.

Avec cette requête, voyons les exemples dans le tableau 8.6.

Dans l'exemple, le document pertinent 00214319 est mieux classé avec VSM_Concept qu'avec VSM_Terme. Il faut noter qu'une fois la pondération des termes d'indexation calculée, coté requête et coté documents, la correspondance se calcule uniquement par le produit scalaire entre le vecteur requête et celui du document, quel que soit le type d'indexation. Pour un document donné, la valeur du produit scalaire est d'autant plus élevée qu'il y a de termes en commun entre la requête et ce document, car les pondérations sont toujours positives ou nulles. On peut donc constater que, la valeur de pondération égale, cette technique de correspondance favorise les larges intersections de termes d'indexation. Avec VSM_Terme, ce document a seulement 1 terme «oral» en commun avec la requête. Sachant wue w est la pondération du terme d'indexation dans la requête ou le document, normalisée par la longueur du vecteur des termes (on l'appelle la pondération normalisée), la similarité document-requête $sim(d,q)$ est :

$$\begin{aligned} sim(d,q) &= w(<< oral >>, q) * w(<< oral >>, d) && (8.7)\\ &= 0.31 * 0.38 && (8.8)\\ &= 0.11 \end{aligned}$$

Avec VSM_Concept, le terme «mouth» du document est synonyme avec le terme «oral cavity» de la requête et ils sont associés à des concepts identiques. Cela permet plus de concepts en commun entre ce document et la requête et donc un meilleur classement du document par rapport à VSM_Terme. La similarité document-requête $sim(d,q)$ est :

$$\begin{aligned} sim(d,q) &= w(<< C0226896 >>, q) * w(<< C0226896 >>, d) + && (8.9)\\ &\quad w(<< C0442027 >>, q) * w(<< C0442027 >>, d) + && (8.10)\\ &\quad w(<< C1278910 >>, q) * w(<< C1278910 >>, d) && (8.11)\\ &= 0.22 * 0.48 + 0.16 * 0.21 + 0.22 * 0.48 && (8.12)\\ &= 0.25 \end{aligned}$$

Limite : Redondance dans l'identification des concepts qui baisse le résultat dans VSM_Concept

Requête 5 : Show me x-ray images of a hip joint with prosthesis.

Pour cette requête, voyons les exemples dans le tableau 8.7.

Dans cet exemple, «X-RAY» est associé avec plusieurs concepts. De ce fait, le document non pertinent qui ne contient que «X-RAY» en commun avec la requête a pourtant plusieurs concepts en commun avec la requête est mieux classé. L'indexation conceptuelle produit donc une sorte de redondance dans index car plusieurs concepts sont identifiés à partir d'un terme.

Avantage et limite du modèle Bayésien avec la prise en compte des relations par rapport à VSM_Terme

Amélioration avec la prise en compte de relation entre concepts dans modèle Bayésien par rapport à VSM_Concept

Requête 30 : Show me images of findings with Alzheimer's Disease.

Interprétation conceptuelle de la requête :

C0150627 image

C0243095 findings

C0002395 Alzheimer

C0012634 Disease

Pour cette requête, voyons l'exemple de document correct mieux classé avec la prise en compte de relations dans le modèle Bayésien dans le tableau 8.8.

Dans l'exemple, le document ne contient que le concept«C0012634 (disease)» en commun avec la requête. Sa similarité document-requête $sim(d,q)$ dans le modèle VSM_Concept est :

$$sim(d,q) = w(C0012634,q) * w(C0012634,d) = 0.15 * 0.21 = 0.0315$$

(où w est la pondération normalisée du concept dans document ou requête).

Il n'est pas retrouvé dans les 1000 premiers documents retrouvés. Par contre, dans le modèle Bayésien avec la prise en compte de la relation hiérarchique entre «Dementia» et «Alzheimer», la similarité document-requête $sim(d,q)$ est :

$$\begin{aligned} sim(d,q) &= w(C0012634,q) * w(C0012634,d) + && (8.13)\\ &\quad w(C0497327,d) * SimSem(C0497327,C0002395) && (8.14)\\ &= 0.15 * 0.21 + 0.55 * 0.05 && (8.15)\\ &= 0.06 && (8.16) \end{aligned}$$

où SimSem est la similarité sémantique entre deux concepts calculée avec mesure de Leacock. Cette correspondance est plus élevée et permet de mieux classer ce document (rang : 590).

Limite de la prise en compte de relation entre concepts dans modèle Bayésien par rapport à VSM_Concept

Requête 27 : Show me images containing a Budd-Chiari malformation.

Interprétation de la requête :

C0150627 Image
C0332256 Containing
C0003803 Chiari malformation
C0000768 Malformation

Pour cette requête, voyons les exemples dans le tableau 8.9.

Dans cet exemple, le document non pertinent, qui ne contient que le concept «C0000768 (malformation, deformity)» en commun avec la requête, n'est pas retrouvé dans les 1000 premiers documents avec le modèle VSM de l'indexation conceptuelle. Par contre, le modèle Bayésien avec la prise en compte de la relation hiérarchique entre «malformation» et «skeletal deformity» a mieux classé ce document (rang : 35). Ce document est pourtant non pertinent car la requête cherche «Budd-Chiari malformation», et pas «skeletal deformity». Pour cette requête, d'après l'examen du résultat, il n'y a pas de documents pertinents qui changent leurs RSV par rapport au modèle VSM car il n'y a pas de relation sémantique entre leurs concepts et ceux de la requête trouvés (cela correspond à ce qu'on prouve dans la section 5.4 de la thèse : quand il n'y a pas de relation sémantique, le résultat de notre modèle Bayésien est identique au modèle VSM). Nous pensons donc que ce sont des documents non pertinents mieux classés qui ont influencé le rang des documents pertinents, et puis le MAP. Cet exemple montre que la prise en compte des relations sémantiques n'améliore pas toujours le MAP, mais peut même le baisser dans certain cas. Ce problème pourra donc faire partie du travail futur.

TAB. 8.5 – Résultats (évalués par MAP) du modèle VSM de l'indexation par terme (VSM_Terme), par concept (VSM_Concept) et le modèle Bayésien (Bayes RB5) sur 30 requêtes de ImageCLEFMed2006.

Requête	VSM_Terme	VSM_Concept	BAYES_RB5
1	0.3016	0.3953	0.4308
2	0.1936	0.4114	0.4136
3	0.0121	0.0119	0.012
4	0.0005	0.0002	0.0002
5	0.001	0.0007	0.0008
6	0.0081	0.0113	0.0113
7	0.0224	0.1029	0.1247
8	0.0002	0.0004	0.0008
9	0.4713	0.5145	0.5154
10	0.0699	0.0516	0.0481
11	0.0643	0.0856	0.1045
12	0.0113	0.0143	0.0143
13	0.006	0.0346	0.0335
14	0.0013	0.0004	0.0006
15	0.0684	0.0952	0.1011
16	0.0377	0.1229	0.122
17	0.358	0.4401	0.4454
18	0.4051	0.5252	0.5366
19	0.6274	0.6445	0.6712
20	0.1673	0.2032	0.2413
21	0.222	0.2656	0.2819
22	0.0708	0.0255	0.0228
23	0.1083	0.1249	0.1298
24	0.1382	0.2014	0.2403
25	0.1007	0.1832	0.1841
26	0.0184	0.0261	0.0271
27	0.0758	0.0479	0.0005
28	0.6301	0.8333	0.8333
29	0.3806	0.3933	0.4255
30	0.433	0.4983	0.6552
Moyenne	0.1668	0.2089	0.221

TAB. 8.6 – Exemple d'avantage de l'indexation conceptuelle

	VSM_Terme	VSM_Concept
Interprétation de la requête	Cavity Image Tooth Gum Tissue Oral	C0150627 Image C0150627 Image C0226896 Oral cavity C1278910 Oral cavity C0011334 Cavity NOS C0333343 Cavity C0442027 Oral C0332257 Including C0040426 Teeth C0040300 tissue C0017562 Gum C0812395 GUM
Rappel sur 213 documents	118	140
20 premiers documents retrouvés P : pertinent N : non pertinent	1 Peir/Images/00246340 -> P 2 Peir/Images/00095042 -> P 3 Peir/Images/00205940 -> P 4 Peir/Images/00095041 -> P 5 Peir/Images/00095003 -> P 6 Peir/Images/00095034 -> P 7 Peir/Images/00250787 -> P 8 Peir/Images/00248742 -> P 9 Peir/Images/00248743 -> P 10 Peir/Images/00219437 -> P 11 Peir/Images/00218613 -> P 12 Peir/Images/00219435 -> P 13 Peir/Images/00219436 -> P 14 Peir/Images/00214401 -> N 15 Peir/Images/00214404 -> N 16 Peir/Images/00214407 -> N 17 Peir/Images/00214410 -> N 18 Peir/Images/00214413 -> N 19 Peir/Images/00095110 -> N 20 Peir/Images/00095111 -> P	1 Peir/Images/00246340 -> P 2 Peir/Images/00250787 -> P 3 Peir/Images/00250784 -> P 4 Peir/Images/00205940 -> P 5 Peir/Images/00214237 -> P 6 Peir/Images/00219429 -> P 7 Peir/Images/00095003 -> P 8 Peir/Images/00095227 -> P 9 Peir/Images/00250714 -> P 10 Peir/Images/00214319 -> P 11 Peir/Images/00213406 -> N 12 Peir/Images/00250786 -> P 13 Peir/Images/00205285 -> P 14 Peir/Images/00212733 -> P 15 Peir/Images/00212642 -> P 16 Peir/Images/00212364 -> P 17 Peir/Images/00212511 -> P 18 Peir/Images/00212014 -> P 19 Peir/Images/00212361 -> P 20 Peir/Images/00212604 -> P
Exemple de document retrouvé	Document correct retrouvé : 260 Peir/Images/00214319 -> P	Document pertinent mieux classé : 10 Peir/Images/00214319 -> P le document 00214319 apparaît en position 10 alors qu'il est classé 260 dans indexation par terme
Intersection de termes d'indexation	" MOUTH : LUPUS, DISCOID, ORAL.« Oral	" MOUTH : LUPUS, DISCOID, ORAL." C0226896 Oral cavity C0442027 Oral C1278910 Oral cavity

TAB. 8.7 – Exemple limite de l'indexation conceptuelle

	VSM_Terme	VSM_Concept
Interprétation de la requête	X-Ray Image Hip Prosthesis	Joint C0034571 X-ray image C0034571 X-ray C0043299 X-ray C0043309 X-ray C0043309 Xray C1306645 X-ray C0885876 X-rays C0086894 ray C0019558 Hip Joint C1279144 Hip joint C0019552 Hip C0022417 Joint C1269611 Joint C0175649 Prosthesis C0525024 prosthesis
Rappel sur 19 documents	3	3
5 premiers documents retrouvés P : pertinent N : non pertinent	1 Peir/Images/00248780 -> N 2 Peir/Images/00248782 -> N 3 Peir/Images/00248784 -> N 4 Peir/Images/00248786 -> N 5 Peir/Images/00247272 -> N	1 Peir/Images/00235234 -> N 2 Peir/Images/00006248 -> N 3 Peir/Images/00219575 -> N 4 Peir/Images/00235150 -> N 5 Peir/Images/00248598 -> N
Exemple de document retrouvé		Document non pertinent mieux classé : 31 Peir/Images/00006249 "LUNG : CARCINOMA, METASTATIC TO LUNG, X-RAY" il apparaît en position 31 alors qu'il est classé 136 dans indexation par terme
Intersection de termes d'indexation		C0034571 X-ray image C0034571 X-ray C0043299 X-ray C0043309 X-ray C0043309 Xray C1306645 X-ray C0885876 X-rays C0086894 ray

TAB. 8.8 – Exemple de l'avantage de la prise en compte de relation entre concepts

	VSM_Concept	Bayes_RB5
5 premiers documents retrouvés	1 PathoPic/Images/004973 -> P 2 PathoPic/Images/005437 -> P 3 Peir/Images/00234469 -> P 4 Peir/Images/00237276 -> P 5 Peir/Images/00237278 -> P	1 PathoPic/Images/004973 -> P 2 PathoPic/Images/005437 -> P 3 Peir/Images/00234469 -> P 4 Peir/Images/00237276 -> P 5 Peir/Images/00237278 -> P
Rappel sur 80 documents	53	54
Exemple de document retrouvé		Document pertinent mieux classé : 590 Peir/Images/00234418 Ce document est mieux classé avec la prise en compte des relations : il est en position 590 alors qu'il n'est pas retrouvé dans les 1000 premiers documents retournés avec VSM_Concept. Le contenu de ce document est : "BRAIN : DISEASE DEMENTIA ; A117-83, LATERAL HEMISPHERE, ATROPHY" Ses concepts : C0006104 Brain C0012634 disease C0205093 lateral C0236642 atrophy C0497327 Dementia C1269537 Brain

TAB. 8.9 – Exemple de limite de la prise en compte de relation entre concepts

	VSM_Concept	Bayes_RB5
5 premiers documents retrouvés	1 Peir/Images/00233008-> N 2 Peir/Images/00233011-> N 3 Peir/Images/00237702-> N 4 Peir/Images/00243808-> N 5 PathoPic/Images/002839 -> P	1 Peir/Images/00243808 -> N 2 Peir/Images/00233011-> N 3 Peir/Images/00237702-> N 4 Peir/Images/00233008-> N 5 Peir/Images/00253713-> N
Rappel sur 22 documents	4	4
Exemple de document retrouvé		Document non pertinent mieux classé : 35 Peir/Images/00218249 " BONE : SYNDROME, SKELETAL DEFORMITY, PELVIS" Concepts dans ce document : C1266908 bone C1266909 bone C0262950 bone C0391978 bone C0039082 syndrome C0521324 skeletal C0000768 deformity C0241052 skeletal deformity C1279864 pelvis C0030797 pelvis

Liste des publications

Revue

Caroline Lacoste and Joo-Hwee Lim and Jean-Pierre Chevallet and Diem Le Thi Hoang, Medical Image Retrieval based on Knowledge-Assisted Text and Image Indexing, IEEE Transactions on Circuits and Systems for Video Technology, vol17, no7, pp889-900, July, 2007.

Conférence internationale

1. Joo-Hwee Lim and Jean-Pierre Chevallet and Diem Thi Hoang Le and Hanlin Goh, Bi-Modal Conceptual Indexing for Medical Image Retrieval, in The 14th International Multimedia Modeling Conference MMM2008 : 456-465, Kyoto Japan, 9-11 January, 2008.
2. Thi Hoang Diem Le and Jean-Pierre Chevallet and Thi Bich Thuy Dong, Thesaurus-based query and document expansion in conceptual indexing with UMLS, in Research, Innovation and Vision for the Future (RIVF) IEEE International Conference on (2007), pp. 242-246.
3. Joo-Hwee Lim and Caroline Lacoste and Jean-Pierre Chevallet and Diem Le, Knowledge-Assisted Medical Image Retrieval, in ICME 2007, International Conference on Multimedia & Expo, Beijing China, 2007, p.p 791-794.
4. Jean-Pierre Chevallet and Joo Hwee Lim and Thi Hoang Diem Le, Domain Knowledge Conceptual Inter-Media Indexing, Application to Multilingual Multimedia Medical Reports, in ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal , November 6-9, 2007, pages 495-504.
5. LE Thi Hoang Diem and Jean-Pierre CHEVALLET, Extraction et structuration des relations multi-types à partir de texte , in quatrième conférence internationale en informatique dédiée à la Recherche, l'Innovation et la Vision pour le Futur (RIVF'06) , Ho Chi Minh Ville, Viêt-Nam, pp53-58, 12-16 Febuary, 2006.

Conférences nationales

1. Thi Hoang Diem Le, Exploitation des connaissances d'UMLS pour la recherche d'information médicale. Vers un modèle bayésien d'indexation, Rencontres Jeunes Chercheurs en Recherche d'Informations (RJCRI'07), 2007.
2. Saïd Radhouani and Loïc Maisonnasse and Joo-Hwee Lim and Thi-Hoang-Diem Le and Jean-Pierre Chevallet, Une Indexation Conceptuelle pour un Filtrage par

Dimensions, Expérimentation sur la base médicale ImageCLEFmed avec le méta thésaurus UMLS, in COnférence en Recherche Information et Applications CORIA'2006, Lyon France, pp257-269, 15 - 17 mars, 2006.

Campagnes d'évaluations

1. Caroline Lacoste and Jean-Pierre Chevallet and Joo-Hwee Lim and Xiong Wei and Daniel Raccoceanu and Diem Le Thi Hoang and Roxana Teodorescu and Nicolas Vuillenemot, IPAL Knowledge-based Medical Image Retrieval in ImageCLEFmed 2006, in Working Notes for the CLEF 2006 Workshop, 20-22 September , Medical Image Track, Alicante, Spain, 2006.
2. Diem Thi Hoang Le, Jean-Pierre Chevallet, and Joo Hwee Lim , Using Bayesian Network for Conceptual Indexing : Application to Medical Document Indexing with UMLS Metathesaurus, in Lecture Notes In Computer Science, Advances in Multilingual and Multimodal Information Retrieval : 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, pages 631–636.

Rapport

Thi Hoang Diem LE, Extraction et structuration de connaissances pour la recherche d'information, rapport de DEA, Groupe MRIM - CLIPS-IMAG, Juin, 2003.

Bibliographie

[1] Aronson Alan A. Effective mapping of biomedical text to the UMLS metathesaurus : The MetaMap program. In *AMIA*, pages 17–21, 2001.

[2] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from retomness. *ACM Trans. Inf. Syst.*, 20(4) :357–389, 2002.

[3] Alan R. Aronson. Metamap : Mapping text to the umls metathesaurus. http ://mmtx.nlm.nih.gov/docs.shtml, July 2006.

[4] Alan R. Aronson and T.C. Rindflesch. Query expansion using the UMLS metathesaurus. In D.R. Masys, editor, *Proceedings of the 1997 AMIA Annual Fall Symposium*, pages 485–489, 1997.

[5] Alan R. Aronson, Rindflesch C. Thomas, and Browne C. Allen. Exploiting a large thesaurus for information retrieval. pages 197–216, 1994.

[6] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[7] Jean-Pierre Chevallet Bao-Quoc Ho and Marie-France Bruandet. Recherche d'information bilingue français - vietnamienne. In *RIVF*, pages 179–186, 2004.

[8] Mustapha Baziz. *"Indexation Conceptuelle Guidée par Ontologie pour la Recherche d'Information*. PhD thesis, 2005.

[9] Mustapha Baziz, Mohet Boughanem, and Nathalie Aussenac-Gilles. A Conceptual Indexing Approach based on Document Content Representation . In F. Crestani and I. Ruthven, editors, *CoLIS5 : Fifth International Conference on Conceptions of*

Libraries and Information Science , Glasgow, UK, 04/06/05-08/06/05, pages 171–186, Berlin Heidelberg, juin 2005. Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag.

[10] Mustapha Baziz, Mohet Boughanem, and Nathalie Aussenac-Gilles. Conceptual indexing based on document content representation. In *CoLIS*, pages 171–186, 2005.

[11] Mustapha Baziz, Mohet Boughanem, and Nathalie Aussenac-Gilles. Evaluating a Conceptual Indexing Method by Utilizing WordNet . In Carol Peters, Fredric C. Gey, Julio Gonzalo, and Gareth J.F. Jones, editors, *Accessing Multilingual Information Repositories : 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers , Vienna, Austria, 21/09/05-23/09/05*, pages 238–246. Lecture Notes in Computer Science, Vol. 4022, septembre 2005.

[12] Mustapha Baziz, Mohet Boughanem, Nathalie Aussenac-Gilles, and Claude Chrisment. Semantic Cores for Representing Documents in IR . In *SAC'2005- 20th ACM Symposium on Applied Computing , Santa Fe, New Mexico, USA, 13/03/05-17/03/05*, pages 1011–1017, New York, NY, USA, mars 2005. ACM Press.

[13] Bodo Billerbeck and Justin Zobel. Document expansion versus query expansion for ad-hoc retrieval.

[14] Bodo Billerbeck and Justin Zobel. Techniques for efficient query expansion. In *SPIRE*, pages 30–42, 2004.

[15] B C Brookes. Jason Farradane and relational indexing. *J. Inf. Sci.*, 12(1-2) :15–18, 1986.

[16] P.D. Bruza and L.C. van der Gaag. Index Expression Belief Networks for Information Disclosure. *International Journal of Expert Systems*, 7(2) :107–138, 1994.

[17] P.D. Bruza and J.J IJdens. Efficient Probabilistic Inference through Index Expression Belief Networks. In *Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence (AI94)*, pages 592–599. World Scientific, 1994.

[18] A. Budanitsky. Semantic distance in wordnet : An experimental, application-oriented evaluation of five measures, 2001.

[19] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.

[20] Jean-Pierre Chevallet. X-iota : An open XML framework for IR experimentation. In *AIRS*, pages 263–280, 2004.

[21] Jean-Pierre Chevallet, Joo Hwee Lim, and Thi Hoang Diem Le. Domain knowledge conceptual inter-media indexing, application to multilingual multimedia medical reports. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal*, November 6–9 2007.

[22] Jean-Pierre Chevallet, Joo-Hwee Lim, and Saïd Radhouani. A structured visual learning approach mixed with ontology dimensions for medical queries. *Accessing Multilingual Information Repositories. Lecture Notes in Computer Science*, pages 642–651, 2006.

[23] De Loupy Claude. Évaluation des taux de synonymie and de polysémie dans un texte. In *Actes de TALN 2002*, 2002.

[24] H. Cui, J. Wen, J. Nie, and W. Ma. Query expansion by mining user logs, 2003.

[25] Stéfan J. Darmoni, J.-P. Leroy, Benoît Thirion, F. Baudic, Magali Douyere, and J. Piot. CISMeF : a structured health resource guide. *Methods of Information in Medicine*, 39(1) :30–35, 2000.

[26] Stéfan J. Darmoni, Benoît Thirion, J. P. Leroy, Magali Douyère, F. Baudic, and J. Piot. CISMeF : a structured health resource guide for healthcare professionals and patients.

[27] David Eichmann, Miguel Ruiz, and Padmini Srinivasan. Cross-language information retrieval with the UMLS metathesaurus. In *Research and Development in Information Retrieval*, pages 72–80, 1998.

[28] Saïd Radhouani et Gilles Falquet. Using external knowledge to solve multi-dimensional queries. In *ISPE CE*, pages 426–437, 2006.

[29] Loïc Maisonnasse et Gilles Sérasset et Jean-Pierre Chevallet. Using the x-iota system in mono- and bilingual experiments at clef 2005. In *CLEF*, pages 69–78, 2005.

[30] Henning Müller et Thomas Deselaers et Thomas Martin Deserno et Paul Clough et Eugene Kim et William R. Hersh. Overview of the imageclefmed 2006 medical retrieval et medical annotation tasks. In *CLEF*, pages 595–608, 2006.

[31] Victoria Fromkin. *An Introduction to language / by Victoria Fromkin ... [et al.].* Holt, Rinehart and Winston, Sydney :, australian ed. edition, 1984. BibTex item created from National Library of Australia Catalogue record nla.cat-vn835923.

[32] Susan Gauch and Jianying Wang. A corpus analysis approach for automatic query expansion. In *CIKM '97 : Proceedings of the sixth international conference on Information and knowledge management*, pages 278–284, New York, NY, USA, 1997. ACM.

[33] Joseph Goguen. What is a concept ? *Lecture Notes in Computer Science : Conceptual Structures : Common Semantics for Sharing Knowledge*, pages 52–77, 2005.

[34] Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP*, pages 38–44, Montreal, Canada, 1998.

[35] Jean-Michel Renders Hervé Déjean, Éric Gaussier and Fatiha Sadat. Automatic processing of multilingual medical terminology : applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2) :111–124, 2005.

[36] G. Hirst and D. St-Onge. Lexical chains as representation of context for the detection and correction malapropisms, 1997.

[37] Bao-Quoc Ho. *Vers une indexation structurée, basée sur des syntagmes nominaux. Impact sur un SRI en vietnamien and la RI multilingue.* PhD thesis, Université Joseph Fourier, 2004.

[38] Finn V. Jensen, F.V. V. Jensen, and F. V. Jensen. *Introduction to Bayesian Networks.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[39] Finn V. Jensen and Thomas D. Nielsen. *Bayesian Networks and Decision Graphs (Information Science and Statistics).* Springer, 2nd ed. edition, June 2007.

[40] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+, September 1997.

[41] Stefan Klink, Armin Hust, Markus Junker, and etreas Dengel. Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts. In *Proceedings of DAS 2002, 5th International Workshop on Document Analysis Systems*, volume 2423 of *Lecture Notes in Computer Science*, pages 376–387, Princeton, NJ, USA, August 2002. Springer.

[42] Caroline Lacoste, Joo-Hwee Lim, Jean-Pierre Chevallet, and Diem Le Thi Hoang. Medical image retrieval based on knowledge-assisted text and image indexing. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(7) :889–900, July 2007.

[43] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *An Electronic Lexical Database*, pages 265–283, 1998.

[44] Douglas B. Lenat. Cyc : a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11) :33–38, 1995.

[45] Dekang Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[46] H. Liu and P. Singh. Conceptnet : A practical commonsense reasoning toolkit.

[47] J.Y. Nie M. Boughanem, W. Kraaij. Modèles de langue pour la recherche d'informations. *Les systèmes de recherche d'informations - Modèles conceptuels.*

[48] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Research and Development in Information Retrieval*, pages 191–197, 1999.

[49] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.

[50] Wenlei Mao and Wesley W. Chu. Free-text medical document retrieval via phrase-based vector space model. In *Proc AMIA Symp*, pages 489–493.

[51] George A. Miller. Wordnet : a lexical database for english. *Commun. ACM*, 38(11) :39–41, 1995.

[52] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6(1) : 1Ũ28*, 6(1) :1–28, 1991.

[53] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Duret. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81, New York, NY, USA, 1999. ACM.

[54] Douglas W. Oard and Paul Hackett. Document translation for cross-language text retrieval at the university of marylet. In *Text REtrieval Conference*, pages 687–696, 1997.

[55] Vossen P. Introduction to eurowordnet. *Computers and the Humanities*, 32 :73–89(17), 1998.

[56] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[57] Helen J. Peat and Peter Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS*, 42(5) :378–383, 1991.

[58] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. Dictionary-based cross-language information retrieval : Problems, methods, and research findings. *Information Retrieval*, 4(3/4) :209–230, 2001.

[59] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia*, pages 275–281, 1998.

[60] HO Bao Quoc, DONG Thi Bich Thuy, Jean-Pierre CHEVALLET, and Marie-France BRUetET. A structured indexing model based on nouns phrase. In *quatrième conférence internationale en informatique dédiée à la Recherche, l'Innovation and la Vision pour le Futur (RIVF'06) , Ho Chi Minh Ville*, pages 81–89, 12 –16 febuary 2006.

[61] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1) :17–30, 1989.

[62] Saïd Radhouani. *Un modèle de Recherche d'Information orienté précision fondé sur les dimensions de domaine*. PhD thesis, Université Joseph Fourier, 2008.

[63] Saïd Radhouani, Loïc Maisonnasse, Joo-Hwee Lim, Thi-Hoang-Diem Le, and Jean-Pierre Chevallet. Une indexation conceptuelle pour un filtrage par dimensions, expérimentation sur la base médicale imageclefmed avec le méta thésaurus UMLS. In *COnférence en Recherche Information and Applications CORIA'2006, Lyon France*, 15 – 17 mars 2006.

[64] Philip Resnik. Semantic classes and syntactic ambiguity. In *HLT '93 : Proceedings of the workshop on Human Language Technology*, pages 278–283, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[65] Berthier A. N. Ribeiro and Richard Muntz. A belief network model for ir. In *SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, New York, NY, USA, 1996. ACM Press.

[66] Ray Richardson and Alan F. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. Technical Report CA-0395, Dublin, Irelet, 1995.

[67] Ray Richardson, Alan F. Smeaton, and J. Murphy. Using WordNet as a knowledge base for measuring semantic similarity between words. Technical Report CA-1294, Dublin, Irelet, 1994.

[68] Bärbel Ripplinger, Spela Vintar, and Paul Buitelaar. Cross-lingual medical information retrieval through semantic annotation.

[69] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94 : Proceedings of the*

17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[70] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. volume 27(3), 1976.

[71] M. Rodriguez and M. Egenhofer. Determining semantic similarity among entity classes from different ontologies, 2003.

[72] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10) :627–633, 1965.

[73] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11) :613–620, 1975.

[74] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, sept 1994.

[75] Amit Singhal and Ferneto C. N. Pereira. Document expansion for speech retrieval. In *Research and Development in Information Retrieval*, pages 34–41, 1999.

[76] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11–21, 1972.

[77] P. Srinivasan. Retrieval feedback in medline. *J Am Med Inform Assoc*, 3(2) :157–67, 1996.

[78] Howard Turtle and Bruce W. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3) :187–222, 1991.

[79] Howard Turtle and W. Bruce Croft. Inference networks for information retrieval. In *SIGIR '90 : Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, 1990.

[80] H.R. Turtle. *Inference Networks for Document Retrieval.* PhD thesis, 1991.

[81] Amos Tversky. Features of similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.

[82] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, and Evangelos E. Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *WIDM '05 : Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16, New York, NY, USA, 2005. ACM.

[83] Spela Vintar, Paul Buitelaar, and Martin Volk. Semantic relations in concept-based cross-language medical information retrieval. In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*, 9 2003.

[84] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[85] Jianqiang Wang and Douglas W. Oard. Clef-2005 cl-sr at marylet : Document and query expansion using side collections and thesauri. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 800–809. Springer, 2005.