# Exploitation des connaissances d'UMLS pour la recherche d'information médicale. Vers un modèle bayésien d'indexation

Thi Hoang Diem Le

# Exploration de connaissances dans UMLS pour la recherche d'information médicale Vers un model Baysien d'indexation

**Diem Le Thi Hoang**

*Equipe MRIM, Laboratoire CLIPS-IMAG*
*38041 Grenoble Cedex 9, France*
*Thi-Hoang-Diem.Le@imag.fr*

RÉSUMÉ. *La recherche d'information à base de connaissances est largement exploitée, mais avec peu de succès. Dans cet article, nous étudions l'impact de l'exploration d'une base de connaissance, nommée meta-thésaurus UMLS, dans la recherche d'information médicale. L'intégration des étiquettes sémantiques des concepts dans une indexation multicouche donne des résultats encouragants dans ImageCLEF 2006 forum d'évaluation. Afin d'atteindre un système de recherche d'information sémantiquement plus riche , nous explorons les relations hiérarchiques entre concepts dans UMLS. Dans ce but, nous proposons donc un modèle basé sur le réseau Bayesien pour capturer les liens général-specifique entre concepts de la requête et ceux des documents.*

ABSTRACT. *Knowledge-based information retrieval is widely exploited, but still not very well successful. In this paper, we aim to study the impact of explorations of a knowledge source, named UMLS meta-thesaurus, in medical domain information retrieval. The integration of semantic labels of concepts in a multi-layer indexing proves quite encouraging results in ImageCLEF 2006 evaluation forum. In order to reach a more semantic rich information retrieval system, we tend to explore the hierarchy relationships between concepts in UMLS. For this purpose, we propose a Bayesian network based model to capture the general-specific links between query concepts and documents concepts.*

MOTS-CLÉS : *Recherche d'information à base de connaissances, indexation conceptuelle, UMLS, indexation multicouche, réseau Bayesien.*

KEYWORDS: *Knowledge-based information retrieval, conceptual indexing, UMLS, multi-layer indexing, Bayesian network.*

e soumission à *Rencontres Jeunes Chercheurs en Recherche d'Informations (RJCRI'07)*, le 5 janvier 2007

## 1. Introduction

L'utilisation des bases de connaissances externes de documents (thesaurus par exemple) dans la recherche d'information (RI) a été largement explorée. Les résultats gagnent pourtant très peu de succès. Les principales approches concernent :

– l'extraction des concepts : cela est nécessaire pour toute indexation conceptuelle. Une bonne qualité d'extraction de concepts suppose de résoudre le problème de variations des termes et synonymie. Une large source de concepts est évidement un élément important. Dans le cas de doute de la couverture de cette source, l'indexation en combinant des concepts et des mots est proposée pour ne pas perdre le contenu dans le document (Baziz *et al.,* 2005), (Aronson *et al.,* 1994). Pourtant, ce travail d'identification de concepts fait face au problème majeur d'ambiguïté du sens des mots qui demande beaucoup d'effort à résoudre. Ce problème est plus important dans le cas des bases de connaissances générales comme Wordnet.

– l'expansion de requêtes : avec le but d'ajouter les concepts sémantiquement liés a la requête, cette méthode ne montre pourtant pas d'amélioration (Gonzalo *et al.,* 1998) ou très peu sur une collection de petite taille (Mandala *et al.,* 1999).

– le calcul des distances sémantiques entre les concepts : cela est basé sur la structure hiérarchique des concepts. Certaines mesures de distances sémantiques sont proposés (Budanitsky, 2001), (Resnik, 1995), mais elles ne prouvent pas d'effets positifs dans la RI. De plus, la base théorique de ces mesures n'est pas très convaincante.

Nous proposons dans cet article l'étude de l'impact d'exploration de connaissances externes sur la performance du système de recherche d'information (SRI). Nous choisissons le meta-thesaurus UMLS [1] (Unified Medical Language System) de NLM (National Librabry of Medicine) pour étudier son application dans la recherche d'information sur la base d'image médicales. Les différentes méthodes d'exploration sont décrites dans les sections 2, 3, 4 ; notre tentative vers un modèle Bayesien basé sur la structure hiérarchique de UMLS est dans la section 5 ; nos résultats positifs d'expérimentation pour CLEF Image Médicale 2006 [2] dans la section 6 et la conclusion dans la section 7.

## 2. La conceptualisation

La conceptualisation est vue comme une étape pour monter à un plus haut niveau d'abstraction du contenu de texte, ainsi comme une méthode pour établir le lien entre différentes formes de surfaces linguistiques et de sens. Une indexation par concept résout d'ailleurs la barrière de la langue dans la RI multilingue. La ressource de concepts que nous utilisons, appelé UMLS, est un meta-thesaurus de taille importante. Il est la fusion des 140 sources terminologiques, avec plus de 1,1 million de concepts et 5,5 millions de termes dans 17 langues. Pour l'identification de concepts dans le texte,

1. http ://umlsinfo.nlm.nih.gov/
2. http ://ir.shef.ac.uk/imageclef/

NLM nous offre l'outil Metamap (Aronson, 2006) qui produit tous les concepts candidats par rapport à toutes les formes textuelles en Anglais. Pour les textes en Français et en Allemand, nous construisons un outil similaire d'extraction de concepts. La conceptualisation offre donc un unique ensemble des concepts à indexer à partir de textes dans différentes langues.

## 3. Exploration des étiquettes sémantiques des concepts - Indexation multicouche

Dans UMLS, les concepts sont classifiés en 135 *types sémantiques* pour une meilleure organisation et structuration. Ces types sémantiques sont encore classifiés en 15 *groupes sémantiques* (par exemple anatomie, affection etc.). Ces dernières considérées comme les abstractions de plus haut niveau des concepts, nous les appelons les *étiquettes sémantiques* des concepts. Avec ces étiquettes sémantiques, nous pouvons gérer et évaluer l'importance des concepts par rapport au contenu général du texte. Par exemple pour les documents qui décrivent les maladies sur une partie du corps, les concepts avec étiquettes sémantiques "anatomie" et "affection" sont plus importantes que les autres dans la contribution au contenu de ces documents. Cette information est donc intéressante dans la RI dans laquelle la compréhension du texte joue un rôle important. Afin de prendre en compte ces informations, nous proposons une indexation multicouche : la couche des index des concept de bas niveau et la couche des index de leur étiquettes sémantiques au plus haut niveau. L'un est la projection de l'autre en se basant sur les assignements dans UMLS. Ce modèle est applicable avec le modèle d'espace vectoriel (Salton, 1991), e.g. chaque document $d_i$ ou requête $q$ est représenté par deux vecteurs correspondants aux deux espaces différents : $d_i = (\overrightarrow{d_{i1}}, \overrightarrow{d_{i2}}), q = (\overrightarrow{q_1}, \overrightarrow{q_2})$

La correspondance entre document et requête dans chaque couche est exécutée séparément. Chacune produit une valeur de pertinence ou RSV (Relevant Status Value) pour chaque document par rapport à la requête. Le RSV final est obtenu par une fonction de combinaison $f$ des RSV des deux couches :

$$\text{RSV} = f(\text{RSV}_1, \text{RSV}_2)$$

Nous proposons différents schémas de pondération dans deux couches : $tf.idf$ pour la couche des concepts et binaire pour l'autre. Comme nous insistons sur l'influence des étiquettes sémantiques, nous définissons la fonction $f$ comme le produit. Nous pouvons donc utiliser la couche des étiquettes sémantiques comme un filtre par diminution de l'espace d'indexation aux certaines étiquettes sémantiques importantes seulement. Quand nous réduisons encore cet espace à la projection des concepts correspondants, nous aurons donc le filtre sur ces concepts. Nous l'appliquons dans CLEF Image Médicale 2006, avec la réduction de cet espace aux étiquettes d'anatomie, d'affection et de modalité. Cette méthode nous permet d'avoir une meilleure performance dans ce forum d'évaluation (Tab. 1).

## 4. Exploration de structure hiérarchique des concepts

### 4.1. *La tentative*

Pour résoudre le problème concernant la relation général-specifique entre concepts dans le document et dans la requête (par exemple *lésion de peau* et *dermatite*), ni la méthode d'expansion de requête dans le modèle de l'espace vectoriel ni les mesures des distances sémantiques ne prouvent des effets positifs. C'est la raison pour laquelle nous étudions une autre approche qui peut prendre en compte des liens entre les index. L'approche Bayesien est un candidat potentiel par sa base de théorie nette et par son efficacité dans la résolution des problèmes concernant l'incertitude, par exemple la RI.

L'application du réseau Bayesien dans la recherche d'information n'est pas nouvelle. Elle représente l'extension du modèle probabiliste et permet l'intégration des connaissances (ou source d'évidence comme les requêtes anciennes, la rétroaction de pertinence, ect.) dans un seul cadre. Les modèles Bayesiens dans la littérature diffèrent dans les points de vue qui entraînent des descriptions différentes des éléments du modèle. Un des premiers travaux réussis, le modèle de *réseau d'inférence* (Turtle *et al.,* 1991), attire beaucoup d'attentions pour ce modèle. Dans leur point de vue, la RI est une inférence ou un processus de raisonnement dans lequel nous estimons la probabilité qu'un document (vue comme une évidence) satisfait le besoin d'information d'utilisateur. Cette approche montre la capacité d'englober les autres modèles (comme le modèle probabiliste, Booléen, schéma de pondération tf.idf), en plus montre la performance de RI en combinant différentes sources d'évidence et différentes formulations de requêtes.

### 4.2. *Proposition du modèle basé sur le réseau Bayesien*

Le *Réseau Bayesien*(RB), appelé aussi *réseau de croyance, réseau graphique* ou *réseau causal*, est un graphe acyclique orienté (GAO) dont les noeuds sont les variables aléatoires et les arcs représentent les relations de cause-effet ou dépendances entre les noeuds qu'ils lient. Le RB peut décrire qualitativement des dépendances entre les variables (via le graphe causal) et quantitativement ces dépendances (via les probabilités conditionnelles).La probabilité conditionnelle d'une variable est la probabilité calculée en prenant en compte des évidences. Par exemple la probabilité conditionnelle de la variable $A$, noté $P(A|B)$ est la probabilité de A sachant B.

Nous adoptons le point de vue de Croft et l'étendons pour construire notre modèle Bayesien : la RI est donc une inférence ou un processus de raisonnement des documents vers requête via les liens sémantiques entre leurs concepts. Comme le modèle Bayesien peut naturellement décrire qualitativement et quantitativement les concepts et leurs relations dont les influences sont de nature incertaine, il est adapté pour notre tentative vers un modèle d'indexation à base des concepts et leurs relations extraites dans UMLS.

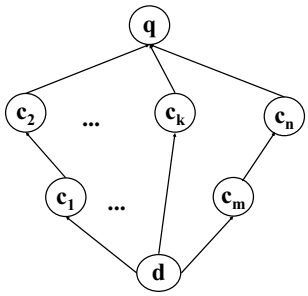


**Figure 1.** *Le modele du reseau Bayesien*

Notre réseau Bayesien contient : l'ensemble des noeuds représentent des documents $D$ ; l'ensemble des noeuds représentent des requêtes $Q$ ; l'ensemble des noeuds représentent des concepts associés aux $D$ ou $Q$. Avec le but de prendre en compte les relations du type général-specifique entre concepts des requêtes vers concepts des documents, nous proposons d'ajouter les liens de concept des documents vers concept de la requête s'il existe ce type de relation dans UMLS. (Fig. 1).

La valeur de pertinence d'un document $d$ par rapport à une requête $q$ est la probabilité conditionnelle du noeud $q$ sachant $d$ :

$$P(q|d) = \prod_{c_k \in pa(q)} P(c_k|d)$$

Avec $c_k$ est un concept dans l'ensemble de concepts associés à $q$, note $pa(q)$, considérés comme les pères de $q$.

Pour les noeuds associés directement au document $d$, quand $d$ est observé ou son état est "vrai ", la probabilité des noeuds associés sera prédéfinie par une constante ou par une fonction de leur distribution globale ou locale.

Pour les autres noeuds $c$ qui a un ensembe de $n$ parents ($pa(c)$), chacun a deux états (vrai, faux), $c$ a donc un ensemble de $2^n$ configuration, noté $\pi(c)$. On a :

$$P(c) = P(c|\pi(c)) = \sum_{i=1}^{2^n} P(c|\pi_i(c)) \times P(\pi_i(c))$$

et en suposant que les variables dans la configuration $\pi_i(c)$ sont independants l'un à l'autre, on a :

$$P(\pi_i(c)) = \prod_{c_k \in \pi_i(c)} P(c_k) \times \prod_{\neg c_l \in \pi_i(c)} P(\neg c_l)$$

**Tableau 1.** *MAP results*

| RUN | MAP(%)-CLEF2005 | MAP(%)-CLEF2006 |
|---|---|---|
| Text | 17.25 | 17.76 |
| Concept | 17.54 | 18.18 |
| Multicouche | **22.01** | **26.46** |

où $c_k \in \pi_i(c)$ signifie que la variable $c_k$ prend la valeur "vrai" dans la configuration $\pi_i(c)$ de $c$ ; $\neg c_l \in \pi_i(c)$ signifie que la variable $c_l$ prend la valeur "faux" dans la configuration $\pi_i(c)$ de $c$. On peut déduire donc :

$$P(c) = \sum_{i=1}^{2^n} P(c|\pi_i(c)) \times \prod_{c_k \in \pi_i(c)} P(c_k) \times \prod_{\neg c_l \in \pi_i(c)} P(\neg c_l)$$

Afin de prendre en compte la relation sémantique entre deux concepts père-enfant dans la formule de probabilité conditionnelle (dans ce cas entre $c$ et son père $c_k$ qui est observé), nous proposons donc :

$$P(c) = \sum_{i=1}^{2^n} \prod_{c_k \in \pi_i(c)} (\alpha \times P(c_k)) \times \prod_{\neg c_l \in \pi_i(c)} P(\neg c_l)$$

où $\alpha$ est une paramètre prédéfini dépendant du type de relation qui lie $c$ avec son père $c_k$. Cette valeur décrit l'influence de la dépendance sémantique entre ces deux concepts et est à régler expérimentalement. Nous pouvons enfin calculer la valeur de pertinence des documents par rapport à la requête $P(q|d)$.

## 5. Expérimentation

La collection CLEF Image Médicale que nous utilisons pour expérimentation contient un total de 50.026 images avec les textes en format XML. Nous travaillons sur 25 requêtes pour CLEF2005 et 30 requêtes pour CLEF2006. Tab. 1 présente les résultats évalués par le MAP(Mean Average Precision) pour l'indexation par termes, par concepts et multicouche. La méthode d'indexation conceptuelle prouve une meilleure performance que l'indexation par termes. De plus, la prise en compte des étiquettes sémantiques des concepts pour l'indexation multicouche et le filtre augmente encore beaucoup plus la précision moyenne. Ces informations sont également intéressantes à intégrer dans le modèle Bayesien.

## 6. Conclusion

Nous avons présenté dans l'article nos études sur l'application de connaissances externes (e.g. UMLS) dans la RI. La conceptualisation est une première étape essentielle et montre que l'indexation conceptuelle dans RI est meilleure que l'indexation par termes. L'intégration des autres connaissances dans UMLS, les classifications des concepts ou les étiquettes sémantiques, nous permet une indexation multicouche performante. Cette méthode facilite la filtration en se basant sur la structure de la requête, ou contexte thématique de la collection. Afin d'explorer les structure hiérarchique des concepts, nous proposons le prototype du modèle basé sur le réseau Bayesien. Les travaux sur d'éventuelles expérimentations et améliorations de ce modèle sont à réaliser prochainement.

## 7. Bibliographie


Aronson A. R., « MetaMap : Mapping Text to the UMLS Metathesaurus », http ://mmtx.nlm.nih.gov/docs.shtml, July, 2006.

Aronson A. R., Rindflesch T. C., Browne A. C., « Exploiting a large thesaurus for information retrieval », *Proceedings of the RIAO 94 : Intelligent Multimedia Information Retrieval Systems and Management*, p. 197-216, 1994.

Baziz M., Boughanem M., Aussenac-Gilles N., « A Conceptual Indexing Approach based on Document Content Representation », *in* , F. Crestani, , I. Ruthven (eds), *CoLIS5 : Fifth International Conference on Conceptions of Libraries and Information Science , Glasgow, UK, 04/06/05-08/06/05*, Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag, Berlin Heidelberg, p. 171-186, juin, 2005.

Budanitsky A., « Semantic Distance in WordNet : An Experimental, Application-oriented Evaluation of Five Measures », 2001.

Gonzalo J., Verdejo F., Chugur I., Cigarran J., « Indexing with WordNet synsets can improve Text Retrieval », *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP*, Montreal, Canada, p. 38-44, 1998.

Mandala R., Tokunaga T., Tanaka H., « Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion », *Research and Development in Information Retrieval*, p. 191-197, 1999.

Resnik P., « Using Information Content to Evaluate Semantic Similarity in a Taxonomy », *IJCAI*, p. 448-453, 1995.

Salton G., « The Smart Project in Automatic Document Retrieval », *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, Chicago, Illinois, United States, p. 356 - 358, 1991.

Turtle H., Croft B. W., « Evaluation of an Inference Network-Based Retrieval Model », *ACM Transactions on Information Systems*, vol. 9, n° 3, p. 187-222, 1991.