Exploitation des connaissances d'UMLS pour la recherche d'information médicale Vers un modèle bayésien 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 étudiée, mais avec peu de succès. Dans cet article, nous étudions l'impact de l'exploration d'une base de connaissance, nommée méta thésaurus UMLS pour la recherche d'information médicale. D'abord, l'indexation par concepts d'UMLS extrait dans des textes ne montre qu'une légère amélioration de MAP(Mean Average Precision) par rapport à l'indexation par termes. Nous intégrons ensuite les étiquettes sémantiques des concepts dans une indexation multicouche qui donne des résultats encourageants pour la collection ImageCLEF 2006. Pourtant, nous notons un problème de rappel dans les cas où le concept de la requête est plus général que celui des documents. Afin de résoudre ce problème, nous proposons un modèle basé sur un réseau bayesien pour capturer les liens générique-spécifique 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 the exploitation of a knowledge source, named UMLS meta-thesaurus, in medical domain information retrieval. First, indexing with UMLS concepts extracted from text shows just a slight increase in MAP (Mean Average Precision) comparing to word-based indexing. We then integrate semantic labels of concepts in a multi-layer indexing which shows quite encouraging results for ImageCLEF 2006 test collection. However, a recall problem is noticed when query concepts are more general than document concepts. In order to resolve this problem, we propose a Bayesian network based model which captures 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.*

CORIA 2007, pp. 443–448
Saint-Étienne, 28–30 mars 2007

1. Introduction

L'utilisation des bases de connaissances externes aux documents dans la recherche d'information (RI) a déjà été largement explorée. L'identification de concepts dans un texte pour l'indexation conceptuelle nécessite de résoudre l'ambiguïté des termes (Baziz *et al.,* 2005), (Aronson *et al.,* 1994). Ce problème est plus sensible pour des bases de connaissances générales comme Wordnet que pour celles d'un domaine spécifique comme le domaine médical. Notre travail montre d'abord que dans le domaine médical, l'indexation par concept est légèrement meilleure que indexation par termes avec relativement peu de désambiguïsation. Une question demeure cependant : "Comment utiliser de manière efficace les connaissances autres que simplement les concepts de la base de connaissances ?" Pour augmenter la performance en terme de précision, nous montrons que la classification des concepts peut être utilisée efficacement dans une indexation multicouche. Une expérimentation est conduite sur la collection CLEF Image Médical 2006[1]. La prise en compte dans la fonction de correspondance des liens sémantiques entre les concepts de la requête et ceux des documents (ex : hyperonymies) a été étudiée pour résoudre des problèmes liés au rappel. D'ailleurs, ces types de relations ont déjà été pris en compte dans des travaux précédents : il s'agit de l'expansion de requête (Gonzalo *et al.,* 1998), (Mandala *et al.,* 1999) ou le calcul de correspondance basé sur distance sémantique entre concepts (Budanitsky, 2001). Malheureusement ces méthodes ne donnent pas de bons résultats. Nous proposons donc un modèle bayesien basé sur la structure hiérarchique des concepts de UMLS [2](Unified Medical Language System) pour prendre en compte les liens sémantiques dans le calcul de la valeur de similarité entre documents et requêtes.

2. Indexation conceptuelle

L'indexation conceptuelle permet d'exprimer le contenu d'un texte à un plus haut niveau d'abstraction. Elle établit aussi un lien entre différentes formes de surfaces linguistiques et le sens. Une indexation par concept permet de s'affranchir de la barrière de la langue pour un SRI multilingue. La ressource conceptuelle que nous utilisons, appelée UMLS, est un méta thésaurus de taille importante. Ce dernier est la fusion de 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, NLM nous offre l'outil Metamap (Aronson, 2006) qui produit tous les concepts candidats correspondants à toutes les formes textuelles en Anglais. Pour les textes en Français et en Allemand, nous avons utilisé un outil similaire (Radhouani *et al.,* 2006). Ces outils comprennent les étapes suivantes : extraction des syntagmes, générations des concepts candidats correspondant aux composants des syntagmes et proposition des meilleurs concepts. L'indexation conceptuelle nous offre donc un ensemble de concepts servant d'index pour des textes dans différentes langues.

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

3. Exploitation des étiquettes sémantiques des concepts

3.1. *Indexation multicouche*

Dans UMLS, les concepts sont classés en 135 *types sémantiques*. Ces types sémantiques sont encore divisés en 15 *groupes sémantiques* (par exemple anatomie, pathologie etc.), considérées comme les abstractions de plus haut niveau des concepts. Nous appelons ces groupes sémantiques, les *étiquettes sémantiques* des concepts. Ces étiquettes nous permettent de mieux évaluer l'importance des concepts par rapport au thème général du document. Par exemple pour les documents qui décrivent les maladies sur une partie du corps, les concepts avec l'étiquette sémantique "anatomie" et "pathologie" sont plus importants que les autres dans la contribution au contenu de ces documents. Afin de prendre en compte ces informations, nous proposons une indexation multicouches : la couche des concepts et la couche de leurs étiquettes sémantiques. L'un est la projection de l'autre en se basant sur les liens dans UMLS. La pondération et la correspondance entre document et requête dans chaque couche sont exécutées séparément. Chacune produit des valeurs de pertinence RSV (Relevant Status Value) pour chaque document d_i par rapport à la requête q_j. Le RSV final est obtenu par une fonction de combinaison f des RSV des deux couches :

$$RSV(d_i, q_j) = f(RSV_1(d_i, q_j), RSV_2(d_i, q_j))$$

Nous pensons que la correspondance des étiquettes sémantiques est très importante. Nous définissons donc la fonction f comme le produit. La couche des étiquettes sémantiques est alors utilisée comme une re-pondération du résultat de la couche des concepts. La fonction f définit comme un produit peut également s'interpréter comme un filtre sur les étiquettes sémantiques des concepts dans les documents retrouvés par la couche des concepts. Cette approche s'apparente au filtrage par "dimension" présenté dans (Radhouani *et al.,* 2006). La principale différence est dans le modèle utilisé pour la prise en compte des étiquette sémantiques, en particulier leur pondération.

3.2. *Expérimentation*

Une expérimentation a été conduite sur la collection CLEF Image Médicale qui contient un total de 50026 images avec les textes en format XML, avec 25 requêtes pour 2005 et 30 requêtes pour 2006. Nous proposons des schémas de pondération différents dans les deux couches : $tf.idf$ pour la couche des concepts et binaire pour l'autre. En supposant que le triplet "anatomie", "pathologie" et "modalité" sont les étiquettes sémantiques les plus importantes dans les requêtes, nous diminuons l'espace d'indexation des étiquettes sémantiques à ce triplet. Cela permet une re-pondération efficace et un filtre plus fort sur ces étiquettes sémantiques. Avec cette méthode, nous obtenons la meilleure performance dans le forum d'évaluation CLEF2006. Le tableau 1 présente les résultats évalués par le MAP (Mean Average Precision) pour l'indexation par termes, par concepts et par notre modèle multicouche. Les résultats montrent que l'indexation conceptuelle est meilleure que l'indexation par termes. De

Tableau 1. *MAP results*

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

plus, la prise en compte des étiquettes sémantiques des concepts dans l'indexation multicouche augmente de beaucoup la précision moyenne.

4. Exploitation de structure hiérarchique des concepts - Modèle basé sur le réseau bayesien

Dans le cas où le document pertinent ne contient que les concepts sémantiquement reliés avec ceux de la requête (par exemple *lésion de peau* dans la requête et *mélanome cutané* plus spécifique dans le document), cela conduit à un problème de rappel. Pour résoudre ce problème, ni la méthode d'expansion de requête dans le modèle vectoriel ni les mesures des distances sémantiques n'apportent d'améliorations. C'est la raison pour laquelle nous étudions une autre approche qui peut prendre en compte des liens entre les index. Le modèle bayesien est un candidat potentiel par son efficacité dans la résolution des problèmes concernant l'incertitude.

L'utilisation d'un réseau bayesien dans la recherche d'information n'est pas nouvelle. C'est une extension du modèle probabiliste et permet l'intégration des connaissances (requêtes anciennes, la rétroaction de pertinence, etc.) dans un cadre unique. Le modèle de *réseau d'inférence* (Turtle *et al.,* 1991) est un des tout premier modèle et a beaucoup attiré l'attention sur le modèle bayesien pour la RI. Selon le point de vue de Turtle et Croft, 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 de l'utilisateur. Ce modèle montre la capacité d'englober les autres modèles (modèle probabiliste, Booléen, vectoriel) et de meilleures performances en combinant différentes sources d'évidence et différentes formulations de requêtes.

Le *réseau bayesien* (RB), appelé aussi *réseau de croyance, réseau graphique* ou *réseau causal*, est un graphe acyclique orienté (GAO) dont les nœuds sont les variables aléatoires et les arcs représentent les relations de cause-effet ou dépendances entre les nœuds qu'ils relient. 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 correspondance est donc une inférence ou un processus de raisonnement

des documents vers les requêtes via les liens sémantiques entre leurs concepts. Comme le modèle bayesien peut naturellement décrire qualitativement (via le graphe) et quantitativement (via les probabilités conditionnelles) les concepts et leurs relations avec des influences de nature incertaine, il nous semble adapté à un modèle d'indexation base des concepts. Notre réseau bayesien contient : l'ensemble des nœuds représentant les documents D ; l'ensemble des nœuds représentant la requête Q ; l'ensemble des nœuds représentant des concepts associés à D ou à Q. Dans le but de prendre en compte les relations de type générique-spécifique entre les concepts des requêtes vers les concepts des documents, nous proposons d'ajouter un lien partant des nœuds concepts des documents vers les nœuds concepts de la requête s'il existe ce type de relation entre les concepts dans UMLS. Le RSV d'un document d par rapport à une requête q est la probabilité conditionnelle du nœud q sachant d :

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

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

Pour les nœuds concepts associés directement au document d, quand d est observé (i.e. son état est "vrai"), leur probabilité sera prédéfinie par une constante ou par une fonction de la distribution globale ou locale de ces nœuds concepts dans la collection. Pour les autres nœuds c qui ont n parents ($pa(c)$), comme chacun a deux états possibles (vrai, faux), c a un ensemble de 2^n configuration, noté $\pi(c)$, donc :

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

et en supposant que les variables dans la configuration $\pi_i(c)$ sont indépendantes les unes des autres, on a :

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

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 :

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

Afin de prendre en compte la relation sémantique entre concepts père-enfant dans la formule de probabilité conditionnelle, nous proposons donc :

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

où α est un paramètre prédéfini dépendant du type de relation qui lie c avec son père c_k qui est observé. 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. Conclusion

Nous avons présenté l'état actuel de notre travail 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 le filtrage en se basant sur la structure de la requête, ou contexte thématique de la collection. Afin d'exploiter les structures hiérarchiques des concepts, nous proposons un modèle basé sur le réseau bayesien. Nos premières expérimentations ne sont pas encore satisfaisantes et doivent conduire à des améliorations de ce modèle.

6. 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., Chrisment C., « Semantic cores for representing documents in IR », *SAC '05 : Proceedings of the 2005 ACM symposium on Applied computing*, ACM Press, New York, NY, USA, p. 1011-1017, 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.

Radhouani S., Maisonnasse L., Lim J.-H., Le T.-H.-D., Chevallet J.-P., « Une Indexation Conceptuelle pour un Filtrage par Dimensions, Expérimentation sur la base médicale ImageCLEFmed avec le méta thésaurus UMLS », *COnférence en Recherche Information et Applications CORIA'2006, Lyon France*, 15 – 17 mars, 2006.

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.