
Une Indexation Conceptuelle pour un Filtrage par Dimensions

Expérimentation sur la base médicale ImageCLEFmed avec le méta thésaurus UMLS

Saïd Radhouani*** — Loïc Maisonnasse** — Joo-Hwee Lim* — Thi-Hoang-Diem Le** — Jean-Pierre Chevallet*

* IPAL, CNRS and Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613

** Equipe MRIM, Laboratoire CLIPS-IMAG
38 041 Grenoble cedex 9, France.

*** Centre universitaire d'informatique - Université de Genève
24, rue Général-Dufour, CH-1211 Genève 4, Suisse

Said.Radhouani@cui.unige.ch

RÉSUMÉ. Dans le but de résoudre des requêtes multi-dimensions, nous proposons une indexation conceptuelle à l'aide d'un méta thésaurus médical (UMLS). Nous étudions l'impact de cette indexation par rapport à une indexation à base de mots. Nous montrons que l'usage du méta thésaurus est délicat à mettre en oeuvre mais peut donner des résultats supérieurs à une indexation par mots. Nous définissons ensuite la notion de dimensions des requêtes. En utilisant une organisation hiérarchique des concepts du méta thésaurus, nous proposons une technique simple pour filtrer le corpus en fonction des dimensions de la requête. Les expérimentations, menées sur la collection ImageCLEFmed 2005, montrent que l'indexation conceptuelle avec la prise en compte des dimensions de la requête améliorent les résultats de 16%.

ABSTRACT. In order to resolve multi-dimensional queries, we propose a conceptual indexing based on a medical meta thesaurus (UMLS). We study the impact of this indexing compared to a words-based indexing. We show that using a meta thesaurus is delicate to set up, but can give better results than words-based indexing. Then, we define the notion of query dimensions. Exploiting the hierarchical organization of concepts of the meta thesaurus, we propose a simple technique to filter the corpus according to the query dimensions. We have performed evaluation on the ImageCLEFmed 2005 collection. Using conceptual indexing and taking into account the query dimensions, we improve results about 16%.

MOTS-CLÉS : Indexation conceptuelle, Requêtes multi-dimensions, Thésaurus, UMLS.

KEYWORDS: Conceptual indexing, multi-dimensional queries, Thesaurus, UMLS.

1. Introduction

Dans les domaines spécifiques tels que la médecine, les documents sont exprimés dans un vocabulaire technique spécifique précis et peu ambigu. Le but d'un praticien est d'encoder des informations précises, en minimisant le risque d'erreur d'interprétation à l'aide de termes fortement porteurs de sens. Pour un domaine technique restreint, la reconnaissance des termes et éventuellement leur *conceptualisation*¹, doit alors provoquer un gain de précision pour la tâche de Recherche d'Information (RI). Nous proposons donc d'étudier l'impact de l'utilisation de connaissances externes sur la précision des réponses du système de recherche d'information (SRI). Nous validons cette approche sur la collection ImageCLEFmed 2005 qui contient des images annotées en trois langues et des requêtes multimédias : des images-exemples (positives et/ou négatives) et une courte description textuelle. La figure (1) est une des 25 requêtes de cette collection.



Figure 1. Exemple d'une requête extraite de la base ImageCLEFmed 2005

Dans cette requête, nous supposons que l'on recherche des images dont la *modalité* est *x-ray*, l'*anatomie* est un *fémur*, et dont la *pathologie* recherchée est *fracture*. Nous appelons ces catégories (modalité, anatomie et pathologie) les **dimensions** de la requête. Nous supposons alors qu'un document pertinent, par rapport à une requête avec des dimensions, est celui qui décrit correctement ces dimensions. Pour résoudre ce type de requêtes, nous proposons de réaliser une indexation par concepts et d'identifier leurs dimensions pour filtrer les documents. Pour réaliser cette indexation, voici les principaux problèmes auxquels nous nous attaquons :

- L'extraction des concepts d'un texte à indexer ;
- La définition de la notion de dimensions ;
- L'identification des dimensions dans les documents et les requêtes ;
- La prise en compte des dimensions dans un SRI.

Dans la section suivante, nous présentons l'état de l'art, puis (section 3) l'utilisation des ressources externes pour l'indexation conceptuelle et le filtrage par dimensions. Avant de conclure et présenter nos perspectives (section 5), nous décrirons nos évaluations expérimentales et analysons nos résultats (section 4).

1. transformation en concepts

2. L'usage de ressources externes aux documents

L'idée d'utiliser des ressources externes aux documents pour la RI a été largement explorée mais avec assez peu de succès. Les principales propositions concernent l'expansion de requêtes comme par exemple Voorhees [VOO 94] qui réalise cette expansion en utilisant des relations lexicales sémantiques ("synsets" de WordNet). Le problème majeur concerne l'ambiguïté, à savoir le choix des synsets qui contiennent les significations correctes des termes. Les évaluations expérimentales montrent une baisse du résultat même avec une désambiguïsation manuelle. Qiu et Frei [QIU 93] obtiennent des résultats positifs en choisissant les concepts qui sont sémantiquement liés à la requête entière plutôt qu'à chacun des ses termes. Bodner [BOD 96] combine des connaissances précises sur un domaine, et des connaissances universelles. L'impact de l'ambiguïté des termes a été étudié par exemple par Gonzalo et al. [GON 98] avec une désambiguïsation manuelle et l'introduction volontairement d'erreurs de désambiguïsation. Ils montrent ainsi que le système fonctionne mieux avec une indexation conceptuelle ayant moins de 30% d'erreurs de désambiguïsation. Baziz dans [BAZ 05], montre qu'une indexation basée sur une combinaison de concepts et de mots, améliore la qualité contrairement à une indexation basée seulement sur des concepts. Les auteurs expliquent cet échec par le trop faible recouvrement par la ressource (WordNet) sur le vocabulaire du corpus. Dans le domaine médical, plusieurs travaux ont utilisé UMLS pour l'indexation de documents médicaux [HER 01]. Le bénéfice d'une telle indexation n'est pas très nette, et c'est parfois en combinant encore une fois avec une indexation par mots qu'une légère amélioration peut être obtenue [ARO 94].

Nous pensons tout de même que l'utilisation de connaissances externes est une bonne solution pour disposer d'une représentation précise des documents et des requêtes, mais une organisation transversale est nécessaire pour permettre de définir des *dimensions* comme des points de vue de l'utilisateur pour structurer sa requête. C'est cette structuration qui peut rendre les résultats encore plus précis. Dans la suite, nous présentons l'utilisation d'une ressource externe pour l'indexation conceptuelle et la prise en compte des dimensions.

3. Indexation conceptuelle et filtrage par dimensions

Une indexation conceptuelle² de documents techniques n'est pas une idée nouvelle. Par exemple, le système RIME [BER 90] utilisait une analyse sémantique des comptes rendus médicaux produisant des structures de dépendances pour indexer les documents. La mise en place d'une approche de ce type butte sur plusieurs problèmes : la disponibilité des ressources, la complexité des traitements d'extraction, et de la fonction de correspondance. Une indexation sémantique nécessite la construction d'un

2. On peut parler "d'indexation sémantique" bien que ce terme soit plus général : une indexation conceptuelle se limite à l'ensemble des concepts décrivant le document, une indexation sémantique peut proposer une structure à ces concepts.

dictionnaire sémantique : il n'en existe pas couvrant tout le domaine de la médecine. L'analyse de toutes les phrases d'un document et leur transformation en structures sémantiques est actuellement une étape difficile et coûteuse en temps machine. Finalement, les techniques d'appariement entre structures sémantiques, pour résoudre l'appariement entre document et requête, sont à notre avis, actuellement dans un état non satisfaisant pour être utilisées à grande échelle, et leur supériorité pratique sur des structures plus simples³ n'a pas été clairement démontrée.

Pour toutes ces raisons, nous pensons que travailler sur une indexation par concepts est actuellement plus crédible et plus fructueux. Notre travail montre que l'indexation par les concepts d'un méta thésaurus, et la prise en compte de la notion de dimensions, surpassent une indexation basée sur les mots.

3.1. Ressources externes

Pour réaliser une indexation conceptuelle et manipuler des dimensions, nous avons besoin de ressources externes qui possèdent au moins une organisation par concepts et une structure (par exemple hiérarchique). Nous pouvons donc utiliser des dictionnaires terminologiques, des thésaurus ou des ontologies. Nous proposons alors de définir la notion de *dimensions* d'une requête relativement à la structure hiérarchique d'une ressource externe de la manière suivante : une dimension d'une ressource est un ensemble de sous arbres de la hiérarchie de concepts de cette ressource. L'union des sous arbres nous permet de re-définir un ensemble de concepts en utilisant la classification proposée. En effet, cet ensemble de concepts décrit une partie du domaine décrit par la ressource entière.

Il faut bien noter qu'il est illusoire qu'une classification hiérarchique contente tous les praticiens d'un domaine de spécialité, car toute classification contraint la réalité dans un point de vue toujours discutable. Ce point est important, car l'usage d'une classification pour l'indexation revient à forcer un point de vue à tout utilisateur d'un SRI. De manière pragmatique, si la structure est assez vaste, nous pensons qu'il nous sera très souvent possible de définir manuellement ces dimensions pour un large éventail de requête. Nos expérimentations vont actuellement dans ce sens.

3.2. Usage des concepts et des dimensions

Un concept est une entité abstraite qui unifie et résume un ensemble d'objets concrets ou mentaux par abstraction de traits communs pertinents. Par exemple, le concept de *siège* représente tout objet physique utilisé pour s'asseoir. En pratique, l'utilisation du mot "concept" n'est pas aussi précise et désigne plus simplement une signification particulière d'un terme. De son côté, un terme est un assemblage de mots ayant une signification précise dans un domaine particulier. Par exemple, le terme

3. comme des vecteurs

"moniteur" en informatique désigne le dispositif d'affichage. Les SRI se basent généralement sur la simple notion de *mots*. Lorsque l'on dispose d'une base terminologique, on peut alors se placer au niveau du terme. Remarquons qu'une indexation manuelle est toujours une indexation par termes, car elle fait référence à une liste terminologique qui, pour l'occasion, se nomme *liste d'autorité*.

Lors d'une *indexation conceptuelle*, on s'abstrait des mots et des termes pour ne garder que les concepts. En pratique, la mise en oeuvre d'une telle indexation est très difficile à atteindre, d'une part, par le manque de ressources disponibles pour associer à des textes non plus des mots, mais des concepts, et d'autre part, par les processus complexes qu'il faut automatiser pour déduire les bons concepts à partir des textes. En pratique, dans les ressources utilisées en RI (ex : WordNet, MeSH, UMLS), les distinctions entre concepts et leurs instances, et entre les différents types de relations sont rarement explicites. Par exemple, UMLS contient et décrit "officiellement" des concepts mais en fait, il s'agit plutôt d'une fusion de thésaurus et de lexiques avec regroupement de sens. Nous pouvons tout de même supposer que ces concepts explicites dans UMLS et absent des ressources qui la composent, est un plus. Mais relativisons le problème car, en RI, une notion de concept plus lâche est, peut être, suffisante pour parler d'indexation conceptuelle. En pratique, le simple fait de généraliser un ensemble de termes sous le même "concept", peut être suffisant pour apporter un plus à l'indexation. On s'affranchit à la fois du phénomène de la variation, de la synonymie et éventuellement, du multilinguisme (si la ressource utilisée est elle même multilingue).

Dans notre travail, nous faisons donc l'hypothèse qu'une structure telle qu'UMLS est tout de même suffisante pour augmenter la précision de l'indexation. De plus, une hiérarchie a été définie et se place au dessus de tous les thésaurus d'UMLS. C'est cette hiérarchie que nous utilisons pour la notion de dimensions. Pour une indexation conceptuelle avec dimensions, il nous faut alors :

- Une ressource externe qui associe un ensemble de termes à un concept ;
- Un outil de sélection des concepts à partir des textes. Cet outil doit résoudre les ambiguïtés, mais aussi tenir compte des variations des termes ;
- Une structure sur la ressource pour fonder la notion de dimensions ;
- Un modèle de correspondance au niveau des concepts, et un système de pondération qui tiennent compte des dimensions.

Notre indexation conceptuelle consiste à sélectionner, relativement à la ressource choisie, un ensemble de concepts pour chaque document. Cette extraction est décrite dans la partie 4.2. Cette étape substitue à certains mots des documents un identifiant de concepts. Les mots outils ou les mots non reconnus sont simplement supprimés. Le document résultant est alors une séquence d'identifiants de concepts. Nous proposons d'utiliser le modèle vectoriel et de constituer comme index des *vecteurs de concepts*.

3.3. Interrogation par filtrage de dimensions

Même si l'indexation conceptuelle nous permet d'avoir une description plus précise des documents et des requêtes, nous pensons qu'elle n'est pas suffisante pour résoudre des requêtes multi-dimensions. En effet, le modèle vectoriel considère les documents (requêtes) comme des "sacs de concepts", et par conséquent, ne permet pas de prendre en compte les dimensions de la requête. Pour dépasser cette limite, nous proposons d'utiliser les dimensions pour effectuer un filtrage des documents lors de l'interrogation. La séparation en dimensions de la requête consiste à construire différentes sous-requêtes, chacune correspondant à une dimension. Les dimensions sont définies à partir d'une partition de la ressource externe. La connaissance de l'identificateur d'un concept suffit pour déterminer sa dimension. Soit une requête $Q = c_1, \dots, c_n$ constituée d'un ensemble de concepts. L'extraction des dimensions à partir de Q consiste à répartir chaque concept c_i dans une sous requête Q_i en fonction de son appartenance à la dimension i . Le processus d'interrogation est basé sur une extension du modèle que nous avons proposé dans [GUY 05]. La pertinence d'un document par rapport à une requête Q est calculée en deux étapes :

- *Le filtrage* sélectionne les documents qui contiennent les dimensions de la requête.
- *Le classement* organise les documents filtrés par dimensions, en ordre de pertinence.

Le filtrage par dimensions consiste à ne conserver que les documents répondant à un critère booléen de présence de dimensions. Nous l'effectuons en filtrant toute la base des documents par les dimensions de la requête. Ainsi, pour chaque dimension i de la requête Q , nous construisons une requête booléenne Q_i par disjonction de tous les concepts de la dimension i présents dans la requête Q . Par interrogation booléenne de la base par chaque Q_i , nous obtenons les sous ensembles D_i du corpus de la collection où chaque document d de D_i contient au moins un concept de la dimension i de la requête Q . Les documents appartenant à ces sous ensembles sont considérés précis parce qu'ils contiennent explicitement les concepts des dimensions de la requête initiale. Deuxièmement, pour résoudre la requête multi-dimensions d'origine, nous combinons ses dimensions à l'aide d'une expression booléenne sur les dimensions i de la requête. Une conjonction de dimensions force ces dimensions à être présentes ensembles dans le document, c'est à dire, assure qu'au moins un concept de chaque dimension est présent. Une disjonction impose simplement la présence d'un concept d'une des dimensions disjointes. Ces calculs sont réalisés en pratique sur les ensembles D_i .

Par exemple, pour une requête Q contenant trois dimensions, les sous-requêtes Q_1 , Q_2 , et Q_{d3} produisent les ensembles D_1 , D_2 et D_3 . Si nous décidons qu'un document pertinent doit inclure la dimension 1 et la dimension 2 ou la dimension 3, nous calculons le sous-ensemble de documents final filtré D_f en utilisant la formule $(D_1 \cap D_2) \cup D_3$. Après ce filtrage, il ne reste plus qu'à ordonner le sous-ensemble de document D_f .

Nous organisons l'ensemble de documents D_f par ordre de pertinence à l'aide d'un modèle vectoriel appliqué sur la représentation conceptuelle des documents et des requêtes. L'ordre⁴ est alors calculé classiquement comme la similarité entre la requête Q et chaque document de D_f

4. Evaluations expérimentales

4.1. Le corpus et les données

Nous présentons ici les résultats que nous avons obtenus sur la collection ImageCLEFmed 2005. Nous nous sommes servis du système expérimental d'indexation XIOTA [CHE 04]. En tant que partie du "*Cross Language Evaluation Forum*" (CLEF), la piste ImageCLEF 2005, concernant la recherche multilingue d'images, inclut une tâche de recherche d'images médicales (MedIR). La collection de test contient 50.026 images avec des annotations en format XML. La majorité des annotations sont en anglais mais un nombre significatif est en français et en allemand, avec quelques cas sans aucune annotation. Les 25 requêtes de la base ImageCLEFmed 2005 ont été formulées avec des images-exemples et de courtes descriptions textuelles.

Nous avons utilisé le méta thésaurus UMLS à la fois pour l'indexation conceptuelle et également comme référence pour la définition des dimensions. UMLS (Unified Medical Language System) est le résultat de la fusion de 140 sources de données terminologiques (UMLS knowledge sources) du domaine médical. Tous ces concepts sont organisés en 135 catégories, appelées *types sémantiques* dans le *Semantic Network*. Cette structure est un ajout dû à la fusion des thésaurus. Elle permet de "couvrir" cette fusion d'une classification hiérarchique. C'est précisément cette structure que nous utilisons pour la notion de dimensions.

4.2. Mise en oeuvre de l'indexation conceptuelle

Nous comparons les résultats des méthodes proposées à une méthode de référence. Cette méthode se base sur une indexation des lemmes obtenus avec TreeTagger⁵ après un filtrage sur leur type syntaxique. Sur l'index, nous appliquons deux pondérations : la première est une variation du tf-idf (l_{tc}), la seconde est la "divergence from randomness" (DFR) [AMA 02]. En utilisant tous les lemmes⁶, les pondérations l_{tc} et DFR fournissent respectivement une précision moyenne (MAP) de 0.1543 et de 0.1797.

4. On peut noter que nous obtenons un résultat identique avec un filtrage booléen par dimensions effectué *après* l'interrogation vectorielle. Le filtrage en amont a simplement l'avantage de réduire le nombre de documents lors de l'interrogation vectorielle.

5. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

6. excepté le mot image qui est plus proche des méta données que du thème

4.2.1. Méthodes et Hypothèses

Dans la littérature, il existe de nombreuses méthodes pour extraire les concepts. Certaines utilisent des traitements de la langue pour détecter les mots et les syntagmes [ARO 01]. D'autres, plus combinatoires [ZOU 03, CHE 01, HER 98], se basent sur la cooccurrence dans les phrases, des mots composants les termes. Nous partons de l'hypothèse que seuls les termes présents dans UMLS et retrouvés, avec seulement des variantes lexicales dans un texte médical, permettent d'identifier un concept. Cette hypothèse est restrictive car nous savons que les données terminologiques d'UMLS ne couvrent pas toutes les formes textuelles possibles. Pour établir, une association entre une chaîne de caractère et un concept, nous nous sommes basés sur le postulat que dans un texte médical, les concepts pertinents sont ceux représentés par les termes les plus long. Cette hypothèse est notamment utilisée dans les travaux de Baziz [BAZ 05]. Par exemple, dans la séquence "Images of right middle lobe", le concept pertinent à extraire doit correspondre à "right middle lobe" et non à "lobe". Ainsi, nous évaluons la présence dans UMLS des groupes de mots de chaque phrase dans l'ordre décroissant de leurs tailles. Avant d'effectuer cette correspondance, nous analysons les textes à l'aide de TreeTagger qui nous fournit des mots segmentés, étiquetés syntaxiquement et lemmatisés.

Comme UMLS ne contient pas toutes les formes textuelles possibles d'un concept, la correspondance stricte ne permet pas d'extraire tous les concepts. Pour dépasser cette limite, nous proposons de prendre en compte deux types de variations : la variation au niveau de la casse, et la variation au niveau lexical. Au niveau de la casse, nous utilisons deux méthodes. Dans la première, nous ne respectons pas la casse (tous les mots sont transformés en minuscule). Dans la seconde, nous respectons la casse dans le but de détecter correctement certains concepts (ex : ceux qui sont représentés par des acronymes). Au niveau lexical, chaque mot est sélectionné selon une liste de priorité : d'abord sous sa forme d'origine, ensuite sous sa forme lemmatisée.

Le fait de ne pas respecter la casse peut introduire du bruit au moment de l'extraction des concepts. En effet, certains mots outils deviennent équivalents à certains acronymes ou abréviations. Par exemple, l'article "a" est mis en correspondance avec la forme textuelle "A" du concept "Autopsie". Pour résoudre ce problème, nous faisons l'hypothèse qu'un mot outils seul ne peut être l'instance d'un concept.

Pour réduire le problème de l'ambiguïté, nous avons adapté le méta thésaurus UMLS à nos besoins. En effet, certains des thésaurus qui constituent UMLS sont trop précis ou bien portent sur un domaine non pertinent par rapport à notre application. Dans ce sens, Huang [LOW 03] a montré que la sélection de certaines ressources par type de rapport médical permet d'améliorer la détection des concepts. Ainsi, nous émettons l'hypothèse que la suppression de certains thésaurus permet de réduire le degré d'ambiguïté au sein même du thésaurus pour les besoins de la tâche globale. Partant de cette hypothèse, nous éliminons les types sémantiques utilisés dans UMLS dont le domaine n'est pas pertinent pour notre tâche. Par exemple, nous supprimons le type sémantique "Geographic Area" qui représente les éléments géographiques ayant une

frontière. Par cette suppression, la forme "CT" est correctement associée au concept "scanner aux rayons X", sans être associée au concept "Afrique centrale".

4.2.2. Résultats expérimentaux

L'objectif de ces expérimentations est d'évaluer nos hypothèses, ensuite faire une comparaison avec les résultats de l'indexation de référence. Les résultats sont présentés dans le tableau 1 où nous nommons :

F1 : le filtrage effectué sur les étiquettes syntaxiques ;

F2 : le filtrage effectué sur les types sémantiques ;

F3 : le filtrage de certains thésaurus.

Pour la variation de la casse, les résultats obtenus en utilisant une liste de priorité sur les formes en majuscule des termes, sont équivalents voir légèrement inférieurs à ceux obtenus sans les variations. Nous notons aussi une forte diminution des résultats obtenus par la pondération DFR. S'il semble difficile de pouvoir dire quelle méthode effectue la meilleure correspondance entre les formes textuelles et leurs concepts, d'un point de vu RI, la suppression de la casse est plus simple à mettre en œuvre et semble donc plus intéressante.

Nous remarquons aussi la non détection, au niveau lexical, de termes qui pourrait être associé à des concepts. Ce type d'erreur provient de l'analyse lexicale de TreeTagger qui ne permet pas de retrouver les lemmes de tous les mots utilisés dans le corpus. Par exemple, le terme "angiograms", qui est présent dans une requête sous la forme pluriel, ne peut pas être associé au concept correspondant car UMLS ne contient que la forme singulier (angiogram) et TreeTagger n'est pas capable de retrouver le lemme correspondant à "angiograms". En effet, l'analyseur treeTagger est un analyseur général et donc non adapté au vocabulaire médical. L'utilisation d'un analyseur spécialisé sur le domaine pourrait améliorer les résultats.

Nous notons que les méthodes proposées pour réduire l'ambiguïté ont peu d'impact sur les résultats. Nous constatons également que lors de l'indexation basée sur les concepts extraits à l'aide des instances les plus longues, les résultats sont inférieurs à ceux obtenus à l'aide des mots. Cette baisse de performance s'explique par l'extrême précision des concepts extraits. En effet, des concepts comme celui correspondant à la forme "Right middle lobe" ou à "Chest CT" sont trop précis, par conséquent, leur utilisation à la place de leurs constituants entraîne une forte baisse du rappel. D'autres problèmes tels que la métonymie influe sur la correspondance entre les concepts.

Pour résoudre ces problèmes, nous avons relâché l'hypothèse qui privilégie les plus longues formes textuelles. Ainsi, à partir de chaque phrase, les concepts correspondants à toutes les formes textuelles sont extraits. Cette méthode a provoqué une amélioration de 50% du résultat, et par conséquent, a permis de surpasser les résultats de référence. Cette amélioration est la conséquence d'une augmentation du taux du rappel qui est dû à l'extraction de certains concepts plus généraux.

Nous avons ensuite amélioré ces résultats en appliquant le processus de filtrage de certains thésaurus. En particulier, nous remarquons une forte augmentation du ré-

Tableau 1. Impact en RI des opérations d'extraction des concepts

	sans casse	mise en majuscule	F1	F2	F3	tf-idf MAP(%)	DFR MAP(%)
Les plus longues formes textuelles						9.95	14.66
	X					10.01	6.34
	X		X			10.26	7.07
	X		X	X		10.13	7.03
	X		X	X	X	10.85	11.69
		X				9.52	6.2
		X	X			9.87	7.06
		X	X	X		9.76	7.06
Toutes les formes textuelles			X			15.66	11.2
	X		X	X		15.52	10.95
	X		X	X	X	15.41	18.27
		X	X			14.73	11.23
		X	X	X		14.67	11.33
		X	X	X	X	14.69	18.19

sultat lors de l'utilisation de la pondération DFR (66%). Enfin, les meilleurs résultats obtenus lors de l'indexation conceptuelle sont légèrement supérieurs aux résultats de référence.

En conclusion, malgré une détection incomplète des concepts, l'indexation conceptuelle permet d'obtenir une amélioration par rapport à l'indexation par mots. Extraire les concepts qui correspondent à toutes les formes textuelles indépendamment de leurs tailles, permet d'obtenir une meilleure couverture des concepts. Le peu de variations des termes dans la collection médicale semble avoir pour conséquence qu'obtenir les bons concepts n'a finalement pas autant d'impact que prévu, si l'erreur d'assignation est systématique. Cependant, cette erreur devient gênante lorsque l'on souhaite utiliser des informations sur ces concepts comme, par exemple, les relations. En particulier, l'assignation de la dimension peut devenir incorrecte.

4.2.3. Filtrage par dimensions

Le but de ces expérimentations est d'évaluer l'impact de la prise en compte des dimensions sur la précision moyenne du SRI. Pour les requêtes de ImagCLEFmed 2005, nous avons utilisé les dimensions Anatomie, Pathologie, et Modalité. Ces dimensions correspondent respectivement aux types sémantiques suivant de UMLS :

Anatomie : "Anatomical Structure", "Body System" , "Body Space or Junction", "Body Location or Region" ;

Pathologie : "Disease or Syndrome", "Finding", "Injury or Poisoning" ;

Modalité : "Diagnostic Procedure" , "Manufactured Object".

Tableau 2. Résultat du filtrage par dimensions sur des documents indexés par des concepts

	ltc		DFR	
	MAP	(%)	MAP	(%)
H1	15.60	+6%	18.15	-0.2%
H2	16.02	+9%	18.93	+4%
H3	16.17	+10%	18.88	+4%
H4	17.07	+16%	19.03	+5%

Les résultats obtenus dans la suite seront comparés à deux résultats de référence obtenus lors de l'indexation conceptuelle. Chacun de ces résultats de référence correspond à un schéma de pondération : 0.1469 de MAP pour le ltc, et 0.1819 pour le DFR. Dans la suite, chacun de ces deux résultats sera appelé *baseline*. Pour effectuer le filtrage par dimensions, nous avons fait quatre hypothèses en utilisant différentes combinaisons booléennes sur les dimensions de la requête. Les résultats obtenus sont présentés dans le tableau 2 où chaque ligne correspond à une hypothèse. Les valeurs représentent les résultats et leur variation par rapport au *baseline* correspondant. Les hypothèses sont les suivantes :

H1 : *Les documents pertinents doivent inclure au moins une des trois dimensions (si elles existent dans la requête).* Cette hypothèse, améliore le résultat pour ltc mais provoque une légère baisse du résultat pour le DFR.

H2 : *Les documents pertinents doivent contenir l'anatomie présente dans la requête.* En forçant uniquement la dimension "anatomie", nous obtenons un meilleur résultat (+9%) en ltc. Le résultat est meilleur qu'en forçant n'importe quelle dimension : les dimensions ne sont donc pas équivalentes. L'anatomie est importante probablement parce qu'elle est discriminante et non ambiguë, alors que la pathologie seule est plus ambiguë (ex : fracture du fémur, fracture du crâne, fracture du doigt, etc.).

H3 : *Les documents pertinents doivent contenir l'anatomie, ou sinon la pathologie, ou sinon la modalité.* Cette hypothèse propose un ordre d'importance sur les dimensions. Nous obtenons encore une augmentation de performance.

H4 : *Les documents pertinents doivent contenir l'anatomie et la pathologie.* L'amélioration dans ce cas est la plus forte (+16%) (ltc) et 5% (DFR).

Notre technique d'extraction des concepts et donc la reconnaissance des dimensions n'est pas totalement fiable⁷. Cela peut expliquer le déséquilibre d'efficacité entre les dimensions. Il se peut aussi que la modalité ne soit pas assez explicitée dans les documents, ce qui paraît normal car le compte rendu décrit une lésion sur un organe et l'information sur le type d'image est souvent implicite. Un calcul de dépendance entre les dimensions et les concepts (une sorte de "concept mining") sur une collection pourrait nous aider à trouver à priori la meilleure combinaison de filtrage. Par exemple,

7. Il faudrait vérifier manuellement l'extraction et estimer un pourcentage de fiabilité

si dans les textes l'anatomie "fémur", implique fortement la pathologie "fracture" et la modalité "radio", alors il est clair que c'est la dimension "anatomie" qu'il faut privilégier. Ces résultats nous permettent tout de même de montrer que la prise en compte des dimensions permet d'augmenter la précision moyenne du SRI. Il s'agit d'un complément d'information qui structure la requête, et donc la précise. Le résultat obtenu suite au filtrage par dimensions est également complémentaire à celui obtenu par l'indexation conceptuelle. En effet, la représentation conceptuelle a facilité l'identification des dimensions des requêtes. Le filtrage booléen sur ces dimensions est donc un moyen très simple de dépasser les limites de modèle vectoriel qui ne prend pas en compte les relations entre les concepts d'un vecteur.

5. Conclusion et perspectives

Dans cet article, nous avons expérimenté avec succès une indexation conceptuelle sur un corpus spécialisé. Même avec une technique assez simple d'identification des concepts, nous avons surpassé le modèle d'indexation à base de mots. Cette réussite face aux nombreux échecs du passé, s'explique à notre avis, par la large couverture de la ressource utilisée pour l'expérimentation (UMLS), qui en volume est même plus importante que la collection des documents. Comme l'identification des concepts est encore améliorable, nous pensons qu'il y a encore matière à de meilleurs résultats. La représentation conceptuelle nous a permis d'identifier les dimensions présentes dans les requêtes et de les prendre en compte au moment de l'interrogation. L'expérimentation montre clairement le bénéfice de la prise en compte de ces dimensions. Cette idée avait déjà été mise en oeuvre lors de la compétition de CLEF 2005 [CHE 05], cependant nous n'avons pas explicitement réalisé une indexation conceptuelle, mais un filtrage à l'aide d'une petite structure hiérarchique. Cette idée nous avait permis d'obtenir les meilleurs résultats lors de la compétition. Dans les expérimentations présentées ici, nous avons confirmé et consolidé cette approche et nous encourageons à continuer dans le domaine de l'indexation conceptuelle.

6. Bibliographie

- [AMA 02] AMATI G., VAN RIJSBERGEN C. J., « Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness », *ACM Transaction on Information Systems*, vol. 20, n° 4, 2002, p. 357-389.
- [ARO 94] ARONSON A. R., RINDFLESCHE T. C., BROWNE A. C., « Exploiting a Large Thesaurus for Information Retrieval », *Proceedings of RIAO*, New York, October 1994, p. 197-216.
- [ARO 01] ARONSON A. R., « Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program », *AMIA 2001 Annual Symposium on biomedical and health informatics*, 2001, p. 17-27.
- [BAZ 05] BAZIZ M., BOUGHANE M., AUSSENAC-GILLES N., « Conceptual Indexing Based on Document Content Representation », FABIO CRESTANI I. R., Ed., *Informa-*

tion Context : Nature, Impact, and Role : 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005, vol. 3507, Lecture Notes in Computer Science, Jan 2005, p. 171–186.

- [BER 90] BERRUT C., « Indexing medical reports : The RIME approach », *Inf. Process. Manage.*, vol. 26, n° 1, 1990, p. 93–109, Pergamon Press, Inc.
- [BOD 96] BODNER R. C., SONG F., « Knowledge-Based Approaches to Query Expansion in Information Retrieval », *AI '96 : Proceedings of the 11th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, London, UK, 1996, Springer-Verlag, p. 146–158.
- [CHE 01] CHEN P. N. R., BRANDT C., « UMLS concept indexing for production databases : a feasibility study. », *Journal of the American Medical Informatics Association : JAMIA.*, 2001, p. 80–91.
- [CHE 04] CHEVALLET J.-P., « X-IOTA : An Open XML Framework for IR Experimentation Application on Multiple Weighting Scheme Tests in a Bilingual Corpus », *Lecture Notes in Computer Science (LNCS), AIRS'04 Conference Beijing*, vol. 3211, 2004, p. 263–280.
- [CHE 05] CHEVALLET J.-P., LIM J.-H., RADHOUANI S., « Using Ontology Dimensions and Negative Expansion to solve Precise Queries in CLEF Medical Task », *CLEF Workhop, Working Notes Medical Image Track, Vienna, Austria*, 21–23 September 2005.
- [GON 98] 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, 1998, p. 38–44.
- [GUY 05] GUYOT J., RADHOUANI S., FALQUET G., « Ontology-Based Multilingual Information Retrieval », *CLEF Workhop, Working Notes Multilingual Track, Vienna, Austria*, 21–23 September 2005.
- [HER 98] HERSH W. R., DONOHOE L. C., « SAPHIRE International : A tool for cross-language information retrieval. », *Proceedings of the American Medical Informatics Association Annual Fall Symposium*, 1998, p. 673–677.
- [HER 01] HERSH W., MAILHOT M., ARNOTT-SMITH C., LOWE H., « Selective automated indexing of findings and diagnoses in radiology reports », *Comput. Biomed. Res.*, vol. 34, n° 4, 2001, p. 262–273, Academic Press Professional, Inc.
- [LOW 03] LOWE Y. H. H., HERSH W., « A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports. », *Proceedings of the conference of the American Medical Informatics Association*, 2003, p. 580-587.
- [QIU 93] QIU Y., FREI H.-P., « Concept based query expansion », *SIGIR '93 : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1993, ACM Press, p. 160–169.
- [VOO 94] VOORHEES E. M., « Query expansion using lexical-semantic relations », *SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1994, Springer-Verlag New York, Inc., p. 61–69.
- [ZOU 03] ZOU Q., CHU W. W., MORIOKA C., LEAZER G. H., KANGARLOO H., « Index-Finder : A Method of Extracting Key Concepts from Clinical Texts for Indexing », *AMIA 2003 Annual Symposium on biomedical and health informatics*, 2003, p. 763–767.