Une approche pour la Représentation Sémantique de Documents

Mustapha BAZIZ, **Mohand BOUGHANEM**, **Nathalie AUSSENAC-GILLES**
baziz@irit.fr , boughane@irit.fr, aussenac@irit.fr

IRIT
Campus Univ. Toulouse III
118 Route de Narbonne
F-31062 Toulouse Cedex 4, France

Mots clefs:
Recherche d'Information, Représentation Sémantique de Documents, Ontologies, WordNet

Keywords:
Information Retrieval, Semantic Representation of Documents, Ontologies, WordNet.

Palabras clave:
el buscar de la información, representación semántica de documentos, Ontologies, WordNet.

Résumé
Cet article traite de l'application des ontologies au domaine de la recherche d'information. L'objectif de l'approche est de représenter le contenu sémantique de documents. L'approche consiste à projeter les documents sur une ontologie linguistique générale, telle que WordNet. Il s'agit d'identifier pour chaque document les *représentants* de concepts de l'ontologie. Ces derniers peuvent être des mots simples ou des groupes de mots. Un critère de cooccurrence (CF.IDF) est utilisé pour extraire les concepts importants. Un deuxième critère qui est la similarité sémantique entre concepts, permet de les désambiguïser via le réseau sémantique de l'ontologie. Le résultat de ce "matching" entre le document et l'ontologie est un ensemble de concepts désambiguïsés (appelés aussi concepts-sens ou nœuds) avec des liens pondérés entre eux, formant ce que nous appelons le *noyau sémantique de document* qui représente au mieux le contenu sémantique du document. L'approche proposée peut être considérée comme une première étape vers l'objectif à long terme qui est l'indexation intelligente et la recherche sémantique.

1 Introduction

Les nouvelles approches de Recherche d'Information (RI) basées ontologie promettent d'améliorer la qualité des réponses puisqu'elles visent à capturer une partie de la sémantique des documents. Deux principales contributions peuvent alors être distinguées: l'expansion de requêtes et la représentation de documents. Dans le premier cas, des concepts d'une ontologie liés sémantiquement à la requête peuvent être ajoutés pour sélectionner plus de documents pertinents, notamment ceux utilisant un vocabulaire différent et traitant du même sujet [1], [2], [3], [4]. Dans la représentation sémantique de document, désigné souvent par le terme indexation sémantique ou conceptuelle [2] et [5], le but principal est d'identifier les concepts[1] caractérisant le contenu du document. Le challenge est de s'assurer que les concepts non pertinents ne seront pas pris et que les concepts pertinents ne seront pas omis.

Dans ce papier, nous proposons une méthode automatique de sélection de ces concepts à partir des documents en utilisant une ontologie linguistique générale, en l'occurrence la base lexicale WordNet [6]. Il s'agit de construire à partir d'un document donné, un réseau sémantique qui représente son contenu sémantique. Le réseau sémantique auquel nous faisons référence ici est celui défini par [7]: "*un réseau sémantique peut être décrit comme toute représentation reliant des nœuds avec des arcs, où les nœuds sont des concepts et les arcs représentent différents types de relations entre concepts*". Les nœuds du réseau auxquels sont affectés des scores, peuvent être utilisés dans le processus d'indexation et de recherche du Système de Recherche d'Information (SRI).

Le papier est organisé comme suit : la section 2 dresse un bref état de l'art sur l'utilisation de la sémantique en RI. La section 3 décrit notre approche consistant à "matcher" un document avec une ontologie : l'extraction de concepts (3.1), le calcul de similarité sémantique entre concepts (3.2) et la construction du noyau sémantique (3.3). l'évaluation de l'approche est décrite en section 4.

2 Utilisation de la sémantique en RI

Récemment, des approches ont été proposées pour utiliser la sémantique en RI. Dans ces approches, des groupes de mots, de noms, de groupes nominaux sont projetés sur les concepts qu'ils encodent [2]. Par ce modèle, un document est représenté comme un ensemble de concepts: pour cela, une étape cruciale serait le passage via une structure sémantique de simples mots clés comme représentants de documents aux concepts. Une structure sémantique peut être représentée en utilisant différentes structures de données : arbres, réseaux sémantiques, graphes conceptuels, etc. Ces structures peuvent être des dictionnaires, thesaurus ou ontologies [3]. Elles sont soit manuellement ou automatiquement générées ou alors[20], [21], elle peuvent pré exister. WordNet et EuroWordNet sont des exemples de thesaurus basés sur des bases de données lexicales. Elles sont assimilées à des ontologies et largement utilisés pour améliorer l'efficacité des Systèmes de RI. Gonzalo et ses collègues [9], proposent une méthode d'indexation basée sur les synsets de WordNet: le modèle d'espace vectoriel est utilisé, en prenant les synsets comme espace d'indexation au lieu des mots clés. Dans [10], Navigli et ses collègues proposent dans leur système (OntoLearn) une méthode appelée *structural semantic interconnection* pour désambiguïser les glossaires de WordNet (les définitions des concepts). Dans [2] et [22], des concepts d'une ontologie ("regions" pour le premier) sont connectés à ceux des documents (texte pour le premier et documents xml pour le deuxième). Pour ce faire, les deux auteurs proposent une méthode pour mesurer la similarité entre concepts via le réseau sémantique d'une ontologie.
La similarité entre concepts dans les réseaux sémantiques a été l'objet de divers travaux, différentes métriques et méthodes ont été proposées dans la littérature. Les principales sont de [11], [12], [13]. Resnik [13] utilise la sous hiérarchie formée par la relation is-a (généralisation) pour calculer la similarité entre deux concepts donnés. La mesure qu'il propose est basée sur le plus spécifique des concepts qui subsume les deux concepts à comparer. Pederson [14] utilise l'algorithme de Lesk [14] adapté, basé sur le recouvrement des glossaires de Wordnet pour désambiguïser les mots dans le

[1] Dans la suite, par abus de langage, nous ne ferons pas la différence entre les termes désignant un concept et le concept faisant partie du réseau conceptuel de référence.

texte(WSD). La plupart de ces mesures sont utilisées pour la désambiguïsation de mots dans le texte, Dans notre cas, la finalité n'est pas spécialement le WSD — pour le lecteur intéressé par ces méthodes, un état de l'art complet peut être trouvé dans [8] — cependant nous nous somme inspiré de la mesure de Pederson [14] pour calculer les similarités entre les concepts extraits du document.

3 Projeter un document sur une ontologie

L'approche proposée comme schématisée dans la Figure1 comprend trois (3) étapes. Dans l'étape (1), il s'agit d'extraire des documents les concepts les plus fréquents qui correspondent à des entrées dans l'ontologie. Dans l'étape (2), on se sert de l'ontologie dans un premier temps pour récupérer les différents sens possibles pour les concepts retenus. Puis, pour calculer les similarités entre les différents sens de ces concepts en se basant sur les relations sémantiques disponibles telles que la généralisation/spécialisation, les liens de composition, et les liens de domaines. Dans l'étape (3), le réseau sémantique est construit comme suit: d'abord, les concepts sont désambiguïsés: pour un concept donné, le sens dont la somme des liens avec les autres (score) est maximale est retenu. Il représentera un nœud dans le réseau. Pour les arcs (les liens), ils sont matérialisés par les valeurs de similarité entre les couples de concepts calculés à l'étape (2).

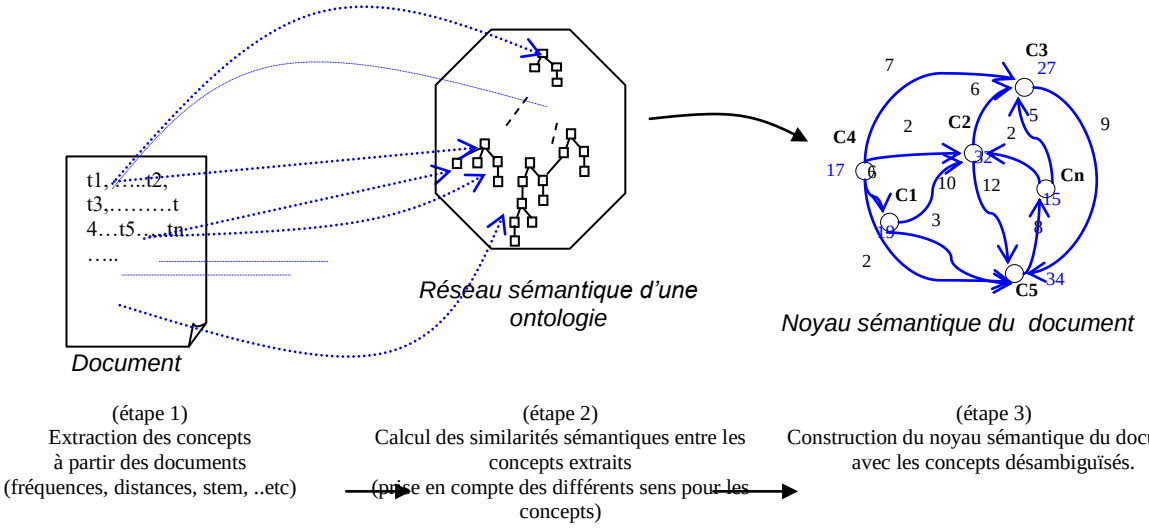

Figure1. *Schéma général de l'approche.*

Nous développerons ces différentes étapes dans les sections suivantes.

3.1 Extraction de concepts

Avant tout traitement, en particulier avant d'élaguer les mots vides du document, un processus important pour le reste des étapes consiste à détecter les représentants de concepts formés par des mots uniques ou des groupes de mots et correspondant à au moins une entrée (ou nœud) dans l'ontologie. Les concepts reconnus, peuvent être des groupes nominaux comme *pull_one's_weight* ou des entités nommées telles que *united_kingdom_of_great_britain_and_northern_ireland*. Pour cela, nous utilisons une technique ad-hoc qui consiste à concaténer des mots adjacents dans le texte pour identifier les concepts multi mots dans l'ontologie. Deux cas peuvent alors être distingués. Le premier consiste à projeter l'ontologie sur le document par l'extraction de tous les concepts multi mots de l'ontologie, puis, identifier ceux qui occurrent dans le document. Cette méthode a l'avantage d'être rapide et permet la construction d'une ressource réutilisable même si le corpus change. Son inconvénient est le risque d'omettre des concepts qui apparaissent dans le texte du document et dans l'ontologie dans des formes différentes. Par exemple, si l'ontologie contient le concept *solar battery*, une simple comparaison ne reconnait pas le même

concept sous sa forme plurielle, *solar batteries*. Le deuxième cas, que nous avons adopté dans ce papier, suit le chemin inverse, c-à-d projeter le document sur l'ontologie. Pour chaque concept candidat formé en combinant les mots adjacents dans le texte, on interroge d'abord l'ontologie en utilisant les mots tels qu'ils se présentent dans le texte, s'ils ne correspondent pas à des entrées dans l'ontologie, on utilise leurs formes de base. Ce qui permet de résoudre partiellement le problème de la morphologie des mots.

Concernant la combinaison des mots, le terme (multi mots) le plus long qui correspond à un concept est retenu. Si nous considérons l'exemple de la Figure2, la phrase contient trois (3) concepts différents qui sont: *abdominal muscle*, *external oblique muscle* et *abdominal external oblique muscle.*

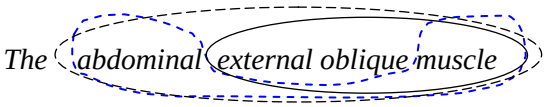

Figure2. *Exemple de texte avec différents concepts.*

Le premier concept *abdominal muscle*, n'est pas identifié car ces mots ne sont pas adjacents. Le deuxième, *external oblique muscle*, et le troisième sont synonymes. Ils appartiennent au même synset de WordNet (que nous désignons aussi par le terme nœud). Leur définition est :

external oblique muscle, musculus obliquus externus abdominis, abdominal external oblique muscle, oblique -- (a diagonally arranged abdominal muscle on either side of the torso)

Le concept sélectionné est associé au plus long terme *abdominal external oblique muscle* qui correspond au sens adéquat dans la phrase. Rappelons que dans la combinaison des mots dans le texte, l'ordre doit être respecté (de gauche à droite) autrement, on pourrait être confronté au problème de la variation syntactique (*science library* est différent de *library science*).

3.1.1 Pourquoi utiliser les concepts multi mots?

La détection de concepts multi mots dans le texte des documents est importante pour réduire l'ambiguïté. Ces concepts sont en général monosémiques même si les mots qui les composent peuvent être ambigus. Si nous prenons par exemple chaque mot séparément, dans le concept *ear_nose_and_throat_doctor*, nous aurons à désambiguïser entre 5 sens pour le mot *ear*, 13 sens (7 pour le nom et 6 pour le verbe) pour le mot *nose*, 3 sens pour *throat* (*and* étant un mot vide n'est pas utilisé) et 7 sens (4 pour le nom et 3 pour le verbe) pour le mot *doctor*. Nous aurons ainsi un nombre de 5x13x3x7= 1365 combinaisons possibles de sens. Mais quand on considère tout ces mots comme formant un seul concept (bien sur, le concept multi mots doit appartenir à l'ontologie), nous aurons seulement un sens. Dans cet exemple, le nœud (synset) de WordNet et sa définition (gloss) sont comme suit:

The noun ear-nose-and-throat doctor has 1 sense
1. ENT man, ear-nose-and-throat doctor, otolaryngologist, otorhinolaryngologist, rhinolaryngologist -- (a specialist in the disorders of the ear or nose or throat.)

Les statistiques que nous avons faites sur WordNet2.0, comme montré dans Table1, montrent que sur un total de 63218 concepts multi mots (composés de 2 à 9 mots), 56,286 (89%) d'entre eux sont monosémiques. 6238 ont 2 sens (9.867%) et seulement 694 (0.506%) concepts multi mots ont plus de 2 sens. Ainsi, plus il y a de concepts multi mots dans le document à analyser, plus facile est leur désambiguïsation.

Nombre de sens	Nombre de concepts multi mots (2-9 mots)	%
1	56286	89,035%
2	6238	9,867%
3	375	0,593%
>=4	319	0,506%
Total	63218	100%

Table1. *Répartition de la polysémie sur les concepts multi mots dans WordNet 2.0*

3.1.2 Pondération des concepts

Les concepts extraits sont alors pondérés selon une variante de TF.IDF que nous appelons CF.IDF. Pour un concept *c* composé de n mots, sa fréquence dans un document dépend du nombre d'occurrences du concept lui-même, et de celui de tout ses sous-concepts dérivés. Formellement:

$$cf(c) = count(c) + \sum_{sc \in sub\_concepts(c)} \frac{Length(sc)}{Length(c)} \cdot count(sc) \qquad (1)$$

Où *Length(c)* represente le nombre de mots dans *c* et *sub_concepts(c)* le nombre de tous les sous-concepts dérivés de *c*: concepts de n-1 mots de *c*, concepts de n-2,.. et tous les mots uniques de *c*. Par exemple, pour le concept *"elastic potential energy"* composé de 3 mots, sa fréquence est calculé comme suit:

f("elastic potential energy") *= count("elastic potential energy") + 2/3 count("potential energy") + 1/3 count("elastic") + 1/3 count("potential") + 1/3 count("energy").*

D'autres méthodes de calcul de fréquences sont proposées dans la littérature. Elles utilisent des analyses statistiques et/ou syntactiques [15], [16]. Elles consistent en général à additionner les fréquences des mots uniques, les multiplier ou multiplier le nombre d'occurrences du concept multi mots par le nombre de mots qu'il contient. Dans notre cas, nous avons utilisé la méthode pondération de l'équation (1). La fréquence globale d'un concept, *idf(C)*, est alors calculée comme suit:

$$idf(c) = f(c).\ln(N / df) \qquad (2)$$

N étant le nombre total de documents et df (document frequency) le nombre de documents où le concept *c* apparaît. Si le concept apparaît dans tous les documents, sa fréquence est nulle. Nous avons utilisé une fréquence seuil égale à 2 pour sélectionner les concepts.

Une fois les concepts importants sont extraits du document, ils sont utilisés pour construire le noyau sémantique de ce document. Comme chaque concept extrait peut avoir plusieurs sens, donc correspondre à plusieurs nœuds (synsets) dans l'ontologie, des mesures de similarité entre les différents sens des concepts sont calculées en vue de sélectionner pour chaque concept, le meilleur sens correspondant (nœud) dans l'ontologie. La mesure de similarité entre deux nœuds (concepts-sens) représente une valeur condensée résultant de la comparaison de deux concepts-sens en utilisant différentes relations de l'ontologie. Cette valeur n'a pas de sens précis mais exprime le degré du lien entre les deux concepts-sens en utilisant le réseau sémantique de WordNet. Nous l'explicitons dans la section suivante.

3.2 Calcul de similarité entre concepts

Étant donné un ensemble de relations de l'ontologie $\mathcal{R}=\{\mathcal{R}_1, \mathcal{R}_2, \ldots \mathcal{R}_n\}$, et deux concepts C_k et C_l Auxquels sont affectés deux sens j_1 et j_2: S_{j1}^{k} et S_{j2}^{l}. La similarité sémantique entre S_{j1}^{k} et S_{j2}^{l} , noté P_{kl} (S_{j1}^{k}, S_{j2}^{l}) ou $overlaps(S_{j1}^{k}, S_{j2}^{l})$ est définie comme suit:

$$P_{kl}(S_{j1}^{k}, S_{j2}^{l}) = \sum_{(i,j)\in\{1,..,n\}} \mathcal{R}_i(S_{j1}^{k}) \cap \mathcal{R}_j(S_{j2}^{l}) \qquad (3)$$

Elle représente l'intersection de Lesk adaptée (nombre de mots en commun) entre les informations retournés prar les relations R_i, quand elles sont appliquées aux concepts-sens S_{j1}^{k} et S_{j2}^{l}.

Exemple: Si on prend S_{j1}^{k}= *applied_science#n#1*, le sens 1 du concept *applied_science* et S_{j2}^{l}= *information_ science#n#1*, le sens 1 du concept *information_science*, et que R_1= *gloss*, R_2=*holonym-gloss* comme des relations connues (les définitions de *gloss* et *holonymie* sont données en bas), la similarité utilisant les deux relations *R1* et *R2* entre les deux concepts est égale à 11:

$\mathcal{R}_1$(applied_science#n#1)= (the discipline dealing with the art or **science** of applying scientific knowledge to practical problems; "he had trouble deciding which **branch of engineering** to **study**")

$\mathcal{R}_2$(information science#n#1)= (the **branch of engineering science** that **studies** (with the aid of computers) computable processes and structures).

P_{kl} (applied_science#n#1, information science#n#1) = 1x "science" +1x "branch of engineering" + 1x "study"
$= 1 + 3^2 + 1 = 11$

Les relations R_i dépendent de celles disponibles dans l'ontologie. Dans notre cas (WordNet), nous avons utilisé les relations suivantes:

– *gloss* qui n'est pas une relation en elle-même [12] mais représente la définition d'un concept avec éventuellement des exemples spécifiques du monde réel. Par exemple, le gloss du sens 1 du mot *car* (car#n#1) est: (*4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work"*). L'exemple spécifique est ce qui est entre *"*. Gloss est utilisée pour exploiter tout le réseau sémantique de WordNet sans tenir compte de la catégorie[2] (Part Of Speech) des mots ;

– la *synonymie*, les synonymes étant associés à la classe *Concept* ;

– l*'hypéronymie***,** la classe des *hyperonymes* contenant les concepts pères pour la relation de généralisation ; sa relation inverse *d'hyponymie* (spécialisation) ;

– la *méronymie* et son inverse *l'holonymie*, contenant respectivement les concepts constituant des parties du concept (… is a part of this concept, … is a member of this concept). Ou dont le concept est une partie (this concept is a part of …, etc.). Exemple : {voiture} a pour meronymes {{porte}, {moteur}}.

– Les relations de *Domaine* sont aussi utilisées pour le calcul de la similarité entre concepts. Elles peuvent dénoter la catégorie, l'usage ou la région :

dmnc: domain – category (exemple: *acropetal* $\xrightarrow{dmnc}$ *botany, phytology*)
dmnu: domain – usage (*bloody* $\xrightarrow{dmnu}$ *intensifier)*
dmnr: domain – region (*sumo* $\xrightarrow{dmnr}$ *Japan*)

Et leurs relations inverses, membre de domaine :

dmtc: member of domain - category pour les noms:
(exemple : *matrix_algebra* $\xrightarrow{dmtc}$ *diagonalization, diagonalisation*)
dmtu: member of domain - usage pour les noms (*idiom* $\xrightarrow{dmtu}$ *euphonious, forward, spang*)
dmtr: member of domain - region pour les noms (*Manchuria* $\xrightarrow{dmtr}$ *Chino-Japanese_War, Port Arthur*).

L'utilisation de ce nombre relativement élevé de relations a pour but de couvrir au maximum les différents types de liens que deux concepts peuvent partager. On peut ainsi détecter des liens non explicites entre concepts. Pour illustrer comment ces relations sont utilisées pour détecter des liens "cachés" entre concepts, prenons comme exemple les deux concepts *Henry Ford* et *car*. A priori, dans WordNet, il n'y a pas de lien explicite entre les deux concepts. Mais quand on utilise le gloss et la

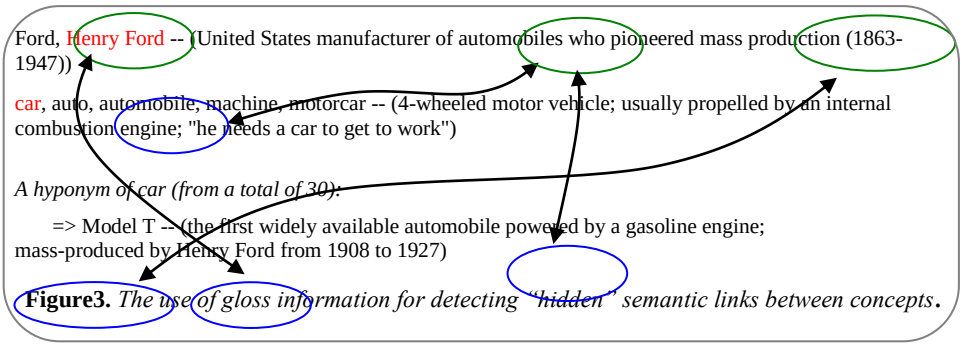

Figure3. *The use of gloss information for detecting "hidden" semantic links between concepts*.

Figure2. *Use of gloss information for detecting "hidden" semantic links between*

[2] WordNet est construit autour de quatre réseaux sémantiques indépendants — un pour chaque catégorie de mots (noms, verbes, adjectives et adverbes). Cette séparation est basée sur l'hypothèse que les relations sémantiques appropriées entre les concepts représentés par les synsets dans les différentes catégories sont incompatibles.

relation d'hyponymie par exemple, nous trouverons un recouvrement (des mots communs) indiquant une forte relation entre les deux concepts (voir Figure2). Il existe d'un coté, un recouvrement entre le gloss du premier concept (*Henry Ford*) et le synset du second (*car*); qui est le terme "automobile", Et un autre entre le synset du premier (*Henry Ford*) et le gloss d'un hyponyme du deuxième (*car*), qui est le terme composé "*Henry Ford".* Et enfin deux recouvrements entre le gloss du concept *Henry Ford* et le gloss d'un hyponyme de *car* qui sont les mots "automobile" et "mass". L'hyponyme est ce qui vient après le symbole "=>" et il est donné dans l'exemple de la Figure3 seulement un hyponyme sur les 30 retournés par WordNet pour le sens 1 de *car*.

3.3 Construction du noyau sémantique

Une fois les concepts importants sont extraits du document, ils sont utilisés pour construire le réseau sémantique. Notons

$$D_c = \{C_1, C_2, \ldots, C_m\} \quad (4)$$

l'ensemble des concepts sélectionnés suivant la méthode décrite en section 3.1. Les concepts peuvent être composés de mots uniques ou multiples et chaque C_i de l'ensemble D_c, peut avoir un certain nombre noté S_i de sens, représentés par les synsets de WordNet:

$$S_i = \{S_1^i, S_2^i, \ldots.., S_n^i\} \quad (5)$$

Le concept C_i *a alors* |Si |=n sens. Si nous choisissons un sens pour chaque concept de D_c, nous aurons toujours un ensemble *SN(j)* de m éléments, car nous sommes sur que chaque concept de D_c a au moins un sens étant donné qu'il appartient à l'ontologie. Nous définissons un réseau sémantique *SN(j)* comme suit:

$$SN(j) = (S_{j1}^1, S_{j2}^2, S_{j3}^3, \ldots, S_{jm}^m) \quad (6)$$

Il représente la jème configuration des sens des concepts de D_c. $j_1, j_2, .., j_m$ désignent des indexes de sens pris entre un et le nombre de sens possibles pour respectivement les concepts $C_1, C_{2,..,} C_m$. Pour les m concepts de D_c, plusieurs réseaux semantiques peuvent être construits en utilisant toutes les combinaisons possibles de sens. Le nombre total de réseaux sémantiques possibles Nb_SN dépend du nombre de sens que chaque concept de D_c peut avoir:

$$Nb\_SN = |S_1| \,.\, |S_2| \ \ldots .|S_m| \quad (7)$$

Par exemple, la Figure3 représente un réseau sémantique possible (S_2^1, S_7^2, S_1^3, S_1^4, S_4^5, S_2^m) résultant d'une combinaison du 2ème sens du premier concept C_1, du 7ème sens de C_2,, .., du 2ème sens de C_m (nous supposons que les liens nuls ne sont pas représentés). Les liens entre les concepts-sens ou nœuds (P_{ij}) représentés dans Figure3 sont calculés comme définis dans la section 3.2.Pour construire le

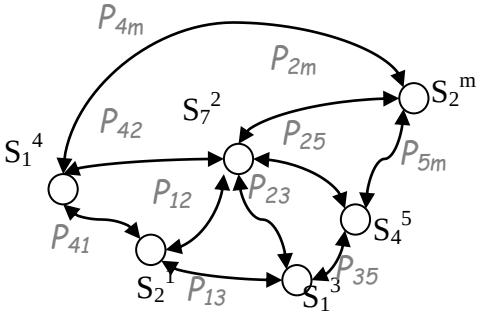

Figure3. *Exemple de réseau sémantique construit à partir d'une configuration ~~d~~de concepts-sens.*

meilleur réseau sémantique qui représente le document, la sélection de chacun de ces nœuds (concepts-sens) passe par le calcul d'un score (*C_score*) pour chaque concept-sens. Le score d'un concept-sens est égale à la somme de toutes les mesures de similarité avec les autres concepts-sens, sauf ceux qui sont dans le même ensemble de sens que le sien : Pour un concept C_i, le score de son kème sens est calculé comme suit:

$$C\_score(S_k^i) = \sum_{\substack{l\in[1..m], l\neq i \\ j\in[1..n]}} P_{i,l}(S_k^i, S_j^l) \qquad (8)$$

Sachant que m est le nombre de concepts de D_c et n représente le nombre de sens (synsets de WordNet) qui est propre à chaque C_i comme défini dans l'équation(5).

Le meilleur concept-sens qui représente au mieux le sens du concept C_i et celui qui maximise *C_score*:

$$Best\_score(C_i) = \underset{k=1..n}{Max}\ C\_score(S_k^i) \qquad (9)$$

Le concept-sens selectionné, permet de désambiguïser le concept C_i. Il représentera un nœud dans le noyau sémantique.

Le noyau sémantique final du document est $(S_{j1}^{\ 1}, S_{j2}^{\ 2}, S_{j3}^{\ 3}, \ldots, S_{jm}^{\ m})$, sachant que les nœuds correspondent respectivement à ($Best\_Score(C_1)$, $Best\_Score(C_2)$, $Best\_Score(C_3)$, $Best\_Score(C_m)$).

4 Evaluation

4.1 Méthode d'évaluation

Nous avons évalué notre approche de représentation sémantique de documents en RI en adaptant la phase d'indexation d'un SRI basé sur le modèle vectoriel [17] pour supporter les concepts-sens ainsi que les formules de calcul de CF.IDF et de *C_score*. Les noyaux sémantiques extraits pour chaque document ont été utilisés pour une indexation sémantique de documents. La collection sur laquelle nous avons travaillé est issue du projet MuchMore[3] [18]. Elle regroupe 7823 documents, 25 topics d'où sont extraites les requêtes ainsi que les jugements de pertinence correspondants faits par des experts. Nous avons opté pour une collection relativement réduite étant donné les contraintes dûes au temps de calcul : le calcul des noyaux sémantiques necessitte en moyenne une minute par document (environ 5 jour pour la collection utilisée). Les documents de la collection traitent du domaine médical, cependant le vocabulaire utilisé est assez général et est largement couvert par WordNet. Voici un exemple de requêtes utilisées (numéro 29):
Query 29: Heparin induced thrombocytopenia, diagnosis and management.

La méthode d'évaluation est faite selon le processus TREC[4] [19]. Dans l'expérimentation, trois cas de figures ont été évalués:

1) Indexation basée mots clés: dans cette méthode classique, les concepts des noyaux sémantiques ne sont donc pas utilisés et les poids des mots clés sont calculés suivant la formule TF/IDF pour les documents et les requêtes.
2) Indexation basée concepts: ici seuls les noeuds (concepts-sens) des noyaux sémantiques des documents sont utilisés pour l'indexation. La formule CF/IDF avec 2 comme seuil est utilisée pour la sélection des concepts. Les valeurs de score (*C_scores*) sont utilisées pour pondérer les concepts dans les documents et les requêtes. Cependant, elles sont passées au log pour atténuer de trop larges variations. Les liens entre les nœuds des noyaux sémantiques ne sont pas utilisés.
3) Indexation basée concepts + mots clés: ici, les deux précédentes méthodes sont combinées. Un dictionnaire est d'abord généré en utilisant de simples mots clés (cas1), puis un noyau sémantique

[3] http://muchmore.dfki.de/ (dernière visite: 15/09/04).
[4]. TREC pour Text REtrieval Conference, un programme international d'évaluation des Systèmes de Recherche d'Information : http://www.trec.nist.

est calculé pour chaque document (cas2). Lors de l'ajout des concepts-sens issus des noyaux au dictionnaire, deux cas peuvent arriver: soit le concept-sens est un multi mots, dans ce cas on le rajoute directement au dictionnaire avec log(*C_score*) comme poids. Soit le concept-sens est un mot unique (donc déjà indexé par la méthode classique), dans ce cas, nous changeons uniquement son poids basé sur TF.IDF par le log de la valeur de son *C_score*.

4.2 Résultats

Dans la Figure6, sont donnés les résultats de la recherche à différents points: 5, 10, 15, 20, 30, 100 et 1000 premiers documents sélectionnées ainsi que Avg-Pr, la précision moyenne à la fin. À chaque point, la précision moyenne pour les 25 requêtes est donnée pour les trois cas: Indexation basée mots clés, indexation basée concepts + mots clés et indexation basée concepts seulement. Les résultats montrent clairement l'avantage d'utiliser l'indexation basée concepts combinée avec la méthode d'indexation par mots clés. La précision est meilleure dans le cas de l'indexation combinée sur tous les points de précision, notamment lorsequ'on considère les premiers documents sélectionnés. Par exemple, dans les 5 premiers documents, la précision est de 0,3440 dans le cas de l'indexation par simples mots clés et de 0.232 dans le cas de l'indexation par concepts seulement. Elle passe à 0,432 (+20%) quand une combinaison des deux méthodes est utilisée. Et la précision moyenne est égale à 0,2581 dans le cas de l'indexation combinée, elle est de 0,2230 dans le cas de l'indéxation classique et seulement de 0.105 quand l'indexation est basée sur les concepts seulement. Nous pouvons conclure donc que l'indexation basée sur les concepts issus des noyaux sémantiques quand elle est combinée à la méthode d'indéxation classique permet d'améliorer la précision dans les réponses du SRI, alors que l'indéxation basée uniquement sur les concepts n'est pas suffisante pour retourner un résultat considérable.

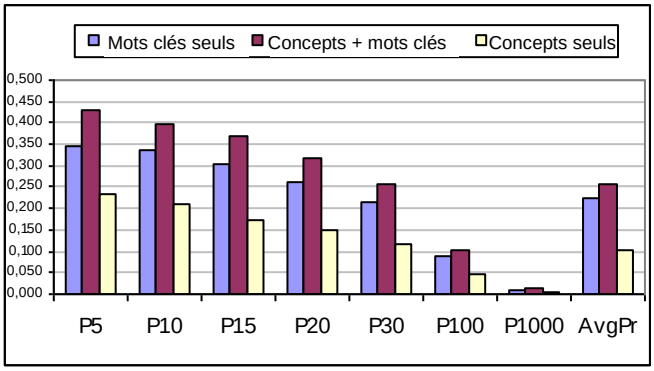

Figure6. Les résultats de la recherche dans les cas de l'indexation par mots clé, concepts + mots clés et concepts seuls.

Nous pouvons aussi remarquer que la différence entre les résultats de l'indexation basée concepts+mots clés et ceux de l'indéxation basée mots clés diminue quand on considère un nombre plus importants de documents. Ceci peut être expliqué par le nombre total de documents pertinents retournés par le SRI qui est relativement réduit, même s'il varient d'une requête à une autre, étant donné la taille de la collection.

5 Conclusion

Le travail que nous avons présenté dans ce papier rentre dans le cadre de l'application des ontologies à la recherche d'information. Nous avons introduit une approche pour la représentation du contenu sémantique des documents sous forme de réseau sémantique où les nœuds représentent des concepts désambiguïsés et les arcs des liens de similarité sémantique calculés à partir de relations présentes dans WordNet. Cette démarche peut se généraliser à toute ontologie présente sous forme d'un ensemble de concepts et de relations sémantiques entre ces concepts.

Une perspective possible à ce travail serait d'exploiter les *c_scores* des noyaux sémantiques pour la thématisation. Pour un document, les nœuds appartenant à une même "échelle" de scores peuvent être regroupés sous un même sous-thème. De même, à un niveau inter documents, les documents dont les noyaux sémantiques intersectent suffisamment, peuvent être vus comme traitant du même sujet.

6 ~~5.~~ Bibliographie

[1] OntoQuery project net site: http://www.ontoquery.dk

[2] Khan, L., and Luo, F.: Ontology Construction for Information Selection In Proc. of 14th IEEE International Conference on Tools with Artificial Intelligence, pp. 122-127, Washington DC, November 2002.

[3] Guarino, N., Masolo, C., and Vetere, G. "OntoSeek : content-based access to the web". IEEE Intelligent Systems, 14:70-80, (1999).

[4] Baziz, M., Aussenac-Gilles N., et Boughanem M. Désambiguïsation et Expansion de Requêtes dans un SRI : Etude de l'apport des liens sémantiques. Revue des Sciences et Technologies de l'Information (RSTI) série ISI, Ed. Hermes, V. 8, N. 4/2003, p. 113-136, décembre 2003.

[5] Mihalcea, R. and Moldovan, D.: Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000.

[6] Miller, G. Wordnet: A lexical database. Communication of the ACM, 38(11):39--41, (1995).

[7] Joon Ho Lee, Myong Ho Kim, and Yoon Joon Lee. "Information retrieval based on conceptual distance in IS-A hierarchies". Journal of Documentation, 49(2):188{207, June 1993.

[8] A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, in: Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, 2001, pp. 29–34.

[9] Gonzalo, J., Verdejo, F., Chugur I., Cigarrán J.: Indexing with WordNet synsets can improve text retrieval, in Proc. the COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing, 1998.

[10] Cucchiarelli, R. Navigli, F. Neri, P. Velardi. Extending and Enriching WordNet with OntoLearn, Proc. of The Second Global Wordnet Conference 2004 (GWC 2004), Brno, Czech Republic, January 20-23rd, 2004

[11] Hirst, G., and St. Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 305–332. MIT Press, 1998.

[12] Resnik, P., "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research (JAIR), 11, pp. 95-130, 1999.

[13] Banerjee, S. and Pedersen, T.: An adapted Lesk algorithm for word sense disambiguation using Word-Net. In Proc. of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2002.

[14] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In Proc. of SIGDOC '86, 1986.

[15] Croft, W. B., Turtle, H. R. & Lewis, D. D. (1991). The Use of Phrases and Structured Queries in Information Retrieval. In Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, A. Bookstein, Y. Chiaramella, G. Salton, & V. V. Raghavan (Eds.), Chicago, Illinois: pp. 32-45.

[16] Huang, X. and Robertson, S.E. "Comparisons of Probabilistic Compound Unit Weighting Methods", Proc. of the ICDM'01 Workshop on Text Mining, San Jose, USA, Nov. 2001.

[17] Boughanem, M., Dkaki, T., Mothe, J., et Soulé-Dupuy, C."Mercure at TREC-7". In Proceeding of Trec-7, (1998).

[18] Buitelaar, P., Steffen, D., Volk, M., Widdows, D., Sacaleanu, B., Vintar, S., Peters S., Uszkoreit, H.. Evaluation Resources for Concept-based Cross-Lingual IR in the Medical Domain In Proc. of LREC2004, Lissabon, Portugal, May 2004.

[19] The Sixth Text REtrieval Conference (TREC{6). Edited by E.M. Voorhees and D.K. Harman. Gaithersburg, MD: NIST, 1998.

[20] Vossen P., 2001. Extending, Trimming and Fusing WordNet for technical Documents, NAACL 2001 workshop on WordNet and Other Lexical Resources, Pittsbourgh, July 2001.

[21] Maedche A. and Staab S., 2000. Semi-automatic Engineering of Ontologies from Text. Proceedings of the Twelfth International Conference on Software Engineering and Knowledge Engineering.

[22] Zarg Ayouna, H., Salotti, S.: Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Dans *Ingénierie des Connaissances*, IC'2004, Lyon Mai 2004. 249-260.

Annexe :

Si nous considérons un exemple concret :

Le Document:

geophysics is the study of the earth and its atmosphere by physical measurements using a combination of mathematics and physics along with electrical_engineering computer_science and geology and other earth sciences the geophysicist analyzes measurements taken at the surface to infer properties and processes deep within the earths complex interior. with its use of physics mathematics geology and other earth sciences electrical_engineering and computer_science geophysics can be truly called an applied and interdisciplinary science it combines these different sciences to study the earths properties and processes from the atmosphere and oceans to the shallow subsurface and deep interior down to its central core except for boreholes which penetrate only a small fraction of the outer surface the earths interior cannot be directly observed all that_is known about the history continuing changes and distribution of resources of the earths interior must be inferred through detective_work based_on mastering and applying the sciences and technologies listed above because the earth supplies societys material needs and is the repository of its used products and home to all its inhabitants the wide range and importance of this field are apparent on the applied side oil companies and mining firms use the exploratory skills of geophysicists to locate deeply hidden resources geophysicists assess the strength of the earths surface when sites are chosen for large engineering and waste management operations on the theoretical side geophysicists try to understand such
earth processes as heat distribution and flow gravitational magnetic and other force_fields and vibrations and other disturbances within the earths interior between the pure and the applied is seismology the branch of geophysics that studies earthquake causes and predictability

a) Extraction de concepts

Les plus importants concepts (cf.idf >=2) avec leurs fréquences sont comme suit:

arth#n 9	electrical_engineering#n 2.5	science#n 3
geology#n 2	resource#n 2	geophysics#n 3
study#n 3	surface#n 3	apply#v 4
infer#v 2	property#n 2	earth_science#n 8
computer_science#n 3.5	side#n 2	deep#n 2
geophysicist#n 4	process#n 3	mathematics#n 2
distribution#n 2	physics#n 2	measurement#n 2
interior#n 5	atmosphere#n 2	

Par example: Frequency (“earth_science”)= 2 x “earth science” + 9/2 x “earth” + 3/2 x “science” =8

b) Calcul de similarités:

Les similarités sémantiques sont calculées entre tous les concepts-sens en utilisant les 11 relations sémantiques:

earth_science#n#1<14-5>computer_science#n#1 = 40
mathematics#n#1<2-3.5>computer_science#n#1 19
process#n#1<3-3.5>computer_science#n#1 6
geology#n#1<2-3.5>computer_science#n#1 33
science#n#1<3-3>study#n#1 4
atmosphere#n#1<2-2>side#n#2 2
study#n#6<3-3.5>computer_science#n#1 599
distribution#n#3<2-3>study#n#9 = 2
distribution#n#2<2-2>side#n#3 = 4
process#n#6<3-2>measurement#n#1 = 14
process#n#2<3-2>infer#v#5 = 1
process#n#1<3-2>measurement#n#1 = 66
science#n#1<3-2>mathematics#n#1 = 400
electrical_engineering#n#1<3-3>study#n#6 = 598

Par exemple, la 3ème ligne veut dire que la similarité entre le sens 6 du nom *process* ayant une fréquence=3 et le sens 1 de *measurement* qui a une fréquence =2 est égale à 14.

c) Selecting the best semantic net:

Pour chaque concept, son sens ayant le plus grand score cumulé est retenu comme le concept-sens approprié :

deep#n#2 = 88
geophysicist#n#1 = 274
science#n#1 = 1867
earth_science#n#1 = 1335
resource#n#1 = 192
infer#v#4 = 126
computer_science#n#1 = 1588
mathematics#n#1 = 800
earth#n#3 = 779
electrical_engineering#n#1 = 1514
physics#n#1 = 926
study#n#6 = 2351
process#n#2 = 452
geology#n#1 = 1065
surface#n#2 = 3198
measurement#n#1 = 466
apply#v#1 = 442
side#n#5 = 3041
geophysics#n#1 = 1007
interior#n#2 = 2367
distribution#n#2 = 203
property#n#3 = 625
atmosphere#n#3 = 1081

Pour le terme *atmosphere* par exemple, les 6 sens correspondant sont comme suit:

1. (18) atmosphere, ambiance, ambience -- (a particular environment or surrounding influence; "there was an atmosphere of excitement")

2. (10) standard atmosphere, atmosphere, atm, standard pressure -- (a unit of pressure: the pressure that will support a column of mercury 760 mm high at sea level and 0 degrees centigrade)
3. (9) atmosphere, air -- (the mass of air surrounding the Earth; "there was great heat as the comet entered the atmosphere"; "it was exposed to the air")
4. (6) atmosphere, atmospheric state -- (the weather or climate at some place; "the atmosphere was thick with fog")
5. (4) atmosphere -- (the envelope of gases surrounding any celestial body)
6. air, aura, atmosphere -- (a distinctive but intangible quality surrounding a person or thing; "an air of mystery"; "the house had a neglected air"; "an atmosphere of defeat pervaded the candidate's headquarters"; "the place had an aura of romance")

Le sens 3 qui est sélectionné correspond effectivement au sens approprié dans le document. Le problème d'ambiguïté pour les concepts multi mots tels que: earth_science, computer_science et electrical_engineering n'est pas posé car ils ont un seul sens dans WordNet.

Finalement le noyau sémantique résultat qui représente au mieux le contenu du document est comme ci-dessous. Ici, seuls les liens avec ayant une valeur >=30 sont représentés pour la visibilité du schéma. On peut remarquer que les concepts-sens (nœuds) peuvent être regroupés en sous groupes suivant les valeurs de *C_score* qu'ils détiennent:

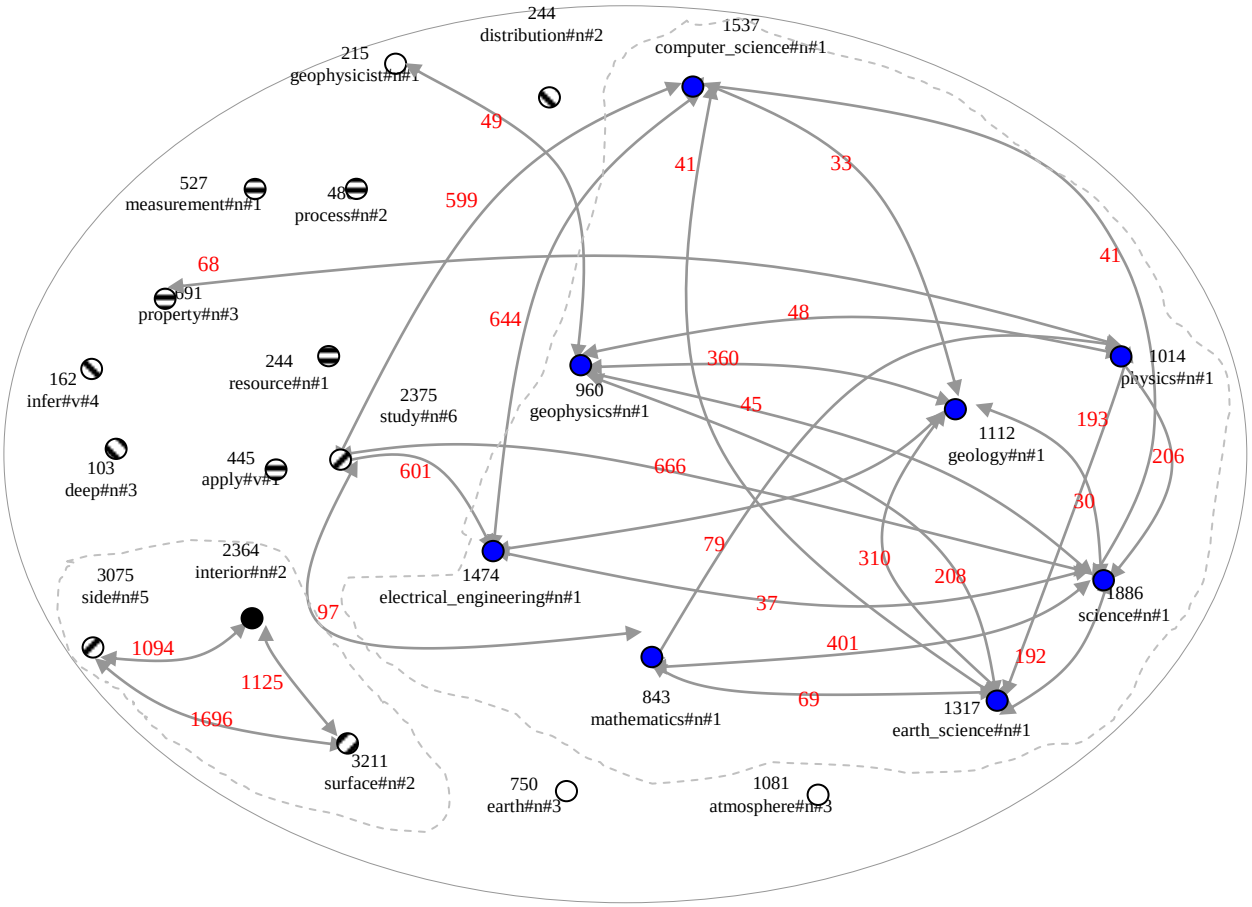

Schéma. Le noyau sémantique résultat pour le document exemple.

Les noeuds du groupe {computer_science#n#1, physics#n#1, science#n#1, earth_science#n#1, geology#n#1, mathematics#n#1, electricalengineering#n, geophysics#n#1} ont des scores qui avoisinent 1000. Ils appartiennent effectivement au même thème (science and applied science). La même remarque peut aussi s'appliquer sur les noeuds {side#n#5, interior#n#2, surface#n#2}.