Evaluation d'associations sémantiques dans une ontologie de domaine

Thabet Slimani[1], Boutheina Ben Yaghlane[2], Khaled Mellouli[2]

ISG de Tunis BP 41 - Rue de la liberté- Bardo 2000
IHEC Carthage, Carthage Présidence 2016, Tunisia
thabet.slimani@issatm.rnu.tn
{boutheina.yaghlane, khaled.mellouli}@ihec.rnu.tn

Résumé : Dans une ontologie de domaine, une association sémantique entre deux entités (concepts, attributs d'un concept) est une représentation d'un chemin ou d'un lien sémantique (LS) indirect entre elles. Un défi prometteur pour le Web sémantique est de développer des méthodes pour découvrir des données fortement liées dans un nombre important d'associations sémantiques rassemblées à partir des sources disparates. Dans ce contexte, cet article présente, en premier lieu, le degré d'un lien sémantique (DLS) pour mesurer une relation directe entre deux entités, et en deuxième lieu le degré d'une association sémantique (DAS) pour mesurer des associations sémantiques extraites à partir d'une ontologie de domaine. Les résultats expérimentaux montrent l'avantage des méthodes proposées et démontrent leur efficacité prometteuse.

Mots-clés : association sémantique, degré d'un lien sémantique, degré d'une association sémantique, ontologie de domaine, concept.

1 Introduction

Aujourd'hui, la croissance massive de données stockées dans le Web est contrôlée par un réseau des sources de données en corrélation, appelé réseau sémantique des liens (RSL), incluant des personnes, des compagnies, des connaissances de domaine, des publications scientifiques, des articles, etc. Un RSL est conçu pour établir des rapports sémantiques parmi diverses ressources dans le Web visant à prolonger le réseau WWW des liens hypertextes à un réseau sémantique riche (Zhuge, 2007). Formellement, un RSL est un réseau contenant des noeuds sémantiques et des liens sémantiques. Un noeud sémantique peut être un concept, un attribut de concept, un schéma d'ensemble de données, un URL, une entité, etc. Les liens entre les noeuds sémantiques (entités) dans une base de connaissance RDF fournissent le type, la signification ou l'interprétation des entités.

La découverte des liens sémantiques est un problème important pour des applications appréhendant des données gérées en réseau. Un lien sémantique (LS) se rapporte à une relation directe entre deux entités. Par contre, une association sémantique (AS) est un chemin connectant deux entités d'une manière indirecte.

L'extraction d'une association sémantique (AS) signifie l'extraction d'un chemin entre deux entités reliées indirectement par des relations concrètes (propriétés) contenues dans un graphe RDF (Aleman-Meza *et al.*, 2003) (Anyanwu *et al.*, 2005) (Aleman-Meza *et al.*, 2006) (Ning *et al.*, 2006). Une AS a un sens différent par rapport à la similarité sémantique. Une similarité sémantique permet de mesurer le degré de proximité de deux entités du point de vue sémantique, alors qu'une AS permet de mesurer le degré de connectivité de deux entités du point de vue sémantique. Par exemple, l'entité "author" a une valeur faible de similarité sémantique avec l'entité "Research-Area" parce qu'elles ne se réfèrent pas au même objet, alors que "author" possède une association sémantique forte avec l'entité "Research-Area".

Dans des applications analytiques telles que la sécurité nationale, la bioinformatique, etc, il est indispensable d'extraire des connaissances significatives à partir des liens et des associations sémantiques déjà extraits. Malheureusement, dans les travaux recensés dans la littérature, les avances remarquables aux niveau des développements actuels montrent un manque de satisfaction pour de nombreuses applications récentes telles que l'exploitation de données, la recherche documentaire, l'extraction des services Web qui exigent l'évaluation des associations sémantiques.

Dans ce papier, nous proposons certaines mesures qui combinent des fonctions sémantiques et statistiques pour mesurer les liens sémantiques directs entre deux entités et les associations sémantiques entre deux entités reliées à travers un chemin en évaluant le rapport entre elles.

Le reste de ce document est organisé comme suit. La section 2 présente les travaux liés et nos contributions principales. La section 3 décrit les spécifications d'un lien sémantique et d'une association sémantique. Les expérimentations réalisées sont présentées dans la section 4. Finalement, la section 5 donne un résumé et une perspective des travaux futurs.

2 Travaux liés

Dans le travail de (Aleman-Meza *et al.*, 2006), les auteurs proposent une approche pour la découverte d'une diversté de liens sémantiques entre les "reviewers" et les "authors" dans une ontologie complète pour déterminer un degré de conflit d'intérêt. La création de cette ontologie a été basée sur les entités et l'intégration des relations de deux réseaux sociaux : réseau foaf (friend-of-a-friend) et réseau "co-author". Dans cette même perspective, (Cao *et al.*, 2005) présentent une approche pour la fouille des communautés cachées par l'exploitation des réseaux sociaux hétérogènes. Peterson *et al.* discutent certaines approches pour exploiter le capital social pour créer un standard réseau sémantique riche (Peterson *et al.*, 2008). En plus, une autre approche qui discute la question de la désambigüisation d'URI dans le cadre des données liées est présentée dans le travail (Jaffri *et al.*, 2008).

Notre travail se situe dans le cadre des travaux permettant d'évaluer les liens et les associations sémantiques. Nous pouvons classifier les approches, dans la littérature, dans deux directions : (1) les approches orientées données (data-driven) qui essayent de capturer la dépendance entre les concepts (dérivés du corpus) par l'information statistique (Cao *et al.*, 2005). Cette approche utilise les co-occurrences des relations binaires

entre les concepts et (2) les approches orientées structures (structure-driven) (Watabe & Kawaoka, 2001) qui exploitent les caractéristiques de la structure d'une ontologie (les classes/concepts d'une ontologie et leurs relations).

Notre approche dérive des approches statistiques basées sur un modèle de langage statistique, puisque les mesures proposées sont basées sur la combinaison des formules de probabilité. Les approches statistiques peuvent être appliquées aux applications dans le domaine de recherche documentaire, de la reconnaissance de forme et la fouille de données. Ces approches utilisent la distribution de probabilité pour mesurer des données liées. Par exemple, dans le domaine de recherche documentaire, la pertinence d'un document avec une requête peut être évaluée par la probabilité de la génération du document vis-à-vis à une requête donnée (Song & Croft, 1999). D'une manière analogue, dans la base de connaissance, deux entités connexes (Classe/Concept, Instance de Classe) peuvent être évaluées, par l'intermédiaire des liens sémantiques, en utilisant des mesures statistiques. Dans le travail de (Tian *et al.*, 2007), les auteurs proposent une approche qui mesure les associations sémantiques dans une ontologie de domaine est bien lié à notre travail. La différence par rapport à notre travail réside au niveau des formules du degré d'un lien sémantique et au niveau de l'évaluation des associations sémantiques.

3 Spécification d'un lien et d'une association sémantique

3.1 Base de connaissance

Une ontologie O est une structure incluant deux ensembles disjoints C et R, dont les éléments s'appellent, respectivement, les classes/concepts et les relations (propriétés/attributs). Au niveau de la partie supérieure de la figure 1, l'ensemble de classes C est défini par : {"Professor", "University", "Course", "Project", "Publication", "Student"}. L'ensemble des relations R est défini par {"Author-Of", "DegreeFrom", "Offers", "Related-to", "Enrolled-In"}. Une base de connaissance est une structure de deux ensembles disjoints incluant les instances de classes et les instances des liens sémantiques (propriétés). Au niveau de la partie inférieure de la figure 1, l'ensemble des instances de classes est défini par : {"Author-Of", "DegreeFrom", "Offers", "Related-to"}. La relation entre l'entité "P0" de type "Professor" et l'entité "PR0" de type "Project", présentée dans la figure 1, constitue une association sémantique (AS) définie par l'expression suivante : $\text{P0} \xrightarrow{DegreeFrom} \text{U0} \xrightarrow{Offers} \text{C0} \xrightarrow{Related-to} \text{PR0}$.

3.2 Signature et degré d'un lien sémantique

Cette section présente des formules qui calculent le degré d'un lien sémantique représenté dans une ontologie, ou des éléments (ressources) représentés dans un schéma. Ces formules de calcul exploitent le fait que les entités (concepts/classes) qui sont comparées peuvent avoir des propriétés (sous format d'attributs) associées entre elles et qui prennent en considération le niveau de généralité (ou de spécificité) de chaque entité dans l'ontologie aussi bien que leurs rapports avec d'autres entités ou concepts.

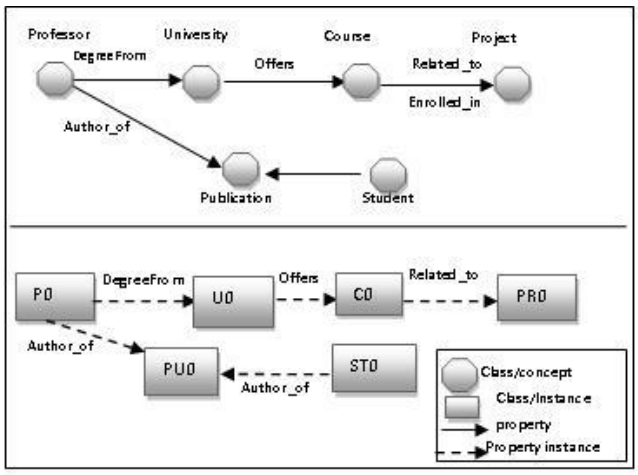

FIG. 1 – Un exemple de base de connaissance décrivant un réseau social.

La signature d'un lien sémantique entre deux entités est définie par les paires de concept/classe source et concept/classe cible. A titre d'exemple, le lien sémantique "DegreeFrom" admet comme signature l'ensemble (Professor, University) ; le lien sémantique "AuthorOf" admet deux signatures : (Student, Publication) et (Professor, Publication). La signature d'un lien sémantique n'est pas symétrique puisqu'il représente une propriété RDF qui se lit à partir d'un seul sens. Les associations sémantiques constituent un chemin qui combine, l'entité source, l'entité cible, les entités intermédiaires (instances de classe/concept) et les propriétés (instances de propriétés). Dans l'exemple de la figure 1, l'association sémantique entre l'entité source "P0" et l'entité cible "PR0" inclut deux entités intermédiaires {"U0", "C0"} et trois propriétés intermédiaires {"DegreeFrom", "Offers", "Related-to"}.

Soit A et B deux concepts/classes d'une ontologie de domaine, nous présentons les définitions suivantes :

- *Lien hiérarchique (LH)* : Si A est défini comme un "SuperClassof" de B, alors B est contenu dans A. Ce lien est représenté par l'expression A$\supset$B.
- *Lien d'équivalence (LE)* : Si A est défini comme "EquivalentClassof" de B, alors A est équivalent à B. Ce lien est représenté par l'expression A$\equiv$B.
- *Lien sémantique (LS)* : Si A est défini comme "PropertyOf" de B ou B est défini comme "PropertyOf" de A, on dit que A et B possèdent un lien sémantique. Ce lien est indiqué par l'expression A$\propto$B ou B$\propto$A.
- *Degré d'un lien sémantique (DLS) :* S'il y a un lien sémantique entre deux concepts/ classes A et B, nous définissons DLS comme le poids d'un lien sémantique, qui mesure le lien entre A et B. Par conséquent, nous pouvons définir l'équation (1) qui calcule $DLS_A(B, l)$, où l représente le lien qui relie les concepts A et B.

$$DLS_A(B, l) = P_A(B|l), l \in \{A \supset B, A \propto B\} \quad (1)$$

La probabilité conditionnelle $P_A(B|l)$ de A et de B avec le lien *l* définie dans l'équation (1) peut être décrite par l'équation (2) comme suit :

$$P_A(B|l) = P_A(B,l)/P_A(l) \tag{2}$$

Où $P_A(l)$ est la probabilité de A ayant *l* comme lien sémantique et $P_A(B,l)$ la probabilité d'apparition de A et de B avec le lien *l*.

En respectant les définitions du langage OWL[1], un concept/classe d'ontologie A peut avoir quelques instances (termes, ressources RDF). En conséquence, $P_A(B,l)$ devrait prendre en compte ces instances. Cependant, l'expression $P_A(B,l)$ dans l'équation (2) est enrichie par l'écriture de l'équation (3) :

$$P_A(B|l) = \sum_{A_i,B_j} P_{A_i}(B_j,l)/(P(B).P(l|B)) \tag{3}$$

Où A_i et B_j sont, respectivement, les instances des concepts A et B. ***P(l\|B)*** signifie la probabilité conditionnelle du lien *l* étant donnée l'entité *B* et ***P(B)*** désigne la probabilité d'apparition de l'entité B. Ensuite, la mesure $DLS_A(B,l)$ est changée pour estimer les 3 expressions : $P_{A_i}(B_j,l)$, ***P(B)*** et ***P(l\|B)***. L'évaluation de ces expressions sera basée sur l'estimation du maximum de vraisemblance. Cette évaluation prend en compte les co-occurrences de A et B avec *l* dans le corpus basé sur une ontologie de domaine.

3.2.1 Estimation de P$_{A_i}(B_j,l)$

Soit T l'ensemble de termes dérivés du corpus basé sur l'ontologie de domaine, et t le terme ayant un lien *l* avec A et/ou B. $P_{A_i}(B_j,l)$ désigne la probabilité d'apparition du concept A et B ensemble avec le lien *l*. Selon le type de *l*, $P_{A_i}(B_j,l)$ doit être prise en compte différemment. Si *l* est un lien sémantique (LS), il faut imposer une portion de texte qui ne dépasse pas une certaine limite (TL : Limite du texte en nombre de mots) et dans laquelle nous pouvons calculer les co-occurrences de A et B. L'estimateur de ce type de lien est obtenu par l'équation (4) :

$$\hat{\mathrm{E}}(P_{A_i}(B_j,l)) = \frac{count_{A_i}(B_j|TL)}{\sum_{t1\in(A_i,B_j)^t\in T} count_t(t1,TL) - count_{A_i}(B_j|TL)} \tag{4}$$

Si *l* est un lien hiérarchique (LH), nous définissons le modèle de co-occurrence TLS comme caractéristique lexique-syntaxique du lien sémantique se produisant dans le texte TL. Par exemple, l'expression "A inclus dans B" dans une portion de texte TL donne une indication d'un lien LH, qui doit être inclus dans la texte de TLS. L'estimateur de ce type de lien est obtenu par l'équation (5) :

$$\hat{\mathrm{E}}(P_{A_i}(B_j,l)) = \frac{count_{A_i}(B_j|TLS)}{\sum_{t1\in(A_i,B_j)^t\in T} count_t(t1,TLS) - count_{A_i}(B_j|TLS)} \tag{5}$$

[1]OWL Reference. http ://www.w3.org/TR/owl-ref/

3.2.2 Estimation de *P(B)*

La distribution de probabilité du concept B possède deux interprétations. De point de vue structure de l'ontologie, l'estimateur de P(B) peut être représenté par l'équation (6), où $|l_B|$ est le nombre des liens de B, c est un concept dans l'ontologie de domaine O, et $|l_c|$ est le nombre de liens du concept c. De point de vue occurrence de termes, P(B) est estimée par l'équation (7). Pour un concept c $\in$ O, fq(c)=$\sum_{c_i} count(c_i, C)$, où C est le corpus donné, c_i est le terme instance du concept c, alors *fq(c)* désigne la fréquence de l'apparition des instances du concept c dans le corpus C et fq_B désigne la fréquence de l'apparition du concept B.

$$P_st(B) = \frac{|l_B|}{Max_{c \in O}|l_c|} \tag{6}$$

$$P_ot(B) = \frac{|fq_B|}{Max_{c \in O} fq(c)} \tag{7}$$

L'estimateur de *P(B)* est représenté par le modèle mixte donné dans l'équation (8), dont λ est un coefficient permettant de combiner la structure de l'ontologie avec le corpus. Ce modèle mixte devient $P_ot(B)$ si λ=0 et $P_st(B)$ si λ=1. Ce coefficient est adopté pour optimiser la performance de recherche de l'information.

$$\hat{\mathrm{E}}(P(B)) \quad = \quad \lambda.P_st(B) + (1-\lambda).P_ot(B), 0 \leq \lambda \leq 1 \tag{8}$$

TL, TLS, et λ peuvent être différents lorsque les ontologies varient d'un domaine à un autre. Nous avons attribué une valeur constante, par intuition, comme valeur pour λ, ce qui ne pourrait pas être la meilleure valeur. Pour TL, nous avons obtenu sa valeur en analysant manuellement le corpus.

3.2.3 Estimation de *P(l|B)*

L'équation (9) donne l'estimation de *P(l|B)*, où $|l_B|$ est le nombre des liens B, et $|l_B|r$ le nombre des relations r ayant des liens avec B :

$$\hat{\mathrm{E}}(P(l|B)) = \frac{|l_B|r}{|l_B|} \tag{9}$$

Pour illustrer la méthode ci-dessus, nous énumérons différents types de liens dans l'ontologie MeSH (M*edical Subject Heading*)[2] comme présenté dans l'exemple du tableau 1.

[2]http ://www.nlm.nih.gov/mesh/

LS	Exemple dérivé de MeSH	Représentation graphique
A$\propto$B	L=Formalin Test$\propto${Pain, Intractable}	$A \xrightarrow{L}$B
A$\supset$B	Headache $\supset$Pain	$A \xleftarrow{is-a}$B
A$\equiv$B	Pain$\equiv$Postoperative	$A \leftrightarrow B$

TAB. 1 – Exemples de liens sémantiques extraits à partir de l'ontologie MeSH

3.3 Signature et degré d'une association sémantique

Une AS a trois types de signatures définis comme suit :

- *Signature des propriétés d'une association sémantique (SPAS)* définie par les propriétés contenues dans l'association sémantique. Dans l'exemple de la figure 1, la signature des propriétés d'une AS qui mène de l'objet source "P0" à l'objet cible "PR0" est définie par l'ensemble de propriétés/liens {"DegreeFrom", "Offers", "Related-to"}.
- *Signature d'instances d'une association sémantique (SIAS)* définie par les entités contenues dans un lien sémantique. A titre d'exemple, la signature d'instances de l'association sémantique reliant "P0" et "PR0" est définie par l'ensemble d'instances/ressources {U0, C0}.
- *Signature de classes d'une association sémantique (SCAS)* définie par les classes des entités intermédiaires dans une AS. Par exemple, la signature de concepts/classes de l'association sémantique entre "P0" et "PR0" est définie par l'ensemble des classes {university, course}.

Dans la section précédente, nous avons discuté comment évaluer un LS de deux concepts directement liées par l'évaluation du degré de connectivité entre elles. Dans cette section nous essayerons de formuler une mesure permettant d'évaluer le degré d'une association sémantique (DAS) entre deux classes reliées par un lien sémantique. L'évaluation à travers la mesure DAS sera basée sur la mesure DLS des classes intermédiaires reliées.

Supposons que A et B sont deux concepts/classes contenus dans l'ontologie et DAS(A, B) représente le degré d'association sémantique entre eux. L'expression qui calcule DAS (A, B) est obtenue par les formules conditionnelles suivantes :

- *Cas 1* : Si A$\equiv$B ou B$\equiv$A, alors DAS(A, B)=1
- *Cas 2* : Si A$\supset$B ou A$\propto$B, alors DAS(A, B) =DLS(A,*l*)
- *Cas 3* : Si Non(B$\subset$A) et Non(A$\propto$B), alors DAS(A, B)=0
- *Cas 4* : Autrement, A et B ont une association sémantique. L'expression DAS(A, B) est définie par l'équation (10) :

$$DAS(A,B) = log(n^2). \prod_{i=A...X, j=Y...B} DLS_i(j,l) \tag{10}$$

Où $l \in$(A$\supset$X), Y$\propto$A, B$\equiv$Y et n est le nombre de concepts/classes intermédiaires qui mènent de A à B (les entités dans l'ensemble SCAS).

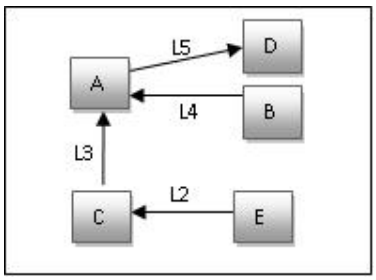

FIG. 2 – Exemples de liens et d'associations sémantiques dans l'ontologie MeSH.

S'il existe un chemin possible entre A et B, la valeur DAS de l'association sémantique entre A et B est calculée par la multiplication des DLS des paires de classes contenues dans ce chemin. Sinon la valeur de DAS est évaluée à 0.

Pour plus de clarification du principe de DAS et DLS, nous énumérons différents types d'associations sémantiques dans l'ontologie MeSH par les exemples de la figure 2 et du tableau 2.

	A	B	C	D	E
A	1	0	0	$\text{DLS}_A(D, L5)$	0
B	$\text{DLS}_B(A, L4)$	1	0	$\text{DLS}_B(A, L4)* DLS_A(D, L5)* Log(4)$	0
C	$\text{DLS}_C(A, L3)$	0	1	$\text{DLS}_C(A, L3)* DLS_A(D, L5)* Log(4)$	0
D	0	0	0	1	0
E	$\text{DLS}_E(C, L2)* DLS_C(A, L3)* Log(4)$	0	$\text{DLS}_E(C, L2)$	$\text{DLS}_E(C, L2)* DLS_C(A, L3)* DLS_A(D, L5)* Log(9)$	1

TAB. 2 – Exemples d'évaluation en utilisant la formule DAS

Une entité dans une base de connaissance RDF est une instance d'une classe spécifique. Si A et B sont deux entités dans la base de connaissance, nous pouvons extraire des informations inattendues à travers l'extraction d'associations sémantiques entre une entité donnée et une autre non spécifiée au départ.

3.4 Exemple d'évaluation d'associations sémantiques : La mesure DAS basée sur la fréquence des propriétés entrantes (DAS-PE)

DAS-PE représente le degré maximum des propriétés entrantes qui relient une entité spécifique d'une association par un lien l déjà connu. Cet exemple de degré est men-

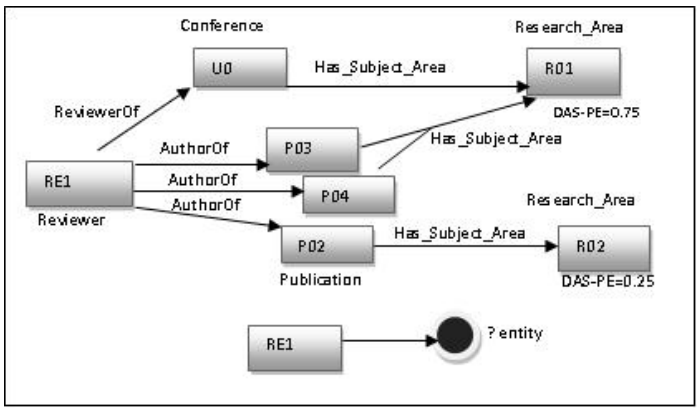

FIG. 3 – Un exemple fictif d'évaluation de DAS-PE.

tionné dans notre précédent travail qui décrit le langage PmSPARQL pour l'extraction des chemins à partir d'un graphe RDF (Slimani *et al.*, 2008). L'exemple présenté dans la figure 3 décrit une évaluation des liens sémantiques à travers la recherche du champ d'expertise d'un "Reviewer" pour une conférence déterminée. Les résultats peuvent être limités à l'identification d'un nouveau lien sémantique (Expert-in) entre l'objet source prédéfini et un objet initialement inconnu.

La valeur de DAS-PE de chaque association sémantique qui mène de A à B est obtenue par la valeur de la probabilité maximale $P_A(B_i|l)$ d'une entité spécifique (B_i) en respectant la théorie des probabilités.

$P_A(B_i|l)$ se réfère à la probabilité de l'entité A étant donnée l'entité B_i et le lien sémantique *l*. La formule de DAS-PE est obtenue par l'expression : $P_A(B_i|l)$=CB_i/NIR ; CB_i désigne le nombre de relations (propriétés) entrantes reliant l'entité spécifiée B_i et *NIR* désigne le nombre total des relations entrantes reliant les instances de B à travers le lien *l*. L'expression de la formule DAS-PE est spécifiée comme suit :

$$DAS - PE(AS) = Max\{P_A(B_1|l), P_A(B_2|l), ..., P_A(B_n|l)\} \quad (11)$$

Dans la figure 3, l'entité "R01" qui désigne l'instance de la classe "Research-Area", possède 3 relations entrantes parmi 4 relations reliant toutes les entités de type "Research-Area" avec le lien "Has-Subject-Area". La valeur de probabilité de "R01" =3/4 =0.75 est la valeur maximale qui sera affectée à DAS-PE. Et finalement le lien sémantique cherché est défini par : (RE1$\xrightarrow{Expert-in}$R01).

Intuition : L'évaluation d'associations sémantiques par DAS-PE est importante dans le cas où la recherche d'un genre de lien sémantique n'est pas explicitement existant dans la base de connaissance. Le nouveau lien sémantique découvert (expert-in) aide à l'enrichissement de la base de connaissance, qui facilite la découverte de nouveaux LS inattendus.

4 Résultats expérimentaux

Au niveau de cette section, nous discutons les évaluations expérimentales, ainsi que les tests appliqués sur la formule DAS discutée dans la section 3. Les tests d'évaluation sont réalisés sur une machine avec un processeur Intel (R) Core (TM) 2 Duo 2.2GHZ, une mémoire de 2GB et Windows Xp.

Nous avons adopté DragonToolkit[3] qui est un paquet de développement Java utile pour l'utilisation académique dans l'extraction sémantique pour évaluer l'efficacité de notre mesure DAS.

Nom du concept		Pain
Liens	Related-To	Intractable
	SubClassOf	Headache
	SynonymOf	Postoperative

TAB. 3 – Les informations de l'ontologie MeSH utilisées dans nos expérimentations.

Pour ce faire, nous avons appliqué le principe de l'expansion des requêtes (Grootjen & T.P.v.d, 2006) appliqué sur l'ontologie MeSH et l'ensemble des données CFC[4] incluant des résumés documentaires. Dans les expériences réalisées, le terme "classe" discutée plus haut est remplacé par le terme "concept" pour des raisons de convenance avec le contenu de l'ontologie MeSH. Les informations utilisées à partir de l'ontologie MeSH sont essentiellement présentées dans le tableau 3. Ce tableau représente le concept "Pain" en terme des liens décrits dans la section 3.2 ("Related-To" c'est un lien LS, "SubClassOf" est un LH et "SynonymOf" est un LE).

Nous avons appliqué une recherche basée sur la sémantique de l'expansion des requêtes pour comparer l'efficacité de la méthode DAS proposée par rapport à DAS dérivé de l'intuition (manuellement). Dans l'étude comparative nous avons utilisé la même valeur proposée dans le travail de (Tian *et al.*, 2007) pour faciliter la comparaison avec d'autres approches. La valeur de DLS avec notre méthode a été fixée à 0.8 (analyse manuelle du corpus) et celle de l'intuition a été fixée à 0.5 (valeur de λ décrite dans la section 3.2.2).

La méthode MAP (*Mean Average Precision*) est l'outil d'évaluation traditionnel qui est adopté dans le domaine de la recherche documentaire. Elle calcule la précision moyenne de toutes les requêtes. MAP (n) est employée dans notre travail pour évaluer les n documents recherchés. La formule *précision (n)* est donc adoptée pour mesurer la précision des n documents cherchés. La mesure de précision est bien connue pour l'examination de la qualité des documents appropriés.

La précision moyenne PM (*Average Precision*) tient en compte la moyenne des scores de précision des documents appropriés parmi top-k documents recherchés par une requête simple. La précision moyenne d'une requête i est définie par la formule suivante :

[3]http ://www.ischool.drexel.edu/dmbio/dragontool/

[4]http ://www.ischool.berkeley.edu/ hearst/irbook/cfc.html

	Document approprié	Précision(30)	MAP(30)
Intuition	22	81.32%	85.12%
DAS	29	91.34%	93.48%
DLS	27	95.33%	97.24%

TAB. 4 – Comparaison de l'efficacité de la recherche sémantique basée sur la mesure DAS avec celle basée sur DLS et l'intuition

$$PM_i(k) = \frac{\sum_{j=1...n_i} j/R_{i,j}}{n_i} \tag{12}$$

Où n_i est le nombre des documents appropriés de l'ième requête, k est le nombre des documents extraits, et R_{ij} est le rang du jème document approprié de l'ième requête.

La fonction MAP(k) est utilisée ici pour calculer le rang des k-premiers documents recherchés avec une précision (n). L'expression de MAP(k) est obtenue par la formule suivante :

$$MAP(k) = \frac{\sum_{i=1...q_n} PM_i(k)}{q_n} \tag{13}$$

Où q_n est le nombre de requêtes exécutées. Les résultats qui figurent dans le tableau 4 montrent l'amélioration apportée par notre méthode DAS permettant d'évaluer les AS par rapport à la méthode DAS discutée par Tian et al (Tian *et al.*, 2007).

L'approche adoptée est appliquée sur une ontologie de grande taille (MeSH) et montre une lenteur au niveau de l'évaluation de la précision des documents. Cependant, un travail conséquent reste à faire d'une part en amont sur l'analyse théorique de ces mesures, et d'autre part sur leur implémentation à grande échelle.

5 Conclusion

Dans cet article nous avons proposé une approche qui se préoccupe par l'évaluation des associations sémantiques dans le cadre de la recherche de l'information basée sur une ontologie de domaine. À ce propos, nous avons présenté une mesure DLS pour mesurer les liens sémantiques directs entre des entités contenues dans une base de connaissance. Nous avons présenté une deuxième mesure qui évalue le degré d'une association sémantique (DAS) entre deux entités qui sont liées d'une manière indirecte. Les mesures proposées pour l'évaluation des liens et des associations sémantiques ont été appliquées sur une ontologie de domaine et du corpus sous format des résumés. Les résultats obtenus montrent une amélioration remarquable au niveau de la précision dans le domaine de l'extraction des documents appropriés.

Références

ALEMAN-MEZA B., HALASCHEK-WIENER C., ARPINAR I. B. & SHETH A. P. (2003). Context-aware semantic association ranking. In *Proceedings of SWDB'03, The first International Workshop on Semantic Web and Databases, Co-located with VLDB 2003, Humboldt-Universität, Berlin, Germany, September 7-8*, p. 33–50.

ALEMAN-MEZA B., NAGARAJAN M., RAMAKRISHNAN C., DING L., KOLARI P., SHETH A. P., ARPINAR I. B., JOSHI A. & FININ T. (2006). Social networks : Semantic analytics on social networks : Experiences in addressing the problem of conflict of interest detection. In *Proceedings of the 15th international conference on World Wide Web, WWW 06*, p. 407–416 : ACM Press.

ANYANWU K., MADUKO A. & SHETH A. (2005). Sem-rank : Ranking complex relationship search results on the semantic web. In *International World Wide Web Conference*, volume 14, p. 117–127, New York, NY, USA : ACM Press.

CAO G., NIE J.-Y. & BAI J. (2005). Integrating word relationships into language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 298–305.

GROOTJEN F. & T.P.V.D W. (2006). Conceptual query expansion. In *Data & Knowledge Engineering*, p. 174–193.

JAFFRI A., GLASER H. & MILLARD L. (2008). Uri disambiguation in the context of linked data. In *Proceedings of the 17th international conference on World Wide Web, WWW08* : ACM Press.

NING X., JIN H. & WU H. (2006). Semrex : Towards large-scale literature information retrieval and browsing with semantic association. In *Proceedings of IEEE International Conference on e-Business Engineering (ICEBE 06)*, p. 602–609 : IEEE Computer Society.

PETERSON D., CREGAN A., ATKINSON R. & BRISBIN J. (2008). Exploiting social capital to create a standards-rich semantic network. In *Proceedings of the 17th international conference on World Wide Web, WWW08* : ACM Press.

SLIMANI T., BEN YAGHLANE B. & MELLOULI K. (2008). Pmsparql : Extended sparql for multiparadigm path extraction. *International Journal of Computer, Information, and Systems Science, and Engineering.*, **2 (3)**, 179–190.

SONG F. & CROFT W. B. (1999). A general language model for information retrieval. In *In CIKM 99 : Proceedings of the eighth international conference on Information and knowledge management*, p. 316321, New York, NY, USA : ACM Press.

TIAN X., LI H. & DU X. (2007). Measuring semantic association in domain ontology. In *IEEE Third International Conference on Semantics, Knowledge and Grid*, p. 515–518 : IEEE Computer Society.

WATABE H. & KAWAOKA T. (2001). The degree of association between concepts using the chain of concepts. In *Systems, Man, and Cybernetics*, volume 2, p. 877–881, Tucson, AZ, USA : Computer Society.

ZHUGE H. (2007). Autonomous semantic link networking model for the knowledge grid. In *Concurrency and Computation : Practice & Experience*, volume 19, p. 1065 – 1085 : John Wiley and Sons Ltd.