Une approche basée sur des relations pour la RI sémantique

Marie-Noëlle Bessagnet[1], Davide Buscaldi[2], Albert Royer[1], Christian Sallaberry[1]

[1] LIUPPA, Université de Pau et des Pays de l'Adour,
F-64000 Pau
{marie-noelle.bessagnet,albert.royer,christian.sallaberry}@univ-pau.fr
[2] LIPN, Université Paris XIII,
F-93430 Villetaneuse
davide.buscaldi@lipn.univ-paris13.fr

Résumé :
Dans cet article, nous comparons trois méthodes de RI basées sur des mots-clés, sur des concepts et, pour la dernière, sur des concepts et des relations sémantiques entre concepts. La dernière méthode met en œuvre un algorithme de calcul de similarité conceptuelle implémenté par un prototype. Notre évaluation démontre que cette méthode améliore la précision par rapport à une RI basée uniquement sur des concepts.
Mots-clés : recherche d'information, ontologie, relation sémantique, mesure de similarité

1 Introduction

Les ontologies et les annotations sémantiques qu'elles permettent ont contribué fortement au développement récent de moteurs de recherche plus efficaces. On peut citer les systèmes KIM [16], Hakia [1] ou encore tous les systèmes cités dans [18].

En accord avec [7], le but de ces systèmes est d'améliorer les performances en dépassant les modèles de recherche basés sur les mots clés.

Pour améliorer les performances des systèmes de recherche d'information (**RI**) sur les documents, la prise en compte de la sémantique des termes

1. http://www.hakia.com

est privilégiée. Ainsi, l'indexation se situe au niveau des concepts (les sens des mots) et permet de mieux décrire le contenu du document et le contenu de la requête. Dans ces approches, des ressources sémantiques telles que des thésaurus ou des ontologies sont utilisées dans les phases d'indexation et de recherche. L'idée d'intégrer des ressources sémantiques dans la recherche d'information a été développée parallèlement à l'évolution du web sémantique. Les premiers travaux proposant une architecture pour la RI sémantique [11] utilisent des ontologies pour l'annotation et l'indexation sémantique des textes dans le contexte du web sémantique.

Cependant, la RI sémantique est plus efficace dans des domaines limités, étant donné les problèmes de production de la ressource sémantique associée.

Les expériences positives ont toujours été menées dans des domaines limités comme le domaine médical [1] et [24] ou le domaine de la biologie [14].

La recherche d'information sémantique (**RIS**) s'appuie donc sur une ontologie du domaine et sur un corpus reflétant ce domaine. Elle est basée sur un processus d'indexation des concepts des documents et de la requête utilisateur, sur un processus d'appariement des documents et de la requête pour fournir la meilleure réponse. Le plus souvent, elle utilise également des formules de calcul de similarité qui exploitent la relation hiérarchique classique. Nous souhaitons étendre cette exploitation des relations structurelles et intégrer dans un système de RIS des relations sémantiques.

Nos travaux contribuent au domaine sur deux points : (1) le processus d'annotation est basé sur une ressource termino-ontologique (cf. 2.1) ; et (2) un algorithme de ranking prend en compte non seulement les concepts mais également les relations sémantiques (cf. 3.3.2).

Dans cet article, nous exposons notre système de RIS et une évaluation de ce dernier. Ainsi, dans la section 2, nous présentons des travaux connexes. Dans la section 3, nous décrivons notre approche pour la recherche d'information sémantique. Dans la section 4, nous détaillons la mise en œuvre de cette proposition dans le cadre du projet MOANO [2]. Dans la section 5, nous rendons compte de l'expérimentation et de l'évaluation menées. Enfin, dans la section 6, nous concluons et nous proposons quelques perspectives.

2. `http://moano.liuppa.univ-pau.fr/`

2 Annotation, indexation et RI sémantique

2.1 Définitions

Aujourd'hui de nombreux systèmes annotent, indexent et supportent la RIS [11] : ces trois processus sont qualifiés de sémantique car cette recherche utilise une ressource termino-ontologique ou **RTO**. La notion de RTO apparaît pour la première fois dans le rapport RTP-DOC [15].

Le but est de modéliser le contenu de documents sélectionnés ou formant une collection sous la forme de réseaux de termes afin d'améliorer l'accès à la connaissance. Ces modèles ou représentations sont connus : des thésaurus, des terminologies, des langages documentaires, des index ou encore des ontologies. Ils sont généralement regroupés sous la notion de ressources terminologiques ou ontologiques [2]. [12] en propose une première définition « *Ressource informatique décrivant le vocabulaire et les concepts spécifiques à un domaine, à une communauté pour le traitement de l'information* ». Cette notion est ensuite concrétisée par les travaux de [21]. Enfin, les travaux récents de Cimiano et al.[3], de McCrae et al.[13], et de Roche et al.[22] ont proposé d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la composante terminologique et la composante conceptuelle. Les RTO [22] sont utilisées dans plusieurs tâches liées à l'ingénierie des connaissances, par exemple, pour l'étiquetage sémantique de corpus, pour l'indexation, pour la recherche d'information ou encore pour la navigation.

Nous distinguons deux types de travaux de recherche en RIS, même si les techniques sont identiques : les systèmes prenant en compte un corpus documentaire spécialisé (notre cas) et les travaux qui s'intéressent à des collections très larges de documents (génériques) comme [11].

Concernant l'**annotation sémantique**, nous pourrions la définir comme le processus qui fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite [11].

Une fois l'annotation effectuée, deux classes d'usages peuvent en découler : annoter pour extraire des connaissances ou annoter pour indexer. Nous nous plaçons dans ce deuxième usage visant l'**indexation sémantique**. Elle permet d'établir une nouvelle représentation de documents d'un corpus à partir des concepts ainsi annotés.

Concernant la **recherche sémantique**, en accord avec [25], il n'existe pas de modèle partagé. Ainsi, plusieurs définitions peuvent être trouvées concernant la RIS. Nous retiendrons celle de [5] qui proposent « ... *La RIS vise à mieux satisfaire les besoins en information des utilisateurs en*

exploitant le sens des termes utilisés. La RIS repose pour cela sur un processus d'indexation destiné à obtenir une représentation sémantique des documents et des requêtes . . . » La RIS cherche à dépasser les limites d'une recherche classique par mots clés.

2.2 Mesures de similarité sémantique

L'utilisation de ressources sémantiques, lors d'une recherche, permet de retrouver les documents qui partagent le maximum de concepts avec la requête. Dans ce cadre, des travaux portent sur le calcul de la similarité sémantique.

Aussi, l'évaluation de la similarité sémantique entre concepts est un problème connu dans le domaine de la RI. Plusieurs méthodes ont été proposées dans ce sens. On peut les classer selon trois catégories :

1. les approches basées sur la distance, c'est-à-dire sur la structure de l'ontologie, que l'on appelle mesures structurelles. Elles sont fondées sur l'analyse et l'exploitation de la structure sémantique des graphes conceptuels où les nœuds représentent les concepts et les arcs représentent la relation *is-a*. D'une manière générale, la distance est caractérisée par le plus court chemin qui fait intervenir un ancêtre commun, le plus petit généralisant, connectant potentiellement deux objets. Parmi les travaux qui implémentent ces mesures on peut citer par exemple : Rada et al. [17], Resnik [19], Wu et Palmer [27] ou encore Dudognon et al [5].
2. les approches utilisant le contenu informatif des concepts, que l'on appelle mesures conceptuelles. Elles stipulent que la distance entre deux concepts est une fonction des instances communes entre eux. La plupart de ces mesures sont fondées sur la notion de contenu informationnel d'un concept, introduite par Resnik [20]. Selon cette approche, le contenu informationnel traduit la pertinence d'un concept en tenant compte de la fréquence de son apparition dans la collection ainsi que de la fréquence d'apparition des concepts qu'il subsume. Ces mesures conceptuelles sont détaillées dans [4].
3. les approches dites hybrides combinent les approches basées sur les arcs et celles basées sur le contenu informationnel qui est considéré comme facteur de décision. On peut citer comme exemple la formule de Jiang et Conrath[10].

Ces mesures sont intégrées dans différentes applications, telles que le

calcul de similarité entre documents, le clustering de documents, la désambiguïsation sémantique, l'indexation, etc.

Les relations hiérarchiques classiques, *part-of*, *is-a* ne suffisent pas à exprimer la sémantique contenue dans les documents et les requêtes. Aussi, il faut modéliser des relations sémantiques et trouver des méthodes permettant d'évaluer la similarité sémantique. C'est ce que nous proposons dans la section suivante.

3 Identification des concepts et des relations sémantiques

Dans cette partie, nous allons présenter notre démarche pour l'annotation de concepts et de relations dans un document ou dans une requête ainsi que l'appariement de ces derniers. La méthode présentée a pour origine les travaux menés dans le groupe MELODI de l'IRIT [3] dans le cadre du projet ANR DYNAMO [6]. Nous avons repris et étendu ces travaux de manière à prendre en compte l'annotation des relations que nous allons détailler.

3.1 Notation

Nous adoptons les notations suivantes pour la formalisation du processus d'annotation de concepts et de relations, d'une part, et du processus d'appariement qui exploite des annotations, d'autre part.

c	pour un concept de l'ontologie,
d	pour un document du corpus,
f	pour un champ (titre, section, paragraphe) d'un document,
r	pour une relation sémantique,
t	pour un terme rencontré dans un document.

3.2 Formalisation des différentes notions

3.2.1 Ontologie, RTO, concepts, relations sémantiques et corpus

Les concepts de notre ontologie sont des classes d'un domaine spécifique ; par exemple, pour le domaine botanique, le concept *gladiolus* est une classe dont le végétal *glaïeul de Colville* est une sous-classe. En plus des relations hiérarchiques classiques comme *part-of* ou *is-a*, sont modélisées des relations sémantiques ; ainsi, nous pourrons noter que le *glaïeul de Colville* se *plante* en *octobre-novembre*.

3. Institut de Recherche en Informatique de Toulouse

C'est pourquoi nous définissons une **ontologie** O par l'ensemble C des concepts du domaine et par l'ensemble R des relations entre concepts et nous écrivons $O = (C, R)$. On note $R = \{r_\nu\}$ avec $r_\nu = (\delta, \nu, \rho)$ où la relation r_ν de nom ν a pour domaine de classe δ et pour co-domaine de classe ρ.

Dans une **RTO**, à chaque concept est associée une liste de termes qui *dénotent* le concept ; on note T l'ensemble des mots pouvant dénoter un concept (c'est-à-dire, des termes dont la présence dans un texte implique automatiquement la présence du concept dénoté). Rappelons que la rédaction de la liste associée à chaque concept de l'ontologie est du ressort d'un spécialiste du domaine.

Sur les **concepts**, plusieurs prédicats sont nécessaires :

$subsumes(c_i, c_j)$,	où $c_i \in C$ et $c_j \in C$, est interprétée c_i subsume c_j ;
$has\_label\_c(c, t)$	pour $c \in C$ et $t \in T$ indiquant que c a pour label le terme t.

Pour les **relations**, les prédicats sont :

$has\_label\_r(r, t)$	pour $r \in R$ et $t \in T$, indique que r a pour label le terme t ;
$relation(c_\delta, t, c_\rho)$	pour $c_\delta \in C, t \in T$ et $c_\rho \in C$ indique une relation révélée par t entre c_δ et c_ρ ;

Le **corpus** est vu comme un ensemble de documents D. Chaque document est composé de plusieurs champs. L'ensemble des n champs d'un document $d \in D$ sera noté $F_d = \{f_0, \ldots, f_n\}$. On note avec $P(F_d) = f_p$ le champ porteur du concept *pivot* c_p pour un document. Le concept pivot correspond à la classe qui caractérise le sujet principal du document (sous l'hypothèse que le document est écrit en style encyclopédique) ; par exemple, la classe *Plante* pour le domaine botanique.

3.2.2 Annotation de concepts

On note $T_{f,d}$ l'ensemble des termes présents dans le champ f du document d. L'annotation se fait par champ. On note $A_C(f) = \{c_0, \ldots, c_m\}$ l'annotation d'un champ f avec les concepts $c_0, \ldots, c_m$.

Nous définissons le prédicat suivant :

$holds\_c(f, c) \iff \exists t \in T_{f,d} \mid has\_label\_c(c', t) \ \wedge \ subsumes(c, c')$

Ainsi, un champ f contient un concept c si et seulement si un terme t dénotant un concept c' existe et si le concept c' est un descendant de c ou si $c = c'$.

3.2.3 Annotation de relations

On note $A_R(f) = \{r_0, \ldots, r_n\}$ l'annotation d'un champ f avec les relations $r_0, \ldots, r_n$.

Nous définissons le prédicat suivant :

$$
\begin{aligned}
holds\_r(f, r) \iff & \exists\, t_r, t_\delta, t_\rho \in T_{f,d} \mid \\
& \quad has\_label\_r(r, t_r) \ \wedge \ relation(c_\delta, r, c_\rho) \\
& \wedge has\_label\_c(c'_\delta, t_\delta) \ \wedge \ subsumes(c_\delta, c'_\delta) \\
& \wedge has\_label\_c(c'_\rho, t_\rho) \ \wedge \ subsumes(c_\rho, c'_\rho) \\
& \bigvee \\
& \exists\, t'_r, t'_\rho \in T_{f,d} \mid \\
& \quad has\_label\_r(r, t'_r) \wedge relation(c_p, r, c_\rho) \\
& \wedge has\_label\_c(c'_\rho, t'_\rho) \wedge subsumes(c_\rho, c'_\rho)
\end{aligned}
$$

Ainsi, un champ f contient une relation r dans deux cas :

- soit que trois termes du champ dénotent, l'un la relation et les deux autres les concepts du domaine c_δ et du co-domaine c_ρ correspondants à cette relation,
- soit que deux termes du champ dénotent la relation et le concept du co-domaine c_ρ correspondant à cette dernière ; le concept du domaine de la relation étant le concept pivot.

3.3 Appariement

3.3.1 Appariement basé sur la similarité de concepts

Soit Q l'ensemble des concepts dans une requête et D l'ensemble des concepts dans un document, et soit $F(C)$ la fonction qui donne le coefficient de dominance d'un concept C (les coefficients sont donnés par l'usager et sauvegardés dans un fichier de configuration), le poids pour un document est calculé de la façon suivante :

$$
w(Q, D) = \frac{\sum_{c_1 \in Q} (F(c_1). \max_{c_2 \in D} s(c_1, c_2))}{\sum_{c_1 \in Q} F(c_1)} \tag{1}
$$

où $s(c_1, c_2)$ est la mesure de similarité.

Cette mesure correspond à une mesure classique de similarité de concepts détaillée dans [5].

3.3.2 Appariement basé sur la similarité de concepts et de relations

Nous prenons en compte les relations pour étendre cette approche fondée sur l'appariement des concepts.

Soit R_Q l'ensemble des relations $r_1, \ldots, r_k$ trouvées dans une requête avec comme ensemble de concepts Q. Chaque relation est un triplet $r = (c_\delta, \nu, c_\rho)$ où c_δ est le concept relatif au domaine, ν le nom de relation et c_ρ le concept relatif au co-domaine. On définit $d(r) = c_\delta$, $n(r) = \nu$ et $e(r) = c_\rho$. Deux relations r_1, r_2 sont comparables uniquement si $n(r_1) = n(r_2)$. On définit l'ensemble R_D comme l'ensemble des relations trouvées dans le document D. Le poids d'un document est calculé comme $w(Q, D) + b(R_Q, R_D)$, où :

$$b(R_Q, R_D) = \frac{\sum\limits_{\substack{r_1 \in R_Q, r_2 \in R_D \\ \text{et } n(r_1)=n(r_2)}} (F(d(r_1)).s(d(r_1), d(r_2)) + F(e(r_1)).s(e(r_1), e(r_2)))}{\sum\limits_{\substack{r_1 \in R_Q, r_2 \in R_D \\ \text{et } n(r_1)=n(r_2)}} (F(d(r_1)) + F(e(r_1)))} \quad (2)$$

De manière transparente pour l'utilisateur, nous qualifions $b(R_Q, R_D)$ de *boost* qui augmente le poids d'un document D qui contient tout ou partie des relations détectées dans R_Q et, par conséquent, replace de tels documents en début de la liste résultat. Ici, pour chaque couple de relations $r_1 \in R_Q, r_2 \in R_D$ de même nom, nous calculons la similarité des concepts relatifs aux domaines et aux co-domaines de ces relations, respectivement. La somme de ces mesures de similarité, normalisée par les coefficients de dominance correspondants, détermine le *boost*.

Nous allons décrire, dans la section suivante, la mise en œuvre de notre méthode implémentée dans le cadre du projet MOANO.

4 Mise en œuvre dans le cadre de MOANO

Notre processus de RIS prend en compte :

– la collection de documents (fiches xml) et la requête de l'utilisateur (expression de son besoin),
– les différentes opérations d'annotation, d'indexation et d'appariement dont le but est la sélection des documents à présenter à l'utilisateur.

À travers l'étape d'indexation (1), le système organise la collection de documents sous la forme d'une représentation sémantique (concepts et relations). L'interrogation du fonds documentaire à l'aide d'une requête nécessite également la représentation de celle-ci sous une forme compatible (concept et relation) avec les documents (2). L'appariement requête-document (3) permet de sélectionner la liste des documents en s'intéressant à la similarité des concepts. Cette liste de documents est réordonnée selon le boost donné par les relations dans une deuxième étape d'appariement (3bis) qui tient compte de la similarité des relations. Ainsi les documents sont proposés à l'utilisateur par ordre de pertinence.

Dans le cadre du projet, le prototype développé annote et indexe les concepts et les relations. Nous avons mené une expérimentation décrite ci-après.

5 Expérimentation

Afin d'évaluer le système de RIS défini dans le cadre du projet MOANO, nous avons mis en place une expérimentation basée sur les deux versions d'appariement proposées : $ThemaSTream_1$ (similarité des concepts) et $ThemaStream_2$ (similarité des concepts et des relations).

5.1 Cadre d'évaluation de systèmes de RI sémantique

Nous nous sommes inspirés des travaux de [25], des campagnes TREC [26], ainsi que de [8] pour la préparation du protocole d'analyse et de la collection de test.

La tâche évaluée est une recherche qualifiée de *ad-hoc* dans TREC : le système de RI (**SRI**) répond à un besoin d'information par une liste de documents ordonnée par pertinence décroissante. L'évaluation vise à mesurer l'efficacité relative des SRI suivants :

– $Lucene$, SRI classique basée sur les mots clés ;
– $ThemaSTream_1$, SRI thématique basée sur les concepts ;
– $ThemaStream_2$, SRI thématique basée sur les concepts et renforcée par les relations.

Pour un *topic* donné, chaque SRI fournit une liste de couples (d, s) représentant le score s de chaque document d restitué. Classiquement, l'efficac-

ité d'un SRI est évaluée grâce aux mesures *Average Precision* (**AP**) pour chaque *topic* et *Mean Average Precision* (**MAP**) globalement. Pour faciliter la phase de jugement de pertinence des documents, nous avons choisi les métriques *Mean Relevance Rank* (**MRR**) et *Precision à 10* (P@10) qui, selon [23], correspondent à des mesures comparables de la qualité des réponses d'un SRI.
À l'image du protocole d'expérimentation de TREC, nous proposons deux niveaux de granularité d'évaluation d'un SRI : le premier niveau *topic* en calculant P@5, P@10 et MRR et le second niveau global en calculant la moyenne arithmétique des n valeurs de P@5, P@10 et de MRR, fournissant ainsi la mesure globale de performance du SRI.
À chaque niveau, les n différences observées $\langle m_i^1 - m_j^1, \ldots, m_i^n - m_j^n \rangle$ sont rapportées en pourcentage (d'amélioration ou de détérioration), où m_s^t représente la valeur de la mesure m obtenue par le système s pour le *topic* t. La significativité des tests statistiques calculée pour les différences observées est également rapportée : les p-valeurs de significativité sont calculées avec le test t de *Student* apparié (la différence est calculée entre les paires de valeurs m_i^t et m_j^t). Lorsque $p < \alpha$ avec $\alpha = 0{,}05$ la différence entre les deux échantillons testés est qualifiée de statistiquement significative [9].

Notre collection de test est composée :

- de 25 « topics » (besoin d'information) issus de questions posées sur le site *Yahoo Answers* décrivant des besoins d'information dans le domaine de la botanique. Les questions ont été sélectionnées de manière à ce qu'un concept et une relation soient présents dans notre ontologie (cf. 5.2.1) ;
- d'un « corpus » comprenant un échantillon de 1 000 fiches-plante du guide Clause Vilmorin dont certaines sont pertinentes pour les *topics* proposés ;
- de « qrels » (jugements de pertinence) désignant, pour chacun des 25 *topics*, l'ensemble des documents pertinents du corpus. Nous nous sommes ici limités à l'évaluation des dix premiers résultats restitués par chaque SRI pour chacun des topics ;
- de ressources ontologiques, décrivant un point de vue relatif au domaine de la botanique sous forme de concepts et de relations sémantiques.

5.2 Expérimentation et évaluation des prototypes ThemaStream

Notre objectif est ici de comparer le SRI $Lucene$ (système de référence) avec les SRI sémantiques $ThemaStream_1$ et $ThemaSream_2$.

5.2.1 Observation quantitative des résultats

Les résultats de la figure 1 confirment l'hypothèse selon laquelle, quand, dans les ressources ontologiques mobilisées, il existe des concepts et des relations spécifiant le contenu du besoin exprimé, une approche de RI sémantique permet d'obtenir de meilleurs résultats qu'une approche classique de RI basée sur des mots-clés.

	Moyenne arithmétique			
25 topics	P@5	P@10	MRR	nombre de résultats
Lucene	0,43	0,46	0,56	405
ThemaStream$_1$	0,53	0,58	0,65	910
Gain//Lucene	23,26%	26,09%	16,07%	
ThemaStream$_2$	0,74	0,74	0,83	910
Gain//Lucene	72,09%	60,87%	48,21%	
Gain//ThemaStream$_1$	39,62%	27,59%	27,69%	

FIGURE 1 – Analyse globale des résultats.

En effet, $ThemaStream_1$ donne de meilleurs résultats que $Lucene$ et $ThemaStream_2$ donne de meilleurs résultats que $ThemaStream_1$ et que $Lucene$. Dans le cas de la comparaison de $ThemaStream_2$ avec $Lucene$ le test t de *Student* apparié donne $p = 0{,}001$ pour P@10 et $p = 0{,}005$ pour MRR, ce qui valide la significativité des résultats obtenus.

Une analyse par topic montre une forte variabilité des résultats, la première fiche pertinente étant mieux classée par $ThemaStream_2$ (ou ex æquo) dans vingt-deux cas sur vingt-cinq et $ThemaStream_2$ assurant une meilleure précision à 10 (ou ex æquo) dans vingt-trois cas sur vingt-cinq. L'observation du nombre de documents pertinents distincts, sur les dix

premiers restitués respectivement par $Lucene$ et $ThemaStream_2$, montre la complémentarité potentielle des deux systèmes. En moyenne, seul un document pertinent sur les dix premiers restitués est commun à $Lucene$ et à $ThemaStream_2$, tous les autres étant distincts. De plus, dans certains cas $Lucene$ pourrait améliorer le MRR correspondant aux résultats de $ThemaStream_2$. Notons toutefois que la non exhaustivité de la taxonomie décrite dans l'ontologie peut être un biais à cette analyse de complémentarité.

5.2.2 Observation qualitative des résultats

Il est nécessaire de rappeler que, pour chacun des 25 *topics* expérimentés, il existe au moins un concept et une relation correspondante spécifiés dans notre ontologie botanique. La présence de plusieurs relations améliore systématiquement la qualité des résultats. La seule exception constatée correspond à la présence d'une négation dans un document restitué.

6 Conclusion et perspectives

Nous avons relaté dans cet article les travaux qui contribuent au domaine de la RIS sur deux points : (1) le processus d'annotation est basé sur une ressource termino-ontologique ; et (2) un algorithme de ranking prend en compte non seulement les concepts mais également les relations sémantiques.

De nombreux autres systèmes de RIS utilisent une ontologie et travaillent sur des collections très larges de documents. Nous nous intéressons à un corpus documentaire spécialisé et donc à une ressource terminologique dédiée. Nous avons inclus dans cette ressource des relations sémantiques pour obtenir une annotation plus fine.

L'algorithme de classement des documents développé dans notre projet tient compte de ces relations sémantiques par le biais d'un calcul de boost. Ainsi, les documents sont classés de manière plus précise.

Cependant notre expérimentation présente certaines limites. En effet, la non exhaustivité de l'ontologie qui ne décrit pas toutes les plantes du corpus, peut conduire à un biais dans les résultats entre la recherche classique et la recherche sémantique. Ainsi, Lucene, par les mots clés, annote tous les documents alors que notre SRI annote les documents avec les seuls taxons existants dans l'ontologie. Aussi, pour lever cette ambiguïté, il faudrait mettre en place deux scénarios d'évaluation :

– une mesure en prenant comme hypothèse que l'ontologie décrit toutes les plantes du corpus (scénario utilisé dans ce papier),
– une mesure en éliminant les documents annotés par Lucene et non annotés par nos systèmes du fait de l'absence de taxons dans l'ontologie.

Nos perspectives sont d'améliorer les prototypes $ThemaStream$ de manière à permettre à un utilisateur de bâtir des requêtes plus expressives et d'obtenir ainsi des résultats encore plus fins.

Remerciements

Cette recherche a été réalisée dans le cadre du projet Moano « Modèles et Outils pour Applications NOmades de découverte de territoire » (`http://moano.liuppa.univ-pau.fr/`), en partie financé par l'Agence Nationale de la Recherche (ANR-2010-CORD-024-01).

Références

[1] ABASOLO J. M. & GOMEZ M. (2000). Melisa. an ontology-based agent for information retrieval in medicine. In *In : Proceedings of the First International Workshop on the Semantic Web (SemWeb2000*, p. 73–82.

[2] AUSSENAC-GILLES N. & CONDAMINES A. (2004). Documents électroniques et constitution de ressources terminologiques ou ontologiques. *Information - Interaction - Intelligence*, **4**(1). Article publié dans le numéro thématique de la revue i3 consacré au document numérique.

[3] CIMIANO P., BUITELAAR P., MCCRAE J. & SINTEK M. (2011). Lexinfo : A declarative model for the lexicon-ontology interface. *Web Semant.*, **9**(1), 29–51.

[4] CORLEY C. & MIHALCEA R. (2005). Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, p. 13–18, Stroudsburg, PA, USA : Association for Computational Linguistics.

[5] DUDOGNON D., HUBERT G., MARCO J., MOTHE J., RALALASON B., THOMAS J., REYMONET A., MAUREL H., MBARKI M., LAUBLET P. & ROUX V. (2010a). Dynamic ontology for information retrieval. In *RIAO*, p. 213–215 : CID - Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

[6] DUDOGNON D., HUBERT G. & RALALASON B. J. V. (2010b). ProxiGénéa : Une mesure de similarité conceptuelle (regular paper). In *Colloque Veille Stratégique Scientifique et Technologique (VSST), Toulouse,*

25/10/2010-29/10/2010, p. (support électronique), http ://www.ups-tlse.fr : Université Paul Sabatier - Toulouse.

[7] FERNÁNDEZ M., CANTADOR I., LOPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : An ontology-based approach. *J. Web Sem.*, **9**(4), 434–452.

[8] HARMAN D. K. (2005). The TREC Test Collections. In [26], chapter 2, p. 21–53.

[9] HULL D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR'93 : Proceedings of the 16th annual international ACM SIGIR conference*, p. 329–338, New York, NY, USA : ACM Press.

[10] JIANG J. & CONRATH D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, p. 19–33.

[11] KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *J. Web Sem.*, **2**(1), 49–79.

[12] LORTAL G. (2006). Annotations dans les activités coopératives : élaboration d'un modèle générique multi-points de vue et utilisation des technologies du web sémantique pour sa mise en œuvre. *Doctorat en informatique, Université de Technologie de Troyes*.

[13] MCCRAE J., SPOHR D. & CIMIANO P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web : research and applications - Volume Part I*, ESWC'11, p. 245–259, Berlin, Heidelberg : Springer-Verlag.

[14] MÜLLER H.-M., KENNY E. E. & STERNBERG P. W. (2004). Textpresso : An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, **2**(11), e309.

[15] PÉDAUQUE R. & SALAÜN J. (2006). *Le document à la lumière du numérique*. C&F Editions.

[16] POPOV B., KIRYAKOV A., OGNYANOFF D., MANOV D. & KIRILOV A. (2004). Kim - a semantic platform for information extraction and retrieval. *Natural Language Engineering*, **10**(3-4), 375–392.

[17] RADA R., MILI H., BICKNELL E. & BLETTNER M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, **19**(1), 17–30.

[18] RAMÍREZ R. C. M. & R. V. M. R. (2007). A semantic web approach to enrich information retrieval answers. In *ICEIS (4)*, p. 299–302.

[19] RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, p. 448–453 : Morgan Kaufmann.

[20] RESNIK P. (1999). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural lan-

guage. *J. Artif. Intell. Res. (JAIR)*, **11**, 95–130.

[21] REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de Ressources Termino-Ontologiques en OWL. In F. TRICHET, Ed., *Actes des Journées Francophones d'Ingénierie des Connaissances (IC 2007)*, p. 169–180, Grenoble, France : Cépaduès Editions.

[22] ROCHE C., CALBERG-CHALLOT M., DAMAS L. & ROUARD P. (2009). Ontoterminology : A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development*, p. 321–326, Madeira, Portugal.

[23] SANDERSON M., PARAMITA M. L., CLOUGH P. & KANOULAS E. (2010). Do user preferences and evaluation measures line up ? In *SIGIR*, p. 555–562.

[24] TRIESCHNIGG R., PEZIK P., LEE V., DE JONG F., KRAAIJ W. & REBHOLZ-SCHUHMANN D. (2009). Mesh up : effective mesh text classification for improved document retrieval. *Bioinformatics*, **25**(11), 1412–1418.

[25] UREN V. S., SABOU M., MOTTA E., FERNÁNDEZ M., LOPEZ V. & LEI Y. (2010). Reflections on five years of evaluating semantic search systems. *IJMSO*, **5**(2), 87–98.

[26] VOORHEES E. M. & HARMAN D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA : MIT Press.

[27] WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, p. 133–138, Stroudsburg, PA, USA : Association for Computational Linguistics.