Une plate-forme open-source de recherche d'information sémantique

Ines Bannour, Haïfa Zargayouna

Laboratoire d'Informatique de l'université Paris-Nord
(LIPN) - UMR 7030
Université Paris 13 - CNRS
99, avenue Jean-Baptiste Clément 93430 Villetaneuse, France
Email: prenom.nom@lipn.univ-paris13.fr

RÉSUMÉ. *Les méthodes de RIS visent à s'affranchir des problèmes classiques de synonymie et polysémie via le passage au niveau conceptuel. Elles reposent souvent sur l'utilisation d'une ressource sémantique. La qualité des résultats dépend des fonctionnalités sémantiques mises en place ainsi que de la qualité de la ressource utilisée. Malgré la profusion des propositions, l'apport d'une sémantique explicite reste à prouver. Nous proposons une décomposition des fonctionnalités qui sont communes aux différentes méthodes de RI. Nous présentons ensuite Terrier SIR qui, via la dissociation des modules sémantiques, favorise aussi bien l'implémentation des méthodes de RIS que leurs éventuelles mises à jour ou adaptations. Nous avons effectué des expérimentations pour mettre en évidence l'intérêt d'une telle plate-forme.*

ABSTRACT. *SIR methods propose to overcome the classic problems of synonymy and polysemy by using an external semantic resource that enable to reason on a conceptual level. The quality of the results depends on the proposed semantic functionalities and on the quality of the used resource. In spite of the number of proposed methods, the contribution of an explicit semantic is not proven yet. The proposed methods are hardly comparable because it is usually difficult to dissociate the semantic functionalities that have to be isolated. In this paper, we describe the semantic functionalities that are common to semantic retrieval approaches. Then, we present Terrier SIR which is a platform that enriches Terrier IR by taking into account these functionalities. The proposed platform ensures the implementation of SIR methods and enables their update or evolution which ease to settle on comparative evaluations. Some experiments were carried on, in order to show the interest of this platform.*

MOTS-CLÉS : *Plate-forme, Ontologies, Moteur de recherche, Fonctionnalités sémantiques*

KEYWORDS: *Platform, Ontologies, Search engine, Semantic functionalities*

CORIA 2012, pp. 167–178, Bordeaux, 21-23 mars 2012

1. Introduction

La recherche d'information sémantique (RIS) a pour but de pallier aux lacunes, souvent citées, de la recherche d'information classique, à savoir les problèmes de synonymie et de polysémie des termes. Les méthodes de RIS visent à s'affranchir de ces problèmes via le passage au niveau conceptuel. Elles reposent souvent sur l'utilisation d'une ressource sémantique externe (thésaurus ou ontologie). En 1968, Salton constate déjà que l'utilisation d'un thésaurus (Harris Synonym) permet d'améliorer les performances, à condition que les termes utilisés pour l'enrichissement soient validés manuellement par l'utilisateur. Il constate également que l'expansion automatique utilisant l'ensemble des termes possibles, dégrade ces performances (Salton, 1968).

La qualité des résultats dépend des fonctionnalités sémantiques mises en place ainsi que de la qualité de la ressource utilisée. Malgré la profusion des propositions, l'apport d'une sémantique explicite reste à prouver. En effet, les méthodes de RIS ne sont pas comparables parce qu'il est souvent difficile de dissocier les fonctionnalités sémantiques pour pouvoir les évaluer séparément (Zargayouna, 2011). Nous proposons une décomposition des fonctionnalités qui sont communes aux différentes méthodes de recherche d'information sémantique et leur intégration modulaire au sein d'une plate-forme de recherche d'information.

Nous commençons dans un premier lieu par décrire l'intérêt de l'utilisation des ressources sémantiques externes en RIS (section 2), puis par recenser l'ensemble des fonctionnalités sémantiques que nous isolons dans notre plate-forme (section 3). En second lieu, nous présentons brièvement la plate-forme *Terrier IR* ainsi que l'architecture générale de *Terrier SIR* : la nouvelle plate-forme mise en place (section 4). Finalement, nous montrons l'intérêt d'une telle plate-forme et la manière dont elle peut être utilisée via l'implémentation et le test de certaines méthodes de RIS (section 5).

2. Exploitation des ressources sémantiques en recherche d'information

Les travaux en recherche d'information sémantique peuvent être classés en deux catégories (non exclusives) selon la phase de prise en compte du niveau sémantique : (i) pendant la phase d'indexation ou (ii) pendant la phase de recherche.

2.1. *Utilisation des ressources sémantiques en phase d'indexation*

La phase d'indexation consiste à identifier dans chaque document certains éléments significatifs qui serviront de clés pour retrouver ce document au sein d'une collection. Elle produit une représentation des documents (l'index). Pour faciliter le processus d'interrogation, le même processus d'indexation s'applique généralement aux requêtes.

En RIS, l'indexation sémantique ou conceptuelle consiste à intégrer dans la représentation des documents, les entités sémantiques représentées dans la ressource.

Les premiers travaux en indexation sémantique utilisent le thésaurus WordNet. WordNet est un réseau lexical qui a l'avantage d'être une ressource très fournie, le grand nombre de *synsets* ainsi que sa facilité d'accès [1] en ont fait une ressource très utilisée en recherche d'information. Néanmoins, l'utilisation de WordNet nécessite une phase de désambiguïsation, un terme pouvant être relié à plusieurs *synsets*. (Stetina *et al.,* 1998) ont effectué des expérimentations sur un petit extrait du *Brown Corpus*. Ils ont montré que la désambiguïsation des termes a contribué à une augmentation du rappel qui a excédé 80% comparée à une indexation qui sélectionne le premier synset retourné (un rappel de 75,2%). (Gonzalo *et al.,* 1998) ont montré qu'en plus de la désambiguïsation, le choix de l'unité d'index a un impact direct sur la qualité de la recherche. Les résultats retournés par une indexation par les *synsets* obtiennent un rappel de 62%, contre 53,2% pour une indexation par les termes-sens[2] et 48% pour une indexation classique par termes (le système SMART).
Le système OntoSeek (Guarino *et al.,* 1999) propose une indexation conceptuelle qui repose sur *Sensus* : une fusion de WordNet et de l'ontologie Penman. Il permet une formalisation des documents et des requêtes par des graphes conceptuels, mais nécessite une phase manuelle de sélection du sens adéquat des concepts dans le cas de la polysémie. OntoSeek a été testé dans le cadre d'une recherche en ligne des pages jaunes et des catalogues de produits.

Même si les ontologies par définition sont des ressources qui ont été désambiguïsées au préalable, une désambiguïsation peut se révéler utile quand des termes sont assez généraux et peuvent de ce fait faire référence à des sous-domaines différents. Khan (Khan, 2000) propose une indexation conceptuelle guidée par une ontologie dans le domaine du sport. La désambiguïsation des concepts qu'il a effectuée, est fondée sur deux principes : la co-occurrence et la proximité sémantique. Les résultats obtenus prouvent une amélioration notable des performances : l'indexation par les concepts augmente le rappel à 92,25% et la précision à 88,22% comparée à une indexation par mots clés[3].
Les ontologies sont de plus en plus utilisées pour une indexation riche des documents de spécialité et ont montré leur utilité dans certains domaines tels que le domaine médical (Abdulahhad *et al.,* 2011), le domaine juridique (Berrueta *et al.,* 2006), etc.

2.2. *Utilisation des ressources sémantiques en phase de recherche*

L'utilisation des ressources sémantiques pendant la phase de recherche se traduit par l'expansion et la reformulation des requêtes.

---

1. WordNet est accessible en ligne et en téléchargement avec des APIs pour avoir accès à sa structure interne.
2. Les termes sont indexés avec leurs synsets. Ce choix ne permet pas de retrouver les synonymes qui seraient présents dans WordNet mais pas dans les documents.
3. Qui obtient un rappel de 56% et une précision de 67,75%.

L'expansion des requêtes est réalisée en étendant le vocabulaire des requêtes au moyen de termes similaires (généralement des synonymes) ou proches. Le but de l'expansion de requêtes est d'élargir l'ensemble de documents retournés et d'augmenter la précision. Afin d'évaluer l'impact de la taille des requêtes, (Lu *et al.,* 1995) rendent compte de trois expériences sur le corpus TREC3. Ces expérimentations ont permis de constater que l'expansion des requêtes (via un thésaurus) améliore la précision de 33%. Plus récemment, (Baziz, 2005) utilise WordNet afin de récupérer les termes reliés à la requête par des relations de synonymie, généralisation et spécialisation. Ses expérimentations, réalisées sur le corpus CLEF2001, avec un ensemble de 50 requêtes ont montré que le processus d'expansion de requêtes dépend aussi bien du nombre de concepts que des liens sémantiques à considérer.

La reformulation des requêtes, quant à elle, est une technique qui a été utilisée dans les retours arrière sur pertinence par (Harman, 1988) ainsi que (Buckley *et al.,* 1994) qui arrivent à des conclusions similaires à celles de (Salton, 1968).

3. Les fonctionnalités sémantiques

Nous avons présenté quelques travaux qui intègrent les ressources sémantiques dans un processus de recherche d'information. Ces travaux se différencient selon qu'ils intègrent ces ressources lors de la phase d'indexation ou de recherche.

Les ressources utiles en recherche d'information sont des ontologies lexicales ou termino-ontologies. La majorité des travaux exploitent WordNet en utilisant les liens lexicaux et conceptuels (hiérarchiques ou méronymiques).

Les expériences ponctuelles ont montré avec plus ou moins de succès l'apport de ces ressources à la recherche d'information. Il est néanmoins difficile de dresser un bilan. Ceci est essentiellement dû au fait que les expérimentations effectuées sont difficilement reproductibles et ne facilitent pas par ce fait les évaluations comparatives. De plus l'hétérogénéité des approches proposées, nous oblige à considérer les systèmes dans leur globalité. Il est difficile de dissocier l'impact de la qualité de la ressource de la pertinence de son utilisation.

Nous proposons de dégager les fonctionnalités sémantiques qui sont communes à ces approches afin d'avoir une vision modulaire et différenciée.

3.1. *L'annotation sémantique*

Cette fonctionnalité consiste à reconnaître dans les documents des unités d'informations qui seront associés à des entités sémantiques représentées dans la ressource. En recherche d'information cette phase est souvent couplée au choix des unités d'index. Une indexation conceptuelle revient à associer les concepts aux documents. Nous proposons d'isoler l'annotation sémantique du choix des unités d'index. Ceci permet de dissocier les choix liés à la richesse de la ressource et aux difficultés d'ancrage dans

le texte, des choix liés à la richesse de l'index ainsi que sa structuration.
La désambiguïsation des termes sera alors à intégrer dans cette fonctionnalité. Même si l'ancrage sémantique en recherche d'information se restreint au niveau des concepts, la reconnaissance des relations entre concepts peut s'avérer utile pour certaines applications.
Isoler l'annotation sémantique permet aussi de juger de sa qualité séparément. Dans le cas d'annotations formelles par exemple, il est possible de valider sa cohérence en profitant des inférences du langage d'annotation (Ma *et al.,* 2011). Il serait également possible d'évaluer l'impact de l'annotation sémantique par rapport à une annotation manuelle.

Cette fonctionnalité permet aussi de juger de la couverture de la ressource (Hernandez *et al.,* 2005) et de calculer le degré d'ambiguïté.

3.2. *Choix d'unité d'index*

Le choix des unités d'index est un facteur déterminant dont dépend fortement la qualité de la recherche. Tout élément ne figurant pas dans ces unités d'index ne sera pas retrouvé dans le processus de recherche. Il s'agit donc de sélectionner les descripteurs représentatifs et discriminants.

La majorité des méthodes de RIS proposent d'indexer les documents par les concepts qui ont été ancrés dans les textes (résultat de l'annotation). Les termes non ancrés sont jugés non représentatifs. Le choix de ces unités d'index n'est pas trivial, si les concepts sélectionnés sont trop généraux, le système risque de perdre en précision. Si au contraire les concepts sont trop spécifiques (en ne considérant que les feuilles), le système risque de perdre en rappel. Il s'agit donc de choisir une stratégie qui garantie un équilibre entre exhaustivité et abstraction.

(Seydoux, 2006) a proposé une indexation par concepts tout en exploitant les liens hiérarchiques. Cette méthode est fondée sur la notion de Coupe de Redondance Minimale (CRM). La CRM est définie par Seydoux comme étant l'ensemble minimal de sommets couvrant la totalité des feuilles du graphe orienté sans cycle modélisant la ressource sémantique. Elle permet ainsi de garantir une bonne couverture du corpus, tout en minimisant la redondance introduite par la présence de sommets voisins.

(Zargayouna, 2005) propose de choisir les termes comme entités d'index et de les enrichir par leur voisinage sémantique (termes ancrés à des concepts voisins). L'exploitation de la richesse des liens dans la ressource est alors reportée à la phase de pondération des termes.

3.3. *Le choix de la stratégie de pondération sémantique*

La stratégie de pondération est fortement en rapport avec le choix des unités d'index en question. Le fait de dissocier ces deux fonctionnalités permet de tester des stratégies différentes de pondération et de déterminer les meilleures combinaisons.

Plusieurs méthodes de pondération des concepts ont été proposées. Nous citons celle proposée par (Hernandez *et al.,* 2005) qui prend en considération à la fois la représentation statistique et sémantique d'un concept dans un granule g (g pouvant être un document, une partie du document, ou un regroupement de documents). La méthode de pondération de (Baziz, 2005) exprime la pondération d'un concept C dans un document d par **CF_IDF**, une variante de TF_IDF qui prends en compte les liens hiérarchiques entre concepts. En effet, la fréquence d'un concept dans un document dépend du nombre d'occurrences du concept lui-même, et de celui de ses sous-concepts. Formellement[4] le poids global d'un concept C dans un document d, $CF\_IDF(C, d)$, est alors calculé comme suit :

$$CF\_IDF(C, d) = cf(C) * \ln(\tfrac{N}{df}) \quad [1]$$

avec :

– N le nombre de document de la collection. df le nombre de documents contenant C. $cf(C)$ représente la fréquence du concept C. Elle est calculée comme suit :

$$cf(C) = count(C) + \sum_{SC \in SubConcepts(C)} \frac{Length(SC)}{Length(C)} * count(SC) \quad [2]$$

avec :

– $count(C)$ mesure le nombre d'occurrences de C dans le document d (toutes ses occurrences lexicales). $SubConcepts(C)$ l'ensemble des sous-concepts de C. $Length$ calcule le nombre de termes que contient la dénotation lexicale de C.

La pondération des termes par enrichissement sémantique de (Zargayouna, 2005)[5] repose sur la notion de similarité entre concepts. Elle exprime la pondération d'un terme t dans un document d par **SemTF_IDF**, comme suit :

$$SemTF\_IDF(t, d) = SemTF(t, d) * SemIDF(t, d) \quad [3]$$

$SemTF(t, d)$ (*Semantic Term Frequency*) calcule la fréquence d'un terme enrichie par la similarité des termes voisins qui sont pondérés par leurs fréquences :

$$SemTF(t, d) = TF(t, d) + \sum_{t_i \in V} TF(t_i, d) * sim(\varphi(t), \varphi(t_i)) \quad [4]$$

avec :

4. Pour avoir une présentation uniforme, nous avons pris la liberté de changer les notations originales des formules.
5. Les formules ont été adaptées en faisant abstraction de la prise en compte de la structure des documents.

– $\varphi(t)$ représente le concept associé à t (si t n'est pas ancré $\varphi(t) = t$). $sim(\varphi(t), \varphi(t_i))$ mesure de similarité entre les concepts $\varphi(t)$ et $\varphi(t_i)$. V représente l'ensemble des termes qui constituent le vocabulaire de la collection.

$SemIDF(t, d)$ représente l'importance du terme (ou concept si le terme est ancré).

$$SemIDF(t, d) = \ln(\frac{N}{|D^{\varphi(t)}|} + 1) \quad [5]$$

avec :

– N étant le nombre de documents de la collection. $|D^{\varphi(t)}|$ le nombre de documents contenant au moins une occurrence de $\varphi(t)$.

La prise en compte des liens sémantiques explicités dans la ressource est généralement assurée par des mesures de similarité sémantique (Zargayouna *et al.,* 2004). Nous considérons le calcul de ces mesures comme faisant partie du calcul de pondération.

4. Terrier SIR : Une plate-forme de recherche d'information sémantique

Nous proposons d'enrichir un SRI open-source existant (*Terrier IR*) en y intégrant les différentes fonctionnalités sémantiques décrites ci-dessus. *Terrier SIR* (Terrier Semantic Information Retrieval) est une plateforme de recherche d'information sémantique qui favorise aussi bien l'implémentation des méthodes de RIS que leurs éventuelles mises à jour ou adaptations.

4.1. *Terrier IR*

Terrier est une plate-forme dédiée à la recherche d'information. Elle implémente les différents modules intervenant dans le processus de RI classique et offre en plus un cadre pour l'évaluation des résultats de recherche pour différentes applications (Ounis *et al.,* 2006). Terrier a été été largement éprouvée (Middleton *et al.,* 2007). Le choix de cette plate-forme est dû aussi à sa capacité à traiter de grandes collections de documents telles que les collections TREC.

L'architecture de la plate-forme Terrier distingue les deux phases classiques : l'indexation et la recherche. Un corpus documentaire est fourni en entrée au module d'indexation. Les documents de la collection passent par un ensemble de pré-traitements tels que la *tokenisation*. Les tokens sont ensuite injectés dans une chaîne de traitement *TermPipLines*, à savoir le *StopWords PipeLine* pour l'élimination des mots vides de sens, ou encore les *Stemming pipeline* et qui dépendent de la langue en question. La phase d'indexation conduit à la construction de l'index (*Data structures*).

La phase de recherche comprend le *Manager*, un module qui interagit avec l'application, réalise la mise en correspondance à travers les calculs des pondérations (selon le schéma de pondération (*Weighting Model*) choisi : PL2, BM25, etc.) ainsi que les

scores des documents. Le résultat renvoyé à l'utilisateur, est la liste des documents jugés pertinents et leurs scores respectifs.

4.2. *Architecture de Terrier SIR*

Nous avons enrichi la plate-forme *Terrier IR* pour la mise en place des fonctionnalités sémantiques, traduites en un ensemble de modules.

La figure 1 présente l'architecture générale de *Terrier SIR* et montre l'intégration de ces modules. Ces différentes fonctionnalités sont en interaction avec les modules d'indexation et de recherche mais parfaitement dissociées pour une meilleure réutilisabilité de la plate-forme. La plate-forme conserve évidement le fonctionnement de base de *Terrier IR*.

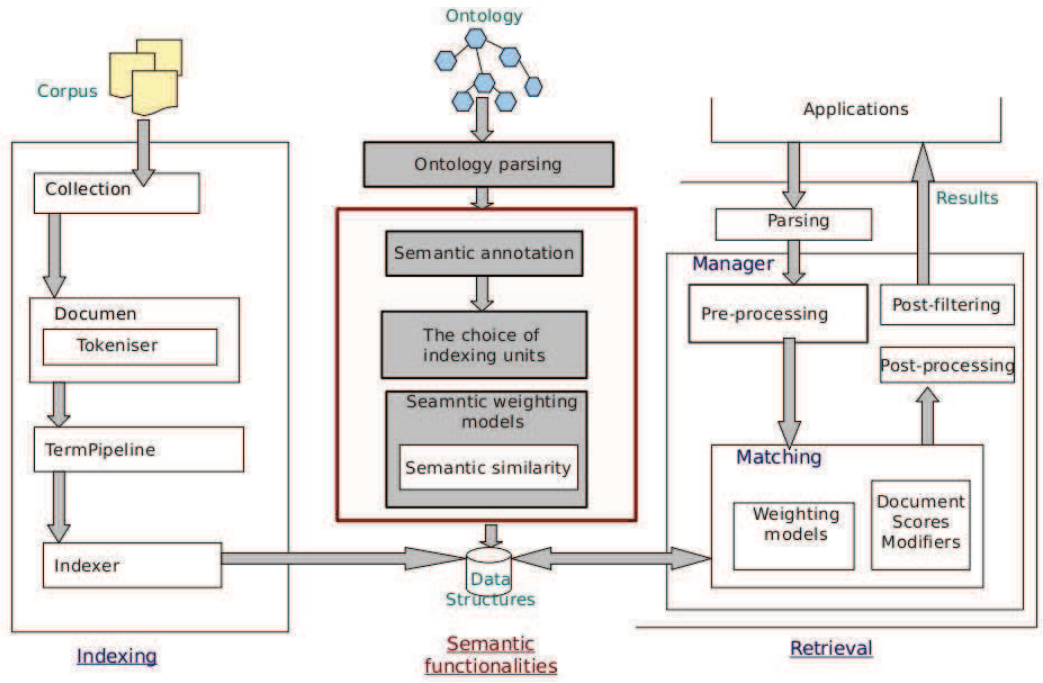

Figure 1. *Architecture-Terrier SIR*

Terrier SIR prend en entrée un corpus documentaire, mais également l'ontologie choisie. L'annotation sémantique permet de mettre en oeuvre différentes techniques de désambiguïsation. Ce module peut aussi servir à récupérer le résultat d'une annotation manuelle et à l'injecter dans la chaîne. L'annotation sémantique est appliquée aux documents et aux requêtes au regard de l'ontologie fournie en entrée.

Le *parsing d'ontologies* est considéré comme un module extérieur mais qui peut avoir un certain impact sur l'annotation sémantique. Nous ne considérons pas ce *parsing* comme une fonctionnalité à part entière mais plutôt comme un pré-traitement de la ressource sémantique.

Pendant la phase de recherche, les mêmes traitements sont réalisés sur l'ensemble des requêtes. La mise en correspondance est réalisée selon *la méthode de pondération sémantique* implémentée.

5. Exemple d'utilisation de Terrier SIR

La plate-forme proposée permet la mise en place et l'évaluation de divers méthodes de RIS. Notre but n'est pas l'évaluation de ces différentes méthodes mais plutôt l'élaboration d'un cadre favorisant l'instanciation des différentes fonctionnalités présentées auparavant et la prise en compte de l'ontologie dans un SRI. Cependant, et à titre d'exemple nous présentons les premiers résultats de l'implémentation de certaines méthodes de RIS.

5.1. *Environnement de test*

Dans le but d'évaluer la plate-forme mise en oeuvre, nous avons implémenté certaines des méthodes de RIS citées dans l'état de l'art. Nous utilisons pour ce fait l'API *Jena* pour le *parsing d'ontologies* ainsi que *Yatea* (Hamon *et al.,* 2006) pour *l'extraction des termes*[6] (simples et composés).

Le module d'annotation sémantique implémenté consiste à comparer termes extraits aux étiquettes des concepts après lemmatisation. Les concepts ancrés sont ceux qui sont dénotés par des termes qui ont été mis en correspondance.

Nous avons mis en place : (i) une indexation par termes simples (**TS**)[7], (ii) une indexation par termes simples et composés **TS&TC**[8] (méthode de pondération TF-IDF), (iii) une indexation par concepts avec la pondération **CF_IDF** , (iv) une indexation par termes simples (**TS-SemTF.IDF**) et (v) une indexation par les termes simples et composés (**TS&TC-SemTF.IDF**) (méthode de pondération **SemTF_IDF** présentée en section 3.3), la mesure de similarité est celle de (Wu *et al.,* 1994), le seuil est fixé à 0.7.

Nous utilisons le produit scalaire comme mesure de correspondance pour ces différentes méthodes de RI. Les expérimentations sont réalisées sur un corpus de recettes de cuisine et une ontologie de cuisine.

---

6. Yatea nécessite au préalable l'étiquetage morpho-syntaxique de *Treetagger*
7. Pour la méthode **TS** les termes représentent les lemmes simples extraits par TreeTagger (noms, adjectifs et adverbes)
8. Pour la méthode **TS&TC**, les lemmes simples sont les mêmes que ceux utilisées pour **TS** et les lemmes composés sont extraits par Yatea

5.2. *Expérimentations*

La base de test est composée d'un corpus documentaire de 1489 recettes de cuisine fournies dans le cadre de la compétition internationale *Computer Cooking contest* (CCC). Les tests sont effectués sur un ensemble de 4 requêtes dont nous disposons de jugements de pertinence [9]. L'ontologie utilisée est une ontologie de cuisine formalisée en OWL, elle comprend 4552 concepts[10]. Nous avions à notre disposition une version préliminaire, une version améliorée avec un meilleur niveau lexical et une meilleure structuration est en cours de développement. Le but de cette expérimentation n'est pas d'effectuer une évaluation comparative mais plutôt de montrer la faisabilité d'une telle évaluation. Vu la qualité de l'ontologie (Després *et al.,* 2009), le nombre de documents jugés pertinents et le processus d'annotation plutôt "simple" mis en place, nous savons qu'il est prématuré de tirer des conclusions d'une telle expérimentation.

Le tableau 1 présente l'évaluation des résultats obtenus par les différentes méthodes de RI implémentées.

		Méthodes de recherche d'information				
	REQ	TS	TS&TC	TS-SemTF.IDF	TS&TC-SemTF.IDF	CF.IDF
MAP						
	REQ1	0.15	0.16	**0.23**	0.05	0.09
	REQ2	0.46	0.55	0.61	**0.68**	0.32
	REQ3	0.16	0.12	**0.25**	0.24	0.15
	REQ4	**0.51**	**0.51**	0.28	0.15	0.12
MAP	ALL	0.32	**0.34**	**0.34**	0.28	0.17

Tableau 1. *Évaluation de différentes méthodes de RI*

Le tableau 1 montre une légère amélioration de la précision moyenne des méthodes d'indexation par termes simples et composés (TS&TC et TS&TC-SemTF.IDF) comparées aux méthodes d'indexation par les termes simples uniquement (TS et TS-SemTF.IDF). L'indexation par concepts a été plus impactée par la "qualité" de l'ontologie. La mesure de pondération CF_IDF semble être moins adaptée à l'ontologie

---

9. La liste des requêtes est la suivante :

- **REQ1** : Cook an Asian soup with leek (documents pertinents :5).
- **REQ2** : Prepare a fruit salad with kiwano (documents pertinents :8).
- **REQ3** : I would like to cook a pear pancake (documents pertinents :4).
- **REQ4** : Prepare a cake with plum (documents pertinents :7).

10. Cette ontologie a été produite dans le cadre du projet TAAABLE. Une version améliorée est accessible à ($http : //wikitaaable.loria.fr/index.php/Food\_tree$)

utilisée (il n'y a pas beaucoup de liens de composition entre les étiquettes des concepts comme c'est le cas pour WordNet).

6. Conclusion et perspectives

Nous proposons une décomposition de la prise en compte de la sémantique en recherche d'information en fonctionnalités sémantiques modulaires. La traduction de ces fonctionnalités en modules différenciés, favorise aussi bien l'implémentation des méthodes de RIS que leurs éventuelles mises à jour ou adaptations. Elle permet de mettre en place des évaluations comparatives et d'avoir une visibilité sur l'apport de chaque module. Nous avons implémenté quelques méthodes de l'état de l'art et effectué des premières expérimentations pour mettre en évidence l'intérêt d'une telle plate-forme et la manière dont elle peut être utilisée. Nous voudrions à terme pouvoir intégrer des méthodes du Web Sémantique au processus de Recherche d'Information.

Des expérimentations à plus grande échelle sont en cours de développement, l'utilisation d'un extracteur de termes pose problème sur des collections volumineuses (telle que WT10G de TREC), la question d'indexation par plusieurs ontologies est aussi posée. Le parsing d'ontologies s'avère aussi délicat, les dénotations en OWL pouvant varier[11] et influencer la qualité de l'annotation et le calcul de pondération.

Nous comptons aussi mettre en place des tests de robustesse et de fiabilité, le but étant de pouvoir distribuer cette plate-forme pour pouvoir la faire évoluer d'une manière collaborative.

7. Bibliographie

Abdulahhad K., Chevallet J.-P., Berrut C., « Solving Concept mismatch through Bayesian Framework by Extending UMLS Meta-Thesaurus. », *CORIA'11*, p. 311-326, 2011.

Baziz M., Indexation conceptuelle guidée par ontologie pour la recherche d'information, PhD thesis, 2005.

Berrueta D., Labra J. E., Polo L., « Searching over Public Administration Legal Documents Using Ontologies », *Proceedings of the 2006 conference on Knowledge-Based Software Engineering : Proceedings of the Seventh Joint Conference on Knowledge-Based Software Engineering*, 2006.

Buckley C., Salton G., Allan J., Singhal A., « Automatic Query Expansion Using SMART : TREC 3 », *TREC*, 1994.

Després S., Zargayouna H., Bentebibel R., « Quelles connaissances pour se mettre à TAAABLE ? », *Actes de 17ème atelier sur le raisonnement à partir de cas - RàPC-09 (2009)*, 2009.

11. Les étiquettes des concepts peuvent être représentées dans des labels ou directement dans l'uri par exemple.

Gonzalo J., Verdejo F., Chugur I., Cigarrán J. M., « Indexing with WordNet synsets can improve Text Retrieval », *CoRR*, 1998.

Guarino N., Masolo C., Vetere G., « OntoSeek : Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs », *National Research Council, LADSEBCNR*, 1999.

Hamon T., Aubin S., « Improving Term Extraction with Terminological Resources », *in* , T. Salakoski, , F. Ginter, , S. Pyysalo, , T. Pahikkala (eds), *Advances in Natural Language Processing*, vol. 4139 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, Berlin, Heidelberg, chapter 39, p. 380-387, 2006.

Harman D., « Towards interactive query expansion », *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '88, ACM, New York, NY, USA, p. 321-331, 1988.

Hernandez N., Mothe J., Poulain S., « Accessing and mining scientific domains using ontologies : the OntoExplo System », *Poster, Proceedings of The 28th Annual International ACM SIGIR, pp : 607-608*, 2005.

Khan L. R., Ontology-based Information Selection, PhD thesis, Faculty of the Graduate School, University of Southern California, August, 2000.

Lu X. A., Keefer R. B., « Query expansion/reduction and its impact on retrieval effectiveness », *Proceedings of the Third Text Retrieval Conference TREC-3*, p. 231-239, 1995.

Ma Y., Levy F., Ghimire S., « Reasoning with annotations of texts », *24th International FLAIRS Conference (FLAIRS'11) : Track AI, Cognitive Semantics, Computational Linguistics and Logics*, 2011.

Middleton C., Baeza-yates R., A Comparison of Open Source Search Engines, Technical report, October, 2007.

Ounis I., Amati G., Plachouras V., He B., Macdonald C., Lioma C., « Terrier : A High Performance and Scalable Information Retrieval Platform », *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

Salton G., *Automatic Information Organization and Retrieval*, McGraw Hill Text, 1968.

Seydoux F., Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire, PhD thesis, Lausanne, 2006.

Stetina J., Kurohashi S., Nagao M., « General Word Sense Disambiguation Method Based on a Full Sentential Context », *In usage of WorDnet in natural language processing, proceeding of Coling-ACL workshop*, 1998.

Wu Z., Palmer M., « Verbs semantics and lexical selection », *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 133-138, 1994.

Zargayouna H., Indexation sémantique de documents XML, PhD thesis, Université Paris-Sud, December, 2005.

Zargayouna H., « Quelle évaluation pour la Recherche d'Information Sémantique », *Troisième Atelier Recherche d'Information SEmantique RISE@CORIA'2011*, 2011.

Zargayouna H., Salotti S., « Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML », *Actes de la conférence Ingénierie des Connaissances, IC'2004*, 2004.