Recherche d'information sémantique : état des lieux

H. Zargayouna, Catherine Roussey, J.P. Chevallet

Recherche d'information sémantique : état des lieux

Haïfa Zargayouna[*] — **Catherine Roussey**[**] — **Jean-Pierre Chevallet**[***]

[*] *LIPN (CNRS UMR 7030 - Université Paris 13 Sorbonne Paris Cité), Villetaneuse, France*
haifa.zargayouna@lipn.univ-paris13.fr
[**] *Irstea UR TSCF, 9 avenue Blaise Pascal, CS 20085, Aubière, France*
catherine.roussey@irstea.fr
[***] *LIG (CNRS UMR 5217 - Université de Grenoble Alpes), Grenoble, France*
Jean-Pierre.Chevallet@imag.fr

RÉSUMÉ. *Cet article fait le point dans le domaine de la recherche d'information sémantique (RIS), qui est à l'intersection de plusieurs disciplines : recherche d'information (RI), ingénierie des connaissances (IC) et traitement automatique des langues (TAL). Nous présentons les notions de base pour comprendre comment fonctionnent les systèmes de RI sémantiques ainsi qu'une classification des types de ressources sémantiques utilisés en recherche d'information. Nous détaillons le rôle de la phase d'annotation des documents à l'aide d'une ressource, puis la transformation de ces annotations dans l'espace d'indexation. Nous détaillons les trois grandes familles de modèles de recherche d'information et leur prise en compte de ces nouveaux espaces d'indexation.*

ABSTRACT. *This article focuses on the field of Semantic Information Retrieval (IR), which is at the intersection of several disciplines: Information Retrieval, Knowledge Engineering and Natural Language Processing. We present the basics to understand how a semantic IR system works. We also present a classification of the types of semantic resources used in information retrieval. We detail the role of the documents' annotation phase using a resource and the processing of these annotations in the indexing space. We detail the three main types of information retrieval models and how the new indexing space is exploited.*

MOTS-CLÉS : *annotation sémantique, indexation, modèle de recherche d'information, ressource sémantique, ontologie.*

KEYWORDS: *semantic annotation, indexation, information retrieval model, semantic resource, ontology.*

TAL. Volume 56 – n° 3/2015, pages 49 à 73

1. Introduction

Le domaine de la recherche d'information (RI) concerne l'étude des systèmes informatiques utilisés comme assistants pour rechercher de l'information dans des documents. Ces systèmes sont appelés : *systèmes de recherche d'information*. Un système de RI facilite donc l'accès à un corpus de documents à partir desquels l'utilisateur va extraire de l'information. Le système prend en entrée une requête formulée en langue naturelle et il propose en sortie une sous-liste ordonnée des documents du corpus. Pour réaliser la correspondance entre une requête et des documents, le système de RI utilise un index : il s'agit d'une structure spécifique optimisée pour garantir un accès rapide aux documents pertinents. Nous nous intéressons aux systèmes de RI capables de construire leur index à partir du contenu en texte intégral des documents. Cette phase d'indexation est caractéristique des systèmes de RI.

La recherche d'information sémantique (RIS) ajoute à cette définition le fait que le système de RI ne considère pas simplement les documents comme un ensemble de signes, sans liens entre eux. Le système de RIS tient compte de la signification véhiculée par les mots des documents. Ces significations sont formalisées informatiquement dans une ressource. Ce type de ressource est intitulé *ressource sémantique* et peut prendre la forme d'un thésaurus, d'une ontologie, etc. Notons qu'un sous-domaine de l'informatique, intitulé ingénierie des connaissances (IC), est dédié à la formalisation des significations dans le but d'une exploitation informatique des connaissances. Un système de RIS est capable de reconnaître les mots des documents et de les associer à une ressource sémantique. Cette étape, préalable à toute indexation, est appelée annotation sémantique et repose sur des techniques de traitement automatique des langues (TAL).

Cet article fait le point dans le domaine de la RIS, qui est à l'intersection de plusieurs disciplines : RI, IC et TAL. Nous présentons les notions de base pour comprendre comment fonctionnent les systèmes de RIS ainsi que les différentes approches mises en œuvre pour intégrer les ressources sémantiques dans les systèmes de recherche d'information.

1.1. *Qu'est-ce que la recherche d'information sémantique ?*

La recherche est *sémantique* parce que le système est capable de prendre en compte la signification des mots dans les documents et les requêtes. Cette prise en compte permet une meilleure compréhension du contenu des documents et du sens des requêtes et aboutit à une amélioration des performances des systèmes en termes de précision et de rappel.

Les travaux en RIS que nous rapportons s'appuient sur des connaissances explicitées dans des ressources sémantiques. Ces connaissances sont ensuite intégrées aux divers processus de recherche d'information : la représentation des documents, la représentation des requêtes, la correspondance entre les représentations des documents

et des requêtes. Nous ne présenterons pas les travaux reposant sur une sémantique distributionnelle telle que la recherche par *Latent Semantic Indexing* (LSI) (Dumais *et al.*, 1988).

Dans cet article, nous nous intéressons aux systèmes exploitant des ressources sémantiques explicites, modifiables et compréhensibles par un intervenant humain. Les entrées des ressources sémantiques utilisées en RIS sont les termes. Comme nous le verrons dans la section 2, ces ressources contiennent au moins un ensemble de termes. Ce réseau est utilisé pour faire correspondre les termes du corpus avec les entités de la ressource. Cette étape d'annotation sémantique est importante et influence d'une manière directe les modèles de recherche.

1.2. *Organisation de l'article*

La section 2 fait état des différentes ressources sémantiques utilisées en RIS et présente leurs caractéristiques. La section 3 définit l'annotation sémantique comme établissant un lien entre les expressions terminologiques du document et des entités de la ressource. Un processus de désambiguïsation des termes polysémiques du corpus est nécessaire pour déterminer l'entité de la ressource adéquate. Cette section présente également les plates-formes qui existent pour générer des annotations. Les annotations sémantiques participent à la construction des représentations des documents ou des requêtes. Comme nous le verrons dans la section 4, les constituants d'un index sont dépendants des types d'annotations et des types de ressources utilisés. Les différents modèles qui permettent la prise en compte d'une ressource sémantique sont décrits en section 5. Nous concluons avec des perspectives dans la section 6.

2. Les types de ressources sémantiques

Les ressources sémantiques utilisées en RIS portent souvent le nom d'*ontologie*, mais cachent en réalité une multitude de ressources hétérogènes. La définition d'ontologie la plus souvent citée est celle de Gruber (1995) : « spécification explicite d'une conceptualisation partagée ». Malheureusement cette définition très générique peut s'appliquer à tous les domaines des systèmes d'information, comme la recherche d'information, les bases de données, l'ingénierie des connaissances, les outils d'aide à la décision, etc., car ils sont tous fondés sur une conceptualisation du domaine qui sera partagée par les utilisateurs du système d'information.

Cette section clarifie les ontologies potentiellement utiles à la tâche de RI. Pour ce faire nous décrirons les ressources par rapport à leurs composants et les usages prévus par leurs concepteurs. Nous présentons tout d'abord les entités disponibles dans les ressources et les hypothèses inhérentes à leur présence dans une ressource sémantique. La figure 1 présente notre vision des interactions entre ces différentes entités et leurs articulations les unes avec les autres.

2.1. *Les entités des ressources sémantiques*

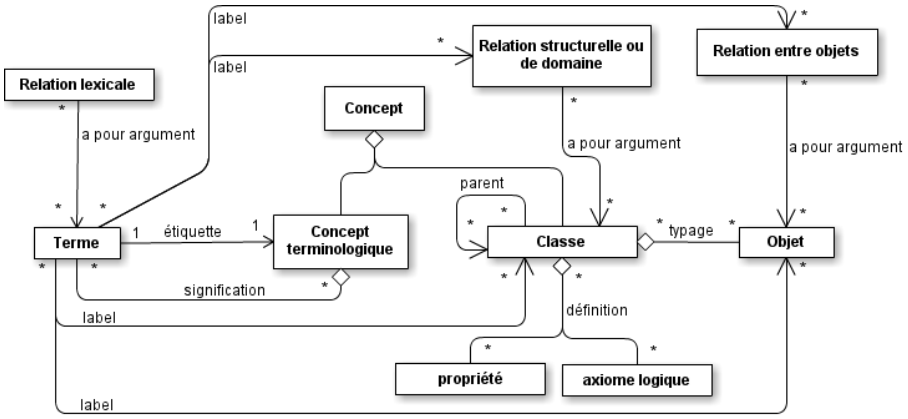

Figure 1. *Schéma UML représentant les différentes entités d'une ressource sémantique*

Le terme : il est constitué d'un mot ou d'une séquence de mots (principalement des syntagmes nominaux) susceptibles d'être choisis par un analyste comme étiquettes de concept (Bourigault et Aussenac-Gilles, 2003). Les travaux sur les ressources sémantiques utilisées en RIS (Bourigault et Aussenac-Gilles, 2003 ; Charlet *et al.*, 2004 ; Desprès et Szulman, 2008) partent tous du postulat que le terme doit avoir une signification stable et consensuelle dans un domaine particulier, identifiable en dehors du contexte d'un document. Au contraire de son occurrence qui peut prendre de nombreuses nuances de sens dans le contexte d'un document (Rastier, 1995)[1]. Par exemple « chemin » est un terme auquel nous associons tous facilement une signification. Alors que la signification du terme « chemin creux » n'est généralement connue que des géomètres.

Le concept : il est défini comme un élément de la pensée (ISO, 2011). Au sein du triangle sémiotique (le terme, le concept, l'objet), le concept est la construction mentale qui représente la signification du terme et qui fait référence à l'objet (Rastier, 1995 ; Daconta *et al.*, 2003). Le concept consiste en un ensemble de caractères que nous reconnaissons comme étant commun à un certain nombre d'objets du monde réel ou imaginaire (Daconta *et al.*, 2003 ; Charlet *et al.*, 2004).

Le concept terminologique : il représente la signification normalisée des termes par le biais d'une définition en langue naturelle. La définition énonce les conditions nécessaires et suffisantes pour que le terme soit pourvu de sa dénotation correcte (Rastier, 1995). Un seul terme, intitulé parfois « label préféré », est choisi comme étiquette du concept terminologique (Desprès et Szulman, 2008 ; ISO, 2011). Cette

1. Les travaux de Rastier (1995) donnent de plus amples explications sur le passage du mot au terme puis au concept et les problèmes sous-jacents.

contrainte est spécifiée dans la figure 1 par l'association « étiquette » de multiplicité égale à 1. Un concept terminologique est aussi une agrégation des termes partageant la même signification, comme le montre la figure 1 avec la relation d'agrégation intitulée « signification ».

La classe : elle représente le concept dans un langage informatique comme un langage objet de représentation des connaissances. La classe est une représentation partielle et orientée du concept (Charlet *et al.*, 2004)[2].

Dans la figure 1, le concept est représenté par l'entité concept terminologique et/ou l'entité classe, car le concept étant une représentation mentale il peut être associé à différentes représentations informatiques (Charlet *et al.*, 2004). Pour résumer notre point de vue, le concept terminologique est associé à une définition en langue naturelle et des termes synonymes, alors que la classe est associée à l'ensemble des propriétés et des relations qui caractérisent le concept.

La propriété : une classe peut se définir par un ensemble de caractéristiques aussi appelées propriétés ou attributs. Les propriétés peuvent être valuées et leurs valeurs varient suivant la classe à laquelle on fait référence. Les valeurs des propriétés sont généralement d'un type dit primitif (entier, chaîne de caractères). Par exemple, la classe $Chemin$ a comme propriété $longueur$ correspondant à la distance entre le début et la fin du chemin.

L'objet ou instance de classe : une classe peut également être définie par l'ensemble de ses instances, c'est-à-dire l'ensemble des objets qui sont typés par cette classe. Il s'agit de définition en extension. Le « chemin des muletiers » qui permet d'atteindre le sommet du Puy-de-Dôme est une instance de la classe $Chemin$ que nous identifierons par $chemin\_des\_muletiers$. Dans les textes, les syntagmes référençant les instances sont principalement des noms propres appelés aussi *entités nommées* dans les processus d'extraction d'information. Par exemple, le nom propre « Puy-de-Dôme » dénote une instance puy_de_dome de la classe $Departement$.

Les relations lexicales : c'est une relation entre deux termes de type synonymie, antonymie, hyponymie, hyperonymie, ou méronymie (Desprès et Szulman, 2008). Un terme se définit aussi par les relations qu'il entretient avec les autres termes. Par exemple, les deux termes « chemin creux » et « chemin de service » partagent la même signification (le même concept terminologique). Ils sont donc synonymes. Jousse (2010) propose une modélisation des relations lexicales et une définition de leurs propriétés.

Les relations entre classes : dans tout système informatique, les classes sont définies dans une arborescence par le biais d'une relation hiérarchique de spécialisation ou de généralisation. La figure 1 modélise l'arborescence des classes par l'association intitulée *parent*. La relation de spécialisation peut prendre plusieurs formes : relation de subsomption en logique (toute instance d'une classe fille est une instance de sa

2. La formalisation du concept dans un langage informatique entraîne plusieurs choix de conception, voir (Charlet *et al.*, 2004) pour plus de détails.

classe mère), relation d'héritage en langage objet (toute propriété déclarée dans une classe mère est héritée par les classes filles). D'autres relations entre classes peuvent être définies en fonction des usages. Par exemple, les classes représentant des objets spatiaux comme $Chemin$ sont associées entre elles par une relation d'inclusion spatiale dite de $localisation$. Cette relation construit une hiérarchie (un chemin est inclus dans un département qui est inclus dans une région, etc.). D'autres relations non hiérarchiques sont aussi nécessaires pour expliciter les connaissances du domaine, nous les appelons *relations de domaine* comme indiqué dans la figure 1.

Les relations entre instances : les instances peuvent être liées entre elles par des relations. Par exemple, $chemin\_des\_muletiers$ est associé au puy_de_dome, instance de la classe $Departement$, par la relation de $localisation$. Ces liens représentent des faits.

Axiome logique : dans les systèmes à base de déduction logique, une classe est définie par un ensemble de conditions nécessaires ou suffisantes que les instances doivent remplir pour être typées par cette classe. On parle de définition en intension. Ces définitions sont composées d'axiomes logiques qui permettent de produire des déductions c'est-à-dire d'inférer de nouvelles connaissances. Par exemple, si l'on définit la classe $CheminPuyDome$ avec l'axiome suivant : si une instance de la classe $Chemin$ a une relation de $localisation$ avec l'instance de département puy_de_dome alors ce chemin est de type $CheminPuyDome$ [3]. D'après cette définition $chemin\_des\_muletiers$ est une instance de la classe $CheminPuyDome$.

2.2. *Les différents types de ressources*

Une ressource est une formalisation informatique de connaissances extérieures au corpus de documents sur lequel la recherche est réalisée. Pour être utilisée en RIS, une ressource doit contenir des ensembles de termes. Ces ressources présentent un ensemble de caractéristiques communes : les connaissances qu'elles contiennent sont consensuelles. Elles présentent également une cohérence terminologique : leur vocabulaire est normalisé (ISO, 2011). Elles sont ainsi utilisées pour lever les ambiguïtés terminologiques.

Nous présentons les différents types de ressources couramment utilisés en RIS, en fonction de leur contenu : du plus simple au plus complexe. Pour ce faire, nous présentons entre parenthèses leurs entités, indiquons pour quel objectif elles ont été construites et donnons un exemple utilisé dans un système de recherche d'information.

Thésaurus (terme + concept terminologique + relation lexicale) : les thésaurus sont utilisés dans le domaine documentaire pour indexer les documents et les requêtes. Ainsi un documentaliste construit la liste de mots-clés qui représentent le contenu documentaire à partir d'un thésaurus. Un thésaurus a une fonction de normalisation des

3. Cet axiome s'écrit en logique de description : $CheminPuyDome \equiv Chemin \sqcap \exists localisation(puy\_de\_dome)$.

mots-clés utilisés dans un système de RI. Il est constitué d'une liste de termes identifiés comme descripteurs ou non-descripteurs. Les termes non descripteurs renvoient aux descripteurs par une relation d'équivalence. Un thésaurus normalise son vocabulaire pour être cohérent : un terme descripteur ne doit être associé qu'à un seul concept terminologique et chaque concept terminologique ne doit avoir qu'un seul terme descripteur par langue. Les descripteurs sont reliés entre eux par au moins deux types de relations : une relation de spécialisation ou de généralisation et une relation associative sans signification explicite (ISO, 2011). Le MeSH (*Medical Subject Heading*)[4] est un thésaurus anglais biomédical qui est utilisé pour indexer le corpus MEDLINE (PubMed). Le projet NOESIS (Gedzelman *et al.*, 2005) a intégré MeSH dans une ressource sur les maladies cardiovasculaires. Cette nouvelle ressource est utilisée dans un système de RIS (Diallo *et al.*, 2006).

Base lexicale (terme + concept terminologique + relation lexicale) : elle ressemble dans son organisation à un thésaurus. Les termes sont regroupés par synonymes et identifient des concepts terminologiques. Les objectifs sont néanmoins bien distincts. Le but d'une base lexicale est de différencier tous les sens possibles qu'une occurrence d'un terme peut prendre dans un texte et non de sélectionner le sens le plus consensuel dans un domaine d'étude. Les bases lexicales sont les versions enrichies des réseaux terminologiques. Les termes sont caractérisés par leurs propriétés linguistiques et grammaticales. Elles sont des outils d'aide à la traduction et à la rédaction de documents. WordNet[5] (Fellbaum, 1998) est une base de données lexicales en anglais très utilisée en RI. Les noms, verbes, adjectifs et adverbes sont regroupés en groupes de synonymes (aussi appelés synsets), chacun exprimant un concept terminologique distinct. Les synsets sont reliés entre eux par des relations lexicales comme par exemple : la relation hyperonymie hyponymie, la méronymie, etc. Un synset représente un concept terminologique et il n'existe pas d'étiquette unique de concept. Les travaux de Dragoni *et al.* (2012) utilisent toutes les entrées possibles de WordNet (nom, verbe, adjectif et nom propre) pour produire automatiquement les index des documents dans un système de RIS.

Ontologie peuplée d'instances (classe + objet + propriété + relation entre objets) : avec l'avènement du Web de données liées ou *Linked Open Data (LOD)*, plusieurs jeux de données sont disponibles. Ils décrivent des faits, qui sont typiquement le contenu d'une base de données. Le jeu de données le plus représentatif du LOD est DBpedia, construit automatiquement à partir des infobox de Wikipédia. DBpedia est utilisé pour la recherche d'entité (Mendes *et al.*, 2011). Ces jeux de données ne vérifient pas le critère de cohérence terminologique, car les identifiants de leurs entités ne sont pas des termes mais des URI. La propriété « rdfs :label » permet d'associer à chaque URI un ensemble de termes, mais aucune vérification de la cohérence terminologique n'est exigée.

4. `http://mesh.inserm.fr/mesh/`
5. `http://wordnet.princeton.edu/wordnet/`

Base de connaissances terminologiques ou BCT (terme + concept terminologique + classe + relation entre classes + relation lexicale) : les BCT peuvent aussi porter d'autres noms en fonction des auteurs. Par exemple, Charlet *et al.* (2012) évoquent des ressources termino-ontologiques (RTO). Elles associent un réseau terminologique (réseau de concepts terminologiques) à une conceptualisation informatique (réseau de classes) riche en relations. Les BCT cumulent les avantages des thésaurus, à savoir la normalisation du vocabulaire et la cohérence terminologique. Elles permettent également de pallier leurs inconvénients. En effet, les BCT ont non seulement plus de relations que les thésaurus mais la signification de ces relations est explicitée et n'est pas laissée à l'interprétation du lecteur. Les BCT sont utilisées pour annoter ou indexer les documents textuels. Le projet Dynamo[6] propose plusieurs cas d'usage de recherche d'information sémantique à base de BCT. Un de ces travaux (Reymonet *et al.*, 2010) construit une BCT pour faciliter la recherche dans une base de rapports de pannes automobiles.

Ontologie axiomatisée (classe + objet + propriété + relation entre classes + relation entre objets + axiomes logiques) : les ontologies axiomatisées sont une spécialisation et un enrichissement des ontologies d'instances. Avec une ontologie axiomatisée, il est possible de spécifier des contraintes d'appartenance d'un objet à une classe. Ces ontologies ont pour but de produire des inférences (Guissé *et al.*, 2012). Pour que les inférences soient valides, il faut que l'ontologie ait une cohérence logique. L'une des premières ontologies volumineuses axiomatisées est CYC (Lenat et Guha, 1989) qui avait pour objectif ambitieux de permettre à des systèmes informatiques d'intelligence artificielle de simuler le raisonnement humain. Une version de CYC a été utilisée pour la désambiguïsation dans les moteurs de recherche comme Lycos (McCuinness, 2005).

3. Annotation sémantique

Le processus d'annotation vise à lier une ressource à une autre ressource. Cette vision très générique est défendue par le groupe de travail du W3C (*Web Annotation Working Group*). Il propose un modèle appelé OADM (*Open Annotation Data Model*) (W3C, 2014) qui peut être étendu à différents usages.

Le terme *annotation* sert à la fois à désigner le processus ainsi que son résultat. Les premiers processus d'annotation mis en place étaient manuels et ont été améliorés en intégrant des processus automatiques dont les résultats peuvent être validés manuellement.

Cette section décrit un type particulier d'annotation : celle qui consiste à apposer à une *entité documentaire* une *entité sémantique*. L'annotation peut porter sur l'ensemble du document sans distinction ou être ancrée dans un document : l'ancre est la partie du document qui porte l'annotation. L'entité documentaire peut être un ou

6. `http://www.irit.fr/dynamo/`

plusieurs mots, une phrase ou un paragraphe. L'entité sémantique peut être un des composants des ressources sémantiques présentés dans la section 2.

Le résultat du processus d'annotation peut être intégré au document ou à l'extérieur du document et constituer ainsi une base d'annotations. Dans certains cas, les annotations servent à enrichir une ontologie axiomatisée ou une BCT, on parle alors de peuplement d'ontologies (Amardeilh et Damljanovic, 2009).

L'annotation sémantique pose des problèmes classiques d'extraction d'information à savoir le repérage des entités documentaires à lier aux entités sémantiques, la délimitation de leurs frontières dans les textes ainsi que la désambiguïsation dans le cas où les entités sont polysémiques.

Nous classons les annotations selon les entités sémantiques auxquelles elles font référence : terme, concept (concept terminologique ou classe), instance et relation (entre classes ou entre instances). Chaque type d'entité pose un problème particulier.

– Terme : l'annotation terminologique équivaut à un processus d'extraction et de reconnaissance des termes. Il faut être capable de retrouver toutes les variantes lexicales (pluriel, singulier) du terme dans les textes, aussi appelées occurrences.

– Concept : l'annotation de concepts enrichit l'annotation terminologique. Un concept terminologique étant une agrégation de termes, il s'agit donc de retrouver les variantes terminologiques (synonymes, abréviations) mais également, de désambiguïser les termes polysémiques et, si possible, de résoudre des anaphores. Les termes associés aux concepts ne sont pas forcément des entités textuelles contiguës et peuvent être exprimés par des entités complexes, qui expriment elles-mêmes des relations (par exemple : « chemin traversant les forêts » fait référence à « chemin forestier »).

– Instance : nous pouvons ramener ce problème à un problème de reconnaissance d'entités nommées. Les entités nommées couvraient traditionnellement des classes restreintes comme Personne, Organisation, Lieu, Temps. Ces classes ont évolué et englobent désormais des classes plus spécialisées organisées en hiérarchie (Sekine *et al.*, 2002). La frontière entre reconnaissance de termes et reconnaissance d'entités nommées devient perméable et plusieurs travaux comme (Ben Abacha et Zweigenbaum, 2012) peuvent être classés dans les deux.

– Relation : l'annotation de relations nécessite de 1) repérer les entités concernées par la relation, 2) d'extraire les relations entre ces entités généralement par des patrons linguistiques. Auger et Barrière (2008) proposent une revue des méthodes d'extraction de relations qui peuvent servir à l'annotation sémantique ou à l'enrichissement d'ontologies.

Le processus d'annotation intègre le plus souvent une fonction de désambiguïsation. En général, cette fonction détermine le concept qui sera lié à un terme selon son contexte d'apparition (le contexte pouvant être la phrase, le paragraphe ou une quelconque fenêtre d'apparition de mots). Navigli (2009) présente un état de l'art détaillé des différentes méthodes de désambiguïsation. Il distingue les méthodes à base de connaissances fondées sur des ressources telles que les dictionnaires, les thesaurus ou

les ontologies, des méthodes à base de corpus qui n'exploitent aucune ressource. Les méthodes de désambiguïsation à base de connaissances reposent sur des mesures de similarité et/ou de proximité entre concepts. Nous pouvons citer la mesure de Wu et Palmer (1994), une des plus utilisées, qui consiste à calculer la similarité entre deux concepts en fonction de leur profondeur dans la hiérarchie pondérée par la profondeur de leur père commun. Le concept retenu étant celui qui maximise les similarités avec les autres concepts dénotés par les termes du contexte d'apparition du terme à désambiguïser.

Ces similarités sont calculées à partir :

– des définitions : le choix du concept repose sur sa définition. La similarité entre deux concepts est calculée en fonction du nombre de mots en commun dans leurs définitions. Il s'agit de garder la définition qui partage le plus de mots avec la définition des autres concepts. Le travail de Lesk (1986) est précurseur, il permet de calculer un score qui prend en compte l'intersection entre les définitions ;

– des relations entre concepts : le choix du concept repose sur sa proximité avec les autres concepts. Cette proximité est calculée *via* des mesures de similarité ou de proximité sémantique en prenant en compte les relations hiérarchiques entre concepts mais aussi, pour certaines mesures, les relations de domaine. Il s'agit de garder les concepts qui maximisent la somme des similarités. Plusieurs mesures de similarité sémantiques existent, Budanitsky et Hirst (2006) proposent une revue et un exemple d'évaluation de ces mesures.

Les travaux de Fernández *et al.* (2011) désambiguïsent une annotation potentielle d'une entité (classe ou instance) de l'ontologie en construisant son contexte, composé des entités qui lui sont liées directement dans l'ontologie. Une annotation n'est validée que si le document contient à la fois une référence à l'entité et au moins une référence à un élément de son contexte.

Les travaux de Baziz *et al.* (2005) sur la désambiguïsation calculent pour un document donné l'ensemble de toutes les combinaisons de concepts terminologiques possibles contenant uniquement une des significations des termes polysémiques. Pour chaque combinaison, un réseau de concepts est construit à partir de la ressource. Chaque réseau est évalué pour déterminer celui qui a le plus de liens entre concepts. Le réseau retenu maximise les liens entre concepts.

Le processus d'annotation rend possible le calcul de la couverture d'une ressource et définit donc son adéquation par rapport au corpus de documents. La couverture peut être calculée par l'ensemble des annotations rapporté à la taille du vocabulaire dans le corpus des documents[7](Ninova *et al.*, 2005).

Pour que le processus d'annotation soit possible, il faut qu'au préalable le lien entre le volet terminologique et conceptuel soit explicité. En effet, le volet terminologique est le point de départ pour tout type d'annotation. Ma *et al.* (2009) proposent

7. Les auteurs Ninova *et al.* (2005) proposent en réalité une mesure plus complexe accompagnée d'autres mesures telles que la densité.

une dissociation entre ce qu'ils appellent l'annotation linguistique (assurée par les outils de TAL) et l'annotation sémantique guidée par une ontologie axiomatisée. Ils proposent une extension de l'ontologie à base de règles d'annotation qui s'expriment sous forme de patrons. Le travail de Bannour (2015) vient compléter ces propositions en mettant en place une approche interactive et itérative qui permet la mise en place par apprentissage des patrons d'extraction.

Plusieurs outils existent qui permettent différents types d'annotations, de l'annotation manuelle par des experts, semi-automatique (Erdmann *et al.*, 2000) à l'annotation automatique. L'annotation automatique repose sur des techniques d'apprentissages supervisées et non supervisées. Une revue des outils et plates-formes d'annotation sémantique est présentée dans (Uren *et al.*, 2006).

Les outils permettent généralement les trois niveaux d'analyse à savoir : l'analyse morphologique, l'analyse syntaxique et l'analyse sémantique. Des plates-formes génériques qui proposent une approche par composition et agrégation ont vu le jour. Elles proposent des solutions aux problèmes d'interopérabilité des outils de TAL ainsi que de leurs formats de sortie. La plate-forme GATE (*General Architecture for Text Engineering*) (Bontcheva *et al.*, 2004 ; Cunningham *et al.*, 2011) est probablement une des plus utilisées. D'autres plates-formes qui reposent sur UIMA (*Unstructured Information Management Architecture*) (Ferrucci et Lally, 2004) sont également proposées.

La plate-forme d'annotation Ogmios (Hamon et Nazarenko, 2008) utilisée dans le moteur de recherche BioAlvis dans le domaine de la microbiologie montre un exemple intéressant d'implémentation d'outils d'annotation qui a également nécessité l'acquisition des ressources sémantiques (BCT, ontologie).

Il existe des outils d'annotation dédiés à des ontologies générales comme DBpedia Spotlight (Mendes *et al.*, 2011) qui permet une annotation à partir de DBpedia. Néanmoins pour les domaines de spécialités, ces outils d'annotation ont une très faible couverture et souffrent des problèmes classiques d'ambiguïté (Ma *et al.*, 2013). MetaMap (Aronson, 2000) est un exemple d'outil d'annotation pour un domaine de spécialité. Il annote des textes biomédicaux avec des concepts du métathésaurus UMLS. Une version UIMA a également été développée[8]. MetaMap, renvoie tous les concepts terminologiques liés à un terme et ne fait donc pas de désambiguïsation.

De plus en plus de plates-formes permettent de produire des annotations sémantiques. Leur exploitation en RIS reste encore limitée. Ceci est essentiellement dû à des questions de passage à l'échelle et de fiabilité de l'annotation, notamment de la désambiguïsation qui peut ajouter du bruit au lieu de le réduire. Beaucoup de travaux reposent encore sur de simples correspondances entre chaînes de caractères.

Les annotations sont la clé de voûte du processus de recherche sémantique, elles permettent de faire le lien entre le contenu des document et la ressource sémantique. Nous décrivons dans la section suivante comment les annotations sont intégrées dans l'espace d'indexation.

8. Disponible à `http://metamap.nlm.nih.gov/UIMA.shtml`

4. Indexation et traitement des annotations dans l'espace d'indexation

L'indexation est définie comme un processus destiné à représenter par les éléments d'un langage documentaire ou libre le contenu d'un document ou d'une requête dans le but d'en faciliter la recherche [9]. Une perte d'information est inhérente à ce processus, car il faut faire des choix de représentation.

Il s'agit dans un premier temps de définir quels sont les éléments du langage qui vont être utilisés pour représenter les documents et les requêtes. Ils peuvent provenir des documents ou de la ressource sémantique. Ils constituent l'espace d'indexation. Dans la figure 2, ces éléments sont représentés par les formes géométriques colorées. Ces éléments s'intitulent terme d'indexation ou descripteur. Pour clarifier notre discours et lever toute ambiguïté, dans la suite de ce document nous les intitulerons les *indexels*. Comme nous l'avons vu précédemment, une annotation n'a pas forcément été produite dans un objectif de RI. Elle complète le contenu documentaire et ajoute de nouvelles données qui sont utiles pour l'indexation. Les annotations produites peuvent donc être intégrées tout ou en partie dans l'espace d'indexation.

Dans les sections suivantes, nous décrivons les différents espaces d'indexation possibles selon l'indexel.

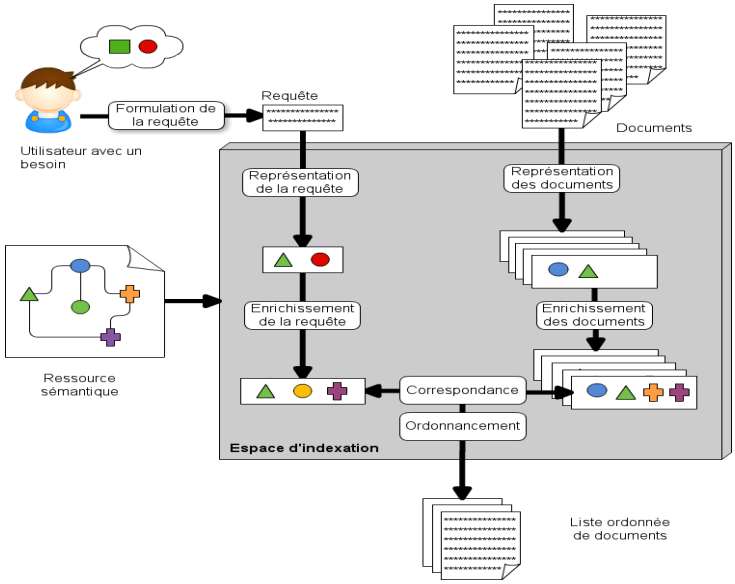

Figure 2. *Espace d'indexation dans une architecture de système de RIS*

9. http://www.adbs.fr/

4.1. *L'espace d'indexation construit à partir des termes*

Considérons le cas où l'espace d'indexation est constitué uniquement de termes. Ces termes peuvent provenir des documents, de la ressource ou de la combinaison des deux. Quand l'indexel est le terme, l'espace d'indexation équivaut à l'ensemble des termes reconnus dans le corpus. Les annotations influencent la construction de l'index et plus particulièrement l'importance d'un terme dans cet index[10]. L'importance d'un terme dans l'index est fonction des autres termes du document annotés par des concepts proches dans la ressource.

Les concepts proches dans la ressource sont déterminés par des mesures de similarité ou de proximité sémantique appliquées sur les concepts dénotés par les termes du document. C'est la même famille des mesures qui serve pour la désambiguïsation (section 3). L'intérêt de considérer les termes des documents comme indexel est de pouvoir prendre en compte des termes du document qui ne sont pas annotés par des entités sémantiques de la ressource (Zargayouna, 2005 ; Fernández *et al.*, 2011).

Ce type d'espace d'indexation est surtout utile quand l'annotation sémantique est appliquée exclusivement à la requête. fr La représentation de la requête est enrichie par des termes qui proviennent de la ressource. Le but de cette extension est soit d'augmenter la couverture terminologique de la requête pour améliorer le rappel, soit de réduire l'ambiguïté des termes de la requête en ajoutant uniquement des termes synonymes ou des termes proches, pour améliorer la précision et le rappel. Bhogal *et al.* (2007) présentent un état de l'art des travaux en extension de requêtes avec des ressources sémantiques.

4.2. *L'espace d'indexation construit à partir des concepts terminologiques*

Le choix de concepts terminologiques comme indexel est lié à la ressource qui a servi à l'annotation. Comme nous l'avons vu dans la section 2, les ressources se composent de concepts terminologiques qui sont une agrégation de termes et dont un terme particulier est choisi comme étiquette de concept.

Le rôle de l'annotation est de donc de traduire le vocabulaire des documents pour que leurs représentations ne contiennent que des termes étiquettes de concepts.

Cependant, le choix des concepts terminologiques à garder pour l'indexation n'est pas évident. En effet, les concepts sont liés les uns aux autres par des relations hiérarchiques ou de domaine. Prendre en compte ou non ces relations est une décision importante qui est guidée par le modèle de RI[10]. Des travaux comme ceux de Dragoni *et al.* (2012) décident de ne garder que les feuilles de la hiérarchie pour construire l'espace d'indexation et éviter les redondances. D'autres, comme Kiryakov et Simov (1999) proposent d'enrichir les concepts terminologiques (synsets de WordNet) retenus par leurs concepts pères (synsets hyperonymes).

10. Cette question sera abordée plus en détail dans la section suivante.

Seydoux (2006) fait état des différents choix possibles et propose une coupe de redondance minimale qui permet de choisir le type des hyperonymes à injecter dans l'espace d'indexation. Un thésaurus est modélisé comme un graphe orienté sans cycle et cette coupe est définie comme étant l'ensemble minimal de sommets couvrant la totalité des feuilles du graphe.

4.3. *L'espace d'indexation construit à partir des relations entre classes et instances*

L'annotation peut produire une relation entre des instances ou des classes de la ressource. Les ressources sémantiques utilisées sont donc des BCT et des ontologies (d'instances ou axiomatisées). L'espace d'indexation stocke toutes les connaissances issues des documents. L'indexel est un triplet composé de deux entités de la ressource liées par une relation. Les systèmes de RIS fondés sur les graphes conceptuels (Sowa, 1984) proposés au début des années 2000 (Ounis et Pasça, 1998 ; Roussey *et al.*, 2001 ; Genest et Chein, 2005) travaillent avec un espace d'indexation qui est un enrichissement de la ressource sémantique, mais sans proposer de construction automatique des graphes à partir des textes. On trouve tout de même des propositions de construction automatique de graphes conceptuels pour l'indexation (Maisonnasse *et al.*, 2007) avec des concepts et des relations du métathésaurus UMLS, mais sans aucune désambiguïsation. Cette vision peut être étendue aux travaux du Web sémantique interrogeant des bases de triplets représentant le contenu des documents (Castells *et al.*, 2007 ; Fernández *et al.*, 2011). La difficulté sous-jacente reste la construction automatique des indexels à partir des textes des documents.

4.4. *Conclusion*

La qualité de l'annotation sémantique touche d'une manière directe l'indexation et le choix des indexels. En effet, les termes qui ne sont pas dans les ressources ne sont pas identifiés dans les documents. En fonction de la couverture de la ressource, on pourrait pencher pour un espace d'indexation par termes ou par concepts ou mixte. Pour pallier cette limite, de plus en plus de travaux combinent plusieurs espaces d'indexation. Par exemple Fernández *et al.* (2011) travaillent d'abord sur un espace d'indexation composé de triplets, puis un autre espace d'indexation composé de termes est construit. En simplifiant, leurs travaux équivalent à construire une BCT riche en termes à partir d'une ontologie et à travailler soit avec l'espace d'indexation des triplets soit avec l'espace d'indexation des termes. L'annotation sémantique a également un impact sur l'importance des indexels dans l'index. La section suivante présente les différentes méthodes de calcul de cette importance et l'intégration des ressources sémantiques au niveau des modèles.

5. Modèles de correspondance pour un système de RIS

Nous avons vu dans la section 4 qu'un indexel est un composant atomique de l'espace d'indexation. Ces indexels sont assemblés dans une structure (index) telle qu'un ensemble, un vecteur, un espace probabilisé, une formule logique, un graphe, etc. L'ensemble formé par les indexels avec leur structure et la fonction de correspondance est appelé le *modèle de recherche d'information*.

Ce qui caractérise les modèles de RIS par rapport aux modèles classiques est la prise en compte de la ressource sémantique pour le choix des indexels ainsi que pour le calcul de la correspondance. Nous discernons trois grands paradigmes de modèles : le paradigme géométrique avec les variantes du modèle vectoriel, le paradigme probabiliste dans lequel les documents et les requêtes sont des événements et la correspondance est une probabilité conditionnelle, et finalement le paradigme logique où la correspondance est interprétée comme une déduction logique.

5.1. *Paradigme géométrique*

Le paradigme géométrique consiste en la représentation des documents et des requêtes dans un espace géométrique. La correspondance est alors calculée par une distance entre des points ou des vecteurs de cet espace. Un représentant célèbre de ce type de modèle est le modèle vectoriel défini par Salton *et al.* (1975).

5.1.1. *Modèle vectoriel classique*

Ce modèle part de l'hypothèse que chaque document est représenté par un vecteur dans un espace vectoriel. Les dimensions de cet espace sont associées aux indexels. Dans la version originale de ce modèle, les indexels sont directement extraits des documents : ce sont les mots des textes, éventuellement filtrés et racinisés. La pondération $w_d(i)$ calculée pour chaque indexel i dans le vecteur $\vec{d}$, représente son importance pour le document d. Elle est obtenue à partir du comptage $\#(i, d)$ du nombre d'apparitions de l'indexel i dans le document d (appelé *frequency*), et du comptage $\#(i, C)$ du nombre de documents du corpus C où apparaît cet indexel (appelé *document frequency*). Une formule qui compose ces deux valeurs, est typiquement :

$$w_d(i) = TF.IDF(i, d) = \#(i, d) * \log \frac{|C|}{\#(i, C)}$$

La valeur $|C|$ est le nombre de documents dans le corpus. Après une normalisation des vecteurs pour les réduire à une norme égale à 1, la fonction de correspondance RSV appelée *Relevance Status Value* qui associe une valeur numérique à chaque couple de vecteurs, est le produit scalaire :

$$RSV(q, d) = \vec{q} \cdot \vec{d} = \sum_i w_q(i).w_d(i)$$

Lorsque les vecteurs $\vec{q}$ et $\vec{d}$ sont normalisés, leur produit scalaire représente le cosinus de l'angle entre ces deux vecteurs.

5.1.2. *Modèle vectoriel étendu*

On a vu dans la section 4 qu'un système de RIS possède une phase supplémentaire avant l'indexation : la phase d'annotation. Le modèle vectoriel s'applique alors, non pas sur les mots des textes, mais sur les entités des annotations.

Par exemple, dans le travail de Baziz *et al.* (2005), les annotations issues d'un document sont transformées en vecteur de concepts terminologiques. La pondération de chaque concept c est la somme des fréquences du terme t le plus long associé au concept c dans la ressource et d'une fraction des fréquences des termes s, inclus dans t :

$$cf(c,d) = \#(t,d) + \sum_{s \in C} \frac{length(s)}{length(t)} \#(s,d)$$

$length$ calcule la taille du terme. Par exemple, pour un concept c dénoté par le terme « chemin de grande randonnée » :

$$\begin{aligned} cf(c,d) &= \#(\text{« chemin grandes randonnées »}, d) \\ &+ 2/3\#(\text{« grandes randonnées"}, d) + 1/3\#(\text{« chemin »}, d) \\ &+ 1/3\#(\text{« grandes »}, d) + 1/3\#(\text{« randonnées »}, d). \end{aligned}$$

Zargayouna (2005) propose une formule de pondération de termes qui intègre une similarité sémantique sim :

$$tf(t,d) = \#(t,d) + \sum_{c'} sim(c,c')\#(t',d)$$

avec t et t' des termes dénotant respectivement les concepts c et c', sim est la mesure de similarité entre concepts de Wu et Palmer (1994).

Dans beaucoup de travaux (Castells *et al.*, 2007 ; Dragoni *et al.*, 2012 ; Boubekeur et Azzoug, 2013 ; Hu *et al.*, 2014) le modèle vectoriel est utilisé avec des concepts terminologiques comme indexels en comptant le nombre d'annotations des concepts dans le document. Les expérimentations menées dans (Baziz *et al.*, 2005) montrent une amélioration de plus de 20 % en précision en fusionnant les sorties d'un modèle vectoriel classique avec les sorties du modèle étendu où les indexels sont les concepts terminologiques.

La fonction de correspondance peut elle aussi tenir compte de la ressource sémantique lors du calcul du produit scalaire. C'est ce que propose Crestani (2000) avec

des modifications de la formule du produit scalaire du modèle vectoriel. Ces travaux proposent d'étendre l'index du document en fonction des indexels de la requête. En effet, si un indexel i de la requête n'est pas dans le document d alors $w_d(i) = 0$, et cet indexel n'influence pas la fonction de correspondance. En revanche, dans la formule suivante revisitée, l'absence de i dans d est compensée par l'indexel i^* du document, le plus proche de i dans la ressource, c'est-à-dire celui qui maximise $sim(i, i')$. Si l'élément i est bien présent dans le document, alors le calcul revient au produit scalaire.

$$RSV(q,d) = \sum_{i \in q/d}^{sim(i,i^*)=\max_{i'} sim(i,i')} sim(i,i^*) w_d(i^*) w_q(i) + \sum_{i \in q \cap d} w_d(i) w_q(i)$$

De cette manière, et grâce à la similarité entre indexels sim calculée à partir de la ressource tous les indexels de la requête peuvent participer à la correspondance. Cette approche permet donc d'étendre simplement le modèle vectoriel pour intégrer une ressource terminologique dans un système de RIS. Les résultats des expérimentations (Crestani, 2000) montrent une augmentation pouvant aller jusqu'à 38 % en précision moyenne par rapport au modèle vectoriel de base mais sur une petite collection de tests.

5.2. *Paradigme probabiliste*

Le paradigme probabiliste considère que la correspondance équivaut à une estimation de probabilité. Il existe deux principales approches : estimer directement la probabilité d'une variable binaire de pertinence R, ou calculer la probabilité d'avoir la requête q sachant le document d.

La première formulation $P(R = 1|d, q)$ conduit aux modèles probabilistes dont, par exemple, le modèle BM25 (Robertson et Walker, 1994).

La seconde formulation $P(d|q)$ conduit aux modèles de langue. Dans ces modèles, le document d et la requête q sont vus comme des événements ayant une probabilité de se produire. La probabilité de trouver l'événement d sachant que nous sommes en présence de l'événement requête q se calcule à l'aide de la formule de Bayes :

$$RSV(q,d) = P(d|q) = \frac{P(q|d)P(d)}{P(q)}$$

Si l'on considère les documents comme équiprobables, et comme on considère une seule requête à la fois, le calcul de pertinence système se réduit à estimer $P(q|d)$. On fait généralement une seconde hypothèse qui considère les indexels i de la requête comme des événements indépendants, ce qui permet d'estimer $P(q|d)$:

$$RSV(q,d) = P(q|d) = \prod_{i \in q} P(i|d)$$

Le reste du calcul consiste à estimer $P(i|d)$, généralement par le maximum de vraisemblance[11] composé avec une opération de lissage. Ce lissage permet d'assurer que $P(i|d) \neq 0$ même dans le cas où i n'appartient pas à d. Par exemple, le lissage de Dirichlet (Zhai, 2008) propose :

$$P_\mu(i|d) = \frac{\#(i,d) + \mu P(i|C)}{|d| + \mu}$$

avec $P(i|C)$ la probabilité de trouver i dans le corpus C, $|d|$ la taille du document d en nombre d'indexels, et μ une constante.

5.2.1. *Comptage des indexels*

Tout comme le modèle vectoriel, les formules de ces modèles sont fondées sur le comptage $\#(i,d)$, mais aussi sur la taille $|d|$ d'un document en nombre d'indexels. Lorsque l'on utilise ces formules pour un système de RIS, les indexels sont alors les concepts terminologiques issus de l'annotation. Si l'on réalise un comptage brut, comme les termes annotés peuvent se chevaucher, le nombre de concepts d'un texte annoté peut dépasser de beaucoup la taille du texte mesuré en mots. Par exemple, en utilisant l'outil d'annotation automatique, MetaMap, un texte qui contient *« lobar pneumonia x-ray »* est annoté par 17 concepts terminologiques de UMLS, qui sont ancrés sur l'une des 6 combinaisons de ces 3 mots. Un comptage simple de ces concepts conduit alors à une mesure qui s'éloigne de l'hypothèse initiale du comptage des mots, qui est que la fréquence est proportionnelle à l'importance des mots.

Des solutions à ce problème de comptage sont proposées (Abdulahhad *et al.*, 2013). L'hypothèse est que la somme des poids de toutes les annotations doit être égale à la taille du document en mots. Cette hypothèse permet de dériver une pondération des concepts qui correspond à une répartition des fréquences (entières) des mots sur les concepts. Ainsi, quel que soit l'indexel utilisé, la taille des documents reste identique. Ce type de pondération introduit dans les modèles probabilistes a permis d'obtenir (Abdulahhad *et al.*, 2013) des gains de plus 30 % de précision moyenne (MAP) sur des collections de tests standard (campagne CLEF), montrant ainsi par la pratique l'importance du choix de la fonction de comptage dans les systèmes de RIS.

5.2.2. *Exploitation de liens entre indexels*

Hormis les travaux sur la fréquence des indexels, des extensions ont été proposées pour tenir compte des liens pondérés entre indexels (Gao *et al.*, 2004). Ces extensions, par exemple, appliquées au modèle de langue, utilisent la dépendance entre les indexels dans une phrase. Le modèle de Gao *et al.* (2004) prend en compte un nouvel élément dans le calcul : le graphe de liens L entre des indexels. Le modèle de langue doit alors calculer :

$$P(q|d) = \sum_L P(q, L|d) = \sum_L P(L|d) P(q|L, d)$$

11. C'est-à-dire la fréquence du terme sur la taille du document.

Le calcul se développe ensuite de manière similaire au modèle de langue présenté ci-dessus, mais en tenant compte de la probabilité de liens entre deux indexels, sachant l'existence d'un graphe de liens particuliers L. Dans le cas particulier où le graphe de liens L est un graphe acyclique, formé de couples (i, i'), où i est le gouverneur de i', et i_h le point d'entrée du graphe L :

$$P(q|L, d) = P(i_h|d) \prod_{(i,i')\in L} P(i|i', L, d)$$

Une autre approche plus simple (Al Masri *et al.*, 2014) étend effectivement le modèle de langue dans le cas du lissage de Dirichlet. Cette proposition est similaire à celle de Crestani (2000) réalisée pour le modèle vectoriel. Elle consiste à intégrer dans la formule de lissage de $P(i|d)$ une valeur de similarité sim avec l'indexel i^* dans le document le plus proche de l'indexel i de la requête. La formule de lissage est alors :

$$P_\mu(i|d) = \frac{\#(i^*, d)sim(i, i^*) + \mu P(i|d)}{|d| + \mu}$$

Cette proposition permet donc d'intégrer dans un modèle de langue une ressource simple définie comme un graphe d'indexel avec des liens de similarité. L'approche de (Al Masri *et al.*, 2014) produit jusqu'à 20 % d'augmentation de MAP dans le domaine médical.

5.3. *Paradigme logique : la correspondance par déduction*

Historiquement, le premier modèle de RI était le modèle booléen, dans lequel les requêtes composaient des mots-clés avec des connecteurs logiques. Actuellement, le Web sémantique utilise des logiques de description pour représenter les connaissances du domaine. Ces deux logiques souffrent du même problème : elles ne font que des calculs binaires. Cependant, la correspondance entre un document et une requête peut être vue comme une déduction incertaine du document vers la requête, tous deux représentés comme éléments d'une logique formelle. Cette idée a été proposée initialement par Van Rijsbergen (1986) qui exprime la fonction de correspondance par une probabilité de déduction du document vers la requête :

$$RSV(q, d) = P(d \rightarrow q)$$

Une logique de description avec une extension probabiliste (Meghini *et al.*, 1993 ; Sebastiani, 1994) a aussi été proposée pour la RIS, mais avec des expérimentations réduites. Des modèles logiques du premier ordre ont également été explorés, notamment avec la représentation en graphes conceptuels (Ounis et Pasça, 1998 ; Genest et Chein, 2005).

Les travaux de Maisonnasse *et al.* (2007) complètent ces graphes par une valeur de confiance. Ces valeurs sont utilisées dans le calcul de la déduction logique et implémentées comme une projection partielle du graphe requête vers le graphe document.

Ce modèle a été testé sur une collection du domaine médical de la campagne CLEF et produit un gain de précision moyenne de plus de 36 % par rapport à une indexation à base de mots.

5.4. *Conclusion*

Les modèles de RIS présentés sont des adaptations des modèles classiques de RI. Souvent, l'usage de la ressource se ramène à une fonction de similarité intégrée dans la fonction de correspondance. Il n'existe pas, à notre connaissance, des modèles dédiés qui prennent en compte toute la complexité des ressources sémantiques.

Nous observons que les approches qui combinent plusieurs modèles de RI ou plusieurs espaces d'indexation donnent des résultats prometteurs. Les travaux présentés dans (Fernández *et al.*, 2011) proposent une approche en deux temps. Un premier modèle logique fondé sur la logique de description permet de sélectionner des instances de l'ontologie (Castells *et al.*, 2007). Ensuite, un modèle vectoriel construit un vecteur pondéré d'instances. Cette approche met ainsi en œuvre deux modèles, dont un seul réalise une correspondance pondérée.

6. Perspectives et points de recherche

La recherche d'information sémantique que nous avons présentée a comme objectif d'améliorer les résultats d'une recherche d'information classique en exploitant une sémantique explicitée dans une ressource sémantique. Nous avons décrit dans cet article les différentes ressources utilisées pour les méthodes de RI, leurs spécificités et leur intégration dans des modèles de recherche d'information. L'annotation sémantique permet de relier les documents à la ressource sémantique. Cette phase reste la clé de voûte pour assurer une bonne couverture du vocabulaire du document. Les avancées en TAL dans ce domaine peuvent être profitables en permettant le passage à l'échelle et une intégration facile dans les systèmes de RIS.

L'évaluation d'un système de RIS reste complexe car il s'agirait d'évaluer en même temps : 1) la qualité de la ressource sémantique, 2) la qualité des annotations, 3) la pertinence du choix de l'espace d'indexation et 4) l'efficacité du modèle. Les évaluations avec une collection de tests ne permettent qu'une évaluation globale et se cantonnent à conclure de l'efficacité (ou de l'inefficacité) de la RIS. La qualité des résultats dépend fortement de la manière dont les ressources sont exploitées ainsi que des annotations qui ne sont possibles que si l'on dispose de ressources suffisamment couvrantes.

La vision que nous avons présentée de la recherche d'information sémantique est amenée à évoluer sur plusieurs critères :

– **le nombre de ressources :** actuellement les travaux de RIS ne travaillent qu'avec une seule ressource associée au corpus. Cette vision monolithique est réductrice. Dans

les faits, même pour un domaine de spécialité il faut que les systèmes de RIS soient capables de prendre en compte plusieurs ressources et sachent gérer des annotations qui peuvent être incohérentes entre elles ;

– **la qualité de la ressource :** le besoin de disposer d'une ressource de qualité validée par l'expert restera toujours présent pour des domaines d'expertise tels que le domaine juridique ou médical. Dans un contexte où l'utilisateur n'est pas expert, on peut exploiter des ressources de qualité dégradée construites, par exemple, par *crowdsourcing* à l'instar de DBpedia issu de Wikipédia. Une telle ressource reflétera un avis moins expert mais permettra par son volume de couvrir un spectre plus large ;

– **l'explicitation de la sémantique de la ressource :** la définition de ressource sémantique que nous avons donnée en introduction est réductrice car elle impose une vision où les ressources sont compréhensibles par un humain. Cependant, la sémantique calculée à partir de ressources non formelles (Wikipédia), est déjà utilisée efficacement dans des systèmes de RIS, à l'instar de *Explicit Semantic Analysis* (Egozi *et al.*, 2011). L'apprentissage automatique de sens sur de très larges collections pourrait à la fois concilier le passage à l'échelle et la couverture tout en se rapprochant de la précision de la sémantique explicite. Le développement, par exemple, de la méthode du *word embedding* (Mikolov *et al.*, 2013) ouvre de nouvelles perspectives pour la RIS.

Nous avons évoqué brièvement les travaux du Web sémantique en section 5. La recherche dans le Web sémantique (WS) s'apparente plus aux systèmes de questions-réponses qu'aux systèmes de recherche d'information. Cela consiste à chercher des faits *via* des requêtes exprimées dans un langage structuré tel que SPARQL (Segaran *et al.*, 2009). Les tendances actuelles vont vers la convergence des travaux des deux communautés (RI et WS). L'engouement de Google pour le *Knowledge Graph* montre que les points de convergence sont nombreux.

Remerciements

Les auteurs remercient les relecteurs anonymes qui ont largement contribué à l'amélioration de l'article. Les auteurs remercient également les participants de l'atelier RISE (Recherche d'Information SEmantique) dont la longévité (Chevallet *et al.*, 2015) témoigne de l'intérêt constant de mener des recherches à la croisée de la recherche d'information, le traitement automatique des langues et de l'ingénierie des connaissances.

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence nationale de la recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

7. Bibliographie

Abdulahhad K., Chevallet J.-P., Berrut C., « Revisiting the Term Frequency in Concept-Based IR Models », *Database and Expert Systems Applications*, vol. 8055 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 63-77, 2013.

Al Masri M., Tan K., Berrut C., Chevallet J.-P., Mulhem P., « Integrating Semantic Term Relations into Information Retrieval Systems Based on Language Models », *Information Retrieval Technology*, vol. 8870 of *Lecture Notes in Computer Science*, Springer International Publishing, p. 136-147, 2014.

Amardeilh F., Damljanovic D., « Du texte à la connaissance : annotation sémantique et peuplement d'ontologie appliqués à des artefacts logiciels », *IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances*, p. 157-168, 2009.

Aronson A., « Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. », *Proceedings of AMIA Annual Symposium*, p. 17-21, 2000.

Auger A., Barrière C., « Pattern-based approaches to semantic relation extraction : A state-of-the-art », *Terminology*, vol. 14, n° 1, p. 1-19, 2008.

Bannour S., Apprentissage interactif de règles d'extraction d'information textuelle, PhD thesis, Université Paris 13, 2015.

Baziz M., Boughanem M., Aussenac-Gilles N., Chrisment C., « Semantic Cores for Representing Documents in IR », *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC '05, ACM, p. 1011-1017, 2005.

Ben Abacha A., Zweigenbaum P., « Une étude comparative empirique sur la reconnaissance des entités médicales », *Traitement Automatique des Langues (TAL)*, vol. 53, n° 1, p. 14, 2012.

Bhogal J., Macfarlane A., Smith P., « A Review of Ontology Based Query Expansion », *Inf. Process. Manage.*, vol. 43, n° 4, p. 866-886, 2007.

Bontcheva K., Tablan V., Maynard D., Cunningham H., « Evolving GATE to Meet New Challenges in Language Engineering », *Natural Language Engineering*, vol. 10, n° 3/4, p. 349-373, 2004.

Boubekeur F., Azzoug W., « Concept-based indexing in text information retrieval », *CoRR*, 2013.

Bourigault D., Aussenac-Gilles N., « Construction d'ontologies à partir de textes », *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues*, p. 27-50, 2003.

Budanitsky A., Hirst G., « Evaluating wordnet-based measures of lexical semantic relatedness », *Computational Linguistics*, vol. 32, n° 1, p. 13-47, 2006.

Castells P., Fernandez M., Vallet D., « An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval », *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, n° 2, p. 261-272, 2007.

Charlet J., Bachimont B., Troncy R., « Ontologies pour le web sémantique », *Revue I3, page 31p*, 2004.

Charlet J., Declerck G., Dhombres F., Gayet P., Miroux P., Vandenbussche P.-Y., « Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation », *23es journées francophones d'Ingénierie des connaissances*, IC 2012, p. 33-48, 2012.

Chevallet J.-P., Roussey C., Zargayouna H. (eds), *RISE 2015 : Recherche d'Information SEmantique (Actes de la septième édition de l'atelier Recherche d'Information SEmaantique)*, 2015.

Crestani F., « Exploiting the Similarity of Non-Matching Terms at Retrieval Time », *Information Retrieval*, vol. 2, n° 1, p. 27-47, 2000.

Cunningham H., Maynard D., Bontcheva K., Tablan V., Aswani N., Roberts I., Gorrell G., Funk A., Roberts A., Damljanovic D., Heitz T., Greenwood M. A., Saggion H., Petrak J., Li Y., Peters W., *Text Processing with GATE (Version 6)*, 2011.

Daconta M. C., Obrst L. J., Smith K. T., *The Semantic Web : a guide to the future of XML, Web services, and knowledge management*, John Wiley & Sons, 2003.

Després S., Szulman S., « Réseau terminologique versus Ontologie », *Toth 2008*, p. 17-34, 2008.

Diallo G., Simonet M., Simonet A., « An Approach to Automatic Ontology-Based Annotation of Biomedical Texts », *Advances in Applied Artificial Intelligence*, vol. 4031 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 1024-1033, 2006.

Dragoni M., da Costa Pereira C., Tettamanzi A. G., « A conceptual representation of documents and queries for information retrieval systems by using light ontologies », *Expert Systems with applications*, vol. 39, n° 12, p. 10376-10388, 2012.

Dumais S. T., Furnas G. W., Landauer T. K., Deerwester S., Harshman R., « Using latent semantic analysis to improve access to textual information », *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 281-285, 1988.

Egozi O., Markovitch S., Gabrilovich E., « Concept-based information retrieval using explicit semantic analysis », *ACM Transactions on Information Systems (TOIS)*, vol. 29, n° 2, p. 1-38, 2011.

Erdmann M., Maedche A., Schnurr H.-P., Staab S., « From manual to semi-automatic semantic annotation : About ontology-based text annotation tools », *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, p. 79-85, 2000.

Fellbaum C., *WordNet*, Wiley Online Library, 1998.

Fernández M., Cantador I., López V., Vallet D., Castells P., Motta E., « Semantically enhanced Information Retrieval : an ontology-based approach », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 9, n° 4, p. 434-452, 2011.

Ferrucci D., Lally A., « UIMA : an architectural approach to unstructured information processing in the corporate research environment », *Natural Language Engineering*, vol. 10, n° 3-4, p. 327-348, 2004.

Gao J., Nie J.-Y., Wu G., Cao G., « Dependence Language Model for Information Retrieval », *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'04, ACM, p. 170-177, 2004.

Gedzelman S., Simonet M., Bernhard D., Diallo G., Palmer P., « Building an ontology of cardiovascular diseases for concept-based information retrieval », *Computers in Cardiology, 2005*, p. 255-258, 2005.

Genest D., Chein M., « A content-search information retrieval process based on conceptual graphs », *Knowledge and Information Systems*, vol. 8, n° 3, p. 292-309, 2005.

Gruber T. R., « Toward principles for the design of ontologies used for knowledge sharing ? », *International journal of human-computer studies*, vol. 43, n° 5, p. 907-928, 1995.

Guissé A., Lévy F., Nazarenko A., « Un moteur sémantique pour explorer des textes réglementaires », *IC 2011, 22èmes Journées francophones d'Ingénierie des Connaissances*, Chambéry, France, p. 451-458, 2012.

Hamon T., Nazarenko A., « Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience », , vol. 49, n° 2, p. 127-154, 2008.

Hu J., Lu X., Guan C., « A Semantic Information Retrieval Approach Based on Rough Ontology », *The Open Cybernetics & Systemics Journal*, vol. 8, p. 399-404, 2014.

ISO, *Norme Internationale ISO 25964-1 :2011. Information et documentation – Thésaurus et interopérabilité avec d'autres vocabulaires – Partie 1 : Thésaurus pour la recherche documentaire*, AFNOR, 2011.

Jousse A.-L., Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales, PhD thesis, Université de Montréal et Université Paris Diderot (Paris 7, 2010.

Kiryakov A., Simov K. I., « Ontologically supported semantic matching », *Proceedings of NoDaLiDa-99 Conference*, 1999.

Lenat D. B., Guha R. V., *Building large knowledge-based systems ; representation and inference in the Cyc project*, Addison-Wesley Longman Publishing Co., Inc., 1989.

Lesk M., « Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone », *Proceedings of the 5th annual international conference on Systems documentation*, p. 24-26, 1986.

Ma Y., Audibert L., Nazarenko A., « Ontologies étendues pour l'annotation sémantique », *20es Journées Francophones d'Ingénierie des Connaissances*, p. 205-216, 2009.

Ma Y., Lévy F., Nazarenko A., « Annotation sémantique pour des domaines spécialisés et des ontologies riches », *20ème conférence du Traitement Automatique du Langage Naturel (TALN 2013)*, p. 464-478, 2013.

Maisonnasse L., Chevallet J., Berrut C., « Incomplete and Fuzzy Conceptual Graphs to Automatically Index Medical Reports », *Natural Language Processing and Information Systems*, vol. 4592 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 240-251, 2007.

McCuinness D. L., « Ontologies come of age », *Spinning the semantic web : bringing the World Wide Web to its full potential*p. 171, 2005.

Meghini C., Sebastiani F., Straccia U., Thanos C., « A model of information retrieval based on a terminological logic », *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, ACM Press, p. 298-307, 1993.

Mendes P. N., Jakob M., Garcia-Silva A., Bizer C., « DBpedia Spotlight : Shedding Light on the Web of Documents », *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., « Distributed Representations of Words and Phrases and their Compositionality », *in* C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds), *Advances in Neural Information Processing Systems 26*, p. 3111-3119, 2013.

Navigli R., « Word sense disambiguation : A survey », *ACM Computing Surveys (CSUR)*, vol. 41, n° 2, p. 10, 2009.

Ninova G., Nazarenko A., Hamon T., Szulman S., « Comment mesurer la couverture d'une ressource terminologique pour un corpus », *TALN 2005*, 2005.

Ounis I., Pasça M., « Effective and Efficient Relational Query Processing Using Conceptual Graphs », *Proceedings of the 20th Annual BCS-IRSG Conference on Information Retrieval Research*, IRSG'98, British Computer Society, p. 8-8, 1998.

Rastier F., « Le terme : entre ontologie et linguistique », *La banque des mots*, vol. 7, p. 35-65, 1995.

Reymonet A., Thomas J., Aussenac-Gilles N., « Ontologies et Recherche d'Information : une application au diagnostic automobile », *21èmes Journées Francophones d'Ingénierie des Connaissances, IC*, p. 283 à 294, 2010.

Robertson S. E., Walker S., « Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval », *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, Springer-Verlag New York, Inc., p. 232-241, 1994.

Roussey C., Calabretto S., Pinon J.-M., « A new conceptual graph formalism adapted for multilingual information retrieval purposes », *Database and Expert Systems Applications*, p. 92-101, 2001.

Salton G., Wong A., Yang C. S., « A Vector Space Model for Automatic Indexing », *Commun. ACM*, vol. 18, nº 11, p. 613-620, November, 1975.

Sebastiani F., « A Probabilistic Terminological Logic for Modelling Information Retrieval », *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, Springer-Verlag New York, Inc., p. 122-130, 1994.

Segaran T., Evans C., Taylor J., *Programming the semantic web*, " O'Reilly Media, Inc.", 2009.

Sekine S., Sudo K., Nobata C., « Extended Named Entity Hierarchy. », *The Third International Conference on Language Resources and Evaluation, LREC*, 2002.

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

Sowa J. F., *Conceptual Structure : Information Processing in Mind and Machine*, The Systems Programming Series, Addison-Wesley, 1984.

Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E., Ciravegna F., « Semantic annotation for knowledge management : Requirements and a survey of the state of the art », *Web Semantics : science, services and agents on the World Wide Web*, vol. 4, nº 1, p. 14-28, 2006.

Van Rijsbergen C. J., « A Non-Classical Logic for Information Retrieval », *Comput. J.*, vol. 29, nº 6, p. 481-485, 1986.

W3C, « Web Annotation Data Model », December, 2014.

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

Zargayouna H., Indexation sémantique de documents XML, PhD thesis, Thèse de Doctorat de l'Université Paris-Sud, 2005.

Zhai C., « Statistical Language Models for Information Retrieval A Critical Review », *Found. Trends Inf. Retr.*, vol. 2, nº 3, p. 137-213, March, 2008.