Recherche d'information s´emantique : Graphe sémantico-documentaire et propagation d'activation

Ines Bannour

UNIVERSITÉ PARIS 13

Groupe Mathématique, Informatique, Signal **Sorbonne Paris Cité**
École doctorale Galilée

Recherche d'information sémantique : Graphe sémantico-documentaire et propagation d'activation

THÈSE

présentée et soutenue publiquement le 9 Mai 2017

pour l'obtention du

Doctorat de l'Université Paris 13 – Sorbonne Paris Cité

(spécialité informatique)

par

Ines Bannour

Composition du jury

Rapporteurs :	Jean-Pierre Chevallet	Maître de conférences HDR, Université Pierre Mendès France
	Nathalie Pernelle	Maître de conférences HDR, Université Paris Sud
Examinateurs :	Lynda Tamine-Lechani	Professeur, Université Paul Sabatier Toulouse
	Thierry Charnois	Professeur, Université Paris 13
Encadrants :	Adeline Nazarenko	Professeur, Université Paris 13 (directrice)
	Haïfa Zargayouna	Maître de conférences, Université Paris 13 (co-encadrante)

Laboratoire Informatique de Paris Nord — CNRS UMR 7030

/

Remerciements

Voilà venu le moment tant attendu, le moment où enfin avec toute confiance je peux remplir ces quelques lignes et exprimer ma gratitude envers tous ceux qui ont pu me pousser vers l'avant, vers la concrétisation de ce rêve d'être Docteur.

Il me serait très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme. Cette thèse n'aurait pas été possible sans mes encadrantes. Adeline NAZARENKO et Haîfa ZARGAYOUNA, je vous remercie grandement pour votre aide inestimable, pour vos soutiens tant scientifiques, que humains, car la thèse c'est avant tout un bout de chemin dans la vie. Merci d'avoir été présentes, pour le temps que vous m'aviez consacré, vos précieux conseils et de m'avoir encouragé à donner le meilleur de moi-même. Merci de m'avoir bien accueillis dans l'équipe RCLN. Je finirai tout simplement par vous dire que j'étais gâtée par la qualité de votre encadrement.

Ma famille, et plus spécialement mes parents, à qui je dédie ce travail, je vous en suis reconnaissante car vous étiez toujours le moteur qui alimentait mes nuits de travail et deadlines. Vous nous aviez appris à viser la lune et voilà qu'aujourd'hui après avoir vu briller Sondes, venu mon tour d'être Docteur et de vous combler de fierté. Merci pour vos ondes positives malgré les distances. Je n'oublie pas bien évidemment, mes soeurs et mes confidentes : Sondes, Fetia et Nesrine qui ont toujours su me booster quand ça n'allait pas et partager mes joies et mes réussites. Plus spécialement, je remercie Sondes, d'être présente, d'être ici, le long de cette thèse, tantôt ma confidente, tantôt mon père, tantôt tout simplement ma chère soeur, en qui j'ai toujours vu mon exemple.

Mes remerciements envers ma famille n'en finirons pas avant d'adresser un "Merci" spécial, à la sève principale qui m'a nourrit d'amour, de soutien et d'énergie le long de ce parcours de thèse, mon mari Hassène. Tu as su faire de ma fragilité un atout et de ma force une gloire. Je te dédie ce travail qu'on peut enfin savourer la réussite ensemble après tant de travail, mais aussi de couches et de biberons partagés. Tu as su être et l'épaule solide qui me retenait, et le papa qui n'hésitais pas à prendre le relais quand il le fallait ou ne fallait pas.

Encore une dernière étoile de cette famille qui illuminait ma thèse, Kaysser, c'est grâce à toi que j'ai eu tant de motivation. Tu es venu au monde pendant cette thèse, pour être ma raison d'être. Tu me m'interdisais d'être triste ou de désespérer, ton sourire et ta gaieté été ma plus grande distraction après une journée de travail.

Je tiens également à remercier tous les membres passés et présents de l'équipe RCLN pour avoir été de bons collègues et pour leur amitié ainsi que tous les membres du laboratoire LIPN. Merci à mes collègues doctorants : Nouha, Sarra, Nada, Aicha, Sondes, Leila, Hanène, Hanane, Wiem, Ehab, Imad, Mohamed, Issam et tous les autres, vous aviez fait de nos petites pauses, un vrai moment de plaisir, de joie et de partage d'évènements heureux pour les uns et les autres, et il n'en manquait pas, dans les longues journées de travail. Je vous souhaite une bonne poursuite et beaucoup de réalisations.

Ines BANNOUR

"Ce que l'on apprend par l'effort reste toujours ancré plus longtemps."
– Luc Lecompte –
A Kaysser ...

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Sommaire

1.1 Contexte et motivations

Depuis l'apparition de l'informatique, le volume d'information disponible, notamment sous la forme textuelle, ne cesse de croître, et le nombre de documents qui stockent ces informations continue de grandir. Trouver dans cette masse d'information ce que l'on cherche devient de plus en plus ardu. Dans ce contexte, permettre aux personnes en quête d'information de localiser rapidement et efficacement une information dans un vaste ensemble de documents électroniques est crucial. C'est l'enjeu majeur de la Recherche d'Information (RI).

Les systèmes de RI (SRI) ont connu un essor fulgurant depuis les années 1990 : ils ont décuplé les moyens d'accéder à l'information, notamment l'information disponible sur le web. Ces SRIs reposent principalement sur des modèles de RI classiques comme le modèle vectoriel [Salton et al., 1975] ou le modèle probabiliste [Robertson and Jones, 1976]. Ces modèles se fondent sur une représentation commune des documents et des besoins ou requêtes des utilisateurs, ce qui permet d'apparier les uns aux autres. Cela suppose naturellement de pouvoir mesurer la proximité d'une requête et d'un document. Ces calculs de proximité se fondent principalement sur les fréquences des mots et sur l'analyse de leurs distributions dans les documents. L'objectif d'un SRI, c'est de présenter à l'utilisateur, non pas une réponse exacte ou directe à son besoin d'information, mais les documents qui vont y répondre en les classant par ordre de pertinence au regard de sa requête.

Ces modèles distributionnels ont été largement éprouvés et on sait qu'ils passent à l'échelle des grandes collections de documents mais la notion de pertinence sur laquelle ils reposent a très vite été critiquée (voir [Gonzalo et al., 1998, Stetina et al., 1998, Khan, 2000], notamment). Elle repose essentiellement sur la base d'appariements de chaînes de caractères et de calculs

statiques de fréquences d'occurrences mais elle ne permet pas de traiter convenablement les problèmes d'ambiguïté et de synonymie qui sont omniprésents du fait de la richesse de la langue et qui provoquent du bruit et du silence dans les résultats des moteurs de recherche (resp. des documents non pertinents retournés par le moteur ou des documents pertinents que le moteur n'a pas retrouvés).

S'attaquer à ces problèmes en RI revient à s'appuyer, au delà des calculs statistiques, sur la signification et la sémantique des mots du vocabulaire, afin de mieux caractériser les documents pertinents au regard du besoin de l'utilisateur et de les retrouver. C'est dans ce but que la recherche d'information sémantique (RIS) a vu le jour.

Pour dépasser ou corriger les approches purement statistiques sur lesquelles se fondent les modèles classiques de RI, on cherche en général, en RIS, à injecter des connaissances comme des couples de synonymes ou les concepts du domaine de manière à désambiguïser le vocabulaire ou enrichir la représentation des documents et des requêtes. L'objectif sous-jacent est de permettre à un SRI de renvoyer un document contenant non pas les mots de la requête mais des mots sémantiquement proches et ainsi réduire le silence, ou, à l'inverse, d'exclure des documents ambigus et d'améliorer la précision. La RIS exploite pour ce faire des connaissances qui sont généralement consignées dans des ressources sémantiques comme les ontologies, les thésauri, les terminologies, etc. Il s'agit en fait de réduire le fossé séparant la représentation des documents et le besoin de l'utilisateur.

Les travaux de RIS se sont multipliés avec l'accroissement du nombre de ressources sémantiques disponibles dans le cadre du Web Sémantique (WS) [Berners-Lee and Lassila, 2001]. Des apports et des améliorations plus ou moins significatifs ont été rapportés. Des campagnes d'évaluation comme TREC [1] ou CLEF [2], ont été lancées : elles donnent des résultats mitigés. Nous constatons que, dans ces compagnes, la RIS semble atteindre un plateau, en dépit de l'intérêt toujours croissant qu'elle suscite et de la richesse des ressources rendues disponibles dans le WS.

On peut en conclure que l'approche consistant à adapter les modèles documentaires classiques par injection de sémantique a atteint ses limites. On a souvent parlé d'approches de "sac de concepts" par analogie avec les modèles classiques qualifiés d'approches "sac de mots". Même si on peut résoudre certains problèmes comme l'ambiguïté linguistique en utilisant des ressources sémantiques, ces approches ne résolvent pas tout. Notamment, elles ne prennent pas en considération les liens implicites entre les mots dans un texte parce qu'elles reposent le plus souvent sur une représentation plate du vocabulaire.

Pour s'adapter à des modèles classiques de RI, les modèles sémantiques, sensés permettre d'aller au delà des mots et raisonner à un niveau conceptuel, sont en fait eux-mêmes aplatis. Ils sont présents, mais de manière algébrique, à travers des mesures de similarité qui sont utilisées dans l'appariement des documents et des requêtes. Les travaux [Zargayouna, 2005] et [Boubekeur and Azzoug, 2013]) en sont des exemples, même s'ils proposent de combiner les modèles documentaires et les modèles sémantiques, en prenant en compte les liens hiérarchiques existants entre les sens des mots.

Ceci ouvre la voie à d'autres approches de RIS. Quand un document évoque un concept donné, le nombre de fois où la dénotation lexicale de ce concept est mentionnée importe moins que l'ensemble des concepts qui lui sont connexes et qui sont eux-mêmes implicitement évoqués. Nous considérons que les analyses sémantiques sont complémentaires aux analyses distributionnelles : elles doivent être prises en comptes en tant que telles et pas seulement comme adaptation ou ajustement de ces dernières.

1. TREC : Text REtrieval Conference
2. CLEF : Conference and Labs of the Evolution Forum

1.2 Vers une véritable combinaison des modèles documentaires et sémantiques

Un modèle de RI est caractérisé par une représentation par les unités lexicales présentes dans le vocabulaire de la collection de documents et leurs occurrences dans les documents. Cette représentation est généralement un index plat, c'est-à-dire une simple liste d'unités lexicales auxquelles des attributs sont associés. Dans le modèle vectoriel, les documents sont des vecteurs d'unités lexicales pondérées en fonction de leur importance dans les documents (tf fréquence d'occurrences) et de leur pouvoir de discrimination dans la collection (idf inverse de la fréquence documentaire). On fait des calculs de pertinence sur la base de ces pondérations qui reflètent la répartition des unités dans les documents.

Les modèles sémantiques, quant à eux, permettent de définir des unités plutôt sémantiques (concepts, instances de concepts, etc.) qui peuvent être liées les unes aux autres. On les représente le plus souvent par des graphes pour traduire les relations qu'entretiennent les unités sémantiques (relations hiérarchiques, rôles des ontologies, etc.). C'est sur la base de ces relations que différents types de raisonnements sur les concepts ou les instances de concepts permettent de déduire de nouvelles connaissances dans le cadre du WS.

Il semble naturel de chercher à combiner ces deux modèles mais on observe au contraire une certaine stagnation des travaux de RIS : les premières expériences ont été décevantes ou ponctuelles pour permettre de mesurer l'apport de la sémantique à la RI. Nous cherchons donc à revisiter ces travaux de RIS et à développer un modèle de RIS qui combine au mieux les deux modèles de base : il s'agit d'intégrer un modèle sémantique dans un modèle documentaire sans pour autant perdre la richesse de la structure des ontologies ou des ressources sémantiques utilisées et sans abandonner non plus les calculs distributionnels qui font la force et la robustesse de la RI classique.

Cette thèse repose sur deux intuitions fortes que nous explicitons ici.

1.2.1 Un modèle unifié

Nous avons cherché à proposer un modèle unifié pour représenter l'ensemble des informations documentaires et sémantiques à mobiliser pour la RI.

Cette intuition repose sur l'hypothèse suivante : si on souhaite combiner efficacement et avec cohérence les modèles documentaires et les modèles sémantiques actuels pour une recherche d'information sémantique, il est nécessaire de masquer leur hétérogénéité tout en préservant le type d'analyse qui caractérise chacun de ces modèles.

Les modèles documentaires et sémantiques sont en effet hétérogènes. Le premier s'appuie sur des analyses statistiques distributionnelles et manipule des informations textuelles brutes. Le second raisonne à un niveau conceptuel, manipule des entités sémantiques, procède par inférence et reflète les connaissances d'un domaine. La représentation même des modèles diffère : faut-il intégrer les connaissances ontologiques dans les représentations vectorielles ou au contraire proposer une représentation en graphes qui préserve la spécificités des calculs distributionnels ?

Avec le Web Sémantique, Tim Berners-Lee propose d'étendre le réseau des hyperliens reliant les pages du web en un réseau de données pour permettre aux machines de comprendre la sémantique et la signification de l'information du web. Ces données sont structurées et explicitées grâce aux modèles sémantiques, des ontologies notamment, ce qui permet aux agents du web d'y accéder et aux utilisateurs d'effectuer des recherches précises. Quand on évoque le WS, on parle donc de connaissances inter-liées mais aussi des efforts de formalisation des connaissances et des langages de modélisation (RDF(S), OWL, etc.) ou d'interrogation des données (SPARQL, etc.)

standardisés par le W3C. En ce sens, le WS propose une alternative intéressante à la recherche d'information avec la formalisation structurée des données dans des bases de connaissances, une "compréhension" sémantique de ces données par les systèmes et la possibilité de raisonner sur ces données.

Les outils du WS sont cependant conçus pour rechercher des données et non pas des documents. Les systèmes de l'état de l'art, comme celui de [Wang et al., 2011], qui ont essayé de faire de la recherche documentaire avec des techniques du web sémantique ne prennent que partiellement en compte le contenu documentaire. Ils ignorent notamment la distribution des unités dans les textes. Ce sont des modèles sémantiques adaptés pour la recherche de documents, mais le document est représenté par les données (ou méta-données) qui lui sont associées.

Le WS ne permet pas de faire une fusion homogène des modèles documentaires et sémantiques : car les analyses distributionnelles y sont dans ce cas absentes, tout comme le sont les analyses sémantiques dans les modèles de RIS fondés sur les modèles classiques de RI.

On a donc affaire à deux extrêmes : des approches purement numériques, où la sémantique est implicite mais où aucun effort de formalisation n'est requis ni pour l'interrogation ni pour la représentation du texte, et des modèles purement sémantiques, où l'on s'abstrait du texte et du contenu documentaire, qui supposent un effort important de formalisation ne serait-ce pour l'interrogation des connaissances, mais où le sens est explicite et peut être étendu par inférence grâce aux ontologies.

1.2.2 Une représentation sous la forme de graphes

Notre seconde intuition concerne l'intérêt des approches à base de graphes, avec l'idée qu'une représentation en graphe de la recherche d'information doit principalement reproduire la recherche d'information classique en préservant toutes les caractéristiques d'un document.

La représentation des connaissances sous la forme de graphes a existé bien avant l'apparition du WS. De même, quand on parle de recherche documentaire sur le Web, on ne peut pas ne pas évoquer l'algorithme *PageRank* [Brin and Page, 1998], qui ordonne les pages web en fonction de leur notoriété calculée sur la structure de graphe hypertexte du web. On pense également aux approches par propagation d'activation utilisées en sciences cognitives et en intelligence artificielle : on a pu montrer la proximité avec des approches de RI classiques [Preece, 1981, Crestani, 1997] mais elles ont été rapidement abandonnées pour des raisons passage à l'échelle et en raison du coût de la modélisation initiale sous la forme de graphe.

La propagation d'activation permet en revanche d'exploiter de manière unifiée des réseaux composés de noeuds et de relations hétérogènes (différents types de noeuds et de relations), ce qui répond à notre première intuition.

Nous proposons dans cette thèse de revisiter les approches de RI par propagation d'activation. Nous montrons que cela permet de formaliser le problème de la RI différemment, de coupler le modèle documentaire et le modèle sémantique dans un graphe pondéré et de rechercher des informations sur ce graphe unifié.

1.3 Contributions : un modèle unifié en graphe par propagation d'activation pour la recherche d'information sémantique

Nous proposons de modéliser les données sémantiques et documentaires sous la forme de graphe pondéré et la recherche d'information comme une propagation d'activation dans ce graphe à partir des nœuds qui ont été activés par la requête de l'utilisateur. Cet algorithme a le mérite

de préserver les caractéristiques largement éprouvées des modèles classiques de RI tout en permettant une représentation adéquate des modèles sémantiques. C'est en ce sens que le modèle que nous proposons est un modèle unifié.

Ce modèle constitue notre contribution et il couvre tous les aspects de la recherche d'information :

Modélisation en graphe : Nous proposons de représenter la collection documentaire et les ressources sémantiques associées sous la forme d'un *réseau sémantico-documentaire*. Ce réseau est implémenté sous la forme d'un *graphe pondéré* qui est exploité pour la recherche d'information. Cette représentation permet d'intégrer facilement de nouvelles connaissances, à la différence des modèles classiques où il faut ré-indexer tous les textes pour ajouter une entrée à l'index. Elle permet également à l'utilisateur de profiter pour la RI de toutes les connaissances qui sont à sa portée, aussi hétérogènes soient elles, et sans changer de modèle ni de système.

Mise en correspondance des requêtes et des documents : La recherche d'information s'effectue par un mécanisme de *propagation d'activation* sur le graphe pondéré. L'algorithme de propagation permet de propager l'information de pertinence de proche en proche sur le graphe, à partir des nœuds associés à la requête de l'utilisateur. Le degré d'activation dépend de l'activation initiale, d'une fonction d'activation associée aux noeuds et des poids des arcs. Des mécanismes de contrôle sont également introduits pour garantir l'arrêt de la propagation. Nous avons testé cet algorithme sur différentes collections documentaires et nous montrons qu'en tirant partie de la structure du graphe, on peut assurer un ensemble des fonctionnalités sémantiques implicites et explicites.

Interrogation par différents points d'entrée : Notre modèle en graphe autorise plusieurs modes d'interrogation : on peut l'interroger par les termes, par les concepts, par les termes et les concepts conjointement, par les documents, etc., sans rien changer au système. Nous n'explorons pas beaucoup cette piste de recherche mais ce mode d'interrogation doit permettre de proposer des modalités d'accès à l'information variées, au-delà de la RI, comme la recherche de données ou la recherche par l'exemple.

Filtrage des résultats : Pour une requête de RI, notre modèle de graphe pondéré retourne comme résultat l'ensemble des nœuds qui ont été activés dans le graphe à partir de la requête utilisateur en les ordonnant en fonction de leurs valeurs d'activation. S'il s'agit de nœuds documents par exemple, ces valeurs correspondent à la fin de la propagation aux scores des documents et on peut exploiter les mesures usuelles d'évaluation de la qualité de recherche (rappel et précision). On peut classer de la même manière les nœuds termes les plus pertinents au regard de la requête.

Selon la recherche effectuée (RI, recherche de données, etc.), on peut filtrer le type des nœuds à sélectionner, ce qui permet par exemple de restreindre les résultats aux documents dans le cas de la RI. On peut cependant filtrer sur plusieurs types de nœuds ce qui permet de justifier des résultats obtenus en première instance (on peut ainsi retrouver les concepts qui ont enrichi et contribué à la recherche documentaire, par exemple).

Ce modèle à base de graphe pondéré et de propagation d'activation assure les fonctionnalités suivantes :

Reproduction de la RI classique : On peut montrer la parenté de notre modèle avec les modèles classiques de RI. Notre modèle prend en effet en compte la distribution des unités d'indexation dans les documents (fréquences d'occurrence, fréquences documentaires, etc.) ainsi que toutes les caractéristiques du document et de la collection sur le graphe (taille

du document, taille du vocabulaire, etc.). Cette parenté se confirme expérimentalement en terme de qualité de la recherche.

Prise en compte implicite de la sémantique La sémantique implicite se manifeste à travers le calcul de *co-occurrence* qui est fait sur le graphe et qui exploite la sémantique latente dans le texte et dans la collection. A travers la propagation d'activation, la co-occurrence permet de mettre en relief les éléments du graphe qui apparaissent ensemble dans le contexte fixé par l'utilisateur. Le fait que les *valeurs d'activation* de ces éléments soient renforcées améliore la qualité de la recherche sans faire appel à des connaissances externes.

Prise en compte explicite de la sémantique : La sémantique explicite provenant des ressources sémantiques externes permet l'exploitation des classes sémantiques et la résolution de certains problèmes classiques :

- la *synonymie* introduit du silence dans les résultats mais elle peut être compensée par l'exploitation des liens terme/concept entre une classe sémantique et les termes ou *labels* qui lui sont attachés ;
- le *term mismatch* pose problème quand les termes de la requête sont absents du vocabulaire documentaire et de l'index mais on peut contourner cette difficulté en exploitant le volet terminologique associé à la ressource sémantique ;
- le défaut de *couverture sémantique* (certains concepts n'ont pas de correspondant terminologique) peut être compensé par l'exploitation de la sémantique à un autre niveau de la propagation d'activation ; prendre en compte les liens directs entre documents et classes sémantiques permet de profiter de l'apport de la sémantique, même quand les liens terme/concept manquent dans le graphe pondéré.

Le modèle de RIS proposé permet donc d'unifier de manière cohérente les informations numériques et symboliques dans un réseau sémantico-documentaire, qui peut se traduire en graphe pondéré riche. La propagation d'activation sur ce graphe est le mécanisme qui assure la mise en correspondance entre le besoin de l'utilisateur formulé sous la forme d'une requête et les documents. Selon qu'on introduit ou pas de la sémantique dans le graphe, cette approche permet de reproduire une RI classique ou assure en sus certaines fonctionnalités sémantiques. Ces fonctionnalités sont validées expérimentalement sur deux corpus de test : un petit corpus jeu dans le domaine de la cuisine et un corpus dans le domaine médical (*Ohsumed87*) qui permet de vérifier le passage à l'échelle de notre modèle.

1.4 Organisation du manuscrit

Après cette première partie introductive qui expose notre problématique et le modèle que nous voulons mettre en place, nous consacrons le chapitre 2 à l'état de l'art sur la RI et la RI sémantique. Nous passons en revue les travaux qui exploitent les ressources sémantiques externes pour l'amélioration de la RI.

Le troisième chapitre présente les travaux d'accès à l'information sur les graphes, allant du *PageRank* au Web Sémantique. Il introduit le principe de propagation d'activation dont nous nous sommes inspirés.

Dans un quatrième chapitre, nous présentons le modèle que nous proposons : la modélisation du réseau semantico-documentaire, sa représentation en graphe et le mécanisme de recherche d'information par propagation d'activation sur ce graphe. Nous expliquons également les fonctionnalités sémantiques auxquelles ce modèle donne accès.

Le cinquième chapitre est consacré aux premières expériences réalisées sur notre modèle, sur un petit jeu de test. Cela permet d'analyser en détail le comportement de notre modèle et d'en identifier les forces et les faiblesses. Nous mettons également l'accent sur les fonctionnalités sémantiques qui permettent de dépasser certaines limites des approches classiques de RI.

Le chapitre suivant présente l'évaluation expérimentale à grande échelle de notre modèle, sur une collection de test du domaine médical, *Ohsumed87*. Ce chapitre donne une évaluation plus fiable de notre modèle mais il permet aussi d'explorer d'autres traductions possibles de notre modèle et de nouvelles fonctionnalités.

Le chapitre de conclusion présente un résumé de nos contributions au domaine de la RI sémantique et évoque également les perspectives qu'ouvre le modèle proposé de propagation d'activation sur un graphe pondéré.

Chapitre 2

Recherche d'information sémantique

Sommaire

La recherche d'information sémantique est une recherche d'information classique, qui, en se fondant le plus souvent sur des *ressources sémantiques externes*, tient compte, en plus, de la signification véhiculée par les mots des documents (et des requêtes) [Zargayouna et al., 2015]. Cette prise en compte de la signification, se manifeste au cours de la représentation des documents et des requêtes (indexation), mais aussi au cours de l'appariement requêtes/documents (recherche), dans le but de résoudre les ambiguïtés linguistiques et d'améliorer la qualité de la recherche.

Pour ce faire, la RIS associe deux sortes d'informations hétérogènes : l'une est présentée par un modèle documentaire, l'autre par ce qu'offre un modèle sémantique pour la recherche de documents.

Un modèle documentaire, permet de définir les unités d'indexation (mots, termes, etc.) et de les relier aux documents dans lesquels elles apparaissent. Il permet également de définir les liens entre documents ou portions de documents (ex. citation). Ces modèles documentaires, ne sont pas exploités en RI que pour la représentation des documents et des requêtes, mais autorisent par la suite des calculs de pertinence numériques fondés sur la répartition des unités d'indexation dans la collection de documents. Ces modèles documentaires sont appelés le plus souvent en RI par *modèles de RI*, dont le plus connu est le *modèle vectoriel*, qui a fait ses preuves.

Un modèle sémantique, quant à lui, définit les unités sémantiques (concepts, instances de concepts, etc.) qui peuvent être reliées par des relations (relations hiérarchiques, rôles, etc.). Ces modèles sémantiques peuvent être formalisés sous la forme d'un thésaurus, d'une ontologie, ou

d'une terminologie, etc. Ils procèdent par inférence et reflètent les connaissances d'un domaine. En recherche d'information sémantique, ils permettent d'aller au-delà des mots et de raisonner au niveau des concepts ou instances de concepts.

La recherche d'information sémantique est finalement une intégration des connaissances ontologiques explicitées par un modèle sémantique dans un modèle documentaire, qui lui définit la répartition des mots dans les documents, mais aussi la mise en correspondance requêtes/documents.

Nous allons commencer ce chapitre par un retour sur les notions clés de la recherche d'information classique (RI) (section 2.1), présenter notamment le *modèle vectoriel*, ainsi que les limites de ces modèles et les solutions rapportées, sans oublier l'évaluation de la RI. Nous passons par la suite à l'exploitation des *ressources sémantiques externes* pour la RIS (section 2.2), à savoir en indexation et en recherche, avant d'en conclure les défis de la RIS (section 2.3).

2.1 La recherche d'information (RI)

La recherche d'information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information [Salton, 1968a]. C'est le processus qui, à partir des besoins en information des utilisateurs (ensemble des requêtes), permet de retrouver l'ensemble de documents qui leurs sont pertinents.

Un système de recherche d'information (SRI) est un système qui gère une collection de documents organisée sous forme d'une représentation intermédiaire reflétant aussi fidèlement que possible le contenu des documents grâce à un processus préalable d'*indexation*. La recherche d'information désigne alors le processus qui permet, à partir d'une expression des besoins d'information d'un utilisateur, de retrouver l'ensemble des documents contenant l'information recherchée et ce par la mise en œuvre d'un mécanisme d'*appariement* entre la requête de l'utilisateur et les documents ou plus exactement entre la représentation de la requête et la représentation des documents.

La chaîne classique exécutée par un SRI se décompose alors le plus souvent en deux phases primordiales. Ces deux phases sont l'*indexation* et la *recherche* (interrogation). La figure 2.1 illustre ces deux processus :

Le processus d'indexation : il consiste en une représentation commune des documents et des requêtes par les unités significatives extraites, qu'on appelle désormais *termes descripteurs* ou *termes*. Cette formalisation du texte est appelée *index* et servira de base à son identification au sein d'une collection mais aussi à rendre la recherche d'information plus efficace ; ces termes descripteurs peuvent être des mots simples ou composés et subissent un ensemble de traitements morphologiques (radicalisation, lemmatisation, troncature, etc.), exploitant des techniques de traitements automatiques des langues ;

Le processus de recherche : c'est l'expression du besoin de l'utilisateur dans des requêtes et leurs appariements avec les documents de la collection selon un modèle documentaire précis (tel que le modèle vectoriel). Les requêtes sont généralement soumises au même processus d'indexation afin de pouvoir ensuite calculer leurs degrés de correspondance avec les documents de la collection. La réponse présentée par le système est donc la liste des documents triés selon la valeur de similarité calculée.

Un modèle de recherche d'information joue un rôle crucial dans un processus de RI. Il doit donc remplir les deux tâches suivantes :

— Définir la représentation des documents ainsi que des requêtes en fonction des termes d'indexation (les termes descripteurs).

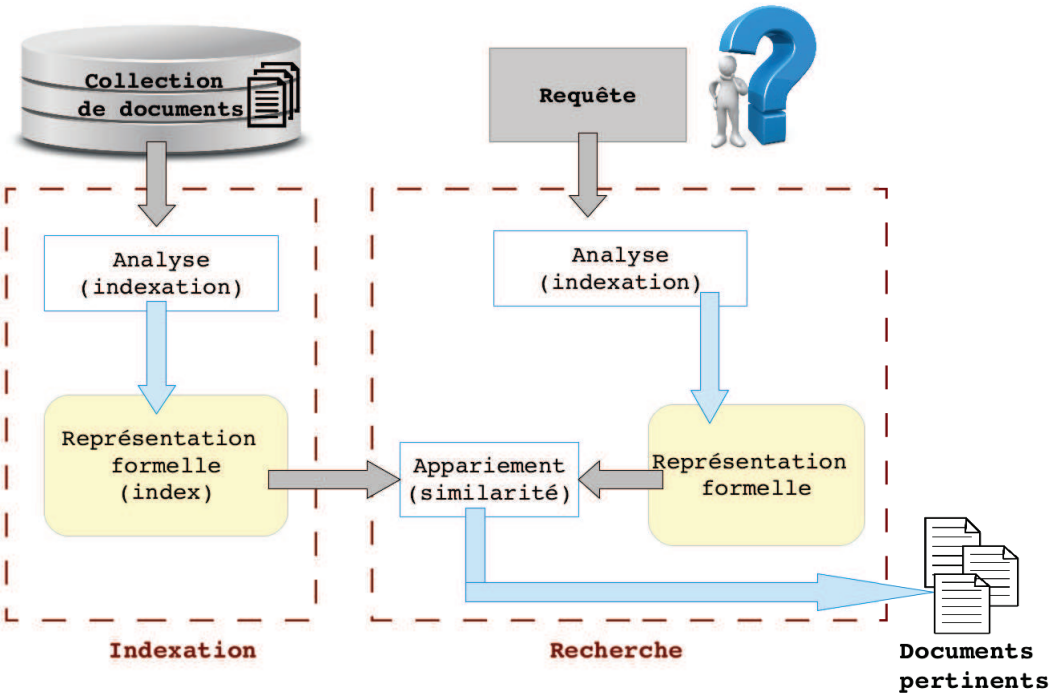

FIGURE 2.1 – Processus de recherche d'information

— Définir la fonction de correspondance RSV (*Retrieval Status Value*) entre chaque document et la requête en question, en vue de mesurer un degré de similarité.

Les modèles de RI manipulent plusieurs variables : les besoins en information de l'utilisateur, la collection des documents, les termes descripteurs, les jugements de pertinence, etc., et se distinguent par le mode d'appariement (exact ou approché) qui dépend du modèle en lui même. [Baeza-Yates and Ribeiro-Neto, 1999] présentent une classification de ces modèles en trois grandes catégories : le modèle booléen, le modèle vectoriel et le modèle probabiliste.

2.1.1 Le modèle vectoriel

Le modèle vectoriel est le modèle dont les fondements sont les plus simples à expliquer en RI. Il s'agit d'un modèle algébrique, tel qu'à chaque terme descripteur t_i est associé sa pondération w_i [Salton and Buckley, 1988b]. Les documents et les requêtes sont représentés par des vecteurs de pondération des termes descripteurs, dans un espace vectoriel composé de tous ces termes d'indexation. Cette représentation, peut être assimilée à une matrice entre termes et documents, qui peut être considérée comme une matrice d'adjacence pondérée d'un graphe. La pertinence d'un document vis à vis d'une requête est définie par des mesures de similarité entre vecteurs.

2.1.1.1 Définition du modèle vectoriel

Vu sa popularité et ses performances reconnues dans la communauté de RI, le modèle vectoriel [Salton et al., 1975] est celui qui nous intéresse. Nous proposons ainsi de détailler ses principes de base :

- Un document est représenté sous la forme d'un vecteur dans l'espace vectoriel composé

de tous les termes d'indexation. Les coordonnées d'un vecteur document représentent les poids des termes correspondants. Formellement, un document d_i est représenté par un vecteur de dimension n :

$$d_i = (w_{i1}, w_{i2}, w_{i3}, ..., w_{in}) \quad i = 1, 2, 3, ...m$$

tels que :
— w_{ij} est le poids du terme t_j dans le document d_i ;
— m est le nombre de documents dans la collection ;
— n est le nombre de termes descripteurs.

- Une requête q est aussi représentée par un vecteur de mots-clés, définie dans le même espace vectoriel que le document.

$$q = (w_{q1}, w_{q2}, w_{q3}, ..., w_{qn})$$

où :
— w_{qj} est le poids du terme t_j dans la requête q ; différentes formules de pondération des termes existent (voir sous-section 2.1.1.2).

- La pertinence du document d_i pour la requête q est mesurée comme étant le degré de corrélation des vecteurs correspondants. Cette corrélation peut être exprimée par diverses mesures de similarités. Ces mesures cherchent à trouver les vecteurs documents qui s'approchent le plus du vecteur requête. Parmi les mesures les plus populaires, on note le *Produit Scalaire* (voir formule 2.1) entre les deux vecteurs, mais également la mesure *Cosinus* (voir formule 2.2), qui mesure le cosinus de l'angle formé par les deux vecteurs, tel que plus le cosinus est grand, plus l'angle est petit et donc plus les deux vecteurs sont similaires.

$$sim(d_i, q) = \sum_{j=0}^{n} (w_{qj} * w_{d_i j}) \tag{2.1}$$

$$sim(d_i, q) = \frac{\sum_{j=0}^{n} (w_{qj} * w_{d_i j})}{\sqrt{\sum_{j=0}^{n} w_{qj}^2 * \sum_{j=0}^{n} w_{d_i j}^2}} \tag{2.2}$$

Dans leurs travaux, Fang et Zhai [Fang and Zhai, 2005, Fang and Zhai, 2006] ont présenté une axiomatisation des fonctions de correspondance en RI. Ces axiomes sont assez intéressant à découvrir, et fixent les bonnes propriétés que doit tout calcul de correspondance en RI pour espérer avoir de bon résultats.

Le calcul des degrés de similarité entre les requêtes et les documents est donc assez déterminant pour la qualité des résultats et une bonne foction de correspondance permet d'améliorer largement le résultat de la recherche. Le modèle vectoriel permet en effet d'ordonner la liste des documents résultats selon les similarités calculées.

2.1.1.2 Pondération des termes descripteurs

La pondération des termes a vu le jour avec le modèle vectoriel, elle consiste en l'attribution de pondérations aux termes descripteurs [Salton and Buckley, 1988b]. Ceci revient à associer un poids discriminant w_{ij} à chaque terme descripteur t_j d'un document d_i. Ce poids reflète le pouvoir de représentativité de t_j dans le document d_i.

De manière générale, les formules de pondération utilisées sont fondées sur la combinaison de :

Un facteur de pondération local : quantifiant la représentativité locale du terme dans le document, et qui correspond généralement à son nombre d'occurrence dans ce document ; plus grande est la fréquence d'occurrence locale d'un terme descripteur, plus grande est la représentativité de ce terme dans ce document ; il est généralement noté tf_{ij}, pour un terme t_j et un document d_i ;

Un facteur de pondération global : quantifiant la représentativité globale du terme vis-à-vis de la collection de documents, et qui correspond à la fréquence d'apparition de ce terme dans cette collection ; il est généralement noté $df(j)$ pour un terme t_j ;

Compte tenu de ces deux facteurs de pondération, l'importance d'un terme est proportionnelle à sa fréquence d'apparition dans chacun des documents (fréquence relative ou locale) et inversement proportionnelle au nombre de documents dans lequel il apparaît (fréquence globale inverse), ce qui se traduit par la formule TF_IDF suivante de [Salton and Buckley, 1988b] (formule 2.3) :

$$w_{ij} = tf_{ij} * \frac{1}{df_i} = tf_{ij} * idf_j \tag{2.3}$$

Où :

— tf_{ij} : la fréquence d'occurrence du terme t_j dans le document d_i ;
— df_j : la fréquence documentaire du terme t_j (les documents qui contiennent t_j) ;
— $idf_j = \frac{1}{df_i}$: la fréquence documentaire inverse du terme t_j ;

Cette mesure a eu en revanche un succès très limité dans les corpus de tailles très variables du fait qu'elle ne tient pas compte ni de la taille de la collection de documents ni de la longueur du document. Les distorsions engendrées par cette hétérogénéité sont corrigées en effectuant une normalisation de la pondération globale. La normalisation de [Jones, 1972] qui prends en compte la taille de la collection lui a été appliquée. En ce qui concerne la longueur du document, le problème posé est que les termes appartenant aux documents longs apparaissent très fréquemment et l'emportent en poids sur les termes appartenant à des documents moins longs. La normalisation de type cosinus de [Buckley, 1996] a été alors proposée.

Plusieurs autres normalisations ont été proposées notamment sous l'impulsion des conférences comme TREC. On peut par exemple citer la normalisation pivotée de *Singhal* [Singhal et al., 1996].

En 1996 [Robertson et al., 1996] Robertson définie sa propre fréquence d'occurrence tf_{ij}, souvent appelée *Robertson_ TF* ou *Okapi_ TF*[3], tel que la longueur de chaque document a été aussi uniformisée dans la collection, en la normalisant par rapport à la longueur moyenne des documents. Cette pondération, a au départ été proposée dans le cadre probabiliste [Robertson et al., 1998], et la mesure de similarité qui prend en compte cette pondération peut être considérée comme un TF_IDF "moderne" prenant mieux en compte la longueur des documents [Claveau, 2012]. Sa formule est donnée dans l'équation 2.4 :

$$w_{BM25_{ij}} = tf_{BM25_{ij}} * idf_{BM25_j} = \frac{tf_{ij} * (k_1 + 1)}{tf_{ij} + k_1 * (1 - b + b * \frac{dl_i}{dl_{avg}})} * \log(\frac{N - df_j + 0.5}{df_j + 0.5}) \tag{2.4}$$

Où :

— k_1 : est une constante qui permet de contrôler l'influence de la fréquence du terme t_j dans le document d_i. Sa valeur dépend de la longueur des documents dans la collection. Dans

3. Cette pondération s'appelle BM-25, mais elle est appelée Okapi pour faire référence au premier système qui l'a mise en place.

la campagne d'évaluation TREC (voir section 2.5.3), les meilleurs résultats sont obtenus pour $k_1 = 1.2$ et 2 ;

— b : est une constante appartenant à l'intervalle [0, 1] qui permet de contrôler l'effet de la longueur du document. Si $b = 1$, c.a.d que les documents sont longs car ils contiennent des termes qui se répètent. Si $b = 0$ alors les documents sont longs car ils sont *multi-topic* (contiennent des termes distincts) ;

— dl_i : est la longueur du document d_i ;

— dl_{avg} : la longueur moyenne des documents dans la collection entière.

[Claveau, 2012] montre récemment, que la formule de pondération Okapi-BM25 est meilleure que TF_IDF, et est devenue une référence grâce aux très bons résultats qu'elle permet d'obtenir sur de nombreuses tâches de RI. Pour cette raison, nous allons utiliser cette formule de pondération, dans les scénarios de RI classique, comme *Baseline* dans notre étude expérimentale.

Le modèle vectoriel présente divers avantages par rapport aux autres modèles, et qui résident en sa simplicité conceptuelle, par opposition au modèle booléen, qui nécessite la connaissance de l'algèbre de Boole, mais aussi en sa capacité de trier les résultats d'une recherche à travers les mesures de similarité. Cet ordonnancement va permettre de mesurer la qualité de la recherche, à travers les métriques usuelles d'évaluation de la RI (voir section 2.1.2).

2.1.2 L'évaluation en RI

L'évaluation est une étape cruciale dans la mise en œuvre d'un SRI. En effet, elle sert de jugement pour sa performance et sa fiabilité. La qualité d'un SRI est déductible selon [Cleverdon et al., 1966] par plusieurs facteurs dont : le temps de réponse, la présentation des résultats, l'effort requis de l'utilisateur pour retrouver parmi les documents retournés, ceux qui répondent à son besoin, et finalement le taux de rappel et de précision.

Toutes les compagnes d'évaluation des SRIs, comme TREC [4] ou CLEF [5], s'intéressent plus à la pertinence des résultats retournés et donc à la qualité du résultat (en terme de rappel et de précision), qu'à la rapidité de la réponse, c'est d'ailleurs pourquoi nous allons mettre l'accent sur ces deux notions de rappel et précision. L'évaluation d'un SRI nécessite : un corpus de test, un ensemble de requêtes, et un ensemble de documents jugés pertinents.

Dans cette section nous proposons, d'introduire les métriques d'évaluation les plus utilisées dans l'évaluation des performances des SRI, à savoir le rappel et la précision, la *MAP* et la *R-PREC*.

Rappel

C'est la capacité du système à fournir en réponse tous les documents pertinents. Elle peut être vue comme une mesure de couverture du système. C'est le rapport du nombre de documents retournés par le système et convenant à l'utilisateur (pertinents) au nombre total de documents convenant à l'utilisateur.

$$Rappel = \frac{|P \cap R|}{|P|} \in [0, 1]$$

4. Test REtrieval Conference
5. Cross Language Evaluation Forum

Précision

C'est la capacité du système à ne fournir que des documents pertinents en réponse. C'est le rapport du nombre de documents pertinents retournés par le système au nombre total des documents retournés par le système.

$$Precision = \frac{|P \cap R|}{|R|} \in [0, 1]$$

Où :
— P : est le nombre de documents pertinents ;
— R : est le nombre de documents retournés ;
— $\cap$: intersection au sens de l'inclusion.

MAP

La *MAP* (*Mean Average Precision*) est la précision moyenne ou encore la moyenne des précisions obtenues à chaque fois qu'un document pertinent est retrouvé. Cette mesure a été introduite dans TREC2 pour sa capacité à résumer les mesures de précision aux 11 points de rappel [Déjean et al., 2010]. Pour une requête q, ayant un nombre P de documents pertinents et un nombre R de documents retournés, elle se calcule comme suit :

$$MAP(q) = \frac{\sum_{k=1}^{R}(P(k) * rel(k))}{P}$$

où :
— $P(k)$: est la précision à k documents retournés (au rang k) ;
— $rel(k)$: est une fonction d'indication, égale à 1 si le document de rang k est pertinent ; nulle sinon ;

R-PREC

La *R-PREC* (*R-Précision*) est la précision après que R documents ont été retrouvés, où $R \leq P$, est le nombre de documents pertinents pour la requête considérée. Cette mesure a été introduite dans TREC2 pour limiter l'influence du nombre de documents pertinents et qui varie en fonction des requêtes [Beitzel et al., 2009]. Pour une requête q, ayant un nombre P de documents pertinents et un nombre R de documents retournés, elle se calcule comme suit :

$$R - PREC(q) = \frac{N_{(R \leq P)}}{P}$$

où :
— $N_{(R \leq P)}$ est le nombre de documents pertinents retournés pour la requête, avant que leurs rangs aient atteint le nombre de documents pertinents de la requête, c-à-d, avant d'atteindre P documents retournés.

Un SRI idéal est un SRI qui restitue tous les documents pertinents (Rappel=1) et rien d'autres que les documents pertinents (Précision=1). Afin de se rapprocher au maximum d'un système idéal, on vise à minimiser le silence ($silence = 1 - rappel$) en vue d'augmenter le rappel, et à réduire le bruit ($bruit = 1 - precision$) pour gagner en terme de précision.

2.1.3 Limites de la RI classique et solutions apportées

La mise en place du processus de RI passe par la spécification d'un modèle de RI intégrant : une représentation des documents, une représentation des requêtes et un appariement entre les deux représentations. Plusieurs modèles ont été proposés dans la littérature ayant pour but de répondre au mieux aux besoins des utilisateurs. Le modèle vectoriel proposé par [Salton et al., 1975] est le modèle le plus simple. Il propose une représentation des documents et des requêtes en vecteurs, dans un espace multidimensionnel composé par les termes descripteurs constituant le vocabulaire de la base documentaire, et une fonction de mise en correspondance. Cette dernière est fondée sur la distribution des termes dans les documents et les requêtes.

Ces modèles classiques de RI, fondés sur des calculs distributionnels, ont été largement prouvés, notamment dans le traitement des grandes collections de documents et passent facilement l'échelle. Cependant, ces modèles algébriques gèrent des informations bas niveaux, et manipulent des fonctions de mise en correspondance permettant un calcul statistique de répartition des termes dans les documents. Parmi les autres lacunes, on cite également le fait que tous les termes sont considérés indépendants, un problème qualifié de *orthogonal models* par [Huang et al., 2012] et le manque d'expressivité des requêtes.

La RI souffre de ce fait de certaines limites qui font accentuer le bruit et diminuer donc le rappel. Ces limites sont dues principalement aux lacunes suivantes :

Une représentation pauvre des documents : Il s'agit d'une représentation en *"sac de mots"* (*Bag Of Words*). Cette représentation ignore les problèmes d'ambiguïté dus à la richesse de la langue. Ceci fait que ces modèles sont incapables de considérer qu'un seul terme peut avoir plusieurs significations (polysémie), ou que plusieurs termes peuvent avoir un même sens (synonymie) [Huang et al., 2012]. On note également le manque d'expressivité des termes simples.
Parmi les conséquences directes d'une représentation pauvre, le problème de *term mismatch* [Crestani, 2000] [Xu and Croft, 2000] [Zhao, 2012]. Ce problème résulte du manque d'expressivité des requêtes, qui sont généralement courtes et expriment le besoin de l'utilisateur différemment des auteurs des documents, pour faire référence aux mêmes notions [Crestani, 2000]. Il s'agit donc d'un *mismatch* entre le vocabulaire de la requête et celui des documents.

Une mise en correspondance statistique : Il s'agit d'un appariement qui ne tient pas compte de la signification des termes descripteurs qu'il manipule et traite les termes de la requête morphologiquement et statistiquement selon un processus de calcul des distributions des termes selon leurs de fréquences d'occurrence. Le terme n'est donc pas considéré comme une entité ayant une sémantique (sens) précise, et se situant dans un voisinage sémantique qui lui est "proche". La pertinence des résultats par rapport aux besoins des utilisateurs est de ce fait numérique, et non fondée sur une compréhension sémantique du besoin.

A partir des années 80, les chercheurs de la communauté de RI ont commencé à exploiter les techniques de traitement du langage naturel (TALN) comme un moyen possible pour pallier à ces limites, tel que l'identification des mots composés. Ceci permet de s'affranchir de quelques problèmes d'ambiguïté dus au manque d'expressivité des seuls mots isolés. Cette problématique a souvent été traitée par les analyses morpho-syntaxiques dont l'enjeu à ce niveau est l'identification de nouvelles unités linguistiques (termes composés, mots composés, etc.). Cette pratique fait augmenter considérablement la précision, vu que les termes composés, sont le plus souvent moins polysémiques que les termes simples. D'après [Jacquemin and Zweigenbaum, 2000], l'utilisation

des multi-termes (termes composés) pour l'indexation des documents est très intéressante afin de pallier l'utilisation restrictive des seuls mots isolés, qui fait partie des conséquences des approches de *"sac de mots"*.

Parmi les autres pistes explorées par les chercheurs en RI, la prise en compte de la sémantique dans un processus de RI, implicitement ou explicitement. Implicitement, à travers la prise en compte de la sémantique latente (*Latent Semantic Indexing* (LSI)) [Deerwester et al., 1990]. La méthode LSI tient compte des potentielles structures sémantiques implicites des termes, représentés par leurs dépendances cachées afin de résoudre les limites de la RI classique, tel que le *term mismatch*. Cette méthode s'est néanmoins révélée décevante pour des documents de petites tailles selon [Rehder et al., 1998] et demande un grand effort d'apprentissage et de calibrage, mais qui continue à être une méthode intéressante pour la prise en compte de la sémantique de nos jours.

La communauté de recherche d'information sémantique, s'est aussi retournée vers la considération explicite de la sémantique dans un processus de RI. Ceci est manifeste à travers l'exploitation des ressources sémantiques externes, allant de la désambiguïsation du vocabulaire par les thésaurus, à l'enrichissement sémantique, par un codage explicite du sens dans le flux d'information lors de la phase d'indexation ou de recherche, moyennant les ontologies ou les taxonomies (voir [Zargayouna et al., 2015] pour différencier entre ces ressources sémantiques externes).

Nous avons présenté dans cette section les différentes limites que nous repérons *a priori* autour des systèmes classiques de RI. Nous nous n'intéressons pas à l'amélioration de la représentation des résultats à l'utilisateur et son acquisition. Notre principal objectif est principalement l'amélioration de la qualité de la recherche (rappel et précision) et donc la résolution des problèmes autour de la représentation pauvre des documents et des requêtes dont souffrent les modèles classiques de RI. Nous réalisons ceci à travers l'exploitation des ressources sémantiques externes. Pour ce faire, nous commençons par faire un état de l'art sur leur utilisation en recherche d'information sémantique (RIS).

2.2 Ressources sémantiques et RI

La RIS a pour but de "caractériser" et "retrouver" un ensemble de ressources documentaires par leurs significations et l'ensemble des connaissances qui y sont contenues. L'exploitation des ressources sémantiques externes en RIS est une solution possible, qui a pour but de permettre aux modèles classiques de RI, tel que le modèle vectoriel ou le modèle probabiliste, etc., d'injecter la sémantique en indexation ou en recherche. Le résultat est le plus souvent une adaptation des modèles classiques pour la prise en compte de la sémantique. Le modèle vectoriel est le modèle qui a été le plus adapté en RIS, mais il existe également des travaux de RIS qui concernent le modèle probabiliste tel que [Maisonnasse, 2008]. Nous nous focalisons plutôt sur l'exploitation de la sémantique dans le modèle vectoriel, étant le modèle le plus simple à adapter. Nous allons commencer tout d'abord par décrire les ressources sémantiques exploitées en RIS, puis par montrer l'apport de ces ressources en indexation et en recherche.

2.2.1 Ressources sémantiques exploitées

Le défi de la recherche d'information sémantique est de développer des systèmes capables d'intégrer plus de sémantique dans leurs traitements. L'objectif étant d'approcher la compréhension de la machine au besoin du demandeur d'information. L'utilisation des ressources sémantiques externes présente une piste active dans ce domaine.

Il existe plusieurs ressources sémantiques exploitées en RIS, tel que les thésaurus, les ontologies, les taxonomies, etc., cependant les connaissances sémantiques utilisées à partir de ces ressources, ne diffèrent pas entre elles. Ces connaissances sont principalement les concepts, les liens hiérarchiques via les calculs de pondérations et de scores, les liens de méronymie/holonymie et de synonymie.

Un concept est défini dans [Chevallet, 2009] comme étant une abstraction généralisée de propriétés communes à plusieurs objets, faits ou événements. Dans un texte, le concept est représenté par un ou plusieurs termes synonymes ayant une entrée dans une ressource terminologique (ontologie, thésaurus, dictionnaire, etc.)[Chevallet, 2009].

Nous ne distinguons pas entre ces ressources sémantiques dans notre présentation de l'état de l'art. En effet, nous appelons "ontologie" la modélisation conceptuelle de toute ressource sémantique externe : thésaurus, taxonomie, etc.

Ces ressources sémantiques doivent uniquement assurer une bonne couverture terminologique du vocabulaire des documents, c-à-d, être *lexicalisées*, afin de faciliter la mise en relation, ou l'*annotation sémantique* entre les deux vocabulaires de la collection et de la ressource et par conséquent l'indexation sémantique/conceptuelle. Elles doivent également garantir un niveau de formalisation suffisant pour faciliter leur *Parsing*.

Ces ressources sémantiques sont soit génériques, soit des ressources de domaine (médical, géographique, juridique, etc.). Les ressources du domaine ont par contre l'avantage de s'affranchir partiellement des problèmes d'ambiguïté de la langue, ceci en ayant restreint l'interprétation des concepts au contexte spécifié par le domaine documentaire.

Les ressources externes les plus utilisées en RI sont notamment *WordNet* pour le domaine générique, et *MeSH*, *UMLS*, *SONMED CT*, *GO*, etc., pour le domaine médical et qui présente l'un des domaines qui ont eu beaucoup de succès étant riche en ressources sémantiques.

WordNet [Miller, 1995] présente une base lexicale largement utilisée dans la littérature de la RI pour les domaines génériques, et qui couvre la majorité des noms, verbes, adjectifs et adverbes de la langue anglaise . Son domaine général lui permet souvent de couvrir les sujets traités dans les collections de test conventionnelles de la RI (telles que TREC, CLEF, etc.) qui sont le plus souvent de type presse. *WordNet* est un réseau de termes organisé en des ensembles de synonymes nommées *synsets*, c-à-d "sens du mot". Chaque ensemble de synonymes désignent ensemble un concept. Ces concepts sont reliés entre eux avec les relations sémantiques suivantes : la synonymie (qui présente en quelque sorte la définition d'un *synset*), les relations de hyponyme-hyperonyme (is-a) ou de générique/spécifique, et la relation de composition ou méronymie/holonymie (partie-tout).

L'utilisation de ces ressources sémantiques externes dans un processus de recherche a pour but de définir notamment de nouvelles représentations des textes (et des requêtes) qui sont plus riches, plus précises et plus efficaces dans la résolution de certains problèmes cités de la RI classique. Des méthodes de RIS fondées sur ces ressources sémantiques ont donc été proposées afin, soit d'enrichir la représentation des documents et des requêtes (indexation), avec ou sans une désambiguisation au préalable, soit de résoudre le problème des requêtes courtes non expressives et de synonymie avec une expansion des requêtes (recherche).

2.2.2 Désambiguïsation sémantique du vocabulaire

Les premiers travaux autour de l'indexation sémantique se sont plutôt focalisés sur la désambiguïsation. La désambiguïsation est le processus qui permet de sélectionner un sens unique pour un "terme" dans le cas de la polysémie. Étant donné l'ambiguïté des ressources sémantiques les plus utilisées en RIS et qui sont notamment le thésaurus *WordNet* et les thésaurus médicaux :

UMLS et *MeSH*, leur utilisation en RIS a souvent été accompagnée d'un processus préalable de désambiguïsation.

Une étude menée par [Torsten et al., 2008] a montré à titre d'exemple d'ambiguïté, qu'il existe 175 termes qui désignent en même temps des espèces et des protéines, 67 termes qui référencent des médicaments et des protéines, 123 termes désignant des cellules et des tissus. Cette ambiguïté peut résulter du fait que dans certains thésaurus, tel que *MeSH*, un même concept peut être localisé à différentes hiérarchies et à différents niveaux dans une hiérarchie de la ressource [Dinh and Tamine, 2010]. A titre d'exemple le concept "pain" de *MeSH* se trouve à 5 branches de 4 hiérarchies différentes, qui sont les 5 concepts les plus génériques suivants : "Nervous system Disease" (C10), "pathological conditions", "Signs and Symptoms" (C23), "Psychological Phenomena and Processes" (F02),"Musculoskeletal and Neural Physiological Phenomena" (G11) [Dinh and Tamine, 2010].

Différentes méthodes de désambiguïsation ont été proposées. Parmi ces méthodes :

Désambiguïsation à partir de la ressource sémantique : [Stetina et al., 1998] utilisent les synonymes de *WordNet* lors de l'identification des *synsets* dans la collection documentaire, et montrent une amélioration dans la qualité de la recherche par rapport à l'indexation ne considérant que le premier sens (*synset*) sélectionné de *WordNet*. [Guarino et al., 1999] incluent également une désambiguïsation supervisée des termes du vocabulaire à partir des *Synsets* de *WordNet* afin de sélectionner le sens adéquat dans le cas de la polysémie. Dans sa méthode de désambiguïsation des concepts de *MeSH*, la proposition de [Dinh and Tamine, 2010] se fonde également sur l'ontologie pour résoudre la polysémie, notamment sur la position dans la hiérarchie, en choisissant le concept le plus à gauche de *MeSH*.

Désambiguïsation à partir du contexte du terme dans le document : [Khan, 2000] réalise une désambiguïsation lors du *mapping* entre la ressource et les documents. L'auteur démarre de l'hypothèse qu'un groupe de mots-clés qui occurrent ensemble dans un même contexte déterminent les concepts appropriés pour désigner ensemble un autre concept. Il utilise alors une nouvelle méthode de désambiguïsation des termes fondée sur deux principes : la co-occurrence et la proximité sémantique [Agirre and Rigau, 1996] au cours de l'identification des concepts. On peut également évoquer la similarité sémantique entre chaque sens du terme et celui des termes voisins évoqués par [Baziz, 2005]. La similarité sémantique est calculée entre les différents sens des termes en vue de sélectionner, pour chaque terme, le meilleur sens correspondant dans la ressource externe. En effet, si un terme donné est dénoté par deux concepts, le concept ayant une meilleur similarité sémantique avec les concepts voisins dans le même document est retenu comme étant le sens unique de ce concept. Cela permet, selon l'auteur, de désambiguïser les termes compte tenu du contexte du document. De même [Dinh and Tamine, 2010] utilisent une méthode de désambiguïsation des concepts de *MeSH* à partir de leurs contextes et se fonde sur (a) l'hypothèse d'un sens unique à un concept dans un document, (b) la corrélation des sens des concepts voisins dans le cadre d'un contexte donné.

Désambiguïsation à partir de quelques choix lors de la sélection du sens d'un mot : [Baziz, 2005] choisit par exemple d'indexer avec les concepts les plus longs [6], car ils sont moins polysémiques, ou encore choisir d'indexer par tous les termes proposés par l'outil d'identification des concepts, tel que dans dans [Koopman et al., 2012], là où on annote

6. Un concept long : le concept ayant le plus de termes qui le composent

par tous les concepts proposés par l'annotateur médical *MetaMap*. [Hamadan et al., 2012] choisissent en revanche d'indexer avec les concepts ayant le plus grand score dans *MetaMap* (la stratégie du *best concept*). Le but de ces choix est, entre autres, d'effectuer un équilibre entre exhaustivité et abstraction du concept ou sens du mot. Si l'on dispose par exemple des concepts `#apple`, `#green_apple` et `#ripe_green_apple`, et du mot composé "green apple", par quel concept peut-on le dénoter : par le concept le plus long (plus spécifique) ou le plus court (le plus générique), par le père ou par les fils, etc.

La plupart des travaux présentés dans ce qui précède montrent certes, l'intérêt des mécanismes de désambiguïsation pour la sélection d'un sens unique des termes du vocabulaire mais leurs impacts sur les performances du SRI n'ont pas beaucoup été prouvés (voir la thèse de [Baziz, 2005] et de [Radhouani, 2008] pour plus de détails sur la désambiguïsation). Ces mécanismes de désambiguïsation diffèrent les uns des autres et sont liés à un nombre de choix à effectuer et qui sont également différents. Ceci rend la tâche d'évaluation de la performance des méthodes de désambiguïsation difficile à mettre en place. [Krovetz and Croft, 1992] et [Sanderson, 1994] ont déjà noté que la principale difficulté pour améliorer les performances de recherche est due à l'inefficacité des outils de désambiguïsation. Pour cette raison, les auteurs de l'état de l'art ont décidé d'agir sur d'autres paramètres de la RI afin d'améliorer la qualité de la recherche.

Parmi ces paramètres on note la qualité de la ressource utilisée et notamment sa couverture par rapport au vocabulaire du corpus. Par exemple Baziz a rapporté que *WordNet* ne couvre pas tout le vocabulaire utilisé dans la collection de documents médicaux constituée d'articles extraits de *SpringerLink* qu'il utilise (le taux de couverture représente 87% du vocabulaire des documents, 77% du vocabulaire utilisé et 32% dans les requêtes) [Baziz, 2005]. Pour pallier ce problème, l'auteur a essayé de faire une expansion de l'espace de recherche en combinant l'espace des mots et l'espace de concepts. Cette constatation a également été faite par[Schütze and Pedersen, 1995] et [Mihalcea and Moldovan, 2000] qui pensent que le majeur inconvénient des outils de désambiguïsation est qu'ils représentent leurs espaces de recherche par des "sac de concepts". En plus du choix des unités d'index, la méthode de pondération des descripteurs choisie pour représenter le contenu des documents et des requêtes a un grand impact sur la qualité de la recherche. Nous allons voir dans ce qui suit l'impact de la prise en compte de la sémantique sur les performances d'un SRI sur ces deux niveaux : l'enrichissement de la représentation des documents (et des requêtes) et la pondération de ces termes descripteurs.

2.2.3 Enrichissement sémantique de la représentation des documents

A travers l'utilisation des ressources sémantiques au cours de l'indexation, l'objectif principal est d'améliorer la représentation des documents et des requêtes, afin de doter l'index classique d'une meilleure expressivité. L'utilisation de ces ressources pour une indexation sémantique revient à s'appuyer sur les concepts d'un modèle et les relations qu'ils entretiennent dans ce modèle pour donner du sens à l'information désambiguïsée exprimée dans les documents (et les requêtes). Ainsi, de manière complémentaire au contenu des textes, on fait appel à des informations absentes des textes (mais explicites dans la ressource) qui aideront à mieux en caractériser le contenu [Aussenac-Gilles, 2008].

Ainsi, le but de l'enrichissement de la représentation des documents, est d'aboutir à une représentation fidèle des documents (et des requêtes) plus riche et précise et moins ambiguë qu'une représentation en "sac de mots".Les étapes d'une indexation sémantique conceptuelle, qui tend à enrichir la représentation des documents sont :

1. Établir un lien entre les documents (et requêtes) et la ressource sémantique externe : ce mécanisme peut s'accompagner des mécanismes de désambiguïsation cités, comme il peut

être un simple ancrage entre les deux vocabulaires ; la limite principale de ce processus, est l'exhaustivité, c-à-d, la difficulté du repérage de toutes les informations possibles (exemple la difficulté de repérage des entités nommées, etc.) ; ce mécanisme de *mapping* est appelé "annotation sémantique".

2. Enrichir la représentation des documents en choisissant des descripteurs susceptibles de bien représenter le contenu informationnel, et de résoudre les problèmes de polysémie et de synonymie. On peut trouver différentes représentations dans l'état de l'art.

L'annotation sémantique en RIS consiste à repérer pour chaque document l'ensemble des informations sémantiques qui y sont rattachées dans le modèle sémantique (thésaurus, ontologie, etc.), et peut s'accompagner de mécanismes de désambiguïsation si cette ressource est ambiguë. L'annotation sémantique est une tâche difficile qui fait généralement appel à plusieurs disciplines telles l'extraction d'information (EI), le Traitement Automatique de la Langue (TAL) et l'Ingénierie des Connaissances (IC), et a pour but d'assurer une meilleure couverture du contenu du document avec un coût minimum. Une fois l'annotation sémantique de la collection documentaire réalisée, son exploitation pour l'enrichissement de l'index peut être réalisée par différentes manières.

Indexation par les concepts

Il s'agit d'une indexation par les seuls concepts reconnus dans le document grâce au processus d'annotation. A titre d'exemple, [Khan, 2000] procède par une indexation moyennant que les concepts qui ont été identifiés à partir d'une ontologie de sport désambiguïsée. 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% comparé à une indexation par mots clés ayant un rappel de 56% et une précision de 67,75%. Ce résultat a ainsi prouvé l'intérêt d'une indexation par les seuls concepts.

De même, les expérimentations effectuées par [Gonzalo et al., 1998] ont prouvé qu'en plus de la désambiguïsation, le choix des descripteurs a un impact direct sur la qualité de la recherche. Les résultats retournés optent pour une indexation par les seuls *synsets* (concepts) de *WordNet*, donnant un rappel de 62%, contre 53,2% pour une indexation par les mots-sens (synonymes) et 48% pour l'indexation classique par mots (le système SMART).

Certes [Gonzalo et al., 1998], [Stetina et al., 1998], [Khan, 2000], ont démontré l'intérêt d'une indexation par les seuls concepts (*synsets* de *WordNet*), cependant des améliorations peu significatives ont été notées face à la lourdeur de la tâche de désambiguïsation qui accompagne le processus d'identification d'un sens unique des termes. A partir de cette constatation [Schütze and Pedersen, 1995] et [Mihalcea and Moldovan, 2000] pensent en effet que le majeur inconvénient de ces travaux c'est la représentation de l'espace de recherche par des "sac de concepts". Ceci revient au fait que si la ressource n'est pas assez couvrante ou si le processus d'annotation n'est pas tout à fait efficace, des informations importantes dans le document peuvent être ignorées lors de la représentation du contenu du document.

Indexation par les termes et les concepts

Étant donné les problèmes cités de l'indexation par les concepts, des travaux ont essayé de présenter le vocabulaire par une combinaison des termes et des concepts.

Sur une collection issue du projet *MucheeMore* contenant 7823 documents qui traitent du domaine médical, constituée d'articles extraits de *SpringerLink*, [Baziz, 2005], a testé l'impact d'une indexation par les seuls concepts, là où il choisit d'indexer par les concepts les plus longs.

Les résultats montrent que cette méthode n'améliore pas la qualité de la recherche par rapport à l'indexation classique par les mot-clés. L'auteur justifie ce résultat par le fait que *WordNet* ne couvre pas tout le vocabulaire utilisé dans la collection. Il procède alors par une indexation combinant et les mots clés et les concepts. Les résultats ont alors montré que l'indexation conjointement avec les concepts de *WordNet* et les termes descripteurs améliore nettement la précision sur tous les points de précision. Par exemple, la précision pour les cinq premiers documents restitués atteint 33% alors qu'elle n'est que de 26% dans le cas de l'indexation classique.

Dans [Dinh and Tamine, 2010], les auteurs ont mis en place une indexation sémantique fondée à la fois sur les concepts de la ressource *MeSH* désambiguïsée et les termes ne correspondant à aucune entrée dans *MeSH* ("termes orphelins"). L'évaluation de cette méthode d'indexation avec le corpus DMP et 25 requêtes, montre une amélioration de 15.67% de la *Mean Reciprocal Rank* MRR [7] de l'approche d'indexation sémantique (par concepts et termes) par rapport à l'indexation classique et prouve l'efficacité de l'indexation conjointement par les concepts et les termes. De même dans [Hamadan et al., 2012], les auteurs se sont également focalisés sur la combinaison des deux espaces de recherches (termes et concepts) pour la représentation des documents.

La plupart des expériences prouvent l'intérêt de l'enrichissement de la représentation des documents par la sémantique, que ça soit pour la résolution des problèmes d'ambiguïté (polysémie et synonymie), en regroupant dans une même "classe" des mots ayant la même signification et en écartant des mots de même morphologie et ayant des sens différents, ou pour la résolution du problème de couverture de la ressource (à travers la combinaison des deux espaces de mots et de concepts). Cependant, en plus de l'enrichissement de la représentation par le choix des unités d'index, l'enrichissement du schéma de pondération des descripteurs joue un rôle important dans l'amélioration de la qualité de la recherche. Plusieurs méthodes de pondération ont alors été proposées.

2.2.4 Enrichissement sémantique des pondérations et scores

A l'issue du choix des descripteurs susceptibles de représenter le contenu des documents ou des requêtes, ces derniers sont pondérés par des poids qui traduisent leur importance dans le document (ou la requête). Les poids des termes du document et de la requête sont combinés dans un score de pertinence associé au document à l'issue de la recherche. De la qualité de la pondération dépend donc la pertinence des scores attribués lors de la recherche. Pour cela, de nombreux travaux en RI ont été consacrés à la définition du schéma de pondération. Ces informations distributionnelles sont d'autant plus importantes que la sémantique des termes. C'est pour cette raison que des adaptations, ayant pour but d'enrichir les schémas classiques ont été proposées afin de pondérer également les connaissances (concepts, etc.).

Pondération des concepts

Parmi les schémas de pondération sémantiques, on cite à titre d'exemple $CF.IDF$ (*Concept Frequency-Inverse Document Frequency*), proposé par [Baziz, 2005] et qui présente une adaptation du schéma classique $TF.IDF$. Cette mesure de pondération prend compte de la longueur des termes composés et des liens hiérarchiques, qui sont censés représenter le degré de spécificité ou

7. MRR (Mean Reciprocal Rank) : La MRR en RI permet d'évaluer le nombre de documents qu'il faut considérer avant de retrouver le premier document pertinent. Elle est égale à la moyenne calculée sur l'ensemble des requêtes, du rang du premier document pertinent [Voorhees, 1999] .

de généricité d'un concept, en pénalisant le poids des concepts les plus génériques (les concepts courts), et renforçant le poids des concepts spécifiques (concepts longs). Baziz prouve que cette pondération améliore la qualité de la recherche pour une représentation de l'espace de recherche par concepts. Pour un document i et un concept c, cette mesure est présentée dans l'équation 2.5,

$$CF.IDF(c,i) = cf(c) * \ln \frac{N}{df} \tag{2.5}$$

tels que :

- $cf(c) = count(c) + \sum_{sc \in SubConcepts(c)} \frac{Lenght(sc)}{Lenght(c)} * count(sc)$ est la fréquence de c ;
- N le nombre de documents de la collection ;
- df est la fréquence documentaire de c ;
- $count(c)$ mesure le nombre d'occurrences de c dans le document i (toutes les occurrences lexicales) ;
- $SubConcepts(c)$ est l'ensemble des sous-concepts de c.
- $Length$ calcule le nombre de termes que contient la dénotation lexicale de c.

[Baziz, 2005] a également essayé d'évaluer l'impact de sa mesure de pondération, en représentant les requêtes et les documents avec les termes et les concepts. Il pondère cette fois les concepts avec la méthode C_score et les termes ne possédant pas de concepts avec $TF.IDF$. C_score est une mesure de pondération qui tient compte du voisinage sémantique d'un concept dans un document. C'est la somme des similarités qu'un concept peut avoir avec les concepts du voisinage, tel que pour un sens k du concept c (noté par S_k^c), le score est donné dans l'équation 2.6,

$$C\_scores(S_k^c) = \sum_{l \neq c} (\phi_{c,l}) \tag{2.6}$$

avec :

- l est un concept différent du concept c ;
- ϕ est une mesure de similarité sémantique.

Le meilleur sens du concept est alors celui qui maximise $C\_scores(S_k^c)$ par rapport à tous les sens du concept c. Les résultats ont montré que cette méthode de pondération améliore plus les performances de la recherche par rapport au schéma de pondération $CF.IDF$. [Baziz, 2005] prouve ainsi qu'en plus de la spécification des entités représentatives des documents, les ressources sémantiques jouent un rôle important dans l'injection de la sémantique au niveau de la pondération des descripteurs choisis.

Le schéma de pondération $CF.IDF$, a été plus tard enrichi dans [Boughanem et al., 2010] avec les notions de "centralité" et de "spécificité". La centralité d'un concept étant définie par [Kang and Lee, 2005] comme étant le nombre de relations avec les autres concepts existants dans le document. Cette définition a été également adapté par [Blanco and Lioma, 2012] pour la pondération des termes dans un graphe.Alors que la spécificité d'un concept est définie comme son degré de spécialité, estimé en fonction de sa profondeur dans la hiérarchie *is-a* de *WordNet*. Ces paramètres ont montré leur efficacité dans des expériences préliminaires établies sur les collections de TREC par [Boughanem et al., 2010].

La notion de centralité lors de la pondération a également été traitée dans l'indexation sémantique des documents biomédicaux. On peut citer par exemple le travail effectué par [Koopman et al., 2012] à *TREC Medical Record Track 2012*, qui met en place une indexation

par les concepts de *SNOMED CT*. Cette méthode d'indexation est enrichie lors de la pondération sémantique par les poids des concepts n'apparaissant pas dans la requête, mais qui sont reliés aux concepts de la requête dans le document par des relations de subsomption. La mesure de pondération d'un concept c_j qui est subsumé, dans un document d, est la racine carrée de la mesure de pondération utilisée pour les concepts originaux (voir équation 2.7).

$$\varphi(w(cj, d)) = \sqrt{w(cj, d)} \tag{2.7}$$

La racine carré dans la formule permet d'éviter que des concepts n'apparaissant pas dans la requête aient la même importance que ceux identifiés dans la requête, et le même impact sur le score final du document. Cette stratégie de pondération minimise, entre autres, le bruit introduit par la prise en compte du voisinage d'un concept dans le document. Les résultats de ce travail montrent que l'indexation par concepts, pondérés par la méthode de pondération ne tenant compte que des liens de subsomption, s'est montrée meilleure à tous les systèmes de *TREC medical Track 2012*. Cependant une perte d'information lors de l'identification des concepts a été constatée dans certaines requêtes.

[Boubekeur and Azzoug, 2013] ont également mis en place une mesure de pondération des concepts, en reprenant également la notion de centralité, et en distinguant entre centralité locale et centralité globale d'un concept. La centralité locale d'un concept est une fonction de sa fréquence d'occurrence ainsi que de sa similarité avec les concepts voisins dans le document (non pas le nombre de relations avec le voisinage) et détermine ainsi l'importance d'un concept c_i dans un document d. Sa formule est la suivante (équation 2.8) :

$$cc(c_i, d) = \alpha * tf(c_i, d) + (1 - \alpha) * \sum_{i \neq l} (Sim(c_i, c_l)) \tag{2.8}$$

tel que $Sim(c_i, c_l)$ est la similarité sémantique entre les deux concepts c_i et c_l, et α est facteur de pondération $\in [0, 1]$ déterminé lors de l'expérimentation.
La centralité globale d'un concept quant à elle, est définie par son pouvoir de discrimination dans la collection : un concept c_i est d'autant plus discriminant que le nombre de documents de la collection là ou il est central est minime, et définie dans l'équation 2.9 suivante :

$$dc(c_i) = \frac{n}{N} \tag{2.9}$$

tel que N est le nombre de documents total de la collection et n est le nombre de documents de la collection là où c_i est central. Au final, le schéma de pondération proposé est le suivant (formule 2.10) :

$$Sem - CC.IDC = cc(c_i, d) * \frac{1}{dc(c_i)} = cc(c_i, d) * idc(c_i) \tag{2.10}$$

Les expérimentations ont été effectuées sur la collection de test *Time*[8], avec le thésaurus *WordNet*. Une évaluation du nouveau schéma de pondération de l'index sémantique ($Sem{-}CC.IDC$) donne de meilleurs résultats par rapport à la pondération classique de l'index sémantique (pondération des termes et concepts par $TF.IDF$ et $BM25$) et également par rapport à l'indexation classique par $TF.IDF$ et $BM25$, en prenant $\alpha = 0,2$.

8. ftp ://ftp.cs.cornell.edu/pub/smart/time/

Pondération des termes

[Zargayouna, 2005] a mis en place une mesure de pondération sémantique des termes, en tenant compte du voisinage sémantique d'un terme. Cette mesure a l'avantage de s'affranchir des problèmes de manque de couverture de la ressource. A travers sa méthode $SemIndex$, l'auteur prouve qu'un terme de la requête peut exister dans un document même si sa fréquence d'occurrence dans ce document est nulle. Sa formule est la suivante :

$$SemTF(t,d) = TF(t,d) + \sum_{t_i \in V} TF(t_i) * sim(\phi(t), \phi(t_i))$$

Tels que :
- $\phi(t_i)$ représente le concept associé à t_i.
- $sim(\phi(t), \phi(t_i))$ est la mesure de similarité entre les concepts $\phi(t)$ et $\phi(t_i)$.
- V représente l'ensemble des termes qui constituent le vocabulaire de la collection.

Cette mesure de pondération sémantique s'est montrée performante, cependant, l'inconvénient majeur est qu'elle augmente le bruit introduit par le voisinage sémantique. D'où l'auteur a eu recourt à la spécification d'un seuil de similarité entre concepts à partir duquel un concept $\phi(t_i)$ est pris en compte dans le voisinage sémantique de$\phi(t)$.

Les connaissances sémantiques exploitées pour l'enrichissement des schémas de pondération sont principalement, les concepts et les liens hiérarchiques pour le calcul des similarités sémantiques. Le résultat de la recherche dépend de la mesure de similarité sémantique exploitée et du schéma de pondération classique adapté. On remarque que uniquement les liens hiérarchiques sont exploités. On voit l'intérêt des notions tels que la centralité d'un concept dans son voisinage local et global, sa spécificité dans la ressource sémantique, etc., l'intérêt des similarités sémantiques dans la prise en compte du voisinage. Cependant certains auteurs, comme [Hamadan et al., 2012], ont montré que les schémas de pondérations classiques $BM25$ et $TF.IDF$ peuvent fonctionner mieux.Ils constatent par ailleurs qu'une meilleure combinaison entre les mesures sémantiques et distributionnelles doit être établie afin d'éviter de dégrader les résultats de la RI classique.

2.2.5 Exploitation des ressources sémantiques en recherche

L'utilisation des ressources sémantiques au niveau de la recherche se traduit par la reformulation ou l'expansion des requêtes. Partant de la constatation que plus la requête est courte, plus il sera difficile au système de trouver les documents pertinents pour celle-ci, l'idée de réaliser une expansion ou une reformulation des requêtes est donc apparue.

L'expansion des requêtes peut être réalisée en étendant le vocabulaire des requêtes au moyen de termes similaires (généralement des synonymes). En 1968, *Salton* constate déjà que l'utilisation du thésaurus *Harris Synonym* permet d'améliorer les performances, à condition que les termes utilisés pour l'enrichissement soient validés manuellement par un documentaliste. Il constate également que l'expansion automatique utilisant l'ensemble des termes possibles, dégrade ces performances [Salton, 1968b].

Dans le même contexte, et afin de confirmer l'apport de l'expansion des requêtes sur les performances du système, [Lu and Keefer, 1995] réalisent 3 expériences sur le corpus TREC3, pour constater que l'expansion des requêtes (via un thésaurus)améliore la précision de 33%, alors que la réduction des requêtes diminue la précision.

[Baziz et al., 2003] utilisent *WordNet* afin de récupérer les termes reliés à la requête par des relations de synonymie, de généralisation et de spécialisation. Ces expérimentations, réalisées sur

le corpus *CLEF2001*, sur un ensemble de 50 requêtes ont montré que le processus d'expansion de requêtes via ce thésaurus, à condition de respecter certaines conditions, améliore les performances d'un SRI. Ils montrent que l'expansion des requêtes dépend aussi bien du nombre de concepts, que des liens sémantiques à considérer. Ils concluent de ce fait, qu'un seul concept par requête améliore la précision de 55,97% contre une amélioration de 32,11% pour l'expansion par un concept par mot, et que la prise en compte de toutes les relations sémantiques (généralisation, spécification et synonymie) améliore la précision de 27,15%.

Plusieurs méthodes d'expansion de requêtes existent dans l'état de l'art. [Bhogal et al., 2007] passent en revue ces méthodes d'expansion de requêtes via les ontologies. Nous n'allons pas détailler ces méthodes car nous ne nous intéressons pas dans ce travail à l'amélioration de la qualité de la recherche via l'expansion des requêtes, mais plutôt via l'indexation.

2.3 Défis de la RIS

La recherche d'information sémantique, à travers la prise en compte des ressources sémantiques externes dans les modèles classiques de RI, s'est montrée performante pour la majorité des expériences ponctuelles citées de l'état de l'art. Cependant, aucun vrai bilan n'a pu être dressé, ces méthodes souffrent encore de certaines limites concernant la prise en compte de la sémantique, dans un processus de RI, mais également concernant leur évaluation. Les travaux de RIS, ne constituent en réalité que des adaptations des modèles classiques de RI. De ce fait, elles se trouvent toujours face à certains défis dont :

La représentation des documents : On est passé d'une représentation en "sac de mots pondérés" à une représentation en "sac de concepts pondérés" ou encore une combinaison entre les deux représentations. Des informations hétérogènes sont aplaties et mises dans un seul "sac", alors qu'elles sont de natures différentes (lexicales statistiques ou sémantiques conceptuelles, etc.). Ces informations sont néanmoins complémentaires : si l'une représente les informations textuelles avec la signification implicite des fréquences d'occurrence et des positions dans le texte et dans les documents, l'autre véhicule le sens et les connaissances rendues explicites ainsi que les relations potentielles entre ces connaissances.

Parmi les limites que peut induire une telle représentation de deux espaces différents, le fait de ne pas pouvoir justifier une amélioration ou dégradation de la qualité de recherche, d'autant plus qu'on appauvrit les connaissances en les traitant statistiquement. En effet ce qui importe pour les informations conceptuelles n'est pas leur fréquence d'occurrence dans le texte comme c'est le cas pour les termes, mais le sens qu'elles véhiculent et les relations qu'elles entretiennent avec les autres concepts.

Une meilleure intégration des schémas de pondération : Les premières tentatives d'intégration des mesures distributionnelles et des mesures de similarité sémantiques à travers la prise en compte du voisinage sémantique d'un terme ou concept donné ont été mises en place. Cependant, les nombreux schémas de pondérations proposés traitent en majorité de la centralité d'un concept ou terme donné dans le document ou dans la collection, à travers la prise en compte de son voisinage sémantique. Cette prise en compte du voisinage est calculée par rapport aux liens hiérarchiques qu'un concept donné a avec son voisinage sémantique. Les autres relations telles que les relations de domaines sont ignorées. La richesse des informations qui peuvent être présentes dans les ontologies est ignorée.

Plus de connaissances et de raisonnements exploités : Les ontologies formelles disponibles maintenant dans le Web Sémantique permettent de faire des inférences et

de déduire de nouvelles connaissances (nous reviendrons sur cette question au prochain chapitre). La réduction des calculs sur les ontologies à des calculs de similarité ampute la recherche d'information de connaissances riches. Des approches qui combinent raisonnement formel et recherche d'information commencent à émerger [Bhagdev et al., 2008],[Bikakis et al., 2010], [Fernandez et al., 2011].

Un modèle de RIS : Les modèles de RIS présentés dans l'état de l'art, sont des adaptations des modèles de RI classique à la RI sémantique [Zargayouna et al., 2015]. Il s'agit d'une intégration de la sémantique dans des modèles classiques, qui par conséquent préservent leurs natures (vecteurs de termes par exemple). La majorité des méthodes de RIS repose donc sur des adaptations des modèles de RI classique. On ne trouve pas de modèles dédiés à la RIS qui proposent une combinaison plus intelligente, des modèles documentaires et des modèles sémantiques, des analyses distributionnelles et des analyses sémantiques. Un modèle dédié de la RIS permettrait de tenir compte de la richesse des deux modèles sans sacrifier la puissance de leurs calculs.

Une évaluation sémantique des méthodes de RIS : Il est difficile de tracer un bilan des expériences menées en RIS. La construction de*benchmarks* dédiés à la RIS permettrait aux travaux de gagner en visibilité et rendrait possible des évaluations comparatives sur des fonctionnalités sémantiques ciblées.

[Zargayouna et al., 2015] donnent les raisons pour lesquelles une évaluation classique est inadéquate pour la RIS. Ils précisent que : *l'évaluation des systèmes de RIS de l'état de l'art 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 test 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.*

Notre objectif dans ce travail est de proposer une meilleur exploitation de la sémantique pour la recherche documentaire. L'avènement du Web Sémantique a encouragé la disponibilité des ontologies et facilité ainsi leur exploitation en RI. Cette exploitation comme nous venons de le voir n'est pas encore optimale et ne tient pas compte de la richesse des ressources sémantiques.

Nous allons par conséquent explorer l'accès à l'information dans le cadre du WS, qui est plutôt une recherche de données, afin de voir une autre manière de l'exploitation des ontologies, ainsi que les premiers travaux d'hybridation entre la recherche documentaire et la recherche de données. La majorité de ces modèles d'accès sur le WS sont des modèles en graphe qui se fondent sur des bases de connaissances contenant l'ontologie et considérant les documents comme des instances.

[Zargayouna et al., 2015] confirment en effet, que les tendances actuelles vont vers la convergence des travaux des deux communautés (RI et WS) et que les encouragements de Google pour le *Knowledge Graph* montrent que les points de convergence sont nombreux.

Chapitre 3

Graphes et accès à l'information

Sommaire

Le formalisme des graphes permet de coder un ensemble d'unités (les nœuds) reliées entre elles par des relations (les arêtes). Il est exploité depuis longtemps en représentation des connaissances et en traitement automatique des langues.

Il faut noter tout d'abord que l'information encodée sous la forme linguistique est de nature relationnelle. Les phrases et les structures syntaxiques sont classiquement représentées sous la forme de graphes de dépendances ou d'arbres syntagmatiques. La sémantique lexicale qui vise à coder le sens des mots et des expressions se représente aussi sous forme relationnelle. Au-delà de la phrase, on connaît également l'importance des relations dans la production linguistique, qu'il s'agisse d'interaction entre différents acteurs (à l'oral ou à l'écrit) ou d'intertextualité (relations entre documents) [Bhatia, 1998]. Plus récemment, avec les blogs et réseaux sociaux du web 2.0, sont apparues d'autres structures de réseaux textuels constitués des réponses que certains envoient à d'autres ou postent en ligne.

Les graphes sont également omniprésents dans les formalismes de représentation des connaissances. Historiquement, on a vu apparaître des réseaux sémantiques pour coder des connaissances de sens communs [Quillian, 1968] mais différents formalismes ont vu le jour comme les graphes conceptuels de John Sowa [Sowa, 1984], autour des logiques de description [Baader et al., 2003] et désormais avec le web de données.

La recherche d'information connaît aussi ce regain d'intérêt pour les approches à base de graphes. Nous en présentons ici quelques unes avant de présenter, au chapitre 4, la méthode que nous proposons et qui s'inscrit dans ce mouvement. Nous montrons ce faisant que les graphes manipulés peuvent être de natures différentes (graphes lexicaux, graphes de connaissances, graphes de citations entre documents) mais que les approches à base de graphes sont souvent les mêmes et que certaines d'entre elles permettent d'appréhender des graphes hétérogènes, comportant différents types de noeuds (termes, documents, concepts, etc.) et de relations (relations lexicales, ontologiques, d'occurrence, de citations, etc.).

3.1 Accès à l'information et graphes de connaissances

L'idée de base du web sémantique (WS) consiste à augmenter le Web de documents avec des méta-données sémantiques qui caractérisent le contenu des documents et sont interprétables par des agents logiciels.

Les méta-données sont représentées par des triplets sémantiques et peuvent reposer sur des ontologies. Un triplet associe un sujet, un prédicat et un objet et un ensemble de triplets qui forme un graphe de connaissances connectant des entités (sujets et objets) entre elles.

Les systèmes d'accès à l'information dans le WS se doivent donc de fournir les moyens nécessaires pour accéder efficacement à ces graphes de connaissances et répondre aux requêtes des utilisateurs. On passe ainsi des approches de RI traditionnelle, où on interroge une base documentaire à l'aide d'un index plat, à des approches sémantiques, où on interroge un graphe représentant une base de connaissances.

3.1.1 Vision et infrastructure du web sémantique

[Berners-Lee and Lassila, 2001] définit le WS comme "un web de données qui peut être directement utilisé par les machines". Le Web sémantique est en effet, une infrastructure qui permet d'exploiter des connaissances formalisées en plus du contenu informel du Web documentaire. Cette formalisation repose sur les ressources sémantiques formelles, plus exactement les ontologies.

Le W3C [9] présente une architecture en couches du Web sémantique (voir figure 3.1). Cette architecture traduit les efforts réalisés notamment en standardisation pour permettre une utilisation plus ou moins unique des outils. Elle est notamment caractérisée par l'ajout d'une couche de connaissances au dessus du Web, qui permet aux applications Web d'interpréter les données sous-jacentes comme l'utilisateur et qui assure l'interopérabilité et le dialogue entre ces applications. Cette architecture est fondée sur une pyramide de langages. Elle repose principalement sur la notion d'URI [10] qui permet de localiser une ressource sur le Web grâce à son identifiant unique. XML et XML-Schema représentent la couche syntaxique de base du web classique, à laquelle vient s'ajouter la couche de connaissances traduite par :

— les langages de représentation RDF [11] et RDF-Schema [12] pour définir le vocabulaire de ces connaissances ;

9. W3C : World Wide Web Consortium qui est un organisme de standardisation
10. URI : Uniform Resource Identifier
11. RDF (Ressource Description Framework) est un modèle de données dédié à la description des ressources et des relations entre ces ressources, en fournissant une sémantique simple. RDF peut être sérialisé en XML, ce qui facilite l'échange de données
12. RDF Schema est un langage de spécification des schémas associés à RDF

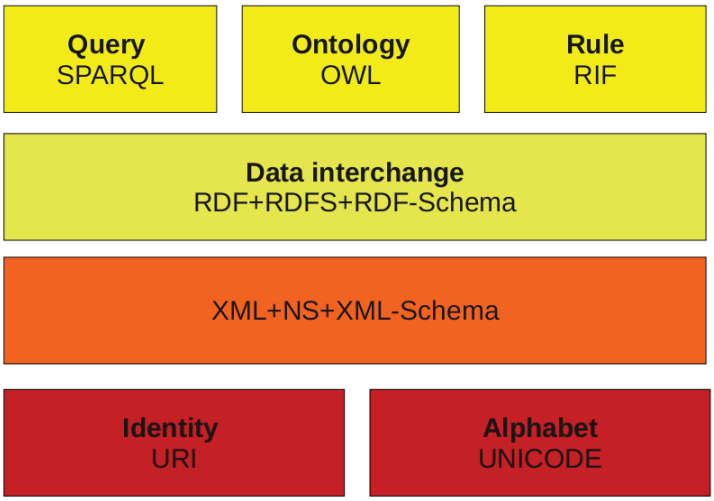

FIGURE 3.1 – Les couches du Web Sémantique selon le W3C

— les langages de représentation des ressources sémantiques (Ontologies en RDF-Schema ou OWL [13] qui permettent aussi de faire des raisonnements sur les connaissances du WS).
— les langages d'interrogation du Web sémantique (SPARQL [14], etc.)

La pyramide du W3C inclut aussi la possibilité de faire des déductions. On peut s'appuyer sur la standardisation d'un langage de règles, pour augmenter l'expressivité des ontologies OWL.

3.1.2 Accès à l'information dans le web sémantique

L'accès sémantique à l'information sur le WS est une recherche de données – et non pas de documents à proprement parler – qui repose sur (i) l'expression d'un besoin d'information sous la forme de requêtes formelles structurées et (ii) la projection de ces requêtes dans l'espace formel et cohérent des triplets sémantiques qui forment la base de connaissance (BC).

Un moteur d'accès sur le WS se compose donc d'un moteur d'inférence et d'un moteur d'interrogation comme le montre le schéma de la figure 3.2. Le moteur d'inférence permet de déduire de nouvelles connaissances à partir de la compilation de règles d'inférences et de la partie axiomatique de l'ontologie puis d'ajouter ces connaissances à la BC. Il permet donc de raisonner sur les données. Le moteur d'interrogation se charge, quant à lui, de réaliser la projection de la requête formelle de l'utilisateur sur cette BC.

Le fait de représenter l'espace de recherche (BC) sous la forme d'un graphe provient de la nature du WS (des données inter-connectées) mais ouvre aussi la voie à des requêtes structurées intégrant des relations entre les éléments, c'est-à-dire à des moteurs d'interrogation et de raisonnement capables d'interroger intelligemment la BC.

3.1.2.1 Représentation en graphe

On peut donc voir une BC dans le WS comme une ontologie peuplée par les annotations des ressources du Web, l'ensemble étant représenté le plus souvent sous forme d'un graphe RDF. Le

13. OWL (Web Ontologie Language) est un langage de représentation d'ontologies sur le Web. Il possède plus de primitives que RDF(S)
14. SPARQL : Protocol and RDF Query Language

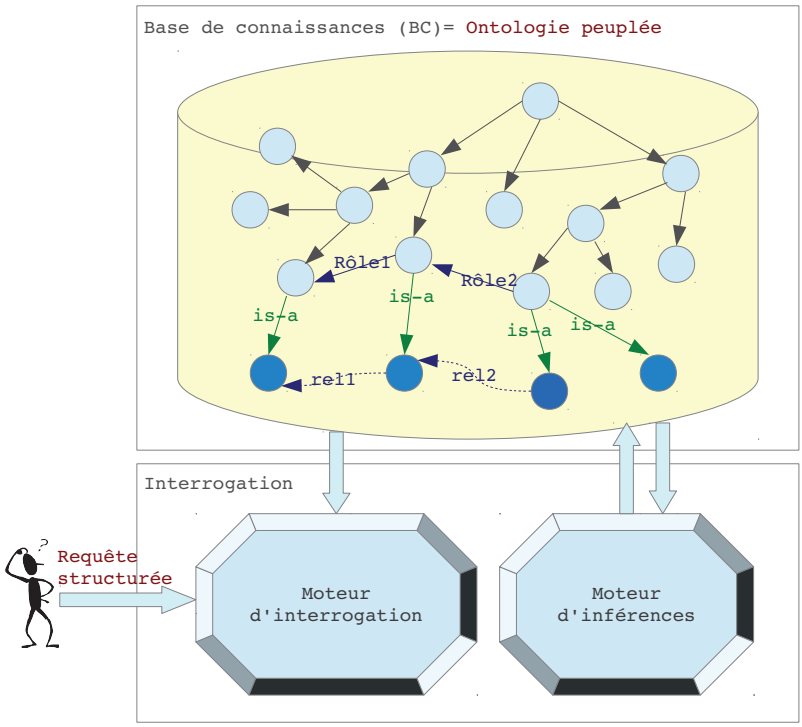

FIGURE 3.2 – Moteur d'accès sur le WS

peuplement de l'ontologie est souvent le résultat d'un processus préalable d'"annotation sémantique", sur lequel nous ne revenons pas ici [15].

L'ontologie, les annotations et les documents associés peuvent être exploités de différentes manières [Fernandez et al., 2011]. Certaines approches exploitent à la fois l'ontologie, les annotations et les documents sous-jacents (voir par exemple les systèmes TAP [Guha et al., 2003], KIM [Kiryakov et al., 2004], K-Serach [Bhagdev et al., 2008]). D'autres considèrent directement les documents comme des instances de l'ontologie, ce qui suppose un couplage fort entre l'ontologie et les documents, c'est à dire les documents présentent des instances de l'ontologie (voir par exemple les annotations du LOD [16], de DBPedia, PowerSet, FreeBase, RDFa, Microformats).

Ces BCs peuvent être stockées de différentes manières : dans des *Triple Store*, sous la forme de triplets RDF, comme pour Jena TDB, Sesame [17] ou Virtuoso, dans des bases de données classiques telle que la base *IBM DB2 V9* qu'utilise [Wang et al., 2011], ou encore sous la forme

15. Dans le domaine du Web sémantique, annoter un document revient à le décrire par des méta-données, en lui associant une description formelle. Dans ce cadre on parle d'annotation sémantique comme étant "Une représentation formelle d'un contenu, exprimée à l'aide des concepts, relations et instances décrites dans une ontologie, et reliées à la ressource documentaire source"[Amardeilh, 2008].

16. LOD : *Linked Open Data*

17. http ://www.openrdf.org/

de représentations internes comme les graphes conceptuels pour le moteur d'interrogation Corese [Corby et al., 2006]. Ces différentes représentations reprennent généralement les primitives de RDF(S).

3.1.2.2 Langages d'interrogation

Des langages d'interrogation formels ont été mis en place pour exploiter ces bases RDF. Ces langages ont la spécificité d'être très expressifs [Castells et al., 2007] et de pouvoir exprimer des requêtes relationnelles, prenant en compte les liens entre ces concepts et instances.

SPARQL [Prud'hommeaux, 2008] est le plus connu de ces langages. C'est un langage de requête destiné à interroger les bases de connaissances, standardisé par le W3C et implémenté en divers langages, notamment en Java avec le framework Jena. Il est utilisé par différents moteurs d'interrogation de données sur le WS, tels que Corese et Sesame. Les résultats des interrogations SPARQL sont des ensembles d'entités sémantiques, de triplets RDF ou de graphes RDF.

Ces langages supposent de connaître le schéma sous-jacent des bases de connaissances et nécessitent un effort d'apprentissage de la part de l'utilisateur pour bien maîtriser leur syntaxe et profiter de leur expressivité. De plus, les utilisateurs ont parfois des besoins d'information trop vagues plus faciles à exprimer par des mot-clés que par des requêtes formelles.

Un système d'accès à l'information efficace doit donc trouver un compromis entre les contraintes d'usage et l'expressivité du langage d'interrogation proposé et retourner des documents pertinents. [Lei et al., 2006, Bhagdev et al., 2008, Wang et al., 2011] proposent des langages d'interrogation hybrides permettant à tout utilisateur d'exprimer son besoin d'information sous la forme soit de mots-clés et de triplets du WS, soit d'une formule en langage naturel qui est ensuite traduite en SPARQL (voir le système SemSearch [Lei et al., 2006]).

3.1.2.3 Moteurs d'interrogation et d'inférence pour le WS

Des moteurs d'interrogation des graphes RDF ont été mis en place sur le WS. Ces moteurs permettent de répondre aux requêtes structurées, mais intègrent également des raisonneurs sur ces graphes. Parmi ces moteurs on peut citer Corese [Corby et al., 2006], les SPARQL endpoints tels que Virtuoso, Sesame, etc. ou encore les systèmes reposant sur les primitives des logiques de description et OWL-DL, comme le système proposé par [d'Amato et al., 2012].

La majorité de ces moteurs d'interrogation intègrent des raisonneurs tels que *Pellet* [Sirin et al., 2007], *RacerPro*, *Fact++*, *CEL* ou *OWLIM*, qui est utilisé avec Jena. Ces moteurs exigent que la BC soit formalisée en OWL-DL, ce qui permet de modéliser des ontologies plus complexes que celles autorisées par OWL Lite [18].

3.1.3 Recherche documentaire et web sémantique

Comme nous mettons l'accent sur la recherche documentaire plutôt que sur la recherche de données, nous nous intéressons aux approches qui reposent à la fois sur une base documentaire et sur une base de connaissances. Nous nous intéressons moins aux systèmes qui proposent de combiner recherche de documents et recherche de données comme *GoNTogle* [Bikakis et al., 2010] ou *K-Search* [Bhagdev et al., 2008], qu'à ceux qui proposent une représentation unique pour les données, les connaissances et les documents, et une approche globale de la recherche d'information sémantique.

18. OWL DL permet d'obtenir un maximum d'expressivité tout en gardant la complétude de calcul (toutes les conclusions sont garanties calculables) et de décision (tous les calculs finiront dans un temps fini).

Le système présenté par [Wang et al., 2011] est un système de recherche hybride qui permet d'interroger aussi bien des données structurées que des données textuelles et sert pour la recherche de connaissances comme pour la recherche de documents. Ce système CE^2 propose une interrogation hybride qui permet d'exploiter les deux types de données de manière homogène et intégrée, qui est efficace et qui passe l'échelle. Il s'agit d'un modèle de recherche en graphe, qui permet de représenter les entités sémantiques (concepts et instances) et les documents ainsi que leurs relations et attributs, comme le montre la figure 3.3.

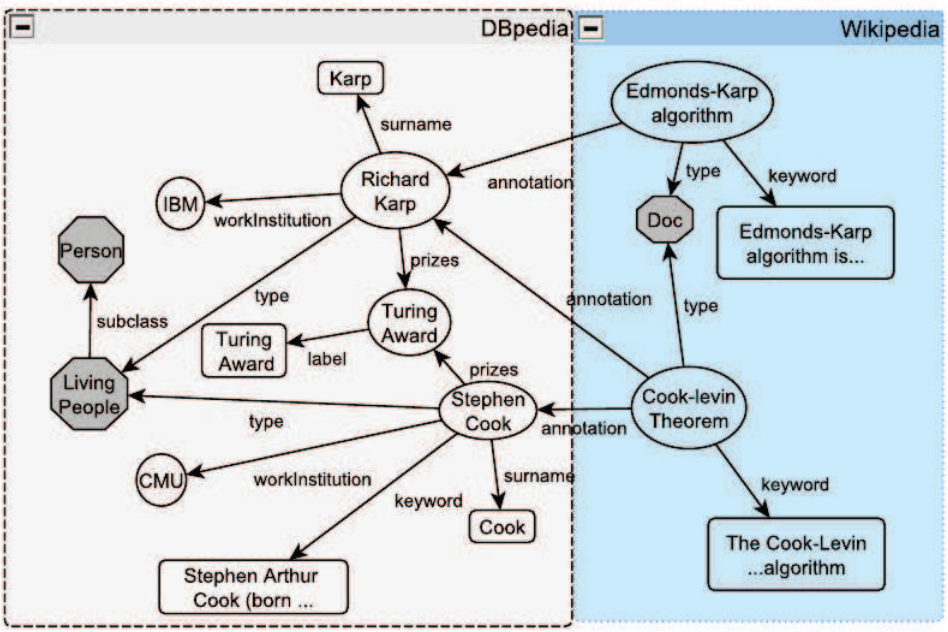

FIGURE 3.3 – Exemple de ressource [Wang et al., 2011]

Une requête est représentée sous la forme d'un graphe du même type. La figure 3.4 montre un exemple de graphe requête, qui permet retrouver un certain type de documents. La recherche

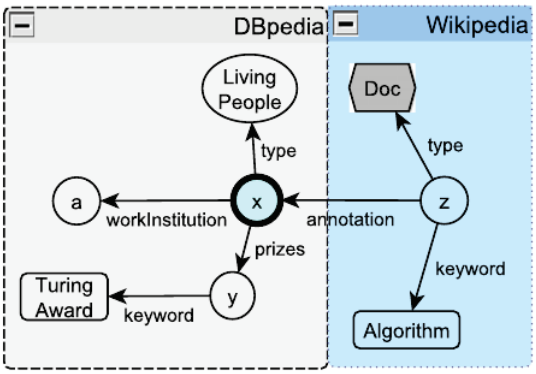

FIGURE 3.4 – Exemple de requête [Wang et al., 2011]

repose sur la projection du graphe requête décomposé sur le graphe de la BC.

Même si ce système permet de retrouver des documents, il ne s'agit pas de la recherche d'information (RI) à proprement parler. On note en effet que les calculs distributionels ne sont

exploités que partiellement pour le classement des documents (*ranking*) ; ils n'interviennent pas pendant la recherche. Or les analyses distributionnelles sont utilisées depuis longtemps en RI classique pour expliciter des informations implicites dans les documents, sur la base des fréquences d'occurrences et de scores de co-occurrence. On est donc ici plus près de la recherche de données que de la recherche d'information à proprement parler. Par ailleurs, les documents n'ont pas le même statut qu'en RI : le document est encodé comme une instance de concept dans le graphe des connaissances. Dans les approches de ce type, les documents retrouvés viennent en appui à la réponse factuelle apportée et ne constituent pas le coeur de la réponse du système.

Ces modèles d'accès sur le WS sont ainsi des modèles de recherche de données qui ont été adaptés pour la recherche de documents, mais dans lesquels le document est une information accessoire marquant la provenance des connaissances. Les graphes de connaissances RDF apparaissent difficilement compatibles à la RI et nous n'avons pas connaissance de systèmes d'accès fondés sur les graphes RDF, faisant de la recherche documentaire à proprement parler.

[Mimouni et al., 2014] ont également proposé d'étendre un modèle de recherche dans les graphes à la recherche documentaire en codant des collections documentaires comme des graphes de textes : les textes peuvent être typés et assortis d'annotations sémantiques mais aussi reliés entre eux par des relations d'intertextualité ou relations sémantiques entre documents (par ex. les liens de transposition, modification ou visa qui sont fréquents dans les textes juridiques). Les auteurs montrent que cela permet d'interroger les collections documentaires avec des requêtes elles-mêmes structurées sous la forme de graphes. On peut ainsi interroger la base documentaire en demandant quelles sont les lois et ordonnances traitant d'un sujet s_1 et transposant une directive européenne prise par le Conseil de l'Union européenne avec le Parlement et concernant un autre sujet s_2. [Mimouni, 2015] montre ainsi l'intérêt des approches à base de graphes pour la RI.

Cette approche relève clairement de la RIS : elle est adaptée au domaine juridique où les approches logiques de RI sont privilégiées mais elle est également difficilement compatible avec le modèle vectoriel traditionnel.

3.2 Ordonnancement de documents à base de graphes

La RI a cependant intégré la notion de graphe bien avant l'avènement du web sémantique. Il s'agissait de prendre en compte la structure de graphe du web documentaire, indépendamment des couches sémantiques pouvant s'y ajouter. Le web est un vaste réseau hypertexte, c'est-à-dire un graphe de documents reliés par des liens de citation encodés sous la forme de liens html orientés. Dans les 1990, l'idée est apparue que la structure de ce vaste réseau de citation pouvait être utilisée pour estimer la notoriété des pages web et améliorer leur classement dans les résultats des moteurs de recherche. Diverses approches ont été proposées même si c'est l'algorithme PageRank qui est le plus connu. Elles reposent sur le même modèle des marches aléatoires.

3.2.1 Modèle des marches aléatoires

Parmi les modèles mathématiques à base de graphes, le modèle de marches aléatoires (ou "marche au hazard", "promenade aléatoire", *random walk* en anglais) est l'un des plus connus[19]. Ce modèle permet de simuler la manière dont on peut parcourir un graphe de manière aléatoire. Il provient du domaine de la théorie des probabilités.

19. La référence principale des marches aléatoires est le livre classique de [Spritzer, 2000], paru initialement en 1964.

Le principe est simple. Il s'agit d'étudier l'évolution "au hasard" d'une variable, dont le mouvement se décompose en une succession de "pas". A chaque pas, on a une certaine probabilité de prendre "au hasard" une direction et une longueur du pas. Ces pas ne dépendent pas les uns des autres : on considère que le système évolue "sans mémoire" et que l'état du présent ne dépend pas du passé. Si on s'imagine sur un graphe, il s'agit de mesurer la probabilité qu'on aurait de se retrouver sur un nœud particulier d'un graphe si on le parcourait de manière aléatoire. L'algorithme vise à trouver un équilibre : on montre que pour des marches suffisamment longues, la distribution de la position finale du marcheur se stabilise, elle converge vers une distribution gaussienne et c'est cette distribution de probabilité qui est utilisée pour ordonner les nœuds du graphe.

3.2.2 Algorithme PageRank

En informatique, le modèle des marches aléatoires est souvent appliqué aux réseaux ou graphes très complexes. En recherche d'information, il est à l'origine de l'algorithme utilisé au départ par le moteur *Google* pour classer les pages retournées en réponse à une requête utilisateur. Dans ce cadre, la probabilité qu'une page web soit visitée représente la notoriété de la page dans le réseau hypertexte du web, ce qui permet de la classer au sein de toutes les pages web disponibles.

En 1997, Larry Page et Sergey Brin proposent un nouvel algorithme appelé *PageRank* qui est rapidement adopté par le nouveau moteur de recherche Google.

PageRank propose de classer toutes les pages du web et donc également celles qui sont retournées par un moteur de recherche en réponse à une requête utilisateur en fonction de leur notoriété relative. Ce classement est d'autant plus important que les utilisateurs ne consultent que les premiers liens proposés par les moteurs. PageRank propose de placer les pages par ordre de notoriété décroissante. Cette notoriété est dépendante des liens de citations entre les pages, mais également de la structure du graphe : l'idée sous-jacente est qu'une page est d'autant plus "célèbre" qu'elle est davantage citée par des pages qui jouissent elles-mêmes d'une bonne notoriété. A chaque page Web est donc associé un score appelé "PageRank".

Le calcul du PageRank repose sur l'intuition que la probabilité d'une page A d'être visitée par un utilisateur qui parcourrait aléatoirement le graphe du web est d'autant plus grande qu'il y a plus de pages qui pointent vers A, que ces pages ont elle-mêmes une probabilité élevée d'être visitées et qu'elles pointent vers peu de pages en dehors de A. Le calcul du PageRank d'une page est défini comme suit par [Brin and Page, 1998] :

$$PR(A) = (1-d) + d * \sum_{i=1}^{n} PR(p_i)/C(p_i) \tag{3.1}$$

tel que :

- $PR(A)$ est le PageRank de la page A,
- n est le nombre de pages pointant vers la page A,
- $PR(p_i)$ est le PageRank d'une page p_i qui a un lien vers la page A,
- $C(p_i)$ est le nombre de liens sortant de la page p_i,
- d est un facteur d'amortissement qui est compris entre 0 et 1, et que [Brin and Page, 1998] propose de fixer à 0,85.

On montre que l'on peut calculer les scores de tous les nœuds d'un graphe hypertexte par itérations successives à partir d'une distribution uniforme de PageRank sur l'ensemble des nœuds.

Si on met à jour les scores de tous les nœuds à chaque itération, on observe que la distribution de PageRank se stabilise au bout d'un certain nombre d'itérations [20].

L'algorithme PageRank a été adapté à de nombreux problèmes de fouille de textes (voir par exemple les algorithmes TextRank [Mihalcea and Tarau, 2004] ou TopicRank [Bougouin et al., 2013]), le principe sous-jacent étant toujours celui des marches aléatoires.

L'une des limites de PageRank pour le classement des pages en RI vient de ce que la notoriété d'une page est une propriété qui ne dépend que de la structure du réseau hypertexte et pas de la requête de l'utilisateur. Autrement dit PageRank classe toujours les mêmes pages dans le même ordre pour toutes les requêtes.

3.2.3 Algorithme HITS

L'algorithme HITS proposé par [Kleinberg, 1999] repose sur la même idée : il s'agit aussi de prendre en compte la notoriété pour classer les documents retournés par un moteur de recherche mais la notoriété d'une page n'est plus une propriété intrinsèque de la page calculée une fois pour toute en fonction d'une structure donnée du Web ; elle dépend aussi de la requête de l'utilisateur.

En pratique, le calcul de notoriété est restreint au sous-graphe du web qui est lié à la requête de l'utilisateur (l'ensemble des pages pré-sélectionnées comme pertinentes pour la requête et les pages situées dans leur voisinage). Cette approche soulève cependant des problèmes de calcul parce qu'à la différence de PageRank qui calcule une fois pour toute la notoriété de toutes les pages du web, HITS doit refaire un calcul de notoriété à la volée et pour chaque requête, ce qui allonge le temps de réponse des moteurs même si le sous-graphe pris en compte pour le calcul de notoriété est très petit au regard de la taille du web (de l'ordre de quelques centaines de pages).

3.3 Propagation d'activation sur les graphes

Les approches à base de marches aléatoires reprennent une idée ancienne issue de la psychologie cognitive et l'intelligence artificielle [Quillian, 1968], selon laquelle les unités lexicales ne sont pas isolées mais prises dans un réseau de relations et que l'activation en mémoire d'une unité active également par association les unités voisines. La propagation d'activation (PA) a été introduite par [Collins and Loftus, 1975] afin de modéliser des phénomènes d'influence qui se propagent dans des graphes. Cette approche a été reprise en RI : de nombreux travaux ont essayé de coder l'information de pertinence sur les graphes, par des mécanismes d'influence, et pas uniquement pour classer les nœuds du graphe comme le font les approches à base de marches aléatoires. On parle dans ce cas d'algorithmes de propagation d'activation.

3.3.1 Notions de base

La propagation d'activation (en anglais *Spreading Activation*) est un algorithme de parcours de graphe qui a été déployé pour (i) les réseaux associatifs et (ii) les réseaux sémantiques. Dans le premier cas, les relations (les arcs du graphe) possèdent uniquement des poids numériques, alors que dans le second, ils possèdent des labels représentant la sémantique de la relation [Rocha et al., 2004].

L'algorithme modélise un processus consistant à activer de proche en proche des nœuds d'un réseau quelconque : réseaux d'inférence, de neurones, réseau d'association, des réseaux sémantiques, etc. Les nœuds représentent des objets du monde réel ou des entités abstraites ; les

20. Naturellement, des méthodes d'approximation sont utilisées pour faire ce calcul pour un graphe de la taille du web, mais seul le principe de base nous intéresse ici.

arcs qui les relient représentent des relations entre ces objets ou entités ; ils peuvent être orientés ou non.

On attribue à certains nœuds des valeurs initiales (activations initiales) et l'algorithme modélise la propagation de cette activation de proche en proche dans le réseau. A la fin de cette propagation, on observe quels sont les noeuds qui ont été atteints par cette "onde" de propagation, ce qui permet de sélectionner les objets ou entités les plus pertinents au regard de l'activation initiale ou les plus "proches" des nœuds sélectionnés au départ.

L'algorithme consiste en une séquence d'itérations. Chaque itération se compose classiquement d'une ou plusieurs impulsions et d'une vérification de terminaison. Chaque impulsion se compose elle aussi de 3 étapes : le pré-réglage, la propagation ("Spreading") et le post-réglage. La première et la troisième étapes contrôlent la rétention ou la diffusion de l'activation issue des impulsions précédentes.

[Preece, 1981] montre qu'il y a plusieurs manières de gérer l'impulsion. Dans sa forme la plus simple, la formule de propagation d'activation telle qu'elle est présentée dans [Crestani, 1997] est la suivante : soit un nœud i relié à des nœuds j avec des arcs allant des nœuds j au nœud i ; l'input I_i du nœud i correspond à la valeur d'activation qu'il "reçoit" de ses prédécesseurs dans le graphe et se calcule comme suit :

$$I_i = \sum_{j=1}^{n} O_j * w_{ji} \tag{3.2}$$

où w_{ji} est le poids de l'arc allant du nœud j au nœud i, n est le nombre de prédécesseurs de i dans le graphe et O_j est l'output du nœud j qui est généralement lui-même fonction de son input :

$$O_j = f(I_j) \tag{3.3}$$

Cette fonction f diffère d'une application à l'autre (fonction linéaire, fonction à pas, fonction sigmoïde, etc.) mais la fonction la plus utilisée pour la propagation d'activation est la fonction de seuil avec un seuil potentiellement différent d'un nœud à l'autre. Dans le cas des valeurs d'activations binaires (0 ou 1), avec un seuil k_j pour le nœud j, cette fonction est la suivante :

$$O_j = \left\{ \begin{array}{ll} 0 & I_j < k_j \\ 1 & I_j > k_j \end{array} \right\} \tag{3.4}$$

Pour la RI par exemple, la fonction identité est la plus utilisée : la somme des inputs est égale à la somme des outputs.

Ainsi, impulsion après impulsion, l'activation se propage dans le réseau jusqu'au point de terminaison. Le résultat de la propagation est la distribution d'activations des nœuds obtenue à la terminaison.

L'algorithme de propagation d'activation dans son modèle initial présente certaines limites que cite [Crestani, 1997], dont principalement :

- le contrôle de la propagation est difficile : à moins de le faire manuellement (par *"user feedback"* pendant la première et troisième étape), il n'est pas évident de contrôler cette activation et éviter la propagation dans tout le réseau ;
- l'exploitation de la sémantique reste limitée : les informations données par les labels associés aux arcs du graphe ne sont pas exploitées, par exemple.

En réponse à ces limitations, une approche de propagation d'activation par contraintes (*Constrained Spreading Activation*) a été proposée. Elle introduit un ensemble de contraintes (ou de règles de terminaison) qui garantissent une terminaison automatique de l'algorithme ;

cela évite que l'activation ne se propage dans tout le réseau et cela permet de prendre en compte la signification des liens quand elle existe. Une bonne application de la propagation d'activation à contraintes pour la RI est rapportée dans [Cohen and Kjeldsen, 1987], avec le système GRANT.

[Crestani, 1997] cite les contraintes les plus utilisées pour la propagation d'activation (elles peuvent être contrôlées au niveau du pré-réglage et post-réglage) :

Contrainte de distance : la propagation d'activation doit s'arrêter quand elle atteint des nœuds qui sont suffisamment loin dans le graphe ; Crestani précise que la force de la relation diminue avec la distance ;

Contrainte *Fan-out* ou de connectivité : la propagation doit s'arrêter au niveau des nœuds fortement connectés aux autres nœuds, car elle possède alors un sens très vague (*broad semantic meaning*) ;

Contrainte de chemin : l'activation doit se propager en utilisant un chemin préférentiel, qui peut être identifié à l'aide de poids sur arcs ou de labels si les arcs possèdent des étiquettes ; le but est d'arrêter la propagation au niveau des chemins non intéressants (*stopping from following less meaningful paths*) ;

Contrainte de seuil : en utilisant une fonction de seuil sur chacun des nœuds du graphe, il est possible de contrôler la propagation dans le réseau.

Dans la suite, nous nous intéressons à la version avec contraintes de l'algorithme. En effet, [Preece, 1981] indique que de meilleurs résultats de RI peuvent être obtenus par une propagation d'activation qui exploite l'une des contraintes mentionnées. Ces contraintes bien qu'elles améliorent la RI, rendent le système dépendant du domaine, notamment quand la sémantique des relations est prise en considération dans les contraintes qui privilégient des chemins préférés pour la propagation.

3.3.2 Propagation d'activation pour la RI

L'exploitation de la propagation sur les réseaux sémantiques pour la RI a suscité de l'intérêt depuis longtemps. Cela a été manifeste dès les années 1980 avec les travaux de [Preece, 1981], [Shoval, 1981], [Cohen and Kjeldsen, 1987], [Croft et al., 1989], etc. L'objectif était d'exploiter les associations entre termes et documents pour la RI. Cette piste a été abandonnée par la suite mais ces premiers travaux montrent que la propagation d'activation sur des réseaux sémantiques donne des résultats significatifs pour la RI et justifient que nous nous y revenions quelques décennies plus tard. Nous montrerons dans la suite que les conditions sont aujourd'hui réunies pour déployer cette approche à une autre échelle.

La modélisation des connaissances sous la forme de graphe et le choix des contraintes à mettre en œuvre diffèrent d'un système à l'autre car les choix de modélisation dépendent des objectifs visés : recherche de documents (IR), expansion de requête, recherche par retour sur pertinence, classification, etc.

3.3.2.1 Modélisation

Quand on modélise un problème de recherche d'information sous la forme de graphe, les nœuds présentent classiquement les documents de la collection documentaire, les termes qui y sont contenus, les auteurs, etc. et les arcs entre ces nœuds peuvent représenter les relations sémantiques existantes entre les termes, les liens de citations entre documents, les liens entre termes et documents, les relations d'auteurs, etc. Un exemple de modélisation présentée dans [Crestani, 1997] est donné dans la figure 3.5 : le graphe comporte différents types de noeuds

(documents, termes, auteurs, etc.) et différents types de relations, certaines étant issues d'une ressource terminologique de type thesaurus.

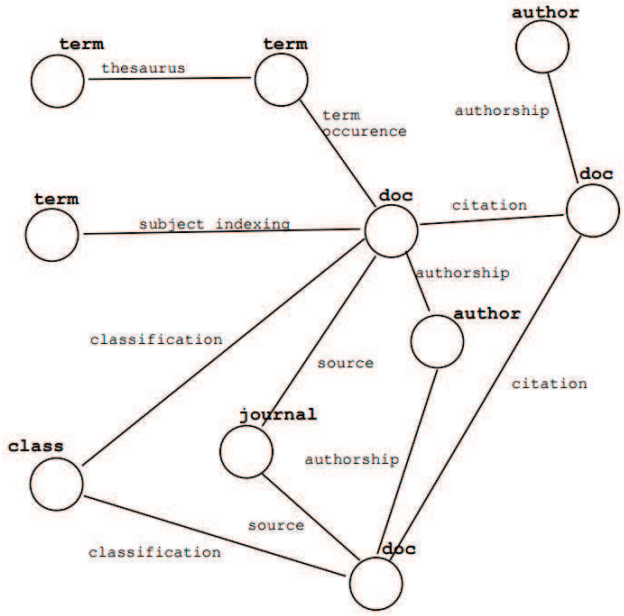

FIGURE 3.5 – Exemple de modélisation sous la forme de graphe pour la propagation d'activation [Crestani, 1997]

Preece présente l'un des premiers travaux réalisant la RI classique par PA sur un graphe de collection documentaire [Preece, 1981]. Il prouve qu'on peut reproduire de la sorte les modèles classiques de la RI, tels que le modèle booléen, ou le modèle vectoriel. Il montre notamment qu'à travers le retour sur pertinence, la PA peut être utilisée pour la classification, l'indexation ou encore la construction de classes sémantiques.

Cohen et Kjeldsen présentent en 1987 le premier système exploitant la propagation d'activation à contraintes pour la RI [Cohen and Kjeldsen, 1987]. Le système GRANT repose sur un graphe simple, composé des documents de la collections et des termes qui y sont associés, mais il obtient des valeurs de rappel et précision intéressantes, meilleures pour l'application visée que les valeurs obtenues par le modèle vectoriel classique.

Shoval a développé une expansion de requête interactive utilisant la PA, sur un réseau sémantique de termes qui a l'avantage d'être construit automatiquement à partir d'un thesaurus [Shoval, 1981]. Il s'agit d'un réseau de termes et les relations modélisées sont celles d'un thésaurus (relations hiérarchiques, de synonymies, relations sémantiques, etc.). Les documents ne sont pas modélisés, seul le vocabulaire de la collection l'est, car l'objectif est de faire de l'expansion de requêtes. Ce travail est considéré comme l'un des travaux les plus performants de l'exploitation de la PA en RI : la construction du réseau de représentation des données est automatique et l'approche n'est pas dépendante du domaine car la ressource terminologique utilisée est générique.

Pour mettre en place un cadre de test et tester les "inférences plausibles" pour la RI, [Croft et al., 1989] utilise le système GRANT sur un réseau sémantique qui comporte des nœuds termes du vocabulaire, des nœuds documents et des nœuds concepts. Aucune distinction entre ces types de nœuds n'est faite mais plusieurs types de contraintes sont déployés. Les points d'entrée de la PA sont les nœuds documents les mieux classés d'un système de RI probabiliste classique. Il s'agit

d'une recherche par l'exemple. La figure 3.6 présente la modélisation du système I^3R présenté par Croft *et al.*.

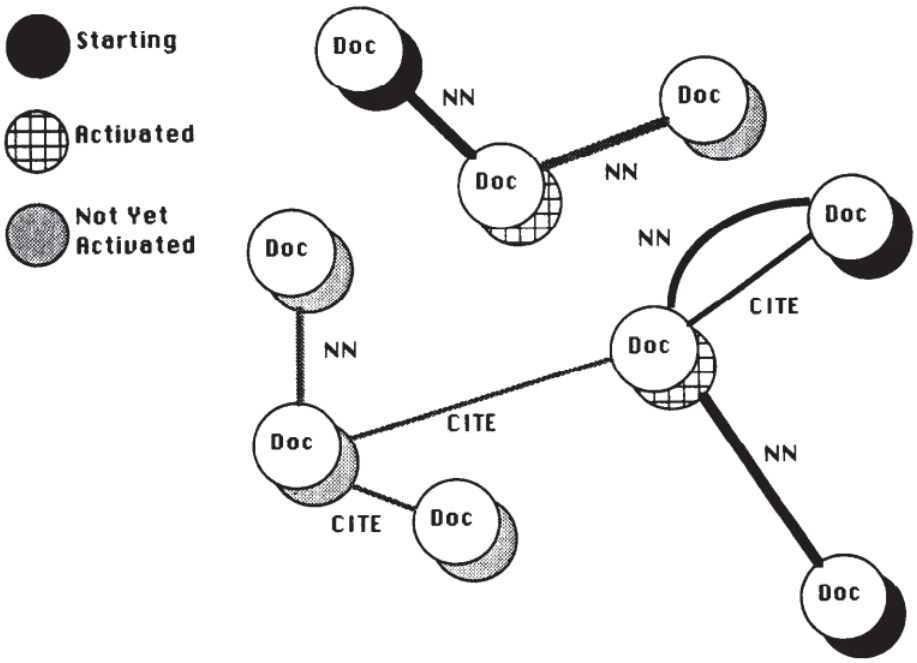

FIGURE 3.6 – I^3R : Représentation du réseau sémantique pour la propagation d'activation [Croft et al., 1989]

[Salton and Buckley, 1988a] comparent la propagation d'activation pour la RI et le modèle vectoriel. Les auteurs utilisent le modèle de PA dans sa forme de base sans contraintes. Le réseau est composé de termes, de documents et des relations entre termes et documents. La figure 3.7 est un exemple du réseau exploité par Salton et Buckley. Les valeurs des arcs terme/document expriment la fréquence d'occurrence (différentes normalisations sont testées). Les auteurs concluent que ce système ressemble beaucoup au modèle vectoriel classique. Leurs évaluations sont en faveur du modèle vectoriel mais ils précisent que des améliorations sont à espérer pour la PA avec l'utilisation de contraintes. La figure 3.8 présente un exemple de propagation d'activation à partir d'une requête du modèle mis en place par Salton et Buckley.

En 1992, Savoy procède par propagation d'activation sur un réseau de documents interconnectés par des liens hypertextes [Savoy, 1992] (voir figure 3.9). Il construit au préalable un arbre bayésien de termes et de documents, afin de choisir les termes correspondant au besoin de l'utilisateur et il leur affecte des poids significatifs. Ces poids servent au calcul des valeurs d'activation des documents. A travers une double propagation sur l'espace d'indexation et l'espace des documents, l'auteur arrive à de bons résultats, qui ne sont cependant pas comparables au modèle de RI classique.

3.3.2.2 Contraintes exploitées pour la PA

Une fois la représentation du réseau sémantique qui abrite la propagation d'activation construite et les nœuds de départ activés, l'activation se propage de proche en proche. Quand il s'agit de PA avec contraintes, la propagation se fait selon certaines heuristiques et règles et en respectant

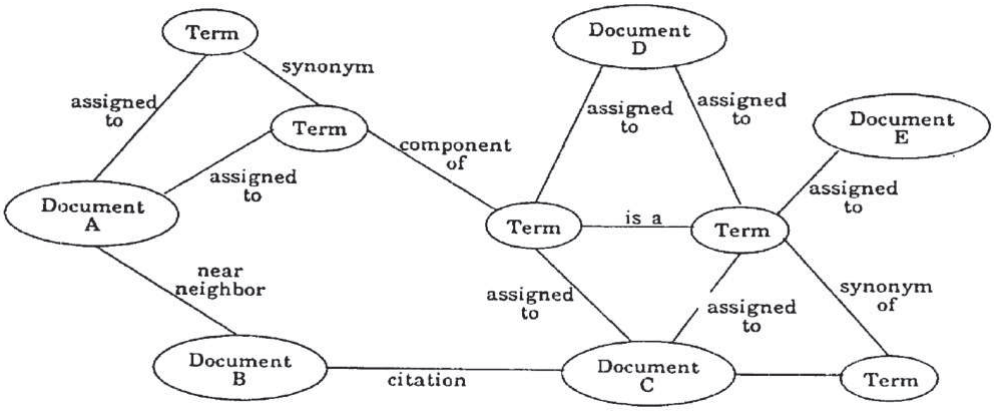

FIGURE 3.7 – Réseau sémantique exploité par [Salton and Buckley, 1988a]

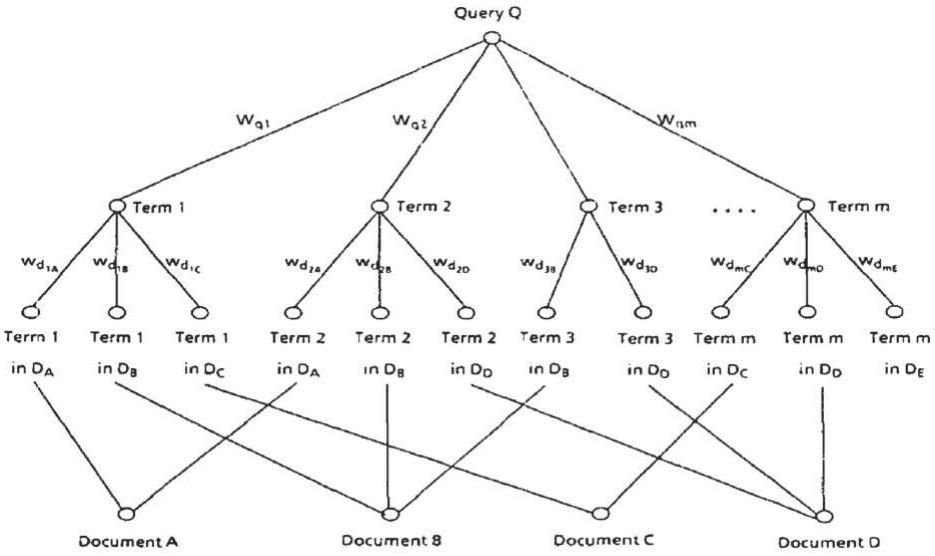

FIGURE 3.8 – Exemple de propagation d'activation [Salton and Buckley, 1988a]

certaines des contraintes de terminaison introduites dans la section 3.3.1. Les poids affectés aux nœuds atteints à chaque itération sont calculés par une fonction d'activation et dépendent, entre autres, des valeurs affectées aux arcs entrant de chaque nœud.

Il faut noter que, pour la majorité des travaux en RI par propagation d'activation, aucune distinction entre les types des nœuds sur le réseau n'est faite [Crestani, 1997]. En revanche, certaines heuristiques exploitent la sémantique des arcs.

Parmi les heuristiques utilisées en PA pour la RI, [Shoval, 1981] propose de "marquer" les termes jugés non pertinents pour la requête lors de son expansion, de manière à bloquer la propagation sur ces noeuds et à éviter leur extraction lors des itérations suivantes.

GRANT [Cohen and Kjeldsen, 1987], qui est considéré comme le premier système de PA avec contraintes, exploite les contraintes classiques, notamment la contrainte de chemin, et procède par renforcement des poids des relations, en attribuant des valeurs positives aux relations préférées et

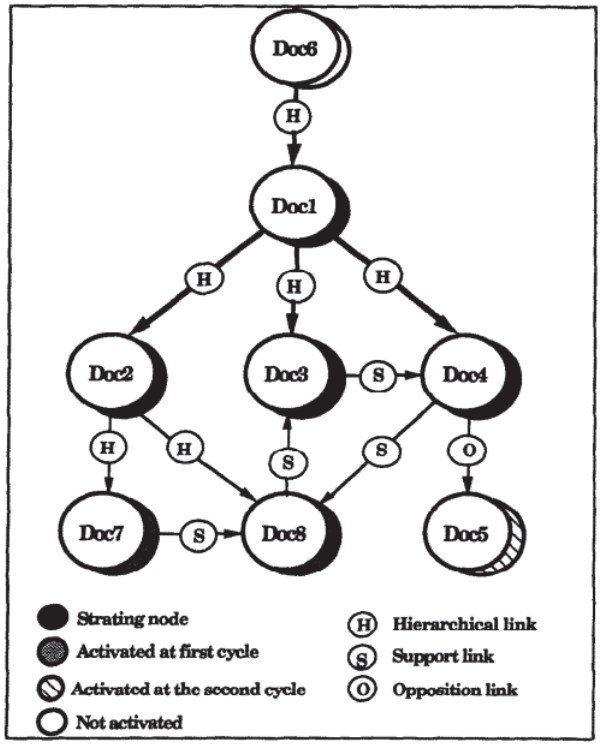

FIGURE 3.9 – Exemple de propagation d'activation sur le réseau hypertexte de [Savoy, 1992]

des valeurs négatives aux chemins non pertinents. Cette contrainte peut être considérée comme une règle d'inférence de la forme : $IF\ x\ AND\ R(x,y)\ \rightarrow\ y$, où $R(x,y)$ est un chemin entre deux nœuds x et y qui peut être représenté par une ou plusieurs relations.

De même le système I^3R proposé par [Croft et al., 1989] a été mis en place afin d'expérimenter quelques règles d'inférences en RI. Il intègre les contraintes suivantes dans la propagation d'activation :

- initialement, seules les relations des plus proches voisins et de citations reliant les documents de départ et d'autres documents sont considérées comme intéressantes ;
- dans la suite de la propagation, seules les relations des plus proches voisins sont exploitées ;
- les poids sur les relations sont utilisés pour le calcul de la valeur d'activation des nœuds ;
- les documents ayant déjà été activés ne sont plus utilisés s'ils sont atteints à nouveau et que leurs valeurs d'activations sont mises à jour.

Savoy exploite également un ensemble de règles pour gérer la propagation sur le réseau des documents [Savoy, 1992] :

- les documents sont typés (*research papers*, *survey*, *articles*, etc.) [21] et cette catégorisation sert à focaliser l'activation sur le type de document qui intéresse le plus l'utilisateur ;
- à la première itération, la propagation peut se faire dans toutes les directions mais, par la suite, l'activation ne parcourt les liens hiérarchiques que dans le sens de la spécialisation (de la catégorie la plus générique à la catégorie la plus spécifique) ;
- le nombre d'itérations d'activation est limité à 5 mais on observe que les meilleurs résultats sont par contre obtenus à la deuxième itération ;
- le système n'autorise pas la réactivation d'un document, si ce document et celui qui l'a

21. Une variable est attachée au document pour spécifier son type.

activé possèdent un père commun.

Ces contraintes sont intéressantes à explorer car elles garantissent une terminaison automatique de la propagation et évitent d'étendre la propagation d'activation à tout le graphe, mais elles sont souvent fortement dépendantes du domaine. Notre objectif étant d'avoir, au moins dans un premier temps, un modèle de RIS le plus générique possible, nous ne creusons pas cette piste dans la suite de ce travail.

3.3.2.3 Limites de la PA pour la RI

L'exploitation de la PA pour la RI a été abandonnée pour de multiples raisons mais nous identifions quatre types de problèmes :

La difficulté de mise en place du réseau sémantique La construction du graphe nécessite un grand effort d'ingénierie des connaissances. En effet, à l'exception de [Shoval, 1981], les travaux de l'état de l'art proposent de construire le réseau sémantique manuellement. C'est la principale limite présentée par [Preece, 1981] : modéliser manuellement la collection documentaire sous la forme d'un graphe prend beaucoup de temps même pour une collection de petite taille. De même [Cohen and Kjeldsen, 1987] évoque la difficulté de la construction du "réseau sémantique". Le système présenté par [Savoy, 1992] a également comme inconvénient principal le fait que l'espace hypertexte est construit manuellement, ce qui est d'autant plus coûteux que la collection est large. Le fait qu'il existe aujourd'hui de nombreux outils de modélisation et de construction des graphes est un argument qui nous a conduit à re-visiter ces approches de RI par propagation d'activation.

La nécessité d'une validation manuelle L'utilisateur doit souvent contrôler manuellement la propagation d'activation. Le système présenté par [Shoval, 1981] est simple. Il a l'avantage de construire le réseau automatiquement à partir d'une ressource indépendante du domaine mais il nécessite en revanche une validation manuelle constante de l'utilisateur pour choisir les termes à partir desquels étendre la requête.

Le réglage difficile des contraintes et règles d'inférences Ce réglage nécessite d'autant plus d'efforts que les règles et les contraintes sont souvent dépendantes du domaine. [Cohen and Kjeldsen, 1987] indique que le réglage des paramètres des chemins préférés pour la propagation d'activation est difficile, mais qu'il est primordial, car sans ce réglage une dégradation de la qualité de la recherche est à prévoir. De même parmi les autres paramètres à régler dans ce système, il y a le nombre d'itérations de propagation qui varie d'une collection à l'autre.

L'indisponibilité des ressources sémantiques Le manque de ressources disponibles telles que les thesaurus et les ontologies dans les années 80 est l'une des principales raisons pour lesquelles l'exploitation de la PA a été abandonnée en RI. Là aussi la situation a évoluée : l'avènement du web sémantique et du web de données change la donne et invite à re-visiter les approches de RI par PA.

3.3.2.4 Modèle vectoriel *vs.* modèle par PA pour la RI [Brouard, 2013]

L'application de la PA en recherche documentaire ne s'est pas faite par hasard. Il s'agit au contraire d'une application bien fondée, qui reproduit les propriétés de base de la RI classique, comme la prise en compte de la distribution des termes dans le document et dans la collection. Nous présentons dans cette section le travail de [Brouard, 2013], dont l'objectif est justement de faire comparer le modèle vectoriel classique et le modèle par PA en RI.

Pour représenter les données, Brouard met en place un réseau à deux couches. La première couche présente les nœuds documents, alors que la deuxième est l'ensemble des nœuds termes du vocabulaire. Les liens n'existent qu'entre les nœuds des couches différentes. Les arcs reliant les termes et les documents du réseau portent un poids w qui dépend du nombre d'occurrences du terme dans le document. Pour tout terme t associé au nœud i et tout document d associé au nœud j, on a $w_{ij} = w_{ji} = \sqrt{tf(t,d)}$.

La propagation d'activation sur un réseau de termes et de documents se fait par itérations successives sachant qu'à chaque itération, il y a une première étape de propagation des nœuds termes de la requête vers les nœuds documents où ils apparaissent et une deuxième étape de rétro-propagation de ces nœuds documents vers les nœuds termes de la requête. Les nœuds initialement activés sont les termes de la requête : le nœud associé au terme t possède une activation initiale $a_0(t)$ qui dépend de son nombre d'occurrences dans la requête q. Partant des nœuds termes initialement activés, l'activation se propage sur le réseau. Les nœuds atteints sont activés et possèdent une valeur d'activation non nulle.

La fonction d'activation est donc différente pour les nœuds termes et les nœuds documents :

— la valeur d'activation d'un document d est égale à

$$a_i(d) = \sum_{t \in d \cap q} a_{i-1}(t) * w_{td} * div(t) \tag{3.5}$$

où

— i correspond à la première ou seconde étape de propagation ($i = 1$ ou $i = 2$) ;

— $div(t)$ est l'inverse du nombre de connections du nœud terme t, ce qui correspond à l'inverse de sa fréquence documentaire ($idf(t)$) utilisée dans le modèle vectoriel de RI ; cela fait référence à la contrainte de connectivité "fan-ou", (voir section 3.3.1) qui contraint la PA aux nœuds (ici les termes) fortement connectés dans le graphe ;

— w_{td} est le poids de l'arc reliant le terme t et le document d : il s'exprime en fonction de $tf(t,d)$, ce qui établit une autre correspondance avec le modèle vectoriel ;

— $a_0(t)$ est l'activation initiale des termes de la requête et correspond à la fréquence d'occurrences de t dans la requête q ($tf(t,q)$) ;

— la fonction d'activation d'un terme t, par rétro-propagation de d vers la couche des termes, est égale à

$$a_i(t) = a_i(d) * a_{i-1}(t) * w_{td} * div(d) \tag{3.6}$$

où $div(d)$ est l'inverse du nombre de connections du document d dans le réseau, ce qui rend compte de la taille du document, un paramètre important dans les mesures de pondérations du modèle vectoriel, comme Okapi-BM25.

Brouard prouve mathématiquement la correspondance entre son modèle de PA et le modèle vectoriel. En effet, en écrivant $a_n(d)$ en fonction de $a_{n-1}(d)$, il arrive à une suite géométrique. Il prouve que lorsque n est grand la suite converge vers sa raison, à savoir :

$$R(d,q) = \sum_{t \in d \cap q} a_0(t) * w_{td}^2 * div(t) * div(d) = div(d) * \sum_{t \in d \cap q} a_0(t) * w_{td}^2 * div(t) \tag{3.7}$$

puisque

$$R(d,q) = idf(d) * \sum_{t \in d \cap q} \sqrt{ft(t,q)^2} * \sqrt{ft(t,d)}^2 * idf(t,q) \tag{3.8}$$

ce qui correspond à la fonction de correspondance cosinus du modèle vectoriel.

Expérimentalement, l'auteur trouve également des résultats très proches en terme de précision moyenne (MAP) entre son modèle de PA et le modèle vectoriel par Okapi-BM25, sur les deux collections TREC3 et CLEF3.

Ce travail a le mérite de montrer qu'on peut simuler le modèle vectoriel classique avec la propagation d'activation et il nous a encouragé à explorer les approches de PA en RI. Le modèle proposé par Brouard est cependant assez complexe parce que la fonction d'activation exploitée est différente pour les nœuds termes et les nœuds documents : cela rend son extension à des graphes plus complexes difficile. Par ailleurs, le modèle ne permet pas d'exploiter la sémantique latente exprimée notamment par la co-occurrence des termes ni d'introduire une sémantique explicite par l'exploitation d'une ressource externe. Nous verrons dans le chapitre suivant comment nous proposons d'introduire la sémantique dans la PA pour en faire un modèle de RIS.

3.3.3 Propagation d'activation pour l'accès aux données

La propagation d'activation est également utilisée pour l'accès aux données dans le cadre du web sémantique (WS). Même si la recherche de données ne fait pas partie de nos objectifs, l'application de cet algorithme à la recherche de données dans les bases de connaissances (BC) du WS nous semble intéressante à explorer parce que nous nous intéressons à l'exploitation des ressources sémantiques externes en RI et au processus de PA comme algorithme de recherche.

Nous montrons dans cette section, à travers les travaux de [Rocha et al., 2004] et [Schumacher et al., 2008], comment on peut modéliser les BCs sous la forme de graphes pour la PA et comment se déroule la propagation dans ce cadre.

3.3.3.1 Approches et choix de modélisation

En recherche de données, ce sont les concepts du domaine, leurs éventuelles instances et les relations qu'ils entretiennent que l'on modélise sous la forme de graphe mais la modélisation présuppose souvent une étape d'annotation sémantique pour intégrer des documents ou la requête à ce graphe.

Dans [Rocha et al., 2004], les auteurs réalisent une recherche de données sur le WS, en exploitant la PA comme mécanisme d'appariement, dans le but de retourner des ressources Web en réponse à une requête formulée à l'aide de mot-clés. Ils s'appuient sur une BC constituée d'une ontologie de domaine peuplée par des ressources Web, qui sont dès lors considérées comme des instances de concept, de même que les pages Wikipedia sont des instances des concepts de l'ontologie DBpedia. La recherche ne s'appuie pas sur le contenu des pages web mais sur l'analyse des méta-données qui leurs sont associées. L'approche proposée est adaptée aux ontologies ayant une riche composante textuelle, où il y a beaucoup d'instances documentaires et où ces instances sont richement décrites à l'aide d'attributs.

La structure de données sur laquelle se déroule la PA est le sous-graphe des instances de la BC. Les nœuds initialement activés sont des instances de concepts ou ressources Web retrouvées par un moteur de recherche classique (Lucene) qui rapproche les mot-clés de la requête et les métadonnées (titre, résumé, etc.) associées aux pages sous la forme d'attributs d'instances.

Dans [Schumacher et al., 2008], les auteurs réalisent une recherche de données sur une BC *PIMO ("Personal Information Model")*. Le graphe de recherche est un réseau sémantique composé des classes (concepts), instances, documents et relations entre instances existants dans la BC. Les points d'entrée au graphe sont les instances et classes résultant d'une recherche de faits antérieure et les documents trouvés par la recherche par mot-clés (moteur de recherche classique). Les instances sont pondérées lors de la recherche de faits alors que les documents reçoivent leurs

scores de RI comme poids initiaux. Les arcs du graphe sont également pondérés de telle sorte que les relations résultats de la recherche de faits sont sur-pondérées.

La figure 3.10 présente un aperçu de l'architecture générale du système de [Schumacher et al., 2008] qui combine recherche de faits, recherche de documents et propagation d'activation sur le réseau sémantique.

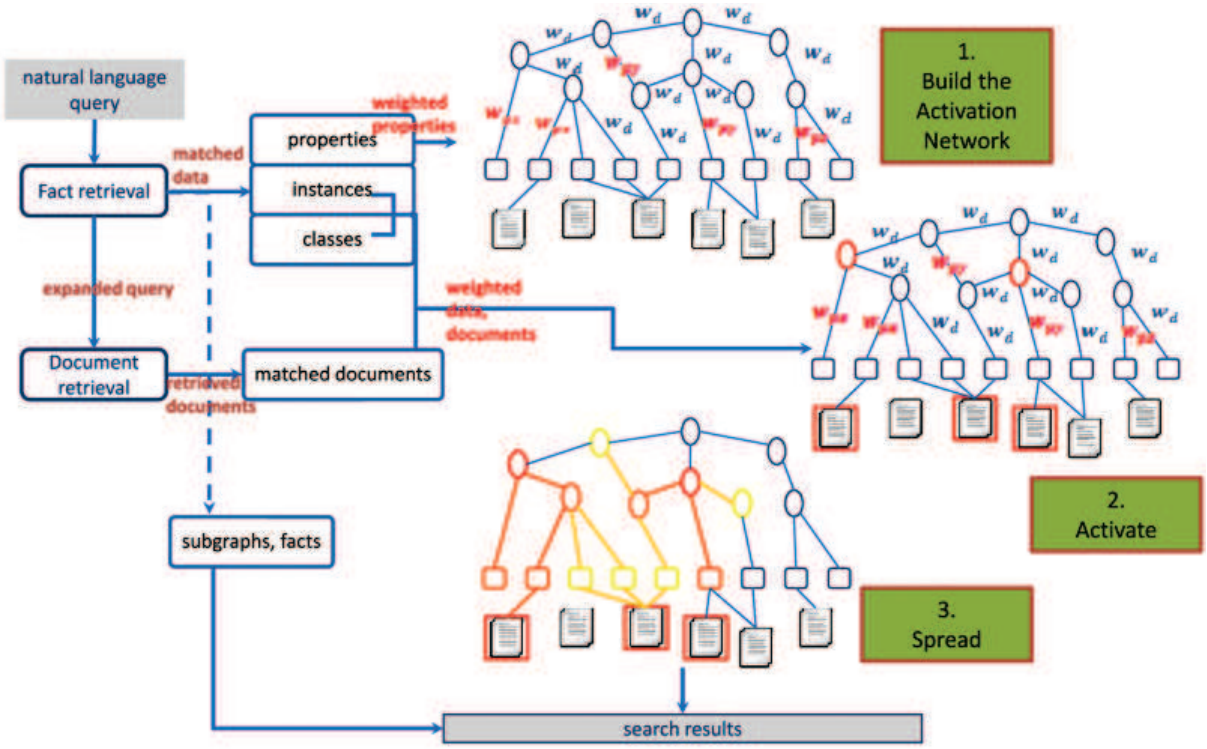

FIGURE 3.10 – Aperçu du système proposé par [Schumacher et al., 2008]

3.3.3.2 Calcul des poids des arcs

Au-delà des poids associés aux points d'entrée au réseau, la pondération des arcs du réseau est primordiale sur des réseaux hétérogènes composés de plusieurs types de nœuds et de relations.

Rocha et *al.* mettent l'accent sur la pondération des arcs du réseau mais tiennent également compte des labels des arcs et donc de la sémantique des liens [Rocha et al., 2004]. Chaque instance de relation a un poids numérique et un label sémantique. Pour le poids numérique, les auteurs introduisent trois mesures :

Mesure de *cluster* : elle donne le degré de similarité entre deux instances de concepts (C_j et C_k) reliées par une relation :

$$w(C_j, C_k) = \sum_{i=1}^{n} n_{ijk} / \sum_{i=1}^{n} n_{ij}$$

où $n_{ij} = 1$ s'il y a une relation entre C_i et C_j, $n_{ij} = 0$ sinon, et où $n_{ijk} = 1$ si C_j et C_k sont reliés à C_i, $n_{ijk} = 0$ sinon.

Mesure de spécificité : $w(C_j, C_k) = 1/\sqrt{(n_k)}$ où n_k est le nombre d'instances de relations ayant C_k comme destination et le même type que l'arc relinant C_j à C_k.

Mesure combinée : elle correspond au produit des deux mesures précédentes et ressemble par conséquent à un $tf.idf$.

Chaque arc ou instance de relation possède en outre un poids relatif, f_{ij}, associé au type symbolique de l'arc, c'est-à-dire à sa sémantique. Cela permet d'empêcher l'exploration des chemins peu pertinents ou de filtrer les résultats, si par exemple l'utilisateur n'est intéressé que par les nœuds de type `professeur` ou `publication`.

Dans [Schumacher et al., 2008], les arcs qui correspondent aux propriétés issues de la recherche de faits reçoivent un poids plus élevé, ce qui permet d'améliorer la précision, en imposant un chemin préféré de propagation. Des poids spécifiques sont également attribués aux autres arcs suivant certaines règles :

— les relations inverses, comme `hasAuthor` et `authorOf`, prennent le même poids,
— les relations de classes à instances sont privilégiées par rapport aux relations des instances à classes (`rdf:type`), ce qui favorise la spécification par rapport à la généralisation et minimise le bruit dans le réseau.

3.3.3.3 Fonction d'activation

La fonction d'activation permet d'attribuer des valeurs d'activation significatives aux nœuds du graphe activés au cours de la propagation. [Rocha et al., 2004] proposent qu'un nœud instance j activé par un nœud instance i reçoive l'activation suivante :

$$I_j(t+1) = O_i(t) * w_{ij} * f_{ij} * *(1-\alpha)$$

où

— $O_i(t) = \sum_{k=1}^{n} I(k)$ si le nœud i admet n relations entrantes,
— w_{ij} est le poids numérique de la relation allant de i à j,
— f_{ij} est le poids sémantique relatif associé au type de la relation allant de i à j (les arcs jugés non intéressants peuvent avoir un poids relatif nul pour interdire la propagation dans certaines branches du réseau),
— α est un facteur d'atténuation qui correspond à la quantité d'activation perdue chaque fois qu'un arc est traversé.

[Schumacher et al., 2008] utilisent une fonction d'activation assez classique :

$$I_j = \sum_i O_i * w_{ij} * (1-\alpha)$$

où α est un facteur d'atténuation.

3.3.3.4 Contraintes de propagation

[Rocha et al., 2004] exploitent certaines contraintes pour l'algorithme de PA qu'il mettent en place :

Contrainte de chemin : elle est utilisée pour empêcher la propagation sur certains types d'arcs et se traduit concrètement par le facteur f ;

Contrainte "fan-out" de connectivité : elle permet d'arrêter la propagation à partir des nœuds qui sont connectés à un certain nombre d'autres nœuds ;

Contrainte de distance : la propagation doit s'arrêter au niveau des nœuds situés à une certaine distance (seuil) de l'ensemble initial de nœuds.

[Schumacher et al., 2008] exploitent également certaines contraintes, notamment la contrainte de seuil sur les valeurs d'activation des nœuds et la contrainte *"fan-out"*.

3.3.3.5 Résultats et limites

Le but de la propagation d'activation proposé par [Rocha et al., 2004] est de trouver toutes les instances de concepts (ressources web) liées à un terme donné de la requête, même si ce terme n'apparaît pas directement dans les attributs de cette instance (métadonnées associées à la ressource). Les auteurs ont testé leur modèle sur deux BC différentes et montrent de bons résultats. Cette approche suppose cependant d'avoir une ontologie avec une bonne composante textuelle. Les auteurs considèrent que la sémantique de la propagation d'activation sur certains chemins n'est pas pertinente et que le retour de l'utilisateur est le seul moyen de pallier cette limite. Par ailleurs, le fait de n'interroger le réseau que par les instances de concepts peut engendrer un certain silence, si l'ontologie manque de couverture ou si certains termes de la requête ne correspondent en première instance à aucune instance.

Le résultat de la propagation d'activation proposée par [Schumacher et al., 2008] est l'ensemble ordonné des instances de la BC et des documents qui sont en relation. Une visualisation sous la forme de graphe est fournie pour expliquer les résultats [22]. Les auteurs montrent l'efficacité de leur modèle sur un jeu de 11 requêtes. L'ordre dans lequel les résultats sont présentés est une combinaison des rangs obtenus par la recherche de faits et par la recherche de documents. On observe que la recherche de faits fournit des informations plus précises concernant les documents et les instances de la BC. Cette approche nécessite elle-aussi d'avoir une BC dotée d'une riche composante textuelle.

Aucune de ces deux approches n'utilise la structure de l'ontologie et les liens hiérarchiques entre concepts : seules les relations entre instances de concepts sont considérées. On note aussi que les auteurs n'expliquent pas comment se déroule concrètement la propagation, par exemple comment les problèmes de boucles sont contrôlés.

La combinaison entre le niveau documentaire et le niveau sémantique n'est pas celle que nous voulons mettre en place car le contenu des documents n'est pas pris en compte (les documents sont vus comme des données ou instances de concept et seules les méta-données qui leur sont associées sont exploitées).

3.4 Discussion : propagation d'activation *vs.* marches aléatoires

Dans la suite, nous optons pour une approche à base de propagation d'activation (PA), mais il est intéressant pour justifier ce choix de comparer cet algorithme de parcours de graphe avec les approches à base de marches aléatoires. Rappelons qu'en recherche d'information (recherche documentaire), la détection de la pertinence par rapport à la requête est aussi importante que l'ordonnancement des résultats pertinents retournés. Il s'agit donc pour nous de déterminer celle des deux approches qui remplit au mieux ces deux fonctions. Nous les comparons sous quatre angle différents.

Histoire : La propagation d'activation provient de l'Intelligence Artificielle (IA) et propose une manière intelligente de parcourir les graphes (réseaux associatifs, réseaux sémantiques et réseaux de neurones) et d'exploiter les associations entre objets. En RI, la propagation d'activation a débuté avec [Preece, 1981] avant d'être reprise par [Crestani, 1997] et s'applique à un graphe composé des documents et des termes qu'ils contiennent. La propagation d'activation a donc été proposée bien avant les marches aléatoires, même si elle n'a pas connu le même succès. Elle a été utilisée pour mesurer la pertinence des

22. JUNG est utilisée comme plate-forme de visualisation

documents au regard de la requête alors que les marches aléatoires ont surtout servi pour ordonner les résultats (*ranking*).

Modèle mathématique : Les approches à base de marches aléatoires, tel que *PageRank*, reposent sur un modèle mathématique rigoureux qui impose de normaliser la distribution de probabilité sur l'ensemble du graphe. Cette contrainte peut être difficile à contrôler, notamment sur des graphes hétérogènes comportant des nœuds de types différents, ce qui rend leur application pour le calcul de pertinence documentaire difficile (on exploite généralement des nœuds termes, documents, concepts, etc.). Rares sont les travaux qui exploitent PageRank en RI, en tenant compte de l'aspect distributionnel des modèles classiques tel que le modèle vectoriel [23]. La propagation d'activation en RI, à l'inverse, est très flexible avec cette contrepartie que le contrôle et la stabilisation du processus ne sont pas garantis et nécessitent un travail d'ajustement expérimental important.

Dépendance à la requête utilisateur : Dans leur modèle de base, les marches aléatoires ne sont pas dirigées par une requête et les adaptations proposées (ex. HITS) sont coûteuses à mettre en œuvre. Le classement d'une page ne dépend pas de sa pertinence par rapport à un besoin de l'utilisateur, mais de sa position dans le graphe documentaire. A l'inverse, les méthodes de propagation d'activation sont plus faciles à mettre en œuvre et à adapter pour la RI : il suffit de déclencher l'activation à partir des nœuds activés par une requête.

Informations modélisées : Même si d'autres utilisations ont été proposées pour les deux modèles [24], généralement, en RI, la propagation d'activation modélise le contenu textuel des documents alors que l'algorithme PageRank ne modélise que les relations de citations, sans prendre en compte le contenu informationnel des pages.

D'autres différences sont citées dans le travail de [Sabetghadam, 2014].

Au vu de cette comparaison, la propagation d'activation paraît adaptée à la RI sémantique et à l'approche que nous souhaitons mettre en place : c'est un algorithme flexible qui permet facilement de modéliser des informations de types différents (le modèle documentaire et un modèle sémantique comme une ontologie), qui est sensible aux requêtes des utilisateurs (*"Query dependent"* pour [Sabetghadam, 2014]), mais qui permet aussi de classer les documents en fonction de leur pertinence distributionnelle et pas seulement de leur notoriété.

Ces raisons associées à l'essor du web sémantique (la disponibilité des ressources sémantiques et l'existence d'outils de construction et de manipulation de grands graphes) sont celles qui nous ont amenée à revisiter les méthodes de PA pour la RI et à proposer une approche de RIS à base de propagation d'activation. Nous suivons en cela la suggestion de [Rocha et al., 2004] qui présente la PA comme une approche permettant d'hybrider des modèles symboliques et numériques :

> With this integration, the hybrid graph is transformed into an associative network where all the edges are sub-symbolic. At first glance, it might appear that from now on the propagation is not hybrid anymore, since the semantics of the relations were transformed in numerical information. This is not really true, since there are various constraints that can be configured in the spread activation which use symbolic information regarding the node instances.

23. La seule tentative, à notre connaissance, est celle de [Blanco and Lioma, 2012] qui essaye de combiner PageRank et BM25 mais PageRank ne sert qu'à la pondération des termes.

24. Voir [Salton and Buckley, 1988a, Rocha et al., 2004, Hussein et al., 2007] pour la PA et [Collins-Thompson and Callan, 2005, Blanco and Lioma, 2012] pour les marches aléatoires.

3.5 Conclusion

Ayant présenté dans un premier chapitre (chapitre 2) un état de l'art sur l'exploitation des ressources sémantiques externes en RI, nous avons conclu que ces modèles ne sont en réalité que des modèles de RI classique adaptés pour intégrer la sémantique sous forme d'informations additionnelles. Ces informations sémantiques sont aplaties pour être en adéquation avec l'index classique de la RI.

Dans ce chapitre, nous sommes passés à une autre représentation des connaissances plus en adéquation avec la nature des ressources sémantiques que sont les ontologies. Il s'agit de la représentation sous la forme de graphes. Nous avons passé en revue les approches d'accès à l'information fondée sur des graphes, qui peuvent être des graphes de connaissances, des graphes documentaires (graphes de citation, web, etc.), des graphes modélisant les relations termes-documentaires ou une combinaison de ces différents types.

Nous avons montré que, dans le domaine du WS, les approches fondées sur les graphes RDF, bien qu'elles aient donné des langages d'interrogation assez expressifs et des moteurs d'interrogation permettant de raisonner sur les BCs, ne nous permettent pas de faire de la recherche d'information à proprement parler en prenant en compte toute la richesse des contenus documentaires. En effet, les modèles d'accès sur le WS font de la recherche de données et ne considèrent les documents que comme des données unitaires.

Nous avons ensuite analysé les modèles de RI qui exploitent le Web comme un graphe documentaire : les marches aléatoires et notamment l'algorithme PageRank. Nous avons montré qu'ils ne permettent pas de modéliser la notion de pertinence par rapport à la requête, car l'exploration du graphe n'est pas ou difficilement dirigée par la requête.

Le processus de propagation d'activation présente quant à lui une méthode d'accès fondée sur les graphes qui est flexible, qui peut prendre en compte des graphes hybrides composés de documents, d'ontologies, de termes, etc. et qui peut être dirigée par la requête. Nous avons rappelé que cet algorithme permet de reproduire les analyses distributionnelles de la RI classique et nous avons présenté les travaux de l'état de l'art exploitant la propagation d'activation pour la recherche documentaire ou l'accès aux données sur le web sémantique.

Les limites des approches par propagation d'activation proviennent de la difficulté de construction automatique des graphes sémantiques pour la recherche, de l'indisponibilité des ressources sémantiques à cette époque, et d'un paramétrage délicat. Il nous semble cependant que les progrès du web sémantique permettent de dépasser certaines de ces limites et qu'il est intéressant d'étendre à la RIS ces approches de propagation d'activation dont [Brouard, 2013] a réaffirmé récemment la valeur pour la RI.

Chapitre 4

Propositions : Modèle unifié et propagation d'activation

Sommaire

Les méthodes de RIS présentées dans l'état de l'art (voir chapitre 2) sont des adaptations des modèles algébriques classiques proposés pour prendre en compte la sémantique dans un processus de RI. Ces méthodes traduisent le plus souvent la pertinence sémantique des connaissances externes en pertinence numérique. Nous avons montré dans le deuxième chapitre que ces adaptations présentent elles aussi leurs limites et que les ressources sémantiques exploitées, notamment les ontologies, le sont de manière appauvries en pratique.

A l'inverse, dans le domaine du Web Sémantique où la richesse des ontologies est mieux exploitée, on a des approches d'accès à l'information purement sémantiques (voir chapitre 3), qui servent surtout à la recherche de données et qui font généralement abstraction du contenu documentaire. Même si ces approches bénéficient de langages d'interrogation expressifs et de moteurs d'interrogation ou d'inférence permettant de raisonner sur les connaissances, elles ne tirent pas profit des analyses distributionnelles qui ont fait leur preuve en RI.

Notre objectif est de proposer un modèle de RIS qui intègre dans une même représentation des informations documentaires et sémantiques – les termes, documents, concepts ainsi que les relations qu'ils entretiennent –, tout en masquant leur hétérogénéité. Nous voulons combiner les

analyses distributionnelles et les analyses sémantiques qui rendent compte respectivement d'une pertinence numérique et d'une pertinence symbolique.

Ce travail repose sur l'intuition qu'on peut modéliser sous la forme de graphes des ressources sémantiques et toutes les caractéristiques d'un document sur lesquelles repose la RI classique. Nous avons alors passé en revue les modèles d'accès à l'information fondés sur les graphes, pour conclure i) que les graphes RDF du WS permettent difficilement de reproduire la recherche documentaire, ii) que les graphes traditionnellement utilisés en RI – qui exploitent le Web comme un graphe documentaire – ne tiennent pas compte du besoin de l'utilisateur mais iii) que la propagation d'activation est un mécanisme flexible qui peut s'appliquer à des graphes hétérogènes. Nous montrons dans ce chapitre que la propagation d'activation permet de reproduire la RI classique tout en intégrant des informations sémantiques diverses.

Dans un premier temps, nous présentons le modèle en graphe que nous avons mis en place pour représenter de manière unifiée l'ensemble des propriétés numériques ou symboliques pertinentes pour la recherche documentaire (section 4.1). Nous montrons ensuite que la *propagation d'activation* peut être utilisée sur ce graphe pour apparier la requête de l'utilisateur et les documents ou les données disponibles. L'information de pertinence se propage en effet de proche en proche, depuis les éléments de la requête utilisateur (section 4.2). On peut ainsi reproduire une recherche documentaire classique mais aussi prendre en compte une sémantique implicite ou explicite, ce qui permet de dépasser certaines limites de la RI(S) (voir les fonctionnalités sémantiques du modèle section 4.3).

4.1 Modèle unifié de Recherche d'Information Sémantique

Le modèle de recherche d'information sémantique que nous présentons va au-delà du modèle vectoriel classique de "sac de mots" et des analyses statistiques sur la distribution des unités d'index dans les documents. Il dépasse aussi les modèles d'accès aux données où le document n'est pas l'objet de la recherche, mais vient seulement en appui à une réponse factuelle.

Notre modèle permet de représenter les données textuelles avec toutes leurs propriétés statistiques (fréquences d'occurrences, etc.) et les connaissances sémantiques des ontologies du WS dans un unique *réseau sémantico-documentaire*. Nous intégrons les relations sémantiques des ontologies et les relations termes-documents de la RI traditionnelle dans un unique modèle de *graphe pondéré* et nous modélisons la fonction de correspondance requête-résultats sous la forme d'un mécanisme de *propagation d'activation* dans le graphe.

4.1.1 Réseau sémantico-documentaire

Nous proposons de représenter la base documentaire et le modèle sémantique qui lui est associé sous la forme d'un unique réseau *sémantico-documentaire*. Cette structure permet d'introduire différents types de nœuds et différents types de relations selon ce qu'on souhaite représenter et de les organiser en trois nivaux de connaissances, les niveaux conceptuel, terminologique et documentaire (voir figure 4.1).

Le réseau *sémantico-documentaire* comporte trois types de nœuds comme le montre la figure 4.1 :

noeuds documents (N_d) : ils représentent tous les documents de la collection documentaire ;

noeuds termes (N_t) : ils représentent le vocabulaire de la collection documentaire ;

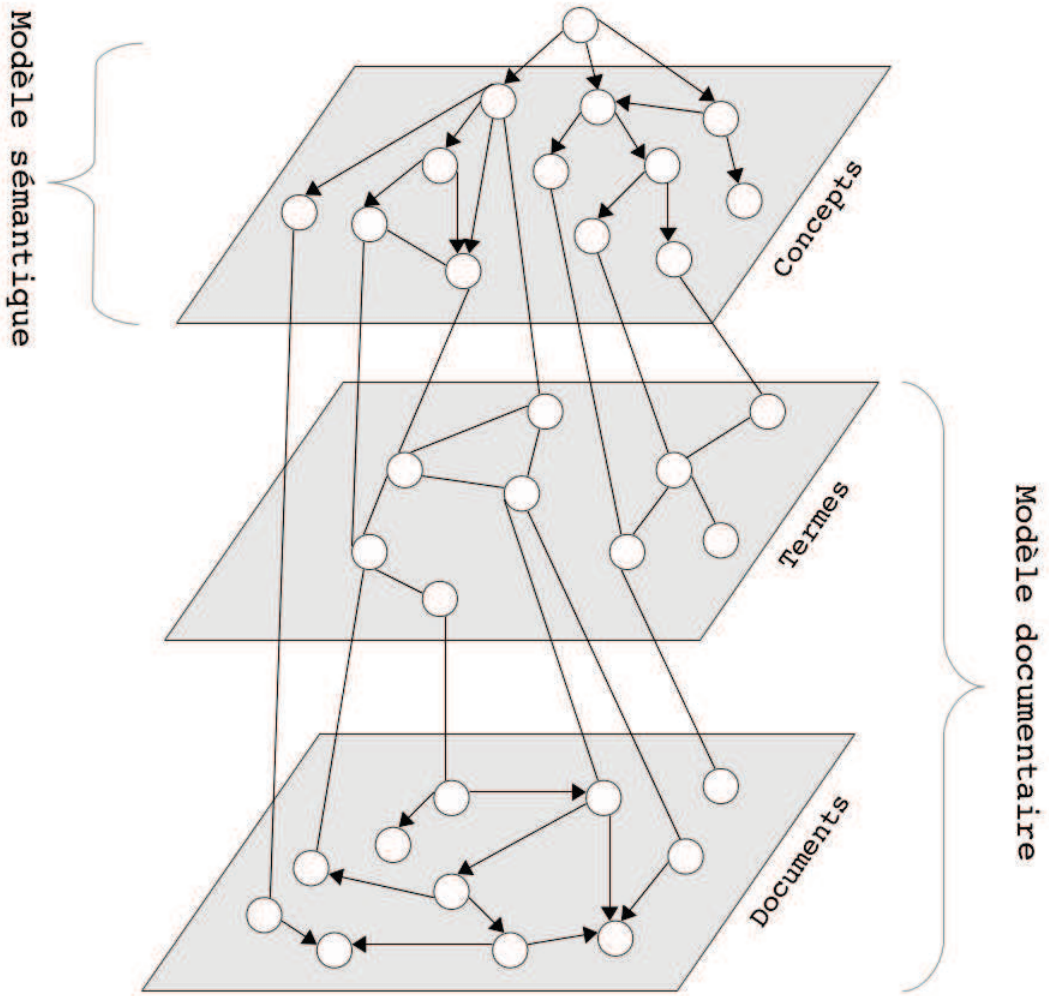

FIGURE 4.1 – Représentation en graphe des trois niveaux

noeuds concepts (N_c) : ils représentent les concepts de l'ontologie associée à la collection documentaire.

Ces trois types de nœuds sont liés les uns aux autres par des relations terme-document, terme-concept et document-concept et entre eux par des relations document-document et concept-concept.

Les nœuds du réseau sont donc reliés par 5 types de relations qui peuvent porter certaines propriétés numériques ou symboliques :

relations d'occurrence (R_{occ}) : ce sont des relations entre termes et documents qui traduisent le fait qu'un terme apparaît dans un document ; une propriété de fréquence peut naturellement être associée à ces relations ;

relations d'intertextualité (R_{int}) : ce sont des relations entre documents, comme par exemple les relations de citation ; ces liens peuvent être typés, les relations de citation n'étant pas les seules à être intéressantes à prendre en compte [25] ;

25. Dans le domaine juridique, [Mimouni et al., 2014] évoquent par exemple la relation de transposition entre une directive européenne et un texte réglementaire ou législatif national.

relations terminologiques (R_{ter}) : ce sont des relations entre termes et concepts qui indiquent quels termes sont les labels de quels concepts ; dans les ressources sémantiques dotées d'une composante terminologique, le fait qu'un terme soit relié à plusieurs concepts traduit son ambiguïté ; un concept peut également avoir plusieurs labels qui le dénotent, certains pouvant être des labels « préférés » ;

relations d'annotation (R_{ann}) : ce sont des relations entre documents et concepts qui associent des concepts ou des catégories comme méta-données à des documents ;

relations ontologiques (R_{ont}) : ce sont des relations entre concepts ou entre concepts et instances qui peuvent représenter aussi bien des rôles que des liens hiérarchiques.

4.1.1.1 Sous-réseaux sémantico-documentaires

On retrouve dans le réseau sémantico-documentaire des sous-réseaux correspondant à des bases de connaissances connues isolément :

Niveaux documentaire et terminologique Mis en relation, ces deux niveaux présentent une *base documentaire* qui contient une collection documentaire et le vocabulaire qui lui est associé. Les documents de la collection peuvent être inter-liés dans un réseau documentaire pour exprimer certaines relations (ex. citations). Le vocabulaire de la collection est représenté quant à lui par un ensemble de termes qui peuvent également être interconnectés par différents types de relations (ex. synonymie) et former un réseau sémantique. Les associations entre ces deux niveaux traduisent différents aspects de la base documentaire qu'on peut retrouver dans une représentation en index classique : un document est composé d'un ensemble de termes, un terme peut appartenir à différents documents, etc.

Niveau conceptuel et terminologique Ces deux nivaux inter-connectés constituent une *base de connaissance*, c-à-d, une ontologie peuplée par des instances de concepts ou une ontologie enrichie d'un volet terminologique. Dans l'ontologie, les concepts peuvent être organisés en hiérarchie mais également abriter des relations sémantiques (rôles) s'il en existe. Les termes du niveau terminologique forment le volet terminologique de l'ontologie, c-à-d les *labels* associés aux concepts ou à leurs instances.

Niveau documentaire et conceptuel L'association entre ces deux niveaux présente une *classification ou catégorisation documentaire*. Les concepts du niveau conceptuel sont dans ce cas les classes sémantiques qui permettent de catégoriser l'ensemble des documents de la collection. Ces liens de catégorisation permettent ici de rapprocher les documents "similaires", qui appartiennent à une même classe.

Différentes configurations du réseau *sémantico-documentaire* sont possibles, selon les connaissances qu'on choisit de représenter et selon les propriétés symboliques ou numériques qu'on décide d'attribuer aux nœuds et aux arcs du réseau. Ce réseau a vocation à être utilisé pour différentes applications (RI, accès aux données, catégorisation, etc.) mais il faut le paramétrer en orientant et en pondérant les liens qui le composent.

C'est pour cette raison que nous introduisons la notion de *graphe pondéré*.

4.1.2 Graphe pondéré

Le réseau *sémantico-documentaire* introduit dans la section 4.1.1 fait référence à différents types d'applications, mais il ne peut pas être utilisé en l'état. Il faut le traduire en graphe et pour cela faire des choix : associer des valeurs aux arcs, les typer et les orienter, mais aussi

introduire les éléments nécessaires au calcul distributionnel de la RI (notamment les fréquences d'occurrences) sans qu'il soit nécessaire de revenir aux documents sources.

En pratique, tout n'est pas utile à représenter pour toutes les applications et tous les types de connaissances ne sont pas toujours disponibles. Les données qu'on possède et l'application que l'on vise déterminent en partie ce qu'on choisit de représenter.

Nous traduisons donc ce réseau *sémantico-documentaire* en un *graphe pondéré* G, sachant que différentes traductions sont possibles.

Ce graphe $G = \langle N, R \subseteq N \times \mathbb{R} \times N \rangle$ est constitué d'un ensemble de nœuds ($N = N_d \uplus N_t \uplus N_c$) et d'arcs qui sont orientés et pondérés ($R = R_{occ} \uplus R_{int} \uplus R_{ter} \uplus R_{ann} \uplus R_{ont}$).

4.1.2.1 Pondération des arcs

Les arcs traduisent les relations qu'entretiennent les différents nœuds et les valeurs associées à ces arcs expriment la force de ces liens. Ces valeurs expriment des propriétés intrinsèques de ces liens, des propriétés qu'on ne peut pas retrouver par calcul ou dériver de la structure du graphe, par exemple. Nous notons par $w(i, j)$ la valeur ou poids de l'arc allant du nœud i vers le nœud j mais l'interprétation de ces valeurs dépend des types des noeuds qui sont mis en relation :

le poids d'une relation d'occurrence représente la fréquence d'occurrence d'un terme dans un document ; il peut être normalisé ou non et permet de garder une trace de l'importance des termes dans un document ;

le poids d'une relation terminologique permet de distinguer, par exemple, le label « préféré » d'un concept par rapport aux autres termes qui lui sont associés ;

le poids d'une relation d'annotation indique si une catégorie est associée à un document ou non (c'est généralement une valeur booléenne) ;

le poids d'une relation d'intertextualité indique si deux documents sont reliés (valeur boolénne) et éventuellement la qualité ou la force de cette relation (valeur numérique) [26] ; ces valeurs peuvent être attribuées par un expert humain au cours de la construction du graphe pondéré ; il faut préciser qu'il ne s'agit pas d'une mesure de similarité documentaire, laquelle n'est pas une propriété intrinsèque des documents reliés car elle se calcule sur l'ensemble de la collection (nous verrons plus loin que cette similarité se calcule sur le graphe) ;

le poids d'une relation ontologique indique s'il y a une une relation hiérarchique ou sémantique entre deux concepts et en donne éventuellement l'importance ; en jouant sur les valeurs de ces liens, on peut activer ou désactiver certains types de liens pour la recherche ; différents choix de modélisation peuvent donner différents paramétrages, même si la similarité entre concepts dépend plutôt de la structure du graphe (la distance dans l'ontologie par exemple). Nous verrons qu'on peut ainsi choisir par exemple de privilégier les liens de spécialisation dans la recherche et au contraire de bloquer les parcours inverses (via des liens de généralisation) en leur attribuant des poids nuls ou négatifs ; de même, certaines relations sémantiques spécifiques au domaine peuvent être considérées plus importantes que d'autres.

Nous n'entrons pas ici plus dans le détail du calcul de ces poids, considérant que différents paramétrages sont possibles, depuis un graphe booléen (sans poids) à un graphe entièrement

26. Ceci peut s'avérer utile dans certains domaines de spécialité où toutes les citations n'ont pas la même valeur ou quand certaines relations ont plus de poids sémantiques que d'autres par ex. lien de transposition *vs.* simple visa dans le domaine juridique.

pondéré, qu'ils reflètent différents choix de modélisation mais qu'ils sont tous compatibles avec le modèle à base de graphe que nous proposons.

4.1.2.2 Orientation des arcs

Les arcs peuvent être orientés ou non :

les relations terminologiques, d'occurrences et d'annotations sont des relations symétriques et ne sont pas orientées (voir la figure 4.1) ;

les relations d'intertextualité peuvent être orientées (par ex. les relations `est_cité_par` et `cite` n'ont pas nécessairement le même poids), ou non ; cela dépend en effet du type de la relation considérée et des choix de modélisation ;

les relations ontologiques sont généralement considérées comme orientées quand il s'agit de relations hiérarchiques mais les rôles peuvent aussi être considérés comme symétriques (par exemple il n'y a pas forcément besoin de différencier les deux sens de la relation d'`#auteur` à `#livre`).

Dans la suite, nous considérons que les arcs sans flèche représentent des relations symétriques et la valeur de pondération de l'arc est la même quel que soit le sens dans lequel on parcourt l'arc. En revanche, un arc orienté ne peut être parcouru que dans le sens de la flèche.

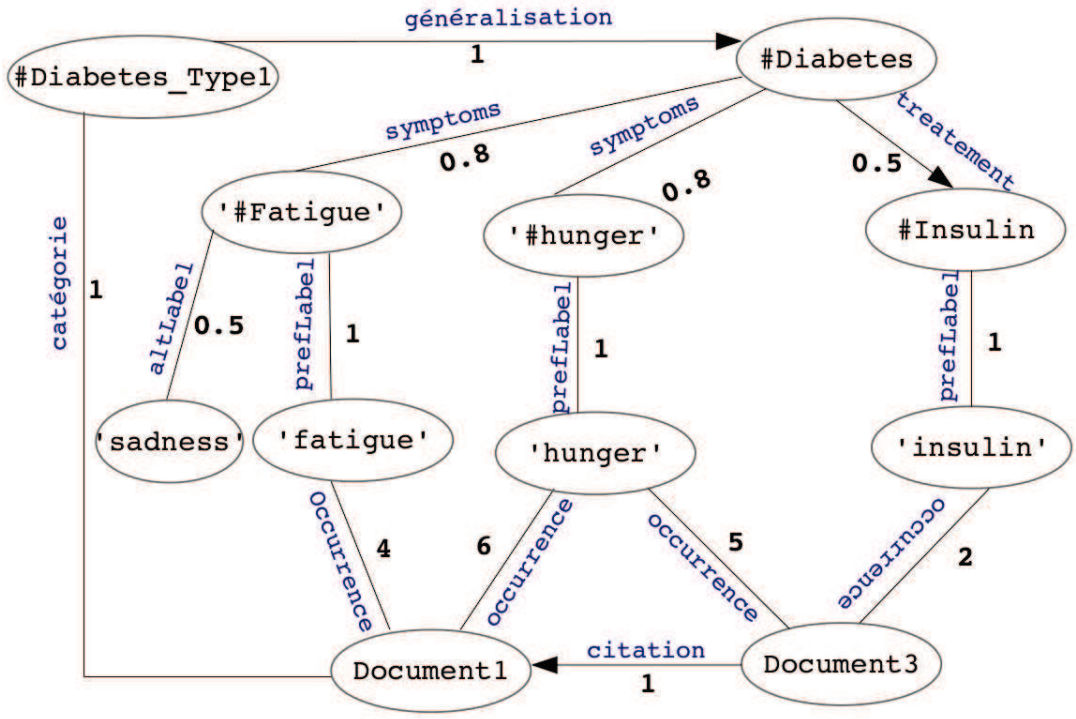

FIGURE 4.2 – Exemple de graphe pondéré G

La figure 4.2 présente une traduction possible d'un réseau *sémantico-documentaire* en *graphe pondéré*. Sur cette figure différents paramétrages ont été faits :

— on peut exploiter les liens hiérarchiques dans le sens de la généralisation, comme entre les deux concepts `#Diabetes_Type1` et `#Diabetes` : ces relations ont un poids de 1 ;

— des liens de citations sont exprimés (le `Document3` cite le `Document1`) et leur poids est de 1 ;

— les liens d'occurrence sont symétriques et ils sont pondérés par la fréquence d'occurrence de chaque terme dans le document auquel il est relié : le poids de l'arc reliant le `Document3` et le terme "hunger" est 5 par exemple ;
— on a représenté les liens terminologiques entre les différents concepts et leurs labels comme des relations symétriques mais en accordant un poids plus élevé aux labels préférés (`prefLabel`, poids de 1) qu'autres labels (`altLabel`, poids de $0,5$) ;
— la relation sémantique *symptoms* est représentée comme symétrique (entre la maladie `#Diabetes` et les symptomes `#Fatigue` et `#Hunger`) et vaut $0,8$, alors que la relation sémantique *treatement* entre `#Diabetes` et `#Insulin` est jugée moins importante (poids de $0,5$) ;
— les liens d'annotation sont symétriques et ont un poids de 1, comme celui qui relie `Document1` et `#Diabetes_Type1`).

Un réseau *sémantico-documentaire* peut être traduit de différentes manières et donner plusieurs graphes pondérés, selon ce qu'on souhaite représenter. Il n'y pas non plus une seule façon d'interroger ces informations unifiées sur le graphe pondéré, ni d'ailleurs un type unique de résultats pour la recherche sur ce graphe.

4.1.3 Interrogation et filtrage des résultats

La représentation unifiée de toutes ces informations sous la forme d'un graphe pondéré permet d'interroger les documents de manière plus riche que par les seuls mots-clés. On peut accéder au graphe par plusieurs points d'entrée selon que la requête comporte des termes (comme en RI traditionnelle), des concepts (comme en RIS), des documents (par ex. pour une recherche à base d'exemples), ou une combinaison de ces différents types d'éléments : $Q = \{t_1, t_2, \ldots, C_1, C_2, \ldots, D_1, D_2, \ldots\}$.

Il est ainsi possible de :
— poser des requêtes en utilisant des termes qui n'existent pas dans le vocabulaire de la collection, ce qui répond au problème de *term mismatch* décrit dans [Crestani, 2000] : si on interroge par exemple le graphe de la figure 4.2 par le terme "sadness" qui n'est pas dans le vocabulaire de la collection, on peut retrouver le document `Document1` même si le terme recherché n'y figure pas ;
— poser des requêtes par des concepts de catégorisation qui ne sont pas forcément associés à un terme du vocabulaire mais qui servent à la catégorisation du domaine et donnent un accès direct aux documents : si on interroge par exemple le graphe pondéré de la figure 4.2 par le concept `#Diatbtes_Type1`, on retrouve directement le document pertinent `Document1`, même si ce document ne contient pas de terme associé au concept `#Diatbtes_Type1` ;
— poser des requêtes par l'exemple, en soumettant comme requête un document et en recherchant les documents similaires : si on interroge également le graphe de la figure 4.2 par le document `Document1`, on peut retrouver `Document3` comme document pertinent.

Le processus de recherche intègre en outre un mécanisme d'enrichissement de la requête qui prend en compte des termes synonymes, des concepts ou des documents reliés et qui retourne un ensemble de nœuds pertinents ordonnés (voir section4.2.4 pour le mécanisme de *ranking* des résultats) sur ce graphe pondéré.

Les réponses attendues peuvent être, comme les requêtes, de différents types : les nœuds pertinents renvoyés sont des documents mais aussi des termes, des concepts ou même une combinaison de plusieurs types de nœuds. Un filtrage peut enfin être effectué sur l'ensemble des nœuds retournés pour sélectionner le/les type(s) de nœud(s) qu'on recherche.

A titre d'exemple, si le graphe de la figure 4.2 est interrogé par les termes "fatigue" et "hunger", ont peut sélectionner des nœuds concepts (`#Fatigue`, `#Hunger`, `#Diabetes`, `#Insulin`, etc.), des nœuds documents (`Document1` et `Document3`) ou encore des nœuds termes ("insulin", "sadness", etc.). Selon ce qu'on recherche, on choisit de filtrer sur un type de nœud ou un autre, mais on peut aussi choisir une combinaison qui permet de justifier la réponse : si on cherche par exemple des documents, on peut choisir de filtrer également sur les nœuds concepts, pour éclairer sémantiquement les documents retournés.

Le modèle permet ainsi de prendre en compte diverses formes de requêtes et de proposer différents types de résultats, sans avoir à changer de système d'accès à l'information ou de langage d'interrogation.

La fonction de correspondance que nous proposons consiste à appliquer une méthode de *propagation d'activation* dans le graphe qui part des nœuds de la requête et active de proche en proche les nœuds voisins dans le graphe. La méthode de propagation est détaillée dans la partie suivante.

4.2 Propagation d'activation

La propagation d'activation est un processus qui permet de propager une information de proche en proche sur un graphe. Ce mécanisme repose sur des *valeurs d'activation* associées aux nœuds du graphe : au départ les nœuds qui correspondent à la requête ont des *valeurs d'activation initiales* positives et les autres nœuds ont une valeur nulle ; le processus de propagation se déclenche à partir des noeuds activés et se propage de proche ne proche sur les noeuds voisins ; quand la propagation s'arrête, les valeurs d'activation obtenues sur les nœuds du graphe [27] déterminent l'ordre de pertinence des nœuds de ce graphe au regard de la requête initiale.

4.2.1 Principe général

La propagation à proprement parler consiste en un ensemble d'étapes de propagation et un ou plusieurs mécanismes de contrôle de la propagation.

4.2.1.1 Étape de propagation

Une *étape de propagation* se décompose en deux phases :

l'activation consiste à sélectionner les nœuds à activer parmi les nœuds dont la valeur d'activation est non nulle et qui n'ont pas encore été activés, puis à déclencher la propagation à partir de ces nœuds-là ;

la propagation transmet l'activité d'un ou plusieurs nœuds sources vers leurs voisins avant de désactiver les noeuds source : cette étape s'accompagne généralement du calcul de la nouvelle *valeur d'activité* des nouveaux nœuds cibles atteints.

L'*activation* s'applique, à chaque étape de propagation n, aux nœuds dont les valeurs d'activation ont été mises à jour par "contagion" à partir de leurs voisins directs à l'étape de propagation précédente $(n-1)$. Ce processus est répété en suivant les même phases d'*activation* et de *propagation* jusqu'à ce que plus aucun nœud ne puisse être activé (sélectionné).

27. Les nœuds peuvent être filtrés pour restreindre le type de résultat.

4.2.1.2 Contrôle de la propagation

La propagation d'activation se termine quand aucun nœud ne peut plus être sélectionné et que la distribution des valeurs d'activation sur le graphe s'est stabilisée.

Ces deux conditions sont contrôlées dans notre modèle par deux mécanismes qui assurent le déterminisme de la propagation d'activation et garantissent la terminaison du processus itératif de propagation d'activation :

— le premier mécanisme est *le calcul des valeurs d'activation* en tant que tel : il garantit la stabilisation du processus, notamment quand le graphe contient des cycles ;

— le deuxième mécanisme est *la modification de l'état des nœuds* du graphe au cours de la propagation d'activation : un nœud est tour à tour *inactif*, *activé* puis *désactivé*.

Dans la suite, nous expliquons tout d'abord comment le processus général de propagation d'activation est contrôlé, sans entrer dans les calculs des valeurs d'activation des nœuds activés à chaque étape (sous-section 4.2.2). Ensuite, nous détaillons le calcul des *valeurs d'activation* sur le graphe (sous-section 4.2.3) et introduisons la *fonction d'activation* qui garantit la stabilisation des valeurs d'activation à chaque étape de propagation (sous-section 4.2.4).

4.2.2 Contrôle de la propagation

Comme les graphes peuvent contenir des cycles, il n'est pas garanti que le processus de propagation itératif se stabilise mais il existe plusieurs façons d'optimiser l'algorithme de propagation d'activation et d'en assurer la terminaison.

Nous avons vu que l'état de l'art présente quelques contraintes classiques qui empêchent l'activation de se propager à tout le graphe et garantissent ainsi une fin plus rapide de l'algorithme (voir section 3.3 du chapitre 3) : les contraintes de distance [Rocha et al., 2004], de chemin [Cohen and Kjeldsen, 1987, Rocha et al., 2004], de "*fan out*" [Rocha et al., 2004, Schumacher et al., 2008] et de seuil sur les valeurs d'activation [Schumacher et al., 2008]. Les travaux de propagation d'activation pour la RI et pour l'accès aux données mettent en avant d'autres contraintes (voir section 3.3.2 et section 3.3.3 du chapitre 3), comme la détermination *a priori* d'un nombre défini d'itérations (étapes de propagation) [Savoy, 1992], la stabilisation d'une forme de monotonie au fil des itérations [Brouard, 2013], ou encore des contraintes qui dépendent de l'application et du domaine. Par exemple, [Savoy, 1992] tient compte des types des liens et nœuds dans les contraintes qu'il met en place dans le contexte d'une recherche par l'exemple.

Dans ce travail, nous n'utilisons pas de contraintes spécifiques pour limiter la propagation sur le graphe, ni de contraintes dépendant du domaine et de l'application.

Le mécanisme de contrôle fait partie intégrante de l'algorithme de propagation d'activation et repose sur *la modification de l'état des noeuds du graphe.*

Au début de chaque étape d'activation n l'ensemble des nœuds du graphe se décompose en trois ensembles disjoints : les nœuds *actifs*, les nœuds *inactifs* et les nœuds *désactivés* (voir figure 4.3) :

un nœud inactif est un nœud dont la valeur d'activation est nulle et qui n'a jamais été activé ;

un nœud actif est un nœud qui a été sélectionné au cours de la phase d'*activation* car (i) sa valeur est non nulle et (ii) il n'a jamais été activé auparavant (il était *inactif* à l'étape $n-1$) : il est prêt à propager sa pertinence à ses voisins à l'étape n mais il ne peut être activé qu'une seule fois ;

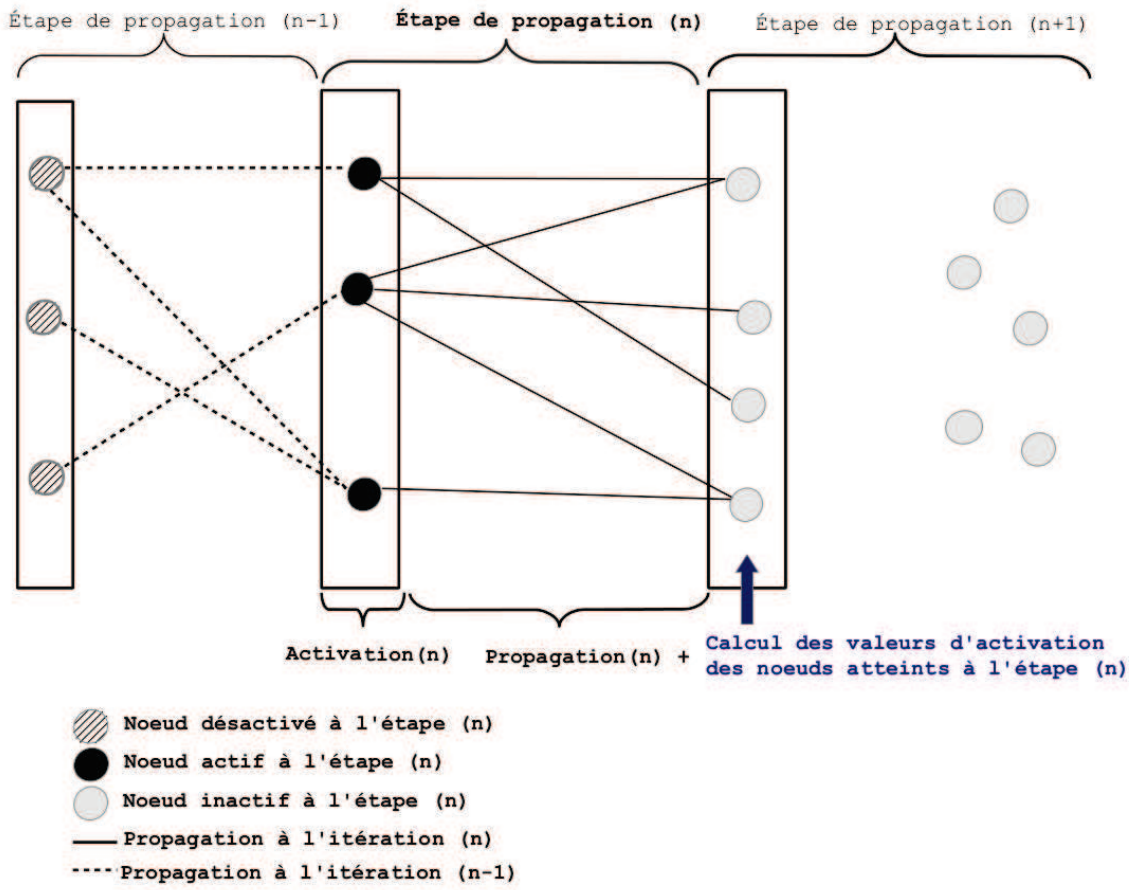

FIGURE 4.3 – Processus de propagation d'activation

un nœud désactivé est un nœud *actif* à une étape antérieure qui a déjà propagé sa pertinence et qui ne peut revenir ni dans l'état *actif* ni dans l'état *inactif*.

Au départ, l'ensemble des nœuds *actifs* contient les nœuds correspondant à la requête, tandis que les autres nœuds apparaissent comme *inactifs*. Le calcul des valeurs d'activation ne s'effectue que sur les nœuds *inactifs* atteints au cours de la propagation (voir figure 4.3). A la fin d'une étape de propagation, les nœuds *actifs* sont *désactivés* tandis que les nœuds *inactifs* dont la valeur est non nulle à l'issue du calcul de propagation constituent les nœuds *actifs* de l'étape de propagation suivante.

Comme un nœud ne peut être activé qu'une seule fois, la terminaison de l'algorithme est assurée : il s'arrête quand l'ensemble des nœuds *actifs* est vide, c'est-à-dire au maximum au bout de $|N|$ itérations, où N est le nombre de noeuds du graphe.

Des mécanismes similaires ont été exploités dans l'état de l'art. Par exemple, [Croft et al., 1989] impose que les documents ayant déjà été activés ne soient plus utilisés ou réactivés s'ils sont de nouveau atteints. De même [Rocha et al., 2004] n'autorise un nœud à s'activer qu'une seule fois, bien que sa valeur d'activation puisse être modifiée.

En résumé, il faut retenir qu'un nœud désactivé ne peut plus être réactivé mais que sa valeur d'activation peut continuer à croître sous l'influence de ses voisins : la propagation d'activation est donc croissante (voir la fonction d'activation section 4.2.4) et elle s'arrête quand il n'y a plus de nœuds *actifs*, qu'il ne reste que des nœuds *désactivés* ou des nœuds *inactifs* qui ne peuvent

pas être atteints par la propagation d'activation.

4.2.3 Activations initiales et valeurs d'activation

La stabilisation globale du processus nécessite cependant pour être atteinte, que le calcul des *valeurs d'activation* soit déterministe pour chaque étape de propagation, y compris quand des boucles existent dans le graphe. C'est en effet la *fonction d'activation* qui assure le calcul des *valeurs d'activation* des nœuds et garantit une distribution déterministe de ces valeurs sur les nœuds du graphe, à partir de configuration de départ identique.

Soient le graphe pondéré $G = \langle N, E \rangle$ et une fonction a sur le graphe G ($a : N \rightarrow \mathbb{R}^+$) qui donne une affectation de valeurs d'activation a sur le graphe G. Soit une suite finie de fonctions d'affectation $\pi = a_0 \dots a_k \dots a_n$, nous notons $a_k(i)$ la valeur d'activation du nœud i par l'affectation a_k.

L'activation initiale des nœuds est fixée *a priori* avant que le processus de propagation d'activation ne soit déclenché. Ainsi, pour une tâche de RI, les nœuds activés peuvent représenter les termes de la requête et l'activation initiale peut être la fréquence d'occurrence de chaque terme dans la requête en question.

L'algorithme de propagation définit une suite finie π d'affectations où a_0 représente l'affectation issue de la requête et donne la restriction ordonnée de a_n sur l'ensemble N_d des noeuds documents du graphe comme résultat de la requête.

Une suite d'affectation dépend du processus de propagation, de la structure du graphe (l'ensemble des nœuds activés à l'affectation précédente, etc.) et de la fonction d'activation responsable des calculs des valeurs d'activation des nœuds à chaque affectation a_k. Pour une tâche de RI, cette fonction d'activation correspond concrètement à la fonction de correspondance.

4.2.4 Fonction d'activation

A chaque étape de propagation, une nouvelle affectation a_k est calculée sur la base des valeurs d'activation a_{k-1} et en fonction de la structure du graphe. C'est cette fonction qui encode réellement le comportement de la fonction de correspondance.

On pourrait introduire plusieurs fonctions de propagation pour les différents types de nœud (concept, terme et document) mais nous avons choisi une fonction de propagation uniforme pour étudier au départ un modèle assez simple.

Soient un nœud i et a_{k-1} l'affectation issue de l'itération $k-1$. La valeur d'activation de i à l'itération k est définie par l'équation 4.1 suivante :

$$a_k(i) = a_{k-1}(i) + \sum_{j \in pred(i) \cap actif(k-1)} \frac{a_{k-1}(j) * w(j,i)}{deg(j)} \tag{4.1}$$

où $pred(i)$ retourne la liste des nœuds qui pointent vers le nœud i, $actif(k)$ est l'ensemble des nœuds actifs à l'itération k, et $deg(j)$ est le degré du nœud j.

On comprend que affectation dépend de *la structure du graphe* et de *l'état des nœuds du graphe.*

4.2.4.1 La structure du graphe

La fonction d'activation définie par l'équation 4.1 est proportionnelle à la somme des valeurs d'activation des nœuds j *prédécesseurs* du nœud i ($j \in pred(i)$) pondérées par les valeurs des arcs reliant ses prédécesseurs au nœud i ($w(j,i)$), et inversement proportionnelle au degré de ces

nœuds *prédécesseurs* ($deg(j)$). Un nœud j est un prédécesseur d'un nœud i, s'il existe un arc entrant de j vers i ($R(j,i)$). Le degré d'un nœud j représente le nombre d'arcs sortant du nœud j.

La prise en compte de la structure du graphe (les nœuds prédécesseurs et leurs degrés) lors du calcul des valeurs d'activation permet de déduire qu'un nœud est d'autant *plus activé* (sa valeur d'activation est d'autant plus grande), qu'il est plus *fortement relié* à des nœuds (le poids w de l'arc est d'autant plus fort) qui lui sont *exclusifs* (les degrés de ces derniers nœuds sont faibles).

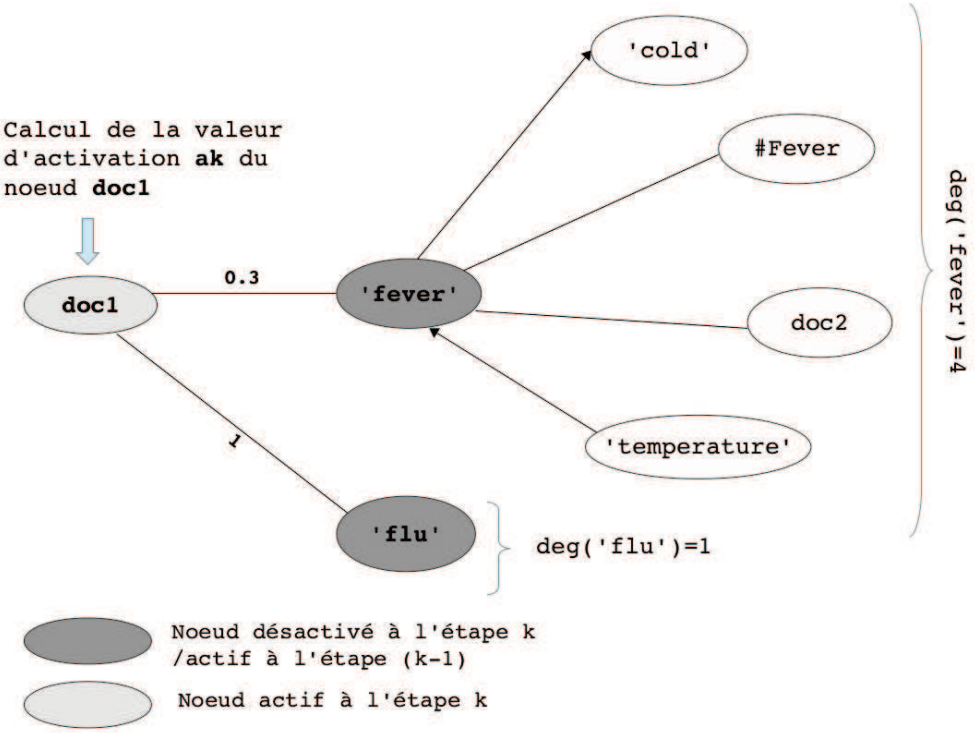

FIGURE 4.4 – Exemple de calcul de la valeur d'activation a_k du noeud `doc1` : impact de la valeur des liens avec les prédécesseurs ("fever" et "flu") et de leurs degrés(deg("fever") et deg("flu")).

La figure 4.4 montre un exemple de calcul de la valeur d'activation a_k du nœud `doc1`. Ce nœud a été activé par ses prédécesseurs "fever" et "flu" à l'étape $k-1$. Le nœud `doc1` est fortement relié au nœud "flu" car le poids de l'arc est élevé ($w(\text{“}flu\text{“}, \texttt{doc1}) = 1$) et plus faiblement relié au nœud "fever" ($w(\text{“}fever\text{“}, \texttt{doc1}) = 0.3$). On remarque aussi que le nœud "flu" est exclusif au nœud `doc1`, car le seul arc sortant de "flu" va vers `doc1`, d'où $deg(\text{“}flu\text{“}) = 1$. Par contre, le nœud "fever" n'est pas exclusif au nœud `doc1`, car il est relié aussi à `doc2`, `#Fever` et "cold", d'où $deg(\text{“}fever\text{“}) = 4$.

Nous intégrons dans le degré des nœuds prédécesseurs tous les arcs de manière uniforme, comme dans l'exemple de la figure 4.4, mais on pourrait différencier les arcs en fonction du type de nœuds reliés. On pourrait ainsi calculer :

le degré documentaire d'un terme : il s'agit de la fréquence documentaire d'un nœud terme, c'est-à-dire le nombre de nœuds documents auquel ils sont reliés (le nombre de documents dans lesquels le terme apparaît) : dans l'exemple de la figure 4.5, le degré documentaire du terme "pain" est égal à 4 ;

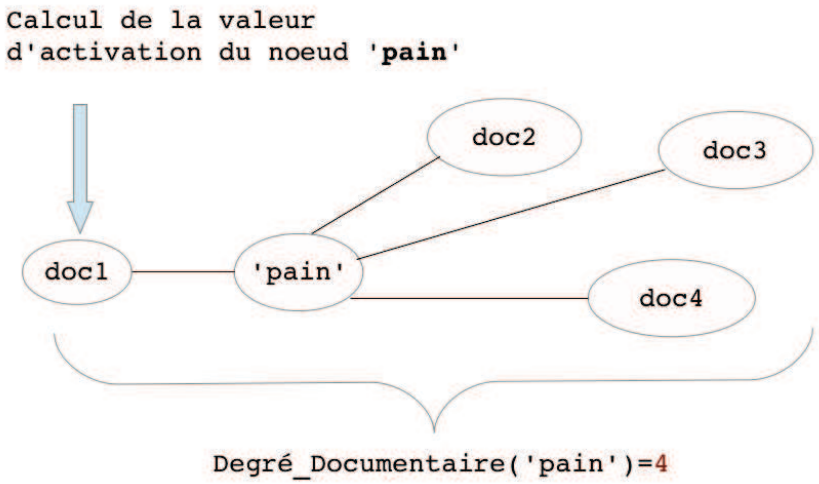

FIGURE 4.5 – Exemple de calcul du degré documentaire d'un terme : "pain"

le degré terminologique d'un document : il s'agit de la taille d'un document, qui peut-être estimée à partir du nombre de nœuds termes reliés au nœud document : le document `doc1` de la figure 4.6 a un degré terminologique de 3 ;

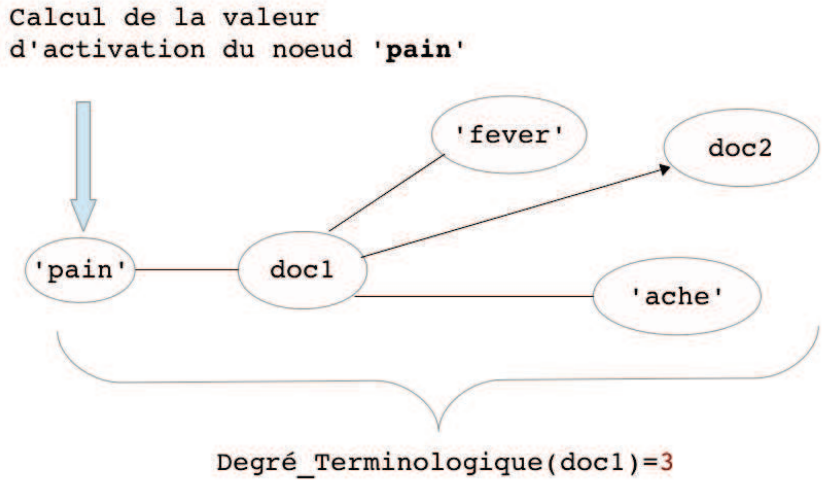

FIGURE 4.6 – Exemple de calcul du degré terminologique d'un document : `doc1`

le degré conceptuel d'un terme : il s'agit du degré d'ambiguïté d'un nœud terme et il dépend du nombre de nœuds concepts auxquels le nœud terme est relié : dans la figure 4.7, le terme terme "pain" est ambigü parce qu'il est le label de deux concepts différents (`#Ache` et `#Distress`) ;

le degré terminologique d'un concept : ce degré, qui correspond au nombre de termes auxquels un concept est relié, traduit la richesse de son volet terminologique du concept mais aussi potentiellement son ambiguïté : la figure 4.8 montre que le concept `#Asthma` a un degré terminologique de 2.

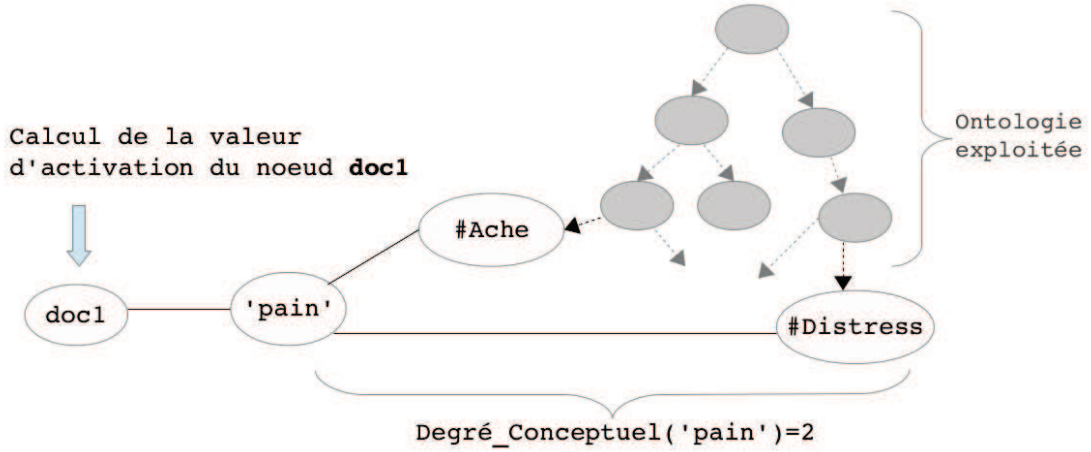

FIGURE 4.7 – Exemple de calcul du degré d'ambiguïté d'un terme : "pain"

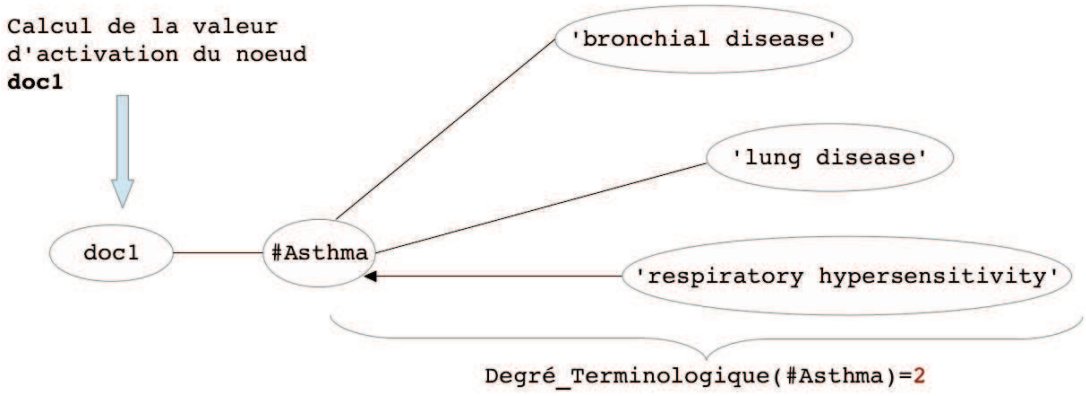

FIGURE 4.8 – Exemple de calcul du degré terminologique du concept : `#Asthme`

4.2.4.2 L'état des nœuds du graphe

La fonction d'activation définie dans l'équation 4.1 ne dépend pas uniquement de la structure du graphe. Elle dépend également de l'état des nœuds : seuls les *actifs* à l'itération $(k-1)$ sont pris en compte dans le calcul de la valeur d'activation d'un nœud à l'itération k ($j \in actif(k-1)$).

4.2.4.3 Le déterminisme

Le fait que le calcul des valeurs d'activation à l'étape k ne dépend que des affectations $k-1$ permet d'avoir un calcul déterministe même si le graphe contient des cycles.

Dès lors que le graphe contient un cycle, on risque en effet d'avoir des valeurs d'activation différentes pour un nœud selon le chemin par lequel on est arrivé à ce nœud i mais le mécanisme de propagation d'activation tel que nous l'avons définit est déterministe. Le calcul des valeurs d'activation à l'étape k dépend des nœuds prédécesseurs actifs à l'étape $k-1$ (qui sont désactivés à l'étape k) et de leurs valeurs d'activation à l'étape $k-1$. Du fait des cycles, les valeurs d'activation de ces prédécesseurs peuvent être réévaluées ultérieurement mais sans que cela n'affecte leurs voisins.

La figure 4.9 montre un exemple de cycle entre les deux nœuds "lung" et `doc1`. Selon l'ordre

dans lequel ces deux noeuds se déclenchent, on n'obtient pas les mêmes valeurs d'activation au final :

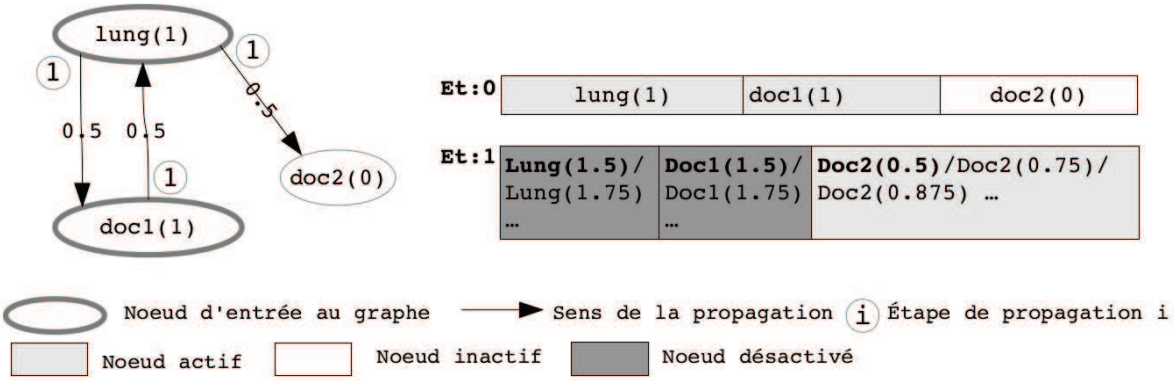

FIGURE 4.9 – Exemple de cycle lors du calcul de la valeur d'activation du nœud `doc2`.Les flèches sur les arcs schématisent le sens de la propagation pour chaque étape et pas l'orientation des arcs.

cas 1 : si le nœud `doc1` se déclenche en premier, il active le nœud "lung", dont la valeur d'activation passe à 1.5 ; le nœud "lung" se déclenche ensuite et active à son tour `doc2` dont la valeur passe à 0.75 ainsi que le nœud `doc1` dont la valeur passe à 1.75 ;

cas 2 : si le nœud "lung" se déclenche en premier, il active le nœud `doc1` et le nœud `doc2` dont les valeurs d'activation passent respectivement à 1.5 et à 0.5 ; lorsque `doc1` se déclenche à son tour, la valeur d'activation de "lung" passe à 1.75 ;

cas 3 : si les nœuds `doc1` et "lung" ont tous les deux une valeur d'activation non nulle au départ comme indiqué dans la figure 4.9 (ce sont deux nœuds d'entrée dans le graphe), ils se déclenchent en même temps et s'impactent mutuellement mais leurs valeurs se stabilisent à 1.5 car la valeur d'activation de chacun à l'étape 1 ne dépend que de la valeur d'activation de départ de l'autre (étape 0) et pas de sa valeur à l'étape 1 ; de même, le nœud `doc2` est atteint par "lung" à l'étape 1 et sa nouvelle valeur d'activation vaut à 0.5 (elle se calcule à partir de la valeur d'activation initiale de "lung", c'est à dire 1).

La valeur d'activation d'un nœud dans un graphe pondéré dépend de la structure du graphe et de l'affectation précédente. Ceci rend la fonction de mise à jour *synchrone*, c'est-à-dire indépendante de l'ordre dans lequel les valeurs d'activation des nœuds sont calculées. L'algorithme de propagation d'activation est de ce fait déterministe (voir un pseudo code de l'algorithme de propagation d'activation dans l'annexe A).

4.3 Fonctionnalités du modèle proposé

Notre modèle unifié de recherche d'information sémantique par propagation d'activation doit permettre de dépasser les limites de la RI et la RIS que nous avons identifiées dans le chapitre 2) :

— l'ambiguïté du vocabulaire est une limite classique de la RI mais nous montrons que l'exploitation des informations encodées dans le graphe pondéré permet d'acquérir des connaissances implicites qui permettent la désambiguisation sémantique et, de ce fait, l'amélioration de la précision de la recherche d'information ;

— la co-occurrence des termes du vocabulaire est rarement exploitée dans les modèles classiques de RI mais, dans notre modèle, elle permet d'améliorer la précision, le rappel et le *ranking* des résultats, par son impact sur les valeurs d'activation des nœuds ;
— le *term mismatch* et la synonymie, qui sont liés à la richesse de la langue, affectent négativement les modèles classiques de RI, mais leur prise en compte dans notre modèle réduit le silence et améliore le rappel dans les résultats ;
— le défaut de couverture de la ressource qu'est le vocabulaire de la collection figure parmi les limites des approches de RI sémantique mais notre modèle résoud ce problème en proposant plusieurs modes d'interrogation sur le graphe pondéré.

Nous montrons dans cette section comment notre modèle répond à ces différents problèmes, sans même exploiter la structure de l'ontologie (les liens hiérarchiques) ni les liens sémantiques (les rôles) : nous considérons les concepts de l'ontologie comme des classes sémantiques indépendantes.

4.3.1 Prise en compte de la co-occurrence

Dans [Manning and Schütze, 1999], les auteurs définissent la co-occurrence entre termes comme *"the fact that two or more terms occur in the same documents more often than chance"*. Deux termes sont co-occurrents lorsque la présence d'un mot dans un texte donne une indication sur la présence de l'autre, sans pour autant qu'ils soient synonymes ou antonymes [Réhel, 2011].

Le fait que deux termes co-occurrent n'explicite pas la sémantique de la relation sous-jacente (on ne sait pas, si on parle dans un contexte donné de "Apple" et "Jobs", que c'est pour désigner la marque "apple inc." qui conçoit et commercialise des produits électroniques et qui est dirigée par son fondateur "Steve Jobs") mais sert néanmoins en Recherche d'Information pour lever l'ambiguïté des termes et rapprocher les différentes désignations lexicales d'une notion (synonymes).

Dans l'exemple précédent, le terme "Apple" peut désigner le fruit "Pomme" ou une société bien connue selon le contexte dans lequel il figure. De même, "Jobs" peut être analysé comme le pluriel de "emploi", mais la co-occurrence entre ces deux termes permet de cibler un contexte particulier (le monde de la téléphonie et des produits électroniques). Sur notre graphe pondéré, la propagation d'activation ne vise pas à identifier toutes les co-occurrences existant entre des termes du vocabulaire de la collection, comme dans [Blanco and Lioma, 2012], mais à repérer les co-occurrences pertinentes au regard de la requête de l'utilisateur.

La figure 4.10 présente un exemple de propagation d'activation sur un graphe. Partant de la requête "Apple et Jobs", on arrive à activer à la première itération les documents qui contiennent un ou plusieurs termes de la requête (`Doc1`, `Doc2` et `Doc3`). A la deuxième itération, ces documents activés propagent leur pertinence aux termes du vocabulaire qui y sont contenus et ces termes sont plus ou moins fortement activés selon les co-occurrences qui entrent en jeu (`Doc1`, `Doc2` et `Doc3`). Les trois nœuds "Steve", "Apple inc" et "iphone" ont des valeurs d'activation plus importantes que les autres termes activés à la deuxième itération parce qu'ils co-occurrent dans des documents (`Doc1` et `Doc2`) qui ont été sélectionnés en fonction des termes de la requête.

Cette co-occurrence permet en effet de :

1. désambiguïser le vocabulaire et diminuer le bruit : les termes co-occurrents propagent leur pertinence à la troisième itération et renforcent les valeurs d'activation des documents activés à la première itération (`Doc1` et `Doc2`), ce qui exclut les documents non pertinents, activés par erreur par les termes de la requête, comme le document `Doc3` ;
2. activer de nouveaux documents pertinents et augmenter le rappel : à la troisième itération les termes co-occurrents activent de nouveaux documents pertinents (`Doc4` et `Doc5`) qui ne contiennent pas exactement les termes de la requête mais des termes qui leur sont sémantiquement proches ("Steve", "Apple inc" et "iphone").

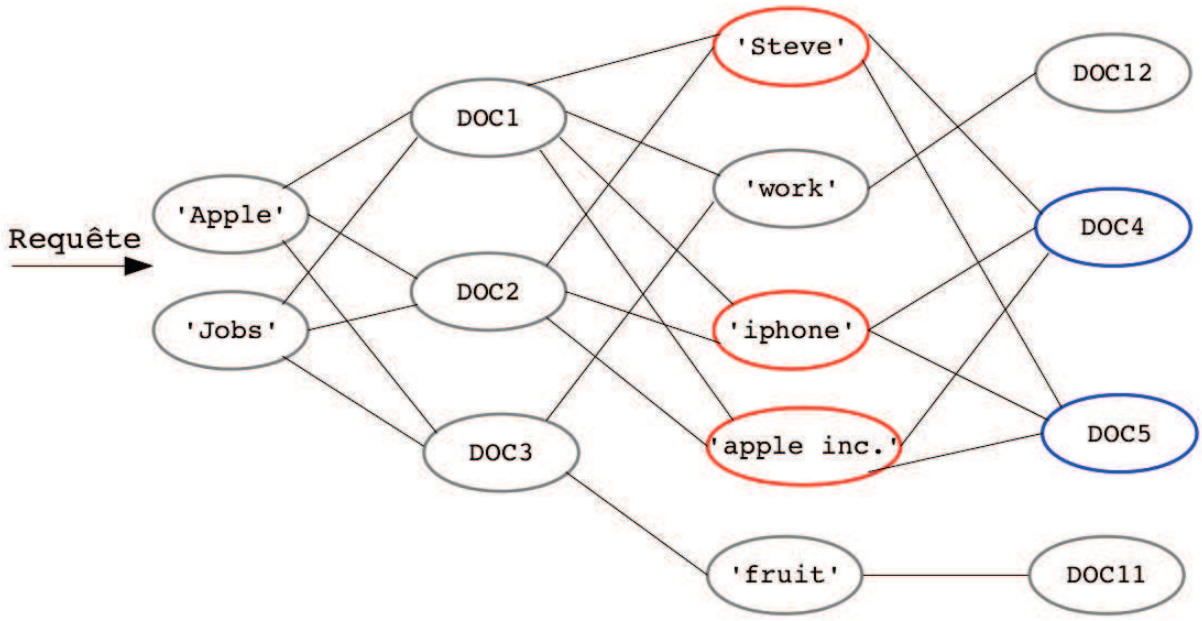

FIGURE 4.10 – Exemple de prise en compte de la co-occurrence. Les termes en rouge sont plus fortement activés que les autre, du fait de la co-occurrence, ce qui permet de retourner comme pertinents des documents (en bleu) qui ne contiennent pas les termes de la requête.

La co-occurrence n'explicite pas la nature des relations qui unissent les deux termes co-occurrents, ni même leur classe sémantique : c'est une sémantique latente, implicite. A l'inverse, l'utilisation des ressources sémantiques externes en RIS permet d'exploiter des connaissances et des relations explicites : des classes sémantiques et des relations comme l'équivalence ou la synonymie.

4.3.2 Traitement de la synonymie

La synonymie est la relation qu'entretiennent entre eux divers termes ou expressions ayant le même sens ou un sens voisin. Dans les modèles classiques de RI, la synonymie pose problème parce que deux termes désignant une même notion sont généralement traités différemment dans les approches par "sac de mots". Pour s'affranchir de cette difficulté, différentes solutions ont été proposées par la communauté de RIS avec l'exploitation de ressources sémantiques externes. On peut notamment affecter une même classe sémantique ou un unique concept aux différents termes ayant une même désignation. Ces termes sont considérés comme synonymes même s'il n'y a pas de lien de synonymie explicite entre eux dans la ressource sémantique. C'est de cette manière aussi que nous prenons en compte la synonymie dans notre graphe : deux termes sont considérés comme synonymes si et seulement si ils sont les *labels* d'un même concept, c'est-à-dire s'il relèvent d'une même classe sémantique.

Cette synonymie joue un rôle important dans notre modèle de RIS à base de graphe.

Amélioration de la précision *via* le *ranking* La figure 4.11 présente un exemple de propagation d'activation où les nœuds "migraine" et "headache" sont synonymes car ce sont deux labels du concept `#Migraine Disorders`. A la première itération, le terme "migraine", qui est un élément de la requête, active les nœuds `#Migraine Disorders` et `Document`. A la deuxième itération ces derniers nœuds activent, à leur tour, le terme

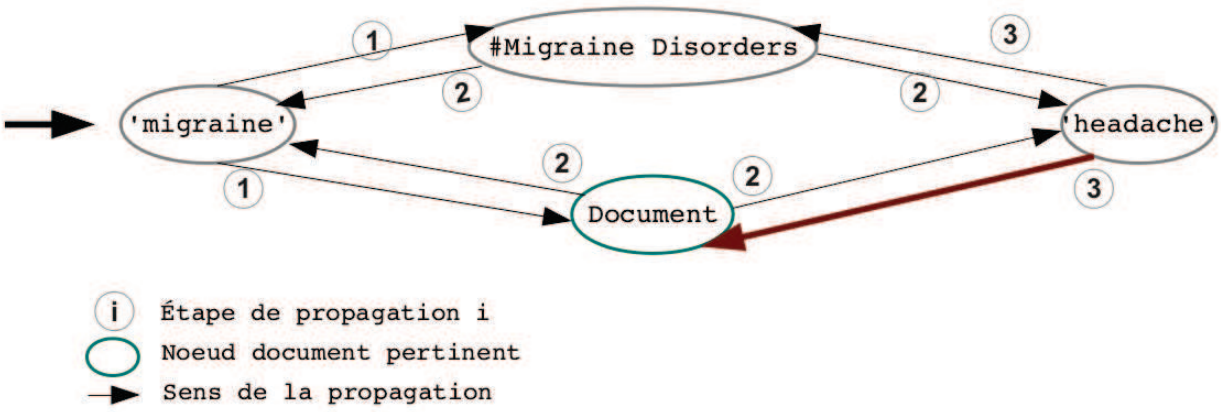

FIGURE 4.11 – Résolution de la synonymie et impact sur la précision

"headache". A la troisième itération, le nœud "headache" est activé et déclenche une propagation en tant que synonyme de "migraine", il contribue au renforcement de la valeur d'activation du document `Document`. Le renforcement de la valeur d'activation de `Document` permet d'améliorer son classement et par conséquent d'améliorer la précision.

Amélioration du rappel avec l'activation de nouveaux documents La figure 4.12 présente un deuxième exemple de propagation d'activation, où les deux termes "flu" et "grippe" sont synonymes. Le terme "flu" appartient au document `Document1`, alors que le terme "grippe" appartient à un autre document `Document2`. L'activation du nœud "grippe" par le concept `#Influenza` à la deuxième itération, permet par "contagion" l'activation d'un nouveau document pertinent pour la requête (`Document2`) mais qui ne pouvait être identifié en première instance parce qu'il ne contient pas le terme de la requête. Prendre en compte la synonymie améliore donc aussi le rappel.

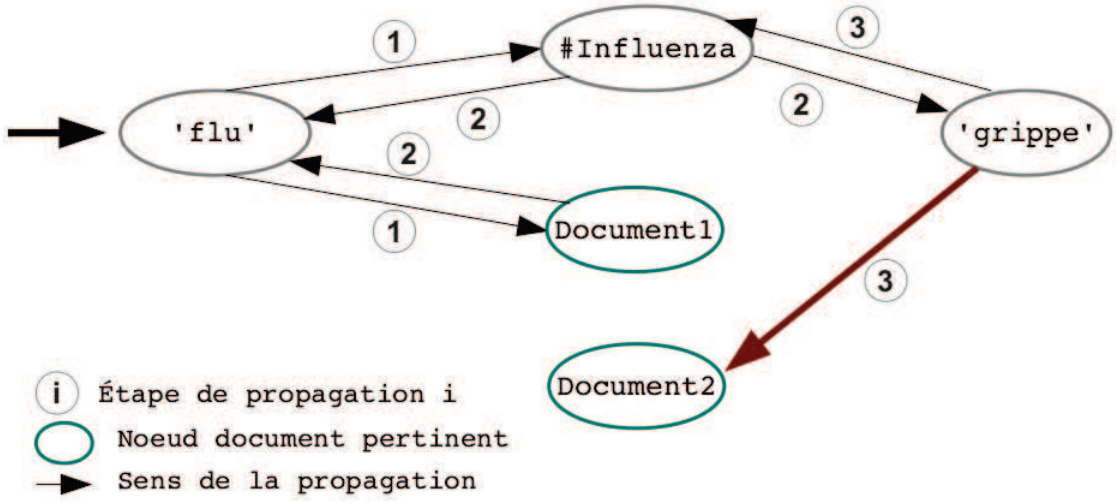

FIGURE 4.12 – Résolution de la synonymie et impact sur le rappel

4.3.3 Résolution du *term mismatch*

Le problème de *term mismatch* est un problème classique de la RI. Ce problème a été souligné au départ par Salton, qui considère que la collection documentaire et les utilisateurs des SRI possèdent des vocabulaires différents pour désigner un même concept [Salton, 1983]. Ce problème résulte principalement du fait que les fonctions d'appariement de la RI reposent généralement sur une mise en correspondance formelle qui ne prend pas en compte le sens des termes de la requête. Or un terme de la requête peut être formellement absent de la base documentaire mais néanmoins présent à travers ses synonymes. Les modèles classiques de RI ne permettent pas de gérer ces cas. Cela engendre d'autant plus de silence que le *term mismatch* est assez fréquent dans les domaines de spécialité, comme les domaines médical et juridique, où l'utilisateur ne maîtrise pas forcément le vocabulaire expert.

Différents travaux de l'état de l'art se sont intéressés à la résolution de ce problème de *term mismatch* [Crestani, 2000, Xu and Croft, 2000, Zhao, 2012], mais la propagation d'activation le résout facilement sans expansion de requêtes et sans enrichissement des documents : en exploitant simplement les relations terminologiques entre termes et concepts dans le graphe pondéré modélisant la collection documentaire.

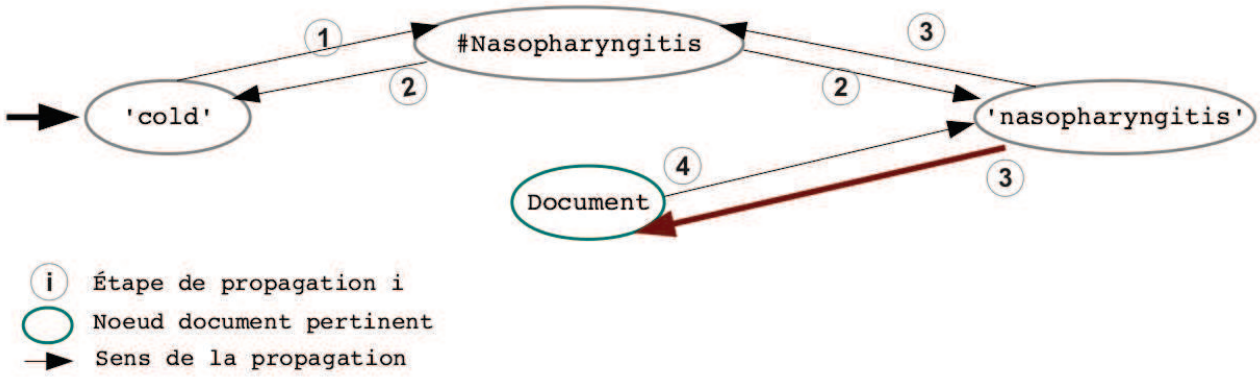

FIGURE 4.13 – Exemple de résolution du problème de *term mismatch*

La figure 4.13 présente un exemple de résolution du *term mismatch* sur un graphe pondéré relevant du domaine médical. Supposons que le terme "cold" exprime le besoin d'information de l'utilisateur et qu'il apparaisse dans le volet terminologique de l'ontologie mais pas dans les documents de la collection. A la première itération, le terme "cold" active le concept `#Nasopharyngitis` qu'il dénote. Ce concept active à son tour à la deuxième itération tous ses labels, qui sont donc synonymes de "cold". Le terme "nasopharyngitis " est activé et propage son activation, à la troisième itération, aux documents dans lesquels il se trouve, comme le document `Document` qui s'avère ainsi pertinent par rapport à la requête initiale. La résolution du *term mismatch* permet ainsi de renvoyer des documents pertinents même s'ils ne contiennent pas les termes de la requête et donc d'augmenter le rappel.

4.3.4 Extension de la couverture sémantique

Le défaut de couverture sémantique est un problème qui est souvent lié aux approches conceptuelles de RI sémantique. Les premiers travaux qui abordent cette question datent de 1986, avec [Croft, 1986]. En RIS, on observe en effet que l'introduction de sémantique peut dégrader les

résultats, notamment quand on interroge par les seuls concepts, si la ressource utilisée couvre mal le vocabulaire de la collection documentaire et le besoin de l'utilisateur (voir chapitre 2).

On parle souvent dans l'état de l'art de *"Knowledge incompletness"* [Fernandez et al., 2011]. Fernandez et *al.* propose de dégrader la RIS en RI (présentent des réponses sans sémantique) quand des connaissances du domaine font défaut. Ils considèrent en effet que la construction et le maintien de ressources sémantiques riches est une tâche très lourde.

Le modèle que nous présentons suit le même principe pour résoudre les problèmes de couverture, étant donné que le graphe pondéré représentant la collection documentaire est facile à interroger par différents points d'entrée (termes, concepts, documents, etc.), sans changer de système ni de formule de pondération, etc.

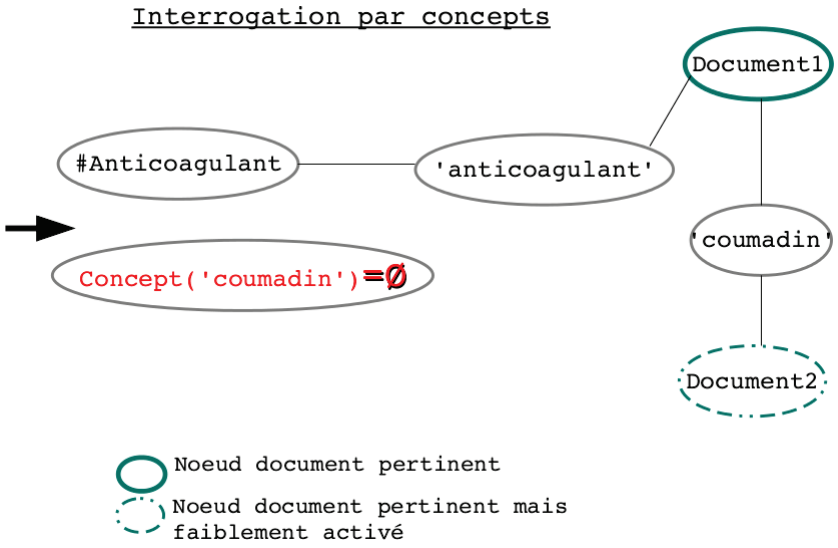

FIGURE 4.14 – Résolution de la couverture sémantique (interrogation par les concepts)

La figure 4.14 présente un exemple de défaut de couverture d'une ressource sémantique médicale. Si on interroge le graphe par les concepts, le terme "coumadin" qui n'est associé à aucun concept ne peut être pris en compte (ce qui revient à déformer l'expression du besoin de l'utilisateur) et le `Document2` ne se retrouve que faiblement activé alors même qu'il contient le terme "coumadin". En effet, sous l'effet de la distance sur le graphe, `Document2` n'est activé que tardivement.

Une solution consiste à interroger le graphe par les termes, comme cela est classique. La figure4.15 montre qu'à partir des deux termes "anticoagulant" et "coumadin", on peut activer les deux documents `Document1` et `Document2` dès la première étape de propagation. A la différence des méthodes de RIS de l'état de l'art cependant, dans notre modèle l'interrogation par les termes ne revient pas à dégrader le système de RIS en un système de RI classique. La sémantique continue à être prise en compte pour l'amélioration de la qualité de la RI mais à un autre niveau sur le graphe (à la deuxième itération). L'exemple de la figure 4.15 montre en effet que des phénomènes comme la co-occurrence ou la synonymie permettent d'activer un nouveau document pertinent, le document `Document3`.

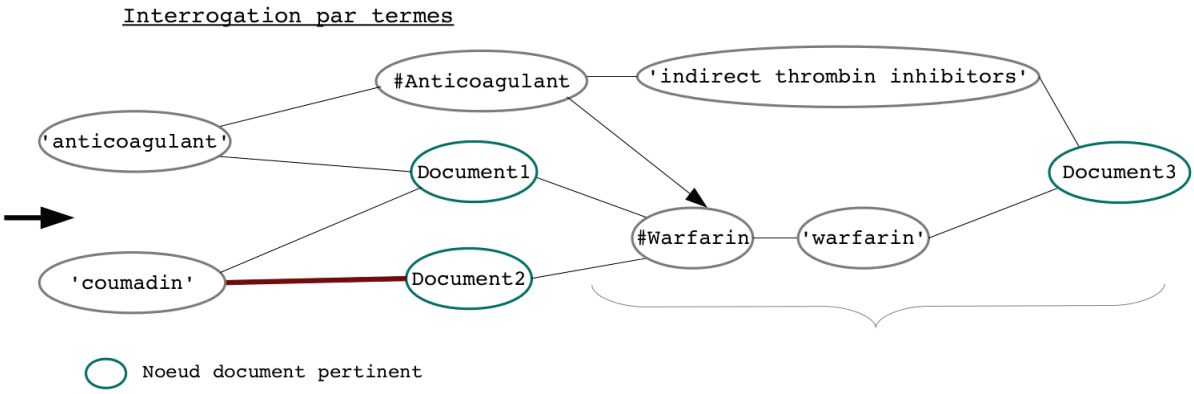

FIGURE 4.15 – Résolution de la couverture sémantique (interrogation par les termes)

4.4 Conclusion

Ce chapitre présente notre modèle de RIS à base de graphes et de propagation d'activation. Ce modèle offre les fonctionnalités de base de la recherche documentaire mais il propose aussi un ensemble de fonctionnalités sémantiques qui permettent de dépasser les limites connues de la RI(S).

De point de vue modélisation, nous proposons de représenter de manière unifiée la collection documentaire et les ressources sémantiques ou base de connaissances associées sous la forme d'un réseau sémantico-documentaire qui prend en compte l'ensemble des propriétés numériques (ex. les fréquences d'occurrences) et sémantiques (ex. les classes sémantiques) qui sont exploitées en RI et RIS. Pour ce faire, notre modèle combine les trois niveaux documentaire, terminologique et conceptuel : le réseau sémantico-documentaire peut comporter 3 types de nœuds (documents, termes et concepts) et 5 types de relations (relation d'occurrence, terminologique, d'annotation, ontologiques et inter-textuelles).

Ce réseau est traduit en un graphe pondéré. Différentes traductions sont possibles, qui correspondent à différentes configurations et modélisations du réseau sémantico-documentaire.

La mise en correspondance requête/documents est réalisée par un mécanisme de propagation d'activation. On propage la pertinence de proche en proche à partir des nœuds de la requête en calculant à chaque étape la valeur d'activation des nœuds atteints. La fonction d'activation que nous avons introduite s'apparente aux fonctions de mise en correspondance de la RI classique : elle tient compte de la distribution des termes dans la collection (*tf*, *df*, etc.) mais aussi des propriétés sémantiques (volet terminologique de l'ontologie, ambiguïté des concepts, etc.) à travers notamment les poids des arcs et la structure du graphe. La fonction d'activation assure aussi le contrôle de la propagation sur le graphe, notamment pour les graphes qui comportent des cycles : un nœud ne peut se propager (activer ses voisins) qu'une seule fois mais la propagation d'activation globale est itérative et synchrone (à l'étape k, tous les nœuds actifs se propagent en parallèle et la valeur d'activation d'un nœud dépend des valeurs d'activation obtenues par ses voisins à l'étape $k-1$). La première contrainte assure la terminaison de l'algorithme de propagation et la seconde le rend synchrone et déterministe.

Le graphe pondéré peut être interrogé de différentes manières (on peut prendre des termes, des concepts et/ou des documents comme points d'entrée) ce qui offre plusieurs formes d'accès à l'information. Pour rechercher des documents, on peut interroger le graphe par les termes ou les concepts et les valeurs d'activation obtenues représentent la pertinence des nœuds documents

au regard de la requête, ce qui permet de trier ou de filtrer les résultats.

Notre objectif était de reproduire la RI classique et de dépasser certaines des limites classiques de la RI(S). La modélisation et la fonction de mise en correspondance proposées reprennent les primitives de base d'une recherche documentaire et nous avons montré qu'elles offrent en outre des fonctionnalités sémantiques intéressantes qui exploitent à la fois une sémantique implicite (phénomènes de co-occurrence) et une sémantique explicite (par la couche conceptuelle du graphe). Nous avons en effet montré théoriquement que la seule prise en compte des classes sémantiques apporte déjà des solutions aux problèmes de la synonymie, du *term mismatch* ou du manque de couverture sémantique.

Notre modèle et ses fonctionnalités sémantiques doivent être testés et analysés en détail mais aussi éprouvés expérimentalement, sachant que l'on peut proposer différentes traductions d'un réseau sémantico-documentaire. C'est l'objet des deux chapitres qui suivent.

Chapitre 5

Premières expériences

Sommaire

Ce chapitre vise à tester le modèle de réseau sémantico-documentaire et le mécanisme d'appariement requêtes/documents par propagation d'activation proposés dans le chapitre précédant et à analyser leur fonctionnement dans le cadre de la recherche d'information. Nous utilisons pour ce faire un jeu de données de taille réduite, ce qui permet d'analyser l'approche proposée en détail, les questions d'échelle n'étant abordées que dans un deuxième temps (voir chapitre 6).

Comme indiqué plus haut, de multiples variantes de notre modèle de RIS peuvent être envisagées : on peut encoder différents types d'information dans le réseau, calculer différentes pondérations, exploiter différents mécanismes de propagation et utiliser différentes formes d'interrogation. Il n'est pas possible de jouer ici sur tous ces paramètres à la fois. Nous en fixons certains (la pondération et le mécanisme de propagation) pour mieux comprendre l'impact des autres et notamment le rôle des informations sémantiques dans le réseau. C'est souvent la richesse des données qui détermine la richesse du graphe (le nombre et le type des nœuds, les relations entre ces nœuds, la collection documentaire en tant que telle, la ressource sémantique et sa richesse, les annotations, etc.), mais il convient aussi de prendre en compte l'application et le besoin de l'utilisateur.

Les objectifs de ce chapitre sont multiples. D'un point de vue technique, il s'agit de comprendre le fonctionnement de notre modèle, notamment la manière dont l'activation se propage dans un graphe pondéré et la parenté entre ce mécanisme de propagation d'activation et un processus classique de RI. De point de vue fonctionnel, nous voulons nous assurer de la qualité de la recherche produite, c'est-à-dire (i) prouver que notre modèle de RIS unifié ne dégrade pas les résultats par rapport à un modèle traditionnel pour des processus de RI classique et (ii) explorer les potentialités de notre modèle en termes de fonctionnalités sémantiques (prise en compte de la co-occurence, gestion de la synonymie, problème de couverture et de *term mismatch*, etc.).

Pour ce faire nous utilisons un petit jeu de données test (documents et requêtes) lié au domaine de la cuisine et nous analysons différents scénarios de RI sur ces données. Ce chapitre présente dans un premier temps la description du jeu de test utilisé (section 5.1) et la mise en place de l'environnement logiciel et expérimental pour nos expériences (section 5.2). Nous comparons ensuite notre modèle à base de propagation avec un modèle classique de RI (section 5.3) et nous montrons les fonctionnalités sémantiques qu'il offre (section 5.4).

5.1 Description des données

De manière générale, un jeu de test (ou *benchmark*) en RI doit comporter une collection documentaire ou corpus, un ensemble de requêtes et les jugements de pertinence, c'est-à-dire, pour chaque requête, la liste des documents de la collection considérés comme pertinents. Un système de RIS intègre en plus une ou plusieurs ressources sémantiques externes telles que les thésaurus ou les ontologies. Les premières expérimentations que nous avons effectuées avaient une visée exploratoire. Elles ont porté sur le corpus de recettes de cuisine exploité dans [Bannour and Zargayouna, 2012, Bannour et al., 2016b, Bannour et al., 2016a], lequel a été enrichi et mis à jour.

5.1.1 Corpus documentaire

Le corpus documentaire que nous exploitons est le corpus fourni par la compétition internationale *Computer Cooking Contest* (*CCC 2009*) [28] pour évaluer la capacité des systèmes à sélectionner et éventuellement adapter des recettes de cuisine en fonction de requêtes d'utilisateurs. La base comporte 1 489 recettes qui sont rédigées en anglais et qui se présentent sous la forme de documents textuels au format XML. Ces documents ne sont que faiblement structurés à l'aide des éléments XML suivants (voir un exemple sur la figure 5.1) :

— le titre de la recette (`<TI>`),
— la liste des ingrédients (`<IN>`),
— la préparation (`<PR>`).

A ce petit corpus, nous avons ajouté 7 autres recettes à des fins expérimentales : il s'agit des documents 1490, 1491, 1492, 1493, 1494, 1495 et 1496. Notre collection documentaire comporte ainsi au total 1 496 documents.

5.1.2 Requêtes et jugements de pertinence

Notre objectif étant différent de celui du concours *CCC*, nous avons sélectionné certaines requêtes pour valider nos scénarios de RI(S) (les requêtes `REQ1` et `REQ4`), mais nous en avons aussi modifié certaines (`REQ2` et `REQ3`) et ajouté de nouvelles (`REQ5` et `REQ6`). Ce faisant, nous avons

28. *CCC* : Cette compétition est organisée par la communauté de Raisonnement à Partir de Cas : http ://www.wi2.uni-trier.de/shared/eccbr/ccc09/

```
<RECIPE>
<ID>23</ID>
<TI>Anadama Bread</TI>
<IN>3/4 c Water</IN>
<IN>1/4 c Molasses</IN>
<IN>3 tb Oil</IN>
<IN>2 1/2 c Flour</IN>
<IN>1/4 c Cornmeal</IN>
<IN>1 tb Brown sugar</IN>
<IN>1 ts Salt</IN>
<IN>2 ts Active dry yeast</IN>
<PR>Warm liquids to 80 to 100 degrees.
Place ingredients into breadmaker per
machines instructions. Use the sweet
mode cycle. Yield 1 med loaf.
</PR>
</RECIPE>
```

FIGURE 5.1 – Exemple de recette du corpus de cuisine

évité les requêtes nécessitant d'avoir recours à des mécanismes d'inférence ou de raisonnement [29], car cela sortait des objectifs que nous nous étions fixés. Chacune des requêtes dispose d'un ensemble de jugements de pertinence [Despres et al., 2009] que nous avons vérifiés ou constitués manuellement. Les requêtes comporte de 1 à 4 termes.

Les 6 requêtes de notre jeu de test sont les suivantes :

`REQ1` : "Cook an *Asian soup* with *leek*" (6 doc. pertinents)

`REQ2` : "Prepare a *fruit salad* with *lemon juice*" (6 doc. pertinents)

`REQ3` : "I would like to cook a *salad* with *kiwano*" (5 doc. pertinents)

`REQ4` : "Prepare a *cake* with *plum*" (6 doc. pertinents)

`REQ5` : "Prepare a *pizza* with *merguez*" (4 doc. pertinents)

`REQ6` : "Prepare *jambalaya*" (6 documents pertinents)

Les jugements de pertinence de cet ensemble de requêtes, ainsi que les termes pertinents dans chacun des documents pertinents sont présentés dans l'annexe B.

5.1.3 Ressource sémantique

Dans le cadre de notre évaluation, nous avons utilisé une ressource sémantique construite dans le cadre du projet TAAABLE [Badra et al., 2008]. Cette ressource est une ontologie formalisée en OWL qui décrit le domaine de la cuisine. Elle a été développée par des chercheurs de l'équipe Orpailleur du LORIA. Cette ontologie a été construite à partir du corpus de recette et du thésaurus de cuisine *Cook's Thesaurus* [30].

L'ontologie de cuisine est constituée de 4 509 concepts, organisés selon des liens hiérarchiques mais sans rôles. Les concepts les plus génériques sont les suivants :

`#dishorigin` : l'origine des recettes (asiatique, italien, etc.),

`#dishrole` : la destination de la recette (petit-déjeuner, diner, dessert, entrée, etc.),

`#dishtype` : le type d'une recette (salade, sauce, boisson, etc.),

29. C'est le cas notamment des requêtes qui portant sur les régimes ou spécifiant des aliments à éviter, qui nécessitent de prendre en compte la négation.

30. http ://foodsubs.com/.

`#ingredient` : les ingrédients utilisés dans les recettes (légume, viande, poisson, etc.).

Nous avons apporté quelques modifications à cette ontologie :

— nous avons étendu la liste de labels du concept `#lemon_juice` avec les labels "lime juice" et "orange juice" ;
— nous avons ajouté "fruit bowl" aux labels du concept `#fruit_salad` ;
— nous avons modifié la liste des labels du concept `#leek` en ajoutant le label "scallion" et en supprimant les labels de `#green_onion` ;
— nous avons supprimé les concepts `#lime_juice` et `#orange_juice` ;
— nous avons ajouté le label "souffle" au concept `#pizza`.

La figure 5.2 présente une représentation hiérarchique de l'ontologie de cuisine à une profondeur 2.

5.2 Environnement logiciel et dispositif expérimental

Les expériences de RI et *a fortiori* de RIS nécessitent la mise en place d'un solide environnement logiciel.

Nous avons enrichi la plate-forme *Terrier SIR* que nous avons proposée dans [Bannour and Zargayouna, 2012] (voir l'architecture dans la figure 5.3) par la plate-forme d'analyse de graphes *JUNG* (*Java Universal Network/Graph Framework*) [31]. *JUNG* permet de modéliser différents types de graphes (orienté, non orienté, etc.) et implémente de nombreux algorithmes connus sur les graphes tels que PageRank et HITS. Ce cadre expérimental nous a permis de mettre en place notre modèle unifié de RIS par propagation d'activation sur le graphe pondéré.

La mise en œuvre de ce modèle passe obligatoirement par la traduction du réseau sémantico-documentaire en un graphe pondéré avant d'y déclencher le processus de propagation d'activation. La construction du graphe pondéré profite quant à elle, de l'index classique de Terrier SIR et d'une annotation sémantique automatique par *SemEx Annotator*[32] au regard de l'ontologie de cuisine que nous exploitons (voir section 5.1.3).

Le graphe sans sémantique est alors composé des nœuds termes et documents, tel que les termes présentent les unités d'index de *Terrier IR* c-à-d, les termes du vocabulaire après avoir subit les analyses classiques (*tokenisation*, élimination des mots vides, analyse lexicale). Comme analyse lexicale on exploite la *racinisation* ou *troncature* (*stemming* en anglais), qui est implémenté dans la plate-forme *Terrier IR*. Quant aux arcs du graphe, on traduit notamment les relations d'occurrence terme/document auxquelles on associe des poids relatifs, qui reflètent la fréquence d'occurrence des termes dans les documents correspondants. Cette fréquence d'occurrence est une propriété intrinsèque qui est également extraite de l'index classique de *Terrier IR*.

Au graphe pondéré sémantique, on rajoute les connaissances explicitées dans l'annotation sémantique par *SemEx Annotator* et l'ontologie associée, à savoir les nœuds concepts et les relations terminologiques dans ce cadre de test.

Toutes ces connaissances sont codées dans une structure de type *Graph* proposée par la plate-forme de gestion des graphes *JUNG*, tel que des attributs sont affectés aux nœuds et arcs du graphe pondéré. Sur ce graphe, la propagation d'activation est ensuite déclenchée par une interrogation de l'utilisateur.

31. JUNG (*Java Universal Network/Graph Framework*) : http ://jung.sourceforge.net/
32. SemEx Annotator : http ://www-lipn.univ-paris13.fr/ szulman/Annotator/annotator.html

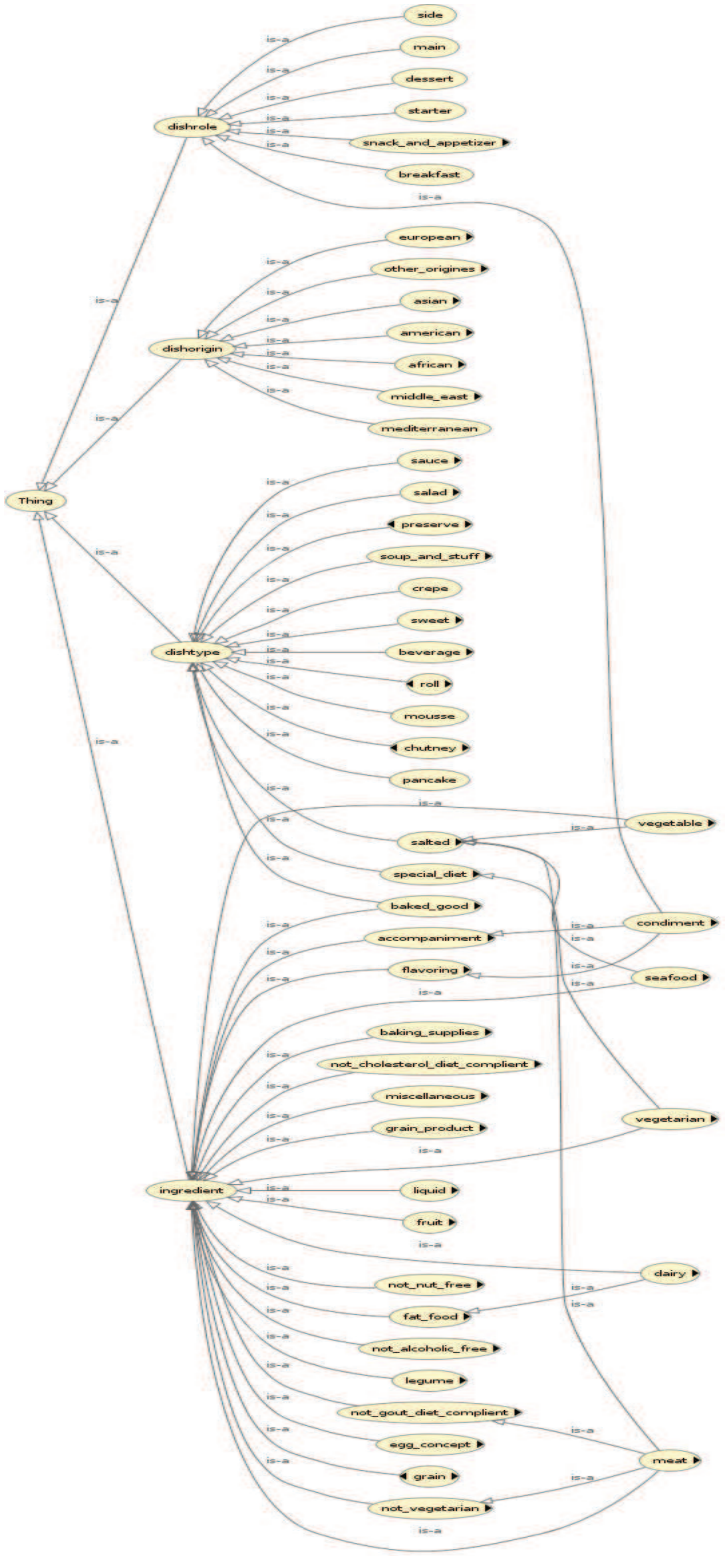

FIGURE 5.2 – Représentation hiérarchique de l'ontologie de cuisine

Représentation sous la forme de graphe

Chaque nœud du graphe possède des attributs statiques qui le caractérisent :

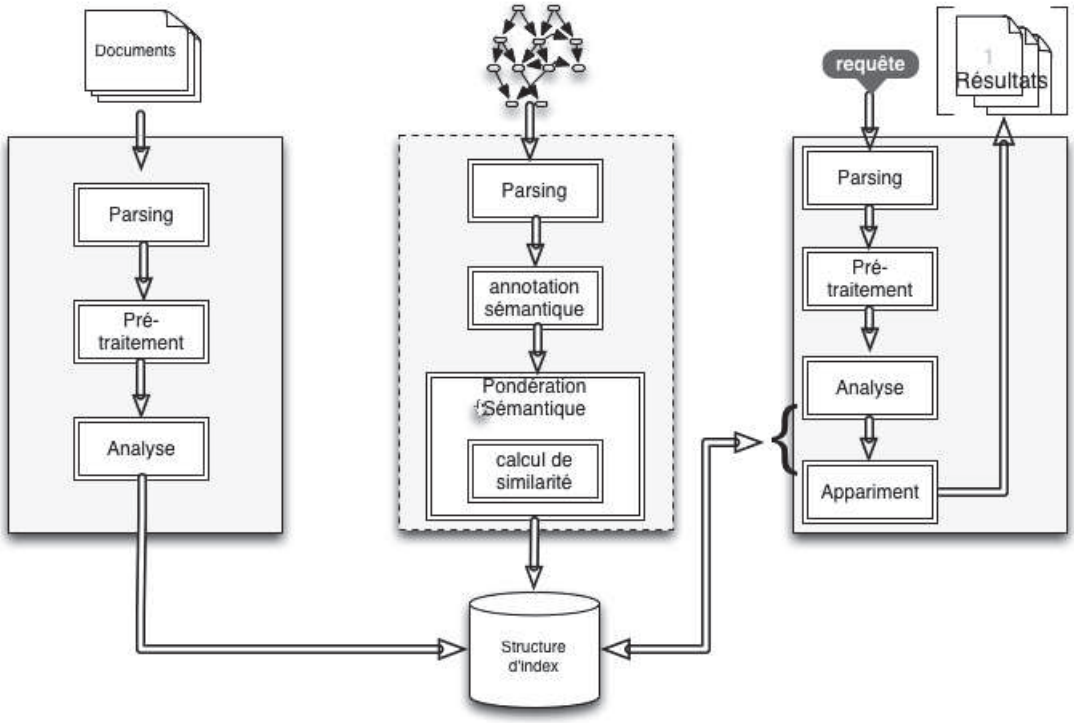

FIGURE 5.3 – Architecture de la plate-forme Terrier SIR

`Id` : identifiant du nœud ;

`Specif` : spécification du type du nœud (ex. `doc` pour les nœuds documents, `term` pour les nœuds termes, etc.) ;

`Name` : intitulé du nœud (ex. le nom du document pour les nœuds `doc`, le terme lui-même pour les nœuds `term`).

D'autres caractéristiques sont attachées aux nœuds mais elles évoluent au cours de la propagation d'activation. Il s'agit d'attributs techniques :

`State` : état du nœud au cours de la propagation d'activation (inactif, actif, désactivé) ;

`InitialWeight` : valeur d'activation initiale, qui est affectée au nœud au moment de l'interrogation du graphe par la requête utilisateur (entrée du graphe) ;

`Weight` : valeur d'activation qui peut être mise à jour à chaque étape de propagation (elle est égale à `InitialWeight` au départ).

Les arcs du graphe sans sémantique traduisent les relations d'occurrences entre les termes et les documents. On affecte à un arc un poids w qui dépend de la fréquence d'occurrence du terme t dans le document d. Deux types de poids sont pris en compte dans les expériences qui suivent selon si on introduit ou pas une normalisation :

poids simple : $w_s(t \to d) = w_s(d \to t) = tf(t,d)$ (sans normalisation)

poids normalisé : $w_n(t \to d) = w_n(d \to t) = \frac{tf(t,d)}{MaxTf(d)}$

où $tf(t,d)$ est la fréquence du terme t dans le document d (le nombre d'occurrences) et où $MaxTf(d)$ est la fréquence maximale de d, *i.e.* la fréquence du terme ayant le plus d'occurrences dans le document d.

Interrogation du graphe

L'interrogation consiste à sélectionner les nœuds du graphe de la collection qui sont activés à partir du graphe de la requête. Pour cela, on projette le graphe requête dans le graphe documentaire et on active les nœuds du graphe documentaire qui correspondent à des nœuds du graphe requête. L'activation initiale des nœuds correspondant aux termes de la requête est donnée par la formule de poids normalisé appliquée cette fois à la requête (et pas à un document). Ces nœuds ont comme valeur d'activation initiale leur fréquence d'occurrences dans la requête normalisée : $w_q(t,q) = \frac{tf(t,q)}{MaxTf(q)}$.

Nous ne nous préoccupons pas *à priori* des temps de construction et de propagation étant donné la taille réduite du corpus jeu de test, ni d'ailleurs de la modularité du système proposé. Ces aspects seront étudiés lors du passage à l'échelle (voir section 6.2 du chapitre 6).

5.3 Comparaison entre la RI par propagation d'activation et le modèle vectoriel

L'objectif de la recherche d'information sémantique est d'améliorer la qualité de la RI, c'est-à-dire de mieux répondre au besoin en information de l'utilisateur en prenant en compte le sens de sa requête au-delà des mots qui la composent.

Il faut donc vérifier dans un premier temps que les méthodes de RIS proposées ne dégradent pas la qualité de la recherche par rapport aux modèles classiques de RI. C'est ce que nous faisons dans cette section : nous nous assurons que notre modèle unifié reproduit la RI classique et préserve les performances d'un modèle algébrique classique, tel que le modèle vectoriel. Nous appliquons notre modèle à propagation d'activation sans introduire de fonctionnalité sémantique et nous comparons ses résultats avec ceux obtenus par le modèle vectoriel pondéré par *BM25*.

L'évaluation que nous effectuons se fonde sur les métriques classiquement utilisées en RI, notamment la *MAP* et la *R-PREC*, en plus du *Rappel* et de la *Précision*.Les mesures présentées dans ce chapitre ne doivent cependant pas être interprétées comme des mesures d'évaluation en tant que telles parce qu'elles sont calculées sur trop peu de données ; elles ne servent qu'à donner des points de comparaison pour confronter les différentes expériences.

5.3.1 Modèle vectoriel pondéré par *BM25* (*Baseline 1*)

Nous prenons comme référence de la RI classique, le modèle vectoriel, avec comme formule de pondération *Okapi* (*BM25*) et comme fonction de correspondance le produit scalaire. Ce modèle est implémenté dans la plate-forme Terrier SIR [Bannour and Zargayouna, 2012].

La pondération d'un terme t de la requête dans un document d par $BM25$ est :

- proportionnelle à la fréquence d'occurrence $tf(t,d)$,
- inversement proportionnelle à la fréquence documentaire $idf(t)$,
- inversement proportionnelle à la taille du document $taille(d)$ rapportée à la taille moyenne des documents de la collection.

Les résultats obtenus pour chacune des requêtes présentées dans la section 5.1.2 à l'aide de cette formule de pondération sont décrits dans le tableau 5.1.

Requêtes	REQ1	REQ2	REQ3	REQ4	REQ5	REQ6	Total
R-PREC	0.33	0.5	0.2	0.5	0.25	0.66	0.41
MAP	0.26	0.53	0.32	0.47	0.37	0.66	0.44
Rappel	(6/6)	(6/6)	(5/5)	(5/6)	(3/4)	(4/6)	29/33

TABLE 5.1 – Résultats de la RI classique (*BM25*) du SRI Terrier IR

5.3.2 Propagation d'activation sans sémantique (*Baseline 2*)

Afin de simuler une recherche d'information classique, nous représentons tout d'abord le réseau sémantico-documentaire sous la forme d'un graphe pondéré tel que :

- les nœuds du graphe correspondent aux termes du vocabulaire [33] et aux documents de la collection ;
- les arcs du graphe représentent les liens terme/document et document/terme et ils sont valués (poids w) ;
- la propagation d'activation sur ce graphe est contrôlée par la condition d'arrêt énoncée dans le chapitre précédent 4, qui empêche qu'un nœud s'active plus d'une fois même si les valeurs d'activation des nœuds peuvent être réévaluées plusieurs fois sous l'influence de leurs voisins.

Comme la RI suppose qu'on représente de la même manière les documents et les requêtes, ces dernières sont elles-aussi représentées comme des graphes mais ces graphes-requêtes sont réduits aux éléments de la requête (nœuds termes de la requête)

Expérimentations

Une première expérimentation, `E1`, a été réalisée en prenant les termes simples du vocabulaire comme nœuds termes (`TS`) et en pondérant les arcs du graphe par w_s. Les résultats sont présentés dans le tableau 5.2. Une deuxième expérience, `E2`, a été menée en considérant les termes simples

Requêtes	REQ1	REQ2	REQ3	REQ4	REQ5	REQ6	Total
R-PREC	0.33	0.0	0.2	0.33	0.0	0.66	0.25
MAP	0.26	0.16	0.2	0.5	0.16	0.82	0.35
Rappel	(5/6)	(6/6)	(5/5)	(6/6)	(3/4)	(**6**/6)	31/33

TABLE 5.2 – Résultats de la RI classique par propagation d'activation (expérience `E1`)

comme précédemment (`TS`), mais en normalisant les poids des arcs ($w_n \in [0, 1]$). Les résultats figurent dans le tableau 5.3. Une troisième expérimentation, `E3`, a été effectuée en gardant les

Requêtes	REQ1	REQ2	REQ3	REQ4	REQ5	REQ6	Total
R-PREC	**0.66**	**0.33**	**0.4**	**0.5**	0.0	0.66	**0.43**
MAP	**0.5**	**0.26**	**0.51**	0.5	0.12	**0.9**	**0.46**
Rappel	(5/6)	(6/6)	(5/5)	(6/6)	(3/4)	(6/6)	31/33

TABLE 5.3 – Résultats de la RI classique par propagation d'activation (expérience `E2`)

conditions expérimentales de `E2`, mais en ajoutant au graphe pondéré des nœuds correspondant aux termes composés du vocabulaire de la collection documentaire et des requêtes (`TC`). Par

33. Union du vocabulaire de la collection documentaire et de celui des requêtes.

exemple, la requête `REQ2`[34] qui est caractérisée par 4 nœuds termes (`fruit`, `salad`, `lemon` et `juice`) dans l'expérience `E2`, est représentée aussi par les nœuds termes `fruit salad` et `lemon juice` dans l'expérience `E3`. Les résultats de `E3` sont donnés dans le tableau 5.4.

Requêtes	REQ1	REQ2	REQ3	REQ4	REQ5	REQ6	Total
R-PREC	0.66	**0.5**	0.4	0.5	0.0	0.66	**0.46**
MAP	0.5	**0.61**	0.51	0.5	0.12	0.88	**0.52**
Rappel	(5/6)	(6/6)	(5/5)	(6/6)	(3/4)	(6/6)	31/33

TABLE 5.4 – Résultats de la RI classique par propagation d'activation (expérience `E3`)

Comparaison des expériences `E1` et `E2`

L'objectif de cette comparaison est de montrer l'impact du choix d'une formule de pondération pour les arcs sur la propagation d'activation, étant donné que les valeurs d'activation des nœuds du graphe dépendent directement de ce facteur (voir la formule de propagation d'activation dans le chapitre 4).

On observe, sur presque toutes les requêtes, une amélioration notable de la qualité de la recherche avec la mesure de poids normalisée de l'expérimentation `E2` :

— on note une amélioration de la *R-PREC* pour les requêtes `REQ1`, `REQ2`, `REQ3` et `REQ4`[35] : certains documents pertinents remontent dans le classement et se retrouvent classés parmi les R[36] premiers documents retournés ;

— la *MAP* augmente pour toutes les requêtes, même pour les requêtes comme `REQ6`[37], où la R-PREC n'est pas améliorée ; de fait, des documents pertinents peuvent remonter dans le classement mais sans pour autant atteindre le rang R qui leur permettrait d'être pris en compte dans la R-PREC.

— le nombre de documents pertinents retournés reste constant.

Sur ce petit jeu de test, la comparaison des expériences `E1` et `E2` montre qu'il est important de normaliser les valeurs de poids des arcs terme/document et document/terme. Cette normalisation a un impact positif sur la qualité de la recherche en termes de *MAP* et *R-PREC*.

Dans la suite, nous introduisons systématiquement une normalisation et nous utilisons la formule suivante : $w_n = \frac{tf(t,d)}{MaxTf(d)}$.

Comparaison des expérimentations `E2` et `E3`

Il s'agit cette fois de montrer l'impact du choix des nœuds d'entrée du graphe (*i.e.* les nœuds initialement activés) sur la propagation d'activation et l'intérêt de prendre en considération les termes composés des requêtes dans cette phase d'initialisation.

On fait les observations suivantes :

— La *MAP* est améliorée pour la requête `REQ2`[38], qui est l'unique requête comportant des termes composés : deux documents pertinents – `doc388` (`fruit salad(1)`) et `doc585`

34. "Prepare a *fruit salad* with *lemon juice*" (6 doc. pertinents)
35. `REQ1` : "Cook an *Asian soup* with *leek*" (6 doc. pertinents)
`REQ2` : "Prepare a *fruit salad* with *lemon juice*" (6 doc. pertinents)
`REQ3` : "I would like to cook a *salad* with *kiwano*" (5 doc. pertinents)
`REQ4` : "Prepare a *cake* with *plum*" (6 doc. pertinents)
36. R est dépendant de la requête : c'est le nombre de documents jugés pertinents pour une requête.
37. `REQ6` : "Prepare *jambalaya*" (6 doc. pertinents)
38. `REQ2` : "Prepare a *fruit salad* with *lemon juice*" (6 doc. pertinents)

(`fruit salad(1)`, `lemon juice(1)`) – sont toujours parmi les R premiers documents retournés mais ils remontent dans le classement parce qu'ils contiennent des termes composés de la requête.
— La *R-PREC* est également améliorée pour la requête `REQ2` : le document `doc904` (`fruit salad(1)`) figure à un meilleur rang et passe parmi les R premiers documents retournés, ce qui améliore la *R-PREC*.

Les nœuds des requêtes initialement activés sont les points d'entrée au graphe à partir desquels la propagation d'activation démarre. Prendre en compte les nœuds associés aux termes composés dans le graphe requête améliore la qualité de la recherche en termes de *MAP* et *R-PREC*.

Dans les expériences suivantes, nous considérons cette dernière expérimentation (`E3`) comme référence pour tous les scénarios de propagation d'activation. Nous confrontons cette 2^e référence (*Baseline 2*) au modèle vectoriel (*Baseline 1*) dans un premier temps (section 5.3.3) puis nous comparons de nouvelles expériences à cette *Baseline 2* pour montrer l'apport de la sémantique (section 5.4).

5.3.3 Prise en compte de la co-occurence (*Baseline 1* vs. *Baseline 2*)

Avant de nous interroger sur l'efficacité du modèle unifié de RIS mis en place, nous comparons notre modèle par rapport à celui de la RI classique. En effet, le modèle vectoriel classique a fait ses preuves : il repose sur un modèle algébrique, il a été largement éprouvé et il passe l'échelle sur les grandes collections de test. Il s'agit donc de montrer que notre modèle peut être aussi efficace que le modèle vectoriel, quand le réseau sémantico-documentaire se ramène à un simple graphe pondéré sans sémantique.

Pour ce faire, nous comparons la *Baseline 1* (voir tableau 5.1) qui représente un modèle vectoriel classique pondéré par *BM25* (voir la sous-section 5.3.1) et la *Baseline 2* (tableau 5.4) où l'on a un scénario de propagation d'activation sur un graphe pondéré sans sémantique (voir la sous-section 5.3.2).

Observations générales

Les résultats sont globalement comparables : la propagation d'activation sur un graphe sans sémantique simule assez bien le modèle vectoriel.

Dans le détail, on note une amélioration du classement de certains documents pertinents pour les requêtes `REQ1` et `REQ3`[39]. Quand les documents atteignent les R premiers documents pertinents, cela améliore la *R-PREC*. De nouveaux documents pertinents peuvent aussi être retrouvés (c'est le cas des requêtes `REQ2`, `REQ4`, `REQ6`[40]). Mais comme ils figurent à un rang inférieur à R, seule la *MAP* augmente (la *R-PREC* reste constante).

On remarque en outre que :
— ces améliorations de classement sont d'autant plus élevées que la taille des documents pertinents est réduite (la taille du document est prise en compte dans notre modèle) ;
— la *MAP* et *R-PREC* peuvent être à l'inverse dégradées pour des requêtes comme `REQ5`[41] auxquelles très peu de documents pertinents sont associés dans les jugements de pertinence (4) et qui se retrouvent de ce fait moins bien classés.

39. `REQ1` : "Cook an *Asian soup* with *leek*" (6 doc. pertinents)
`REQ3` : "I would like to cook a *salad* with *kiwano*" (5 doc. pertinents)
40. `REQ2` : "Prepare a *fruit salad* with *lemon juice*" (6 doc. pertinents)
`REQ4` : "Prepare a *cake* with *plum*" (6 doc. pertinents)
`REQ6` : "Prepare *jambalaya*" (6 doc. pertinents)
41. `REQ5` : "Prepare a *pizza* with *merguez*" (4 doc. pertinents)

— l'ordre des documents retournés change parfois légèrement (ex. REQ2) parce que la distribution des termes de la requête dans le document n'est pas prise en compte de la même manière du fait de la normalisation de la fréquence d'occurrence utilisée (valeur des arcs terme/document et document/terme).

Rôle de la co-occurrence

Les améliorations de classement sont principalement dues au phénomène de *co-occurence*.

La co-occurrence fait aussi parfois gagner en rappel pour certaines requêtes :

Pour la requête REQ4, un document pertinent supplémentaire est retrouvé (doc1323) du fait de la co-occurence de certains nœuds termes tels que "peach", "shortcake", "biscuit" autour de "cake" et "plum". Pour la requête REQ6, deux documents pertinents supplémentaires sont retrouvés grâce à la co-occurrence de certains termes du vocabulaire avec "jambalaya"[42] : ces deux documents, doc1495 et doc1496, sont pertinents par rapport à la requête REQ6 mais ils ne contiennent pas le terme "jambalaya", à la différence des 4 autres documents pertinents. En revanche, ils contiennent des termes comme "rice" et "sausage" qui sont des ingrédients du "jambalaya" : la co-occurrence de ces deux termes, dont les valeurs d'activation deviennent importantes, permet dans notre approche de retrouver les documents doc1495 et doc1496. Dans l'ordre des nœuds termes pertinents sur le graphe, on trouve ainsi "jambalaya", "rice", "pepper", "chicken", "sausage". Mais comme les rangs de ces deux documents (7 et 9) restent inférieurs à R[43], la *R-PREC* est constante, mais la *MAP* augmente, et passe de 0.66 à 0.88. La figure 5.4 schématise la propagation d'activation pour cette requête et met en évidence le rôle de la co-occurrence.

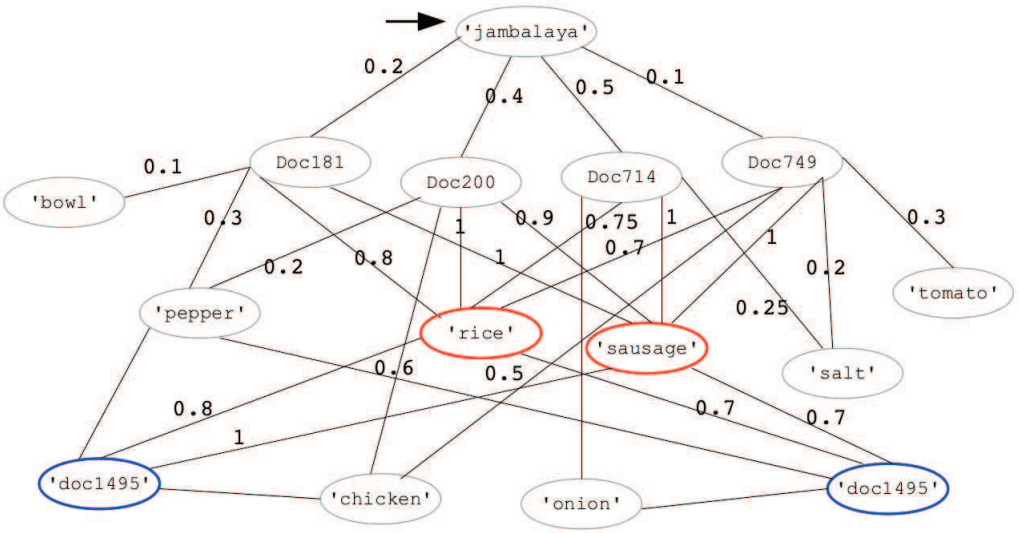

FIGURE 5.4 – Exemple de propagation d'activation et impact de la co-occurrence : les documents doc1495 et doc1496 ne sont pas activés directement par le terme de la requête "jambalaya" qu'ils ne contiennent pas mais indirectement par les termes "rice", "pepper", "sausage" qui co-occurrent avec "jambalaya" dans les documents doc181, doc200, doc714, doc749.

42. Il s'agit d'une préparation de riz et de saucissons.
43. R est égal à 6 pour la requête REQ6

5.3.4 Conclusion

La comparaison du modèle de RI classique (*Baseline 1*) et de notre modèle de base sans sémantique (*Baseline 2*) sur un aussi petit jeu de données ne peut pas permettre de tirer de vraies conclusions, mais elle permet néanmoins de mieux comprendre quels facteurs entrent en jeu dans la propagation d'activation sur le graphe :

- la co-occurrence des termes des documents avec ceux de la requête permet d'améliorer le rappel ou le classement des documents pertinents ;
- la taille des documents tend à faire remonter dans le classement les documents pertinents les plus courts ;
- la valeur de pondération des arcs et la formule de normalisation utilisée influent sur la manière dont la distribution des termes dans les documents est prise en compte.

5.4 Apport des classes sémantiques

Ce qui précède montre que même sur un graphe sans sémantique, le modèle par propagation d'activation prend en compte une part de sémantique : une sémantique latente qui s'exprime *via* la co-occurrence. et qui renforce l'efficacité de notre modèle de graphe pondéré.

On peut aussi introduire explicitement la sémantique dans le graphe en enrichissant le réseau sémantico-documentaire avec une couche sémantique, c'est-à-dire en ajoutant simplement des classes sémantiques au graphe pondéré. On nomme désormais ce modèle par : *modèle document/terme/concept* et on le note DTC. Ce modèle enrichi permet de prendre en charge d'autres fonctionnalités sémantiques (décrites dans le chapitre 2) :

- Prendre en compte la *synonymie* : deux nœuds termes sont considérés synonymes, s'il appartiennent à une même classe sémantique ou dénotent le même concept et il est important de pouvoir retrouver les documents contenant les termes synonymes des termes de la requête.
- Résoudre le problème de *term-mismatch* [Crestani, 2000, Xu and Croft, 2000, Zhao, 2012] : ce problème a été pointé au départ dans [Salton, 1983] et il se pose quand, pour un même concept, le vocabulaire de la collection documentaire et celui des utilisateurs des SRI diffèrent. Les fonctions d'appariement statiques ne permettent en effet pas d'extraire le sens du terme de la requête ni d'atteindre les documents qui "utilisent d'autres mots".
- Résoudre le *manque de couverture sémantique du vocabulaire* : quand un terme de la requête ne possède pas un concept qui le dénote dans l'ontologie "terme orphelin" (un problème pour les scénarios de RIS avec interrogation par les concepts uniquement).

5.4.1 Mise en place du graphe pondéré

Le réseau sémantico-documentaire enrichi (modèle DTC) se traduit en un graphe pondéré composé des éléments suivants :

- les *nœuds* représentent les termes simples (`TS`) et composés (`TC`) du vocabulaire (requêtes et collection) mais aussi les concepts du domaine provenant de l'ontologie de cuisine (classes sémantiques) ;
- les *arcs* traduisent :
  - les relations d'occurrences terme/document et document/terme, avec comme valeur la fréquence d'occurrence normalisée, w_n.

— les relations terminologiques terme/concept et concept/terme, résultant d'une annotation sémantique automatique par *SemEx Annotator* ou les relations entre un concept et les labels qui le dénote (dans ce cas, le poids de l'arc est 1).

Comme précédemment, l'interrogation consiste à activer dans le graphe les nœuds termes qui figurent dans la la requête.

5.4.2 Prise en compte de la synonymie

En exploitant les classes sémantiques d'une ontologie lexicalisée, on peut résoudre les problèmes de synonymie. En effet, si un concept est dénoté par deux labels, on peut considérer que ces deux labels sont synonymes sans pour autant avoir besoin d'établir de lien direct entre eux.

Analysons ce mécanisme sur les requêtes suivantes qui présentent des problèmes de synonymie [44] :

`REQ1` : considérons que "leek" et "scallion" sont deux labels du concept `#leek`.

`REQ2` :

supposons que "fruit salad" et "fruit bowl" sont deux labels du concept `#fruit_salad` mais aussi que "lemon juice", "orange juice" et "lime juice" sont des labels du concept `#lemon_juice` ;

Nous utilisons en outre la requête `REQ4` [45] comme requête témoin pour pouvoir observer un éventuel effet de bord de l'ajout de la couche sémantique sur les requêtes ne nécessitant pas de sémantique.

	Baseline 2			*DTC*		
Requêtes	REQ1	REQ2	REQ4	REQ1	REQ2	REQ4
R-PREC	0.66	0.5	0.5	0.66	**0.66**	0.5
MAP	0.5	0.61	0.5	0.50	**0.86**	0.5
Rappel	(5/6)	(6/6)	(6/6)	(5/6)	(6/6)	(6/6)

TABLE 5.5 – Résultats de la propagation d'activation sur le graphe du modèle *DTC* : `REQ1`, `REQ2`, `REQ4`

Le tableau 5.5 montre une amélioration de la *R-PREC* et de *MAP* pour la requête `REQ2` [46] : le document `doc1322` contient les synonymes "fruit bowl" (1 occ.) et "fruit lime" (2 occ.) qui le font remonter parmi les R premiers documents renvoyés, ce qui augmente la *R-PREC*. Le classement des documents pertinents `doc1467` et `doc544` s'améliore également parce qu'ils contiennent "fruit bowl" (1 occ.), mais comme leurs rangs sont toujours inférieurs au nombre de documents pertinents de `REQ2`, seule la *MAP* s'améliore.

La *R-PREC* et la *MAP* restent constantes pour les autres requêtes :

`REQ1` : il n'y a pas d'amélioration pour cette requête, même si les documents `doc335` et `doc775` contiennent "scallion" (resp. 3 et 2 occ.), qui est un synonyme de "leek". Cela s'explique par la taille de ces deux documents, qui sont assez longs, et la forte connexion du nœud "scallion" dans le graphe [47], ce qui lui confère un facteur *df* élevé.

44. `REQ1` : "Cook an *Asian soup* with *leek*" (6 documents pertinents)
`REQ2` : "Prepare a *fruit salad* with *lemon juice*" (6 documents pertinents)
45. `REQ4` :"Prepare a *cake* with *plum*" (6 documents pertinents)
46. `REQ2` : "Prepare a *fruit salad* with *lemon juice*" (6 documents pertinents)
47. C'est un ingrédient très répandu dans les recettes.

`REQ4` : l'ajout de la sémantique ne dégrade pas le résultat témoin.

La figure 5.5 est un exemple de résolution de la synonymie.

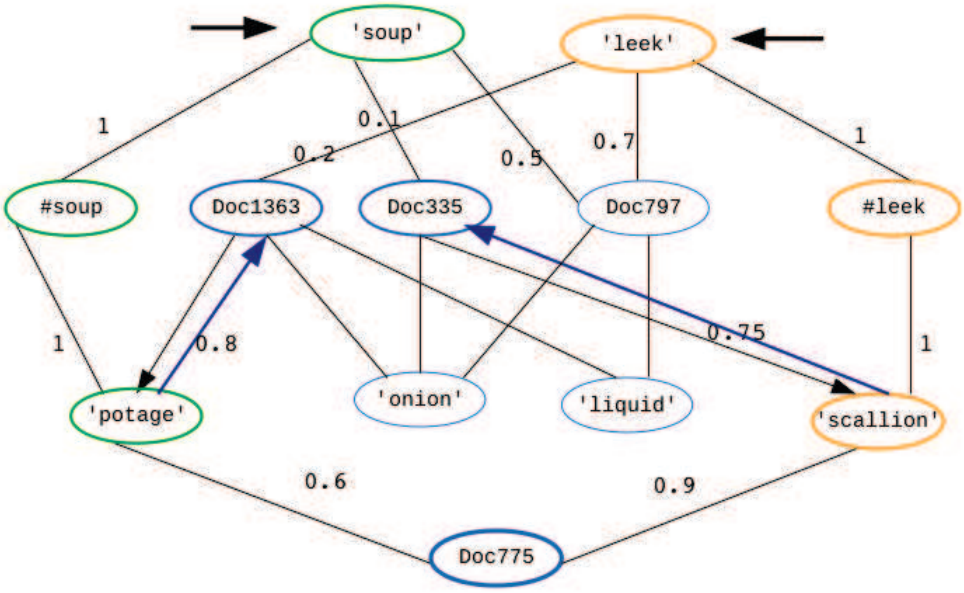

FIGURE 5.5 – Exemple de propagation d'activation et résolution de la synonymie : Les documents `doc1363` et `doc335` sont activés par les termes de la requête, respectivement "soup" et "leek' puis leurs valeurs d'activation sont renforcées respectivement par "potage" synonyme de "soup" et "scallion" synonyme de "leek". Le document `doc775` est activé par les deux synonymes "potage" et "scallion".

5.4.3 Résolution du *term mismatch*

L'introduction des classes sémantiques dans notre modèle vise également à résoudre le problème du *term mismatch*.

Considérons les requêtes suivantes [48] :

`REQ3` : "kiwano" est un terme de la requête qui est absent de la collection documentaire mais qui est associé au concept `#kiwano`, lequel fait partie du modèle sémantique et possède plusieurs labels ("horned melon", "jelly melon", "melano", "African Horned Cucumber", "kiwano melon") dont on trouve des occurrences dans les textes. On est face à un problème de *Term mismatch*.

`REQ5` : le terme "merguez" est également absent dans les documents, mais c'est un des labels du concept `#sausage` qui fait partie du modèle sémantique.

Les résultats qui figurent dans le tableau 5.6, montrent une amélioration de la *R-PREC* et de la *MAP* pour la requête `REQ3`. En effet, le classement des documents pertinents, `doc1490` et `doc1492`, s'améliore et ils atteignent les R premiers documents, car ils contiennent les termes "horned melon" (3 occ.) et "jelly melon" (1 occ.) pour le premier et "african Horned cucumber"

48. `REQ3` : "I would like to cook a *salad* with *kiwano*" (5 doc. pertinents)
`REQ5` : "Prepare a *pizza* with *merguez*" (4 doc. pertinents)

	Baseline 2		*DTC*	
Requêtes	REQ3	REQ5	REQ3	REQ5
R-PREC	0.41	0.0	**0.81**	0.0
MAP	0.5	0.12	**0.8**	**0.19**
Rappel	(5/5)	(3/4)	(5/5)	**(4/4)**

TABLE 5.6 – Résultats de la propagation d'activation sur le graphe du modèle *DTC* : REQ3, REQ5

(2 occ.) et "melano" (2 occ ;) pour le second. Les autres documents pertinents gardent leur rang d'origine car ils contiennent également des labels du concept #kiwano.

La *R-PREC* est constante et la *MAP* est légèrement améliorée pour la REQ5. On constate une amélioration du rang de presque tous les documents pertinents retournés, qui contiennent le terme "sausage" synonyme de "merguez" (doc772, doc1053). Un nouveau document pertinent a été retrouvé (il contient 5 occ. de "sausage"), d'où une augmentation du rappel. La précision s'améliore également et la *MAP* passe de 0.12 à 0.19. Cependant, la *R-PREC* ne s'améliore pas du fait de la forte connexion du nœud "sausage"[49] (*df* élevé) et du fait du petit nombre de jugements de pertinence pour la requête (4 documents pertinents).

Le problème de *Term mismatch* est un problème bien identifié en RI. Il vient du fait que les modèles classiques recherchent uniquement dans les documents les termes identiques à ceux de la requête, sans tenir compte des termes qui ont la même signification sans avoir la même forme. Notre modèle de RIS résout ce problème de *term mismatch* en repérant des termes synonymes ou quasi-synonymes *via* la propagation d'activation dans la couche sémantique. La résolution du *term mismatch* reste cependant sensible aux conditions expérimentales, telles que la taille des documents, la fréquence documentaire des synonymes, le nombre de documents pertinents.

5.4.4 Résolution du problème de couverture sémantique

Les méthodes de RIS sont souvent confrontées au fait que la ressource sémantique qu'elles utilisent ne couvre qu'imparfaitement le vocabulaire des documents et des requêtes qu'elles ont à traiter. On a en quelque sorte affaire à des *termes orphelins*. On rencontre aussi ce problème avec les méthodes d'accès à l'information dans le WS, où les langages formels d'interrogation des bases de connaissances ne tiennent pas compte de tous les termes que l'utilisateur peut employer.

Nous développons ici l'exemple de la requête REQ6[50], où "jambalaya" est un mot-clé sans correspondant conceptuel dans le modèle sémantique. Le tableau 5.7 présente une comparaison des résultats obtenus pour cette requête.

	Baseline 2	Baseline 2 + classes sémantiques : *DTC*
Requêtes	REQ6	REQ6
R-PREC	0.66	**0.83**
MAP	0.88	0.89
Rappel	6/6	6/6

TABLE 5.7 – Résultats de la propagation d'activation sur le graphe du modèle *DTC* : REQ6

49. Ingrédient existant dans plusieurs recettes.
50. REQ6 : "Prepare *jambalaya*" (6 doc. pertinents)

La R-PREC augmente pour la requête `REQ6` parce qu'il y a un document pertinent qui remonte dans le classement. En effet, en plus du phénomène de co-occurence (voir section 5.3.3), qui met en évidence la proximité sémantique des termes "rice" et "sausage" et de "jambalaya", la propagation d'activation met en exergue les concepts `#rice` et `#sausage`, et des concepts proches comme `#andouille` (sorte de saucisson), `#long-grain_white_rice` (sorte de riz), etc.

L'exploitation des classes sémantiques permet donc de retrouver par *co-occurrence conceptuelle*, certains concepts proches de `#jambalaya`. On retrouve ainsi et par ordre de pertinence décroissante les nœuds `#rice`, `#andouille`, `#cajun` (épice qui parfume "jamabalaya"), `#bacon_dripping`, `#sausage`, `#long-grain_white_rice`, `#white_rice`, `#louisiana_hot_sauce`, `#bacon`, `#long-grain_rice`, etc.

Un modèle de RIS reposant sur l'interrogation par mot-clés aurait permis d'extraire les documents contenant "jambalaya" mais pas d'exploiter le modèle sémantique associé. Les méthodes comme celles de [Zargayouna, 2005] et [Boubekeur and Azzoug, 2013] proposent de prendre en compte des termes orphelins à travers la simple interrogation par mot-clé, mais ne tiennent pas compte des synonymes éventuels ou des termes sémantiquement proches (car ces méthodes tiennent compte du voisinage sémantique des termes de la requête ayant des concepts, or "jambalaya" ne possède pas de concept associé dans ce cas). Pour faire sortir les termes proches du terme orphelin, notre modèle de RIS exploite en plus de la co-occurrence des termes, la co-occurrence conceptuelle, c'est-à-dire les concepts associés aux termes concurrents, ce qui permet de retrouver des documents ne contenant pas le terme orphelin ou de les faire remonter dans le classement.

5.5 Conclusion

Ce chapitre avait une visée exploratrice du modèle de RIS que nous avons mis en place. Les expérimentations réalisées sur ce petit jeu de test ne permettent pas de tirer de vraies conclusions. Cependant, elles ont permis de comprendre les facteurs qui impactent la qualité de la recherche sur le graphe pondéré et d'explorer les potentialités en terme de fonctionnalités sémantiques du modèle, quand des classes sémantiques sont prises en compte sur le graphe pondéré.

Ayant fixé la fonction de propagation, ainsi que la condition d'arrêt, nous avons pu recenser quelques facteurs qui influent la distribution finale des valeurs d'activations des documents sur le graphe. Parmi ces facteurs on cite : le mode interrogatoire (les nœuds d'entrée au graphe), la normalisation des valeurs des arcs des relations d'occurrence, la taille des documents, le nombre de connexion d'un nœud terme aux documents dans le graphe (fréquence documentaire), etc. Certains de ces facteurs sont des faits du processus de propagation d'activation et qui prouve la parenté entre ce processus et les modèles classique de RI, la chose qui s'est confirmé expérimentalement par des résultats comparables en terme de qualité de la recherche quand le réseau sémantico-documentaire est traduit en un graphe pondéré sans sémantique. D'autres facteurs sont par contre à choisir et à paramétrer, tel que le choix de la normalisation, qui peut refléter la manière dont la distribution des termes dans le document est prise en compte. L'état de l'art de la RI en regorge, cependant nous avons choisi d'exploiter une normalisation simple et facile à interpréter, même si elle n'est certes pas la meilleure.

Quant aux fonctionnalités sémantiques que notre modèle assure, nous avons pu prouver le rôle de la sémantique latente qui se manifeste à travers le phénomène de *co-occurrence* sur le graphe pondéré. Nous avons montré que ce phénomène permet d'améliorer la qualité de la recherche

notamment en terme de précision, mais également en terme de rappel pour certaines requêtes, en retournant de nouveaux documents pertinents.

Outre la sémantique latente, à partir d'un graphe pondéré avec prise en compte explicite des classes sémantiques, ce modèle permet de dépasser certaines limites autour de l'ambiguïté de la langue. Il prend en compte alors la synonymie, résout le problème de *term mismatch*, et pallie au problème de manque de couverture de la ressource sémantique. Des améliorations en terme de *MAP* et *R-PREC* ont été obtenues pour chacune des fonctionnalités sémantiques citées, mais qui restent confrontées à d'autres facteurs expérimentaux : nombre des documents qui sont dans les jugements de pertinence, taille des documents, fréquence documentaire des termes synonymes, etc. D'autres fonctionnalités restent certes à découvrir, notamment du point de vue gestion des ambiguïtés lexicales et conceptuelles et des erreurs de propagation, etc. Ces aspects nécessiteraient peut-être la modélisation de plus de connaissances dans le graphe pondéré, tel que la prise en compte au delà des classes sémantiques, de la structure de l'ontologie (liens hiérarchiques et rôles).

Une petite réflexion à propos de l'exploitation des liens hiérarchiques sur le graphe pondéré a été entamée, dans le but de pouvoir profiter du voisinage sémantique de la requête utilisateur sans introduire des calculs de similarités conceptuelles, contrairement à ce qui se fait en RIS [Zargayouna, 2005]. Nous avons alors repéré une petite lacune qui nécessiterait peut-être la révision de notre condition d'arrêt. Nous allons détailler un exemple de requête du corpus cuisine qui met en ouvre cette lacune. Soit le graphe pondéré de la figure 5.6, qui modélise cet exemple.

— Modélisation du graphe pondéré : les nœuds sont des termes, documents et concepts, les arcs présentent les relations d'occurrences, les relations terminologiques et les relations ontologiques hiérarchiques de spécialisation.
— Requête : "asian soup" ;
— Interprétation des données sur le graphe : les nœuds d'entrée au graphe sont "asian" et "soup", le concept `#Asian` possède comme fils les 3 concepts `#Thai`, `#Indonesian` et `#Chinese`, les documents `doc714` ("asian", "thai", "liquid", "onion") et `doc115` ("asian", "soup", "indonesian", "onion", "pepper") sont pertinents et les termes respectifs "thai" et "indonesian" doivent également contribuer au renforcement des valeurs d'activation de leurs documents respectifs, le document `doc500` ("liquid", "chinese", "onion", "pepper", "tomato") doit être retrouvé car il contient des termes co-occurrents, mais également le nouveau terme activé "chinese" par la spécialisation du concept `#Asian` de "asian" et qui doit contribuer au renforcement de la valeur d'activation de ce document.

La figure 5.7 montre l'évolution des états des nœuds du graphe au cours des itérations de propagation d'activation :

Itération 0 : nœuds actifs : "asian", "soup".

Itération 1 : nœuds actifs : `doc714`, `doc115`, `doc749`, `doc300`, `#asian`, `#soup`, nœuds désactivés : "asian", "soup".

Itération 2 : nœuds actifs : "thai", "indonesian" sont activés par les documents `doc714`, `doc115` et `doc749`.
`#thai`, `#Indonesian` et `#Chinese` sont activés par le concept `#Asian`.

Itération 3 : nœuds actifs : "chinese" et `doc500`, nœuds désactivés dont les valeurs d'activations sont renforcés : "thai" et "indonesian" sont renforcés respectivement par les deux concepts `#thai`, `#Indonesian`.
Les `doc714` et `doc115` sont renforcés par "thai" et "indonesian", etc.

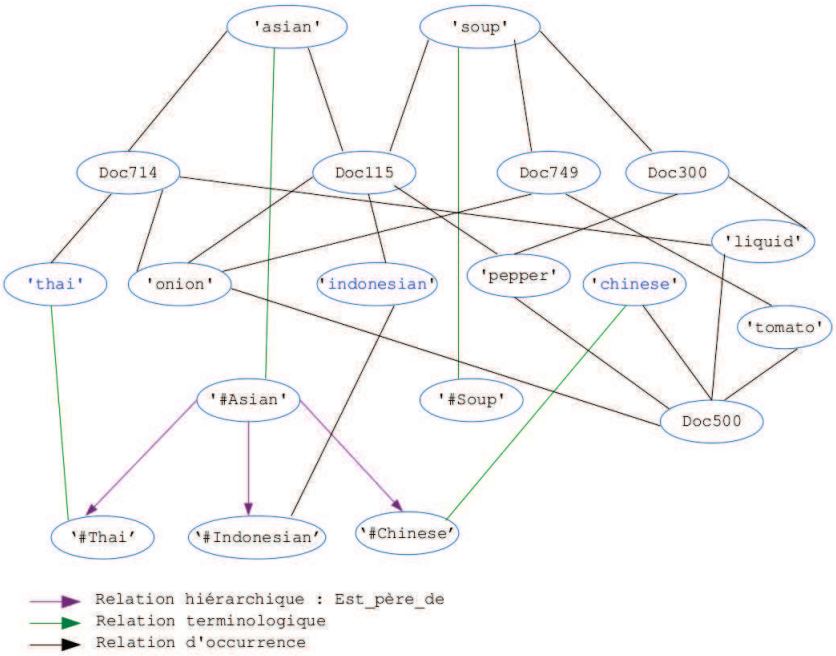

FIGURE 5.6 – Un exemple de graphe avec prise en compte des liens de spécialisation

Lacune : Or comme notre formule de propagation ne tient compte que des valeurs d'activation de l'itération 2 lors du calcul des valeurs d'activation de l'itération 3 (pour éviter les boucles infinis de deux nœuds qui s'activent mutuellement), les valeurs d'activation des documents `doc714` et `doc115` ne sont impactés que par les valeurs d'activations des termes "thai" et "indonesian" à l'itération 2, c-à-d avant d'être renforcés par les concepts `#Thai` et `#Indonesian` fils de `#Asian`. Or comme "thai" et "indonesian" ne peuvent être actifs qu'une seule fois, il ne peuvent pas à l'itération 4 propager ce qu'ils ont reçu à l'itération 3 (c-à-d ce qu'ils ont reçu des concepts `#Thai` et `#Indonesian` par spécialisation).

Itération 4 : nœuds désactivés dont la valeur d'activation se renforce : `doc500` est renforcé par le noeud "chinese" qui a été renforcé à l'itération précédente par le concept `#Chinese`. D'où ce document a pu profiter du lien de spécialisation entre le concept `#Asian` et le concept `#Chinese`, contrairement aux autres documents pertinents.

Le problème détecté à l'itération 3, peut être qualifié par "le non retour du flux de propagation aux documents" étant donné qu'on s'intéresse à la recherche de document dans notre contexte. Afin de bien profiter du voisinage sémantique et des liens hiérarchiques (spécialisation ou généralisation), il faudra spécialement vérifier si le flux de propagation revient aux documents concernés.

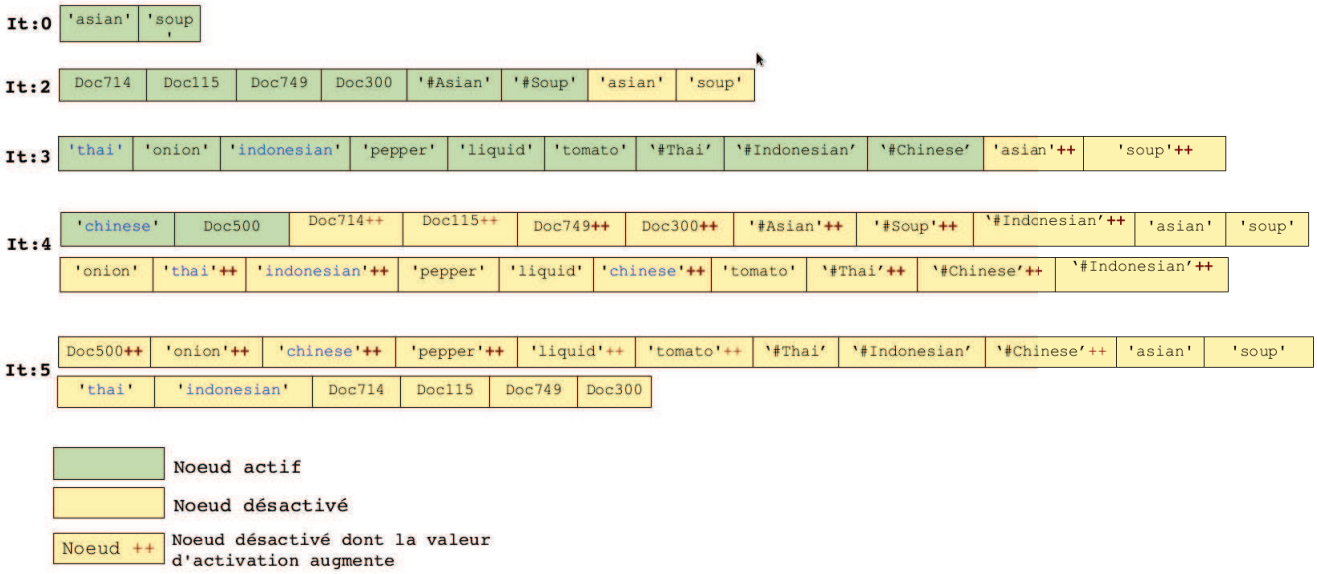

FIGURE 5.7 – États des nœuds au cours de la PA sur le graphe 5.6

Chapitre 6

Expérimentations et passage à l'échelle

Sommaire

Dans ce chapitre, nous décrivons les expérimentations que nous avons conduites pour valider la modélisation en réseau sémantico-documentaire et le mécanisme d'appariement par propagation sur un graphe pondéré que nous avons proposés.

Le but de ces expérimentations est double. Concernant le modèle lui-même, il s'agit de tester son passage à l'échelle, c'est-à-dire de vérifier qu'on peut le déployer sur un jeu de test de grande taille (la collection *Ohsumed*), mais il s'agit aussi d'étudier des variantes du graphe pondéré en

prenant un cadre de test différent de celui présenté dans le chapitre précédent et en enrichissant la base documentaire avec des connaissances. Concernant l'exploitation du modèle pour la recherche d'information, il s'agit naturellement de tester l'efficacité et la viabilité du modèle, c'est-à-dire de s'assurer de la qualité de la recherche en termes de rappel et précision [51] et d'analyser le comportement du système, les résultats obtenus ainsi que certains phénomènes tels que la co-occurrence.

Pour commencer, nous introduisons le cadre de test (section 6.1) et nous décrivons le dispositif expérimental mis en place pour conduire nos expériences (section 6.2), avec les deux modules de construction du graphe et de propagation d'activation. Nous présentons ensuite les expériences faites sur la collection de test *Ohsumed* : certaines sont faites sans annotation sémantique (section 6.3), d'autres avec une annotation manuelle (section 6.4) et d'autres encore avec une annotation automatique (section 6.5).

6.1 Description des données

Dans le domaine médical, on assiste à une explosion des informations et ressources sémantiques et à une diversification des moyens d'accès et de gestion. On peut citer par exemple la collection MEDLINE [52], qui est la base de données médicale de référence en anglais. Elle a été construite par la NLM (*U.S. National Library of Medicine*), c'est la première base de données bibliographique par sa taille : elle est composée de plus de 23 millions de références d'articles en sciences de la vie, et plus précisément en bio-médecine. Les documents sont indexés par le thesaurus *MeSH* (MEdical Subjects Headings). En recherche d'information médicale, *Ohsumed* [Hersh et al., 1994] est l'une des collections de test les plus utilisées. Elle est composée d'un extrait de la base de documents MEDLINE (celui qui est utilisé pour la tâche de filtrage de TREC-9) auquel sont associées des annotations manuelles par *MeSH*, mais également des requêtes et des jugements de pertinences correspondants. Ohsumed présente ainsi un cadre de test idéal pour la RI sémantique et elle est de ce fait largement exploitée par la communauté.

6.1.1 La collection *Ohsumed87*

La collection *Ohsumed* est une sous collection de MEDLINE. Elle constitue un ensemble de 348 566 références de MEDLINE, extraits de 270 journaux médicaux sur une période de 5 ans allant de 1987 à 1991. En général, ces références comportent un titre et un résumé, mais certains d'entre elles peuvent ne comporter qu'un titre. Elles intègrent en plus des annotations manuelles, appelées Medical Subject Headings (*MeSH*).

Nous utilisons dans la suite la partie de la collection qui date de 1987, *Ohsumed87*. Ses principales caractéristiques sont présentées dans le tableau 6.1.

Collection *Ohsumed* (1987)	
Taille de la collection	54 710 documents
Taille du vocabulaire	55 127 termes (*stems*) par Terrier IR
Taille moyenne des documents	74 56 termes (*stems*) par Terrier IR

TABLE 6.1 – Caractéristiques de la collection *Ohsumed87*

51. A noter que le temps d'accès ne fait pas partie des critères que nous prenons en compte à ce stade.
52. https ://www.nlm.nih.gov/pubs/factsheets/medline.html

La collection *Ohsumed87* que nous utilisons est fournie sous format XML. Chaque texte possède un identifiant unique (le contenu de la balise `<DOCID>`), la référence du journal dont il est extrait (balise `<SOURCE>`), les annotations manuelles qui lui sont associées(balises `<ANNOTATION>` et `<CONCEPTS>`), le titre de l'article (balise `<TITLE>`), le résumé s'il existe (balise `<TEXT>`), le type de la revue et les noms des auteurs (balise `<TYPE>`).

Dans la suite, nous n'utilisons que le contenu des balises `<TITLE>` et `<TEXT>` pour l'indexation, ainsi que les balises `<DOC>` et `<DOCID>` pour identifier les documents. La figure 6.1 présente un exemple de document tiré de cette collection.

```
<DOC>
<DOCNO>87049104</DOCNO>
<DOCID>87049104</DOCID>
<SOURCE>Am J Emerg Med 8703; 4(6):552-3</SOURCE>
<ANNOTATION>
<CONCEPT>http://org.snu.bike/MeSH#atropine</CONCEPT>
<CONCEPT>http://org.snu.bike/MeSH#human</CONCEPT>
<CONCEPT>http://org.snu.bike/MeSH#baclofen</CONCEPT>
<CONCEPT>http://org.snu.bike/MeSH#case_report</CONCEPT>
<CONCEPT>http://org.snu.bike/MeSH#adolescent</CONCEPT>
<CONCEPT>http://org.snu.bike/MeSH#bradycardia</CONCEPT>
</ANNOTATION>
<TITLE>
Atropine in the treatment of baclofen overdose.
</TITLE>
<TEXT>
A patient suffering baclofen overdose successfully treated with atropine is reported.
Three hours after admission for ingestion of at least 300 mg baclofen as a single dose,
the patient became comatose and subsequently bradycardic, hypotensive, and hypothermic.
A prompt increase in heart rate and blood pressure followed administration of 1 mg of
atropine sulfate. Atropine appears to be useful in treating cases of baclofen overdose
complicated by bradycardia and hypotension.
</TEXT>
<TYPE>JOURNAL ARTICLE.</TYPE>
</DOC>
```

FIGURE 6.1 – Exemple d'un document de *Ohsumed87*

6.1.2 Les requêtes *Ohsumed*

Dans la collection de test *Ohsumed87*, les requêtes sont organisées en thèmes, qui varient selon la tâche que l'on cherche à tester. Pour la recherche d'information, on trouve 63 thèmes appelés *OHSUMED queries*. Il s'agit d'un ensemble de requêtes utilisateur, qui comportent chacune un titre (balise `<title>`) et une partie narrative (balise `<desc>`), en plus du numéro de la requête (balise`<num>`).

Nous disposons également d'annotations manuelles *MeSH* (balises `<annotation>` et `<concept>`), qui nous ont été fournies par l'équipe de recherche IRIS [53]. Un exemple de requête *Ohsumed* est présenté dans la figure 6.2.

Nous utilisons les balises `<title>` et `<desc>` lors de la recherche, en considérant le contenu de la première comme une requête (la description du patient) et celui de la seconde comme étant sa partie descriptive (les informations recherchées).

Nous avons enfin des jugements de pertinence associés à ces requêtes : de 3 à 22 documents jugés pertinents par requête. Ces jugements de pertinence sont également fournis par TREC-9.

53. L'annotation des requêtes est réalisée par Gilles HUBERT au sein de l'équipe IRIS (*Information Retrieval & Information Synthesis*), à l'IRIT (Institut de Recherche en Informatique de Toulouse)

```
<top>
<num>2</num>
<title>60 yo male with disseminated intravascular coagulation</title>
<desc>pathophysiology and treatment of disseminated intravascular coagulation</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#aged</concept>
<concept>http://org.snu.bike/MeSH#disseminated_intravascular_coagulation</concept>
<concept>http://org.snu.bike/MeSH#blood_coagulation</concept>
</annotation>
</top>

<top>
<num>3</num>
<title>prolonged prothrombin time</title>
<desc>anticardiolipin and lupus anticoagulants, pathophysiology, epidemiology, complications</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#prothrombin_time</concept>
<concept>http://org.snu.bike/MeSH#prothrombin</concept>
<concept>http://org.snu.bike/MeSH#anticoagulant</concept>
<concept>http://org.snu.bike/MeSH#cardiolipin</concept>
<concept>http://org.snu.bike/MeSH#cutaneous_lupus_erythematosus</concept>
</annotation>
</top>
```

FIGURE 6.2 – Exemple de deux requêtes *Ohsumed87* de numéros 2 et 3

6.1.3 La ressource sémantique : *MeSH*

MeSH (MEdical Subject Heading) est une ressource terminologique de référence qui contient le vocabulaire contrôlé de la *National Library of Medecine* (NLM). Il est composé d'un ensemble de termes de médecine organisés hiérarchiquement. Au plus haut niveau de la hiérarchie, on trouve les catégories les plus génériques, qui sont en nombre de 16 : "health care[N]", "Disease [C]", "Anatomy [A]", "Psychiatry and Psychology [F] ", etc. Ce thesaurus est composé de plus que 20 000 termes et est souvent utilisé pour indexer les collections de type MedLine, comme *Ohsumed*.

L'ontologie *MeSH* que nous utilisons est une sous-partie du theraurus qui a été formalisée en OWL : elle comporte 40 247 concepts (termes *MeSH*).

6.1.4 L'annotateur médical : MetaMap

MetaMap [Aronson, 2001] est un outil d'annotation qui a été conçu pour asssocier des concepts du méta-thésaurus UMLS à des documents, *a priori* médicaux. Pour la recherche d'information, cela permet d'améliorer l'indexation ou la recherche à travers l'expansion des requêtes.

Nous ne développons pas ici le processus d'annotation qui est mis en oeuvre (se reporter à [Aronson and Lang, 2010]) mais nous utilisons MetaMap pour les expériences avec annotation automatique (avec l'option de désambiguïsation et en spécifiant *MeSH* comme ressource sémantique).

6.2 Dispositif expérimental

Pour expérimenter le passage à l'échelle, nous avons mis en place une plate-forme logicielle.

Cette plate-forme est modulaire. En effet, pour des raisons de portabilité et pour faciliter la recherche, il est judicieux en RI de séparer le module d'indexation du module de recherche, la construction de l'index pouvant prendre du temps alors que les réponses doivent être données

aux utilisateurs dans des délais très courts [54]. Par analogie, la construction du graphe pondéré doit être faite en amont et doit être isolée de l'interrogation du graphe dans la phase de recherche qui se fait ici par propagation d'activation sur le graphe une fois qu'il est construit.

Nous utilisons le même environnement logiciel que celui qui a été présenté dans la section 5.2. Nous essayons de rendre modulaire l'exploitation de la plate-forme de modélisation de graphe JUNG [55], à l'image de la modularité de la plate-forme de recherche d'information sémantique Terrier SIR [Bannour and Zargayouna, 2012].

Nous avons donc deux modules : le module de construction du graphe pondéré et celui d'interrogation du graphe par propagation d'activation.

6.2.1 Construction du graphe pondéré

La traduction du réseau sémantico-documentaire en un graphe pondéré, et donc la construction de ce graphe pondéré, correspond à la phase d'indexation dans les modèles classiques de RI.

Dans le cadre de la RIS, on peut distinguer l'index documentaire classique et l'index sémantique. L'index classique manipule les documents de la collection, les termes du vocabulaire qui y figurent, ainsi que les relations termes/documents ou documents/documents. L'index sémantique associe à l'index classique des connaissances additionnelles provenant d'un modèle sémantique en s'appuyant généralement sur un processus intermédiaire d'annotation sémantique des documents sources. De même, dans la construction du graphe pondéré, on distingue le graphe minimal sans sémantique (couche terminologique et documentaire et les relations qu'elles entretiennent) et le graphe sémantique qui introduit en plus la couche sémantique.

Le graphe sans sémantique repose sur l'index classique de *Terrier IR*. Nous avons décrit le processus qui a mené à sa construction dans le chapitre précédent. On traduit sous la forme d'arcs dans le graphe les liens d'occurrence termes/documents auxquels on associe des poids relatifs reflétant la fréquence des termes dans les documents correspondants. Cette fréquence d'occurrence est une propriété intrinsèque aux liens termes/documents : elle est également extraite de l'index de *Terrier IR*.

Le graphe sémantique est le graphe minimal précédent enrichi d'une couche sémantique, où figurent les connaissances explicitées lors de l'annotation sémantique, c'est-à-dire l'ontologie associée et les liens documents/concepts. Pour la collection *Ohsumed*, nous disposons d'annotations manuelles qui explicitent pour chaque document les concepts de l'ontologie *MeSH* auxquels ils sont associés. Dans un premier temps, nous n'exploitons que ces relations documents/concepts sans exploiter la structure de l'ontologie en tant que telle (relations concept/concept, par exemple) mais le modèle proposé est compatible avec différents types de graphe pondéré et on pourrait étendre les expériences en y intégrant une couche sémantique plus riche.

Au cours de la construction du graphe pondéré, des tests d'unicité sont réalisés sur les nœuds et les arcs, pour garantir qu'il n'y a pas des nœuds et des arcs de même noms (*name*) avec des identifiants (ID) différents. On vérifie par exemple que le nombre de nœuds termes correspond réellement à la taille du vocabulaire de la collection, que le nombre de nœuds documents correspond aux nombre de documents de la collection, que le nombre de concepts correspond aux concepts de l'ontologie, etc.

L'étape de construction du graphe pondéré prend fin avec le stockage du graphe sur disque dur, ce qui évite d'avoir à le reconstruire pour chaque requête et assure la modularité du

54. Notre implémentation n'est néanmoins pas assez optimisée pour garantir ces temps de réponse (voir section 6.5.4.2)

55. JUNG (*Java Universal Network/Graph Framework*) : http ://jung.sourceforge.net/

système. Nous utilisons des classes spécifiques de la plate-forme *JUNG*, notamment la classe `GraphMLWriter<V,E>`, qui permet de stocker le graphe pondéré dans un fichier *"index"* de format `GraphML`. Il s'agit d'une représentation du graphe construit dans un fichier XML qui conserve les propriétés du graphe (orienté ou non, les nœuds et leurs attributs, les arcs, etc.)

6.2.2 Interrogation et recherche sur le graphe pondéré

L'interrogation du graphe pondéré débute quand un utilisateur interroge sa base documentaire. Trois processus sont alors déclenchés : (1) le chargement du graphe en mémoire à partir du `GraphML` pour chaque requête, (2) le pré-traitement de la requête et l' interrogation du graphe pondéré, (3) la propagation d'activation sur le graphe.

Pour le chargement du `grapheML`, une classe de *JUNG*, `GraphMLReader<V,E>`, est invoquée pour charger le graphe sans avoir à le reconstruire. Ce chargement se fait rapidement, en quelques secondes.

Ensuite, un ensemble de traitements est réalisé sur la requête. S'il s'agit d'une recherche d'information classique sans sémantique par exemple, la requête subit les mêmes analyses que les documents (*tokenisation*, élimination des mots vides, analyse lexicale), ce qui permet d'en extraire les unités lexicales à partir desquelles les nœuds d'entrée au graphe pondéré sont identifiés (nœuds à partir desquels la propagation se déclenche). S'il s'agit d'une interrogation par concepts, on se contente d'identifier les nœuds concepts correspondants dans le graphe.

Une fois les nœuds caractérisant la requête identifiés, des valeurs d'activation initiales $a_0 = 1$ leurs sont attribués. Ces valeurs initiales présentent les valeurs des attributs $InitialWeight$ de chacun des nœuds de départ.

La propagation d'activation est alors déclenchée. Au départ, les nœuds de la requête ont le statut *actif* et possèdent des valeurs d'activation initiales positives, alors que le reste des nœuds est *inactif* et ont des valeurs d'activation nulles. Au cours de la propagation, les nœuds passent de l'état inactif à l'état actif, propagent leur information de pertinence à leurs voisins directs contribuant de ce fait à augmenter leur valeur d'activation et changent de statut (ils deviennent *désactivés*). La propagation se poursuit jusqu'à ce que tous les nœuds activables aient été activés et soient devenus désactivés.

6.2.3 Protocole expériemental

Pour conduire les expérimentations qui suivent, nous avons fait certains choix, qui concernent le type de graphes pris en compte, la forme d'annotation utilisée pour les scénarios sémantiques de RI et le mode d'interrogation du graphe pondéré. Nous distinguons ainsi 3 types de modèles de graphes :

Modèle terme/document (TD) : le graphe possède comme nœuds l'ensemble des unités du vocabulaire et des documents de la collection, et comme arcs les relations d'occurrences entre termes et documents.

Modèle terme/document/concept (TDC) : le graphe possède comme nœuds l'ensemble des unités du vocabulaire, des documents de la collection et des concepts de l'ontologie, et comme arcs les relations d'occurrences entre termes et documents et les relations d'annotations entre documents et concepts.

Modèle terme/concept/document (DTC) : le graphe possède comme nœuds l'ensemble des unités du vocabulaire, des documents de la collection et des concepts de l'ontologie. Les arcs représentent les relations d'occurrences entre termes et documents ainsi que les relations terminologiques entre termes et concepts.

Modèle terme/document/concept/terme (DTCD) : le graphe possède comme nœuds l'ensemble des unités du vocabulaire, des documents de la collection et des concepts de l'ontologie. Les arcs représentent les relations d'occurrences entre termes et documents, les relations terminologiques entre termes et concepts et les relations d'annotation entre documents et concepts.

Nous prenons en compte deux types d'annotation :

Annotation manuelle (Manu) : il s'agit de l'annotation des documents fournie par la tâche TREC-9. Les requêtes sont annotées au sein de l'équipe IRIS à *Toulouse* ; cette annotation n'utilise qu'une sous-partie de l'ontologie *MeSH* et a une couverture moyenne : pour chaque document on dispose des concepts qui lui sont associés, mais on n'a pas les termes auxquels ces concepts font référence dans le document.

Annotation automatique (MetaMap) : il s'agit de l'annotation obtenue avec l'annotateur médical MetaMap au regard de l'ontologie *MeSH* complète ; cette annotation ne pose pas le même problème de couverture que la précédente.

Enfin, nous considérons trois formes d'interrogation différentes :

Interrogation par termes (intT) : la requête utilisateur subit un ensemble de traitements afin d'en extraire les unités lexicales normalisées, qui permettent à leur tour d'identifier les nœuds du graphe devant servir de point de départ pour la propagation.

Interrogation par concepts (intC) : une annotation des requêtes au regard de l'ontologie du domaine permet d'identifier les concepts de chaque requête et par conséquent les nœuds concepts devant servir de point de départ pour la propagation.

Interrogation par termes et concepts (intTC) : il s'agit d'une combinaison des deux modes d'interrogation intT et intC.

La combinaison de ces différents paramètres conduisent aux expérimentations détaillées dans le tableau 6.2.

Nous commençons en comparant les résultats de la RI classique avec *Okapi-BM25* comme schéma de pondération obtenus par *Terrier IR* et ceux de notre modèle TDintT sans sémantique (TD), avec une interrogation par termes (intT) (section 6.3), puis nous prenons en compte l'aspect sémantique, en exploitant les annotations manuelles et nous comparons le modèle TDCintT$_{Manu}$ au modèle sans sémantique TDintT (la sous-section 6.4.1). Dans une troisième série d'expérimentation, nous reprenons le même graphe mais en l'interrogeant par des concepts TDCintC$_{Manu}$, puis par les termes et les concepts TDCintTC$_{Manu}$ (la sous-section 6.4.2 et la sous-section 6.4.3). En deuxième lieu (section 6.5), nous passons à l'annotation automatique de MetaMap : nous interrogeons par les termes un modèle de graphe TDC et un modèle de type DTC [56], nous lançons les deux expérimentations TDCintT$_{MM}$ et DTCintT$_{MM}$ et nous comparons leurs résultats respectifs avec ceux des expérimentations TDCintT$_{Manu}$ et TDintT puis ferons les mêmes comparaisons en interrogeant par les termes et par les concepts (TDCintTC$_{MM}$, DTCintTC$_{MM}$ coomparés respectivement à TDCintTC$_{Manu}$ et DTCintT$_{MM}$). En troisième lieu, de nouvelles expérimentations sont menées toujours en exploitant les annotations de MetaMap, mais en saturant le graphe pondéré par tous les noeuds et relations possibles pour aboutir au modèle DTCD$_{MM}$. Ce dernier modèle est comparé aux autres modèles en fixant à chaque fois le mode interrogatoire.

Notre objectif principal est de voir comment se comporte notre modèle en fonction des ingrédients qui lui sont injectés. Nous cherchons à comparer les approches de RI et de RIS, à voir

56. Rappelons que l'annotateur MetaMap explicite les liens terminologiques entre termes et concepts lors de l'annotation.

	Modélisation de $G = (N, R)$		
Mode d'interrogation	Termes	Concepts	Termes et Concepts
Sans annotation	$\underline{TDintT}$: — $N = N_d \uplus N_t$ — $R = R_{occ}$		
Annotation manuelle	$\underline{TDCintT_{Manu}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{ann}$	$\underline{TDCintC_{Manu}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{ann}$	$\underline{TDCintTC_{Manu}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{ann}$
Annotation automatique	$\underline{TDCintT_{MM}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{ann}$ $\underline{DTCintT_{MM}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{term}$ $\underline{DTCDintT_{MM}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{term} \uplus R_{ann}$		$\underline{TDCintTC_{MM}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{ann}$ $\underline{DTCintTC_{MM}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{term}$ $\underline{DTCDintTC_{MM}}$: — $N = N_d \uplus N_t \uplus N_c$ — $R = R_{occ} \uplus R_{term} \uplus R_{ann}$

TABLE 6.2 – Expérimentations réalisées sur la collection *Ohsumed*

comment les phénomènes sémantiques (comme la co-occurrence) peuvent être pris en compte et à montrer la polyvalence de notre modèle qui utilise un même système pour différents modes d'interrogation (recherche par l'exemple, proximité sémantique, etc.).

6.3 Expériences sur le modèle Terme/Document

Dans cette section, notre but est de comparer la RI par propagation d'activation sur un graphe pondéré sans sémantique, avec la RI classique par le modèle vectoriel, pondéré par *BM25*.

6.3.1 Modèle vectoriel pondéré par *BM25*

Le tableau 6.3 présente les résultats de la recherche par le modèle vectoriel implémenté dans *Terrier IR*, en termes de *MAP*, *R-PREC* et *Rappel*. Cette expérimentation constitue la baseline de recherche d'information pour la suite des expériences.

	R-PREC	MAP	Rappel
$BM25$	0.30	0.30	610/670

TABLE 6.3 – Résultats de la RI classique (BM25) du SRI Terrier IR

L'analyse de ces résultats requête par requête permet de faire certaines constatations. Sur les 63 requêtes du cadre de Test, nous pouvons distinguer :

- 13 requêtes pour lesquelles la *R-PREC* et la *MAP* sont nulles : nous considérons ces requêtes comme difficiles et ces résultats peuvent s'expliquer par le fait que les jugements de pertinences associés à ces requêtes comportent peu de documents (de 2 à 8 documents pertinents).
- 50 requêtes pour lesquelles la R-PREC et MAP ne sont pas nulles, mais avec des valeurs de *R-PREC* extrêmes :
  - des R-PREC supérieures à 0.30 pour 30 requêtes,
  - des R-PREC inférieures à 0.30 pour 20 requêtes.

6.3.2 Modèle Terme/Document

Le modèle terme/document que nous présentons ici est le modèle TDintT du tableau 6.2 avec :

- les nœuds du graphe sont les termes et les documents ($N = N_d \uplus N_t$),
- les arcs correspondent aux relations d'occurrence ($R = R_{occ}$),
- le poids des arcs est donné par la formule suivante pour un terme t et un document d : $w_2 = w(t,d) = w(d,t) = \frac{tf(t,d)}{MaxTf(d)}$,
- l'interrogation se fait par les termes, c'est-à-dire les unités lexicales normalisées extraites des requêtes, qui ont une valeur d'activation initiale a_0 fixée à 1.

Les résultats de la propagation d'activation des 63 requêtes sur le graphe modèle $TDintT$ sont présentés dans le tableau 6.4.

- 12 requêtes ont une *R-PREC* et une *MAP* nulles et avaient également des résultats nuls dans la RI classique par BM25 (voir section 6.3.1).
- 51 requêtes dont la R-PREC et MAP ne sont pas nulles.

	R-PREC	MAP	Rappel
$TDintT$	0.18	0.18	610/670

TABLE 6.4 – Résultats de la propagation d'activation sur un graphe pondéré TD

6.3.3 Comparaison

On peut donc comparer le modèle TDintT et le modèle vectoriel classique pondéré par *Okapi-BM25* (tableaux 6.3 et 6.4). On observe alors :

- une dégradation générale des résultats de 12 points en *MAP* et *R-PREC* quand on passe du premier au deuxième modèle ;
- un nombre de documents pertinents retournés stable, les documents pertinents retournés étant également les mêmes dans les deux expériences ;
- parmi les 13 requêtes ayant des résultats nuls en *R-PREC* et *MAP* avec le modèle vectoriel, seule l'une d'entre elles, la requête `REQ8`, voit la qualité de sa recherche améliorée avec le modèle TDintT ;

— la qualité de la recherche s'est amélioré pour 7 requêtes ;
— la qualité de la recherche est stable pour 6 requêtes.

Sur la base de ces constatations, nous essayons d'expliquer pourquoi la qualité de la recherche s'est dégradée pour certaines requêtes et s'est améliorée pour d'autres en analysant ce qui se passe lors de la propagation d'activation sur le graphe pondéré.

6.3.3.1 Impact de la co-occurrence

L'amélioration de la qualité de la recherche pour certaines requêtes est principalement due au phénomène de co-occurrence entre les termes du vocabulaire.

Prenons l'exemple de la requête `REQ9`[57] (voir AnnexeC). On observe :
— une amélioration considérable des classements des 10 documents jugés pertinents pour cette requête, ce qui améliore la *MAP* ;
— une amélioration de la *R-PREC* qui passe de 0.1 avec le modèle vectoriel à 0.5 avec le modèle TDintT, ce qui donne 4 documents pertinents de plus dans les 10 premiers documents retournés.

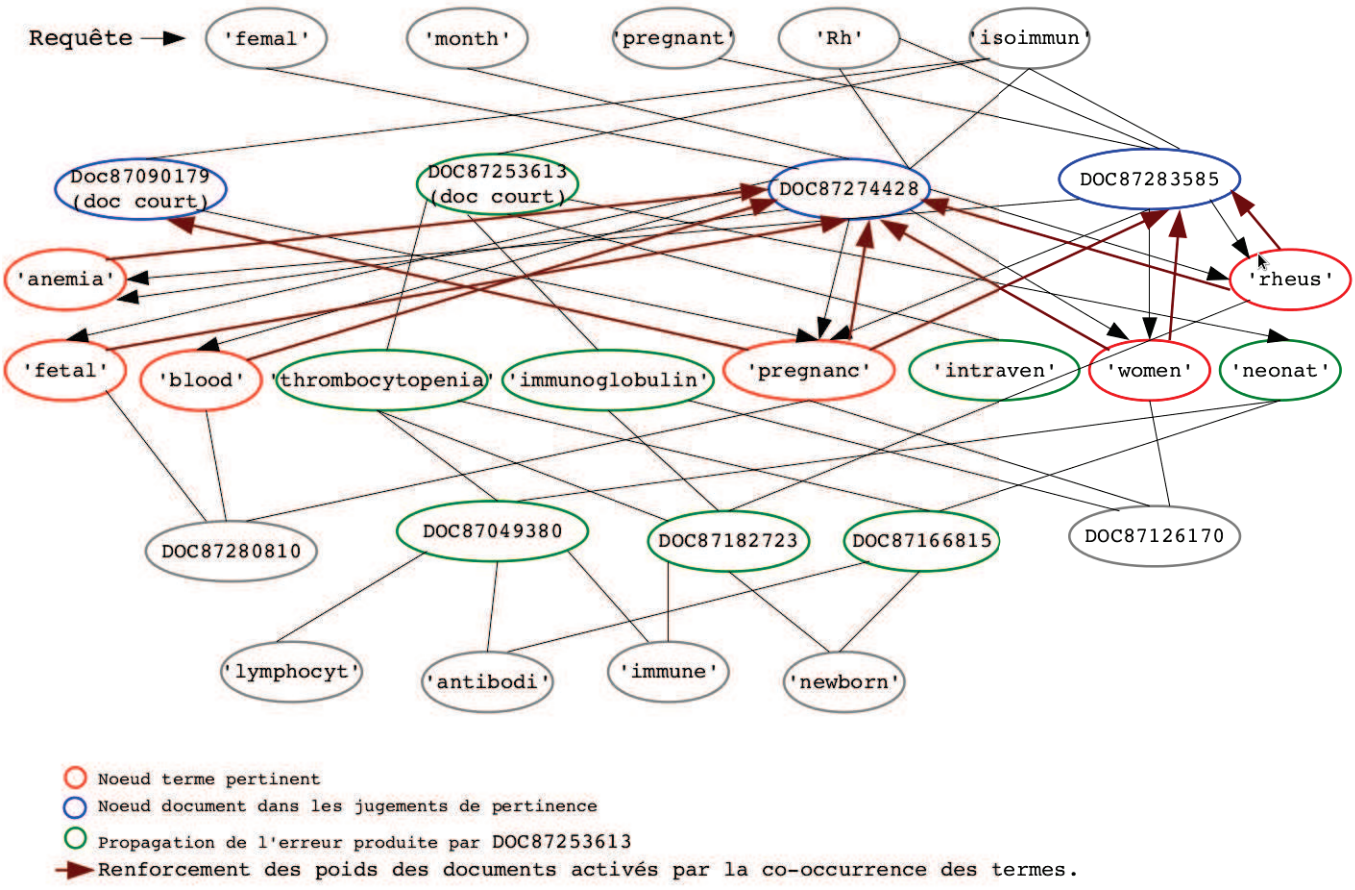

FIGURE 6.3 – Exemple de requête *Ohsumed* (`REQ9`) par TDintT : co-occurrence et propagation d'erreur

57. *REQ*9 : "une femme enceinte de 3 mois, avec une rh isoimmunisation"

Lors de la première itération de propagation, on note que 5 documents pertinents sont activés, et ceux-ci activent à leur tour des termes à la deuxième itération. A la troisième itération, les valeurs d'activation des 5 documents sont renforcées par la co-occurrence des termes activés lors de la deuxième itération (ex. "anemia", "fetal", "women", "pregnace").

La figure 6.3 schématise une simulation de la propagation d'activation dirigée par une partie de la requête `REQ9` sur le graphe pondéré TDintT et montre comment la co-occurrence permet dans ce cas de renforcer les scores des documents pertinents `87090179`, `87274428` et `87283585`. La co-occurrence ne permet pas ici de découvrir de nouveaux documents pertinents mais elle a agit sur les valeurs d'activation (scores) des documents pertinents activés.

6.3.3.2 Impact de la normalisation des valeurs des arcs R_{occ}

On remarque que le modèle TDintT retourne dans les premiers rangs des documents très courts qui comportent peu de variations lexicales et thématiques. Ces documents se réduisent en fait à des titres.

Prenons le même exemple de la requête `REQ9`. On constate que tous les rangs des documents pertinents courts s'améliorent avec notre modèle : le rang du document `87090179` passe du rang 22 avec *BM25* au rang 2 avec TDintT et celui du document `87186146` passe du rang 151 avec $BM25$ au rang 10 avec TDintT.

Le comportement de notre modèle s'explique par la formule de propagation utilisée : 3 éléments entrent dans le calcul de la valeur d'activation d'un document d à partir d'un nœud terme t : la valeur d'activation à l'étape précédente du nœud t, le degré du nœud t et le poids de l'arc reliant t à d que nous avons fixé à $w(t,d) = TF(t)/maxTF(d)$. La normalisation par $maxTF(d)$ permet de tenir compte de la taille du document. En effet, plus un document est long plus l'écart entre le terme t et le terme le plus fréquent tend à être grand et donc plus $w(t,d)$ est faible, ce qui se répercute sur le score du document d. A l'opposé, dans un document court où $TF(t)$ est proche de $max(TF(d))$, $w(t,d)$ est proche de 1 ce qui renforce la valeur d'activation de d. Ce mécanisme tend à placer les documents courts en tête de classement.

La normalisation exploitée dans *Okapi-BM25* permet de réduire l'impact des documents courts et de trouver un compromis entre la pertinence d'un document et sa taille.

Pour montrer l'impact de cette normalisation sur la dégradation de la qualité de la recherche, nous modifions artificiellement les données de l'expérience de nos deux modèles (*BM25* et TDintT) en enlevant les documents courts (ceux qui se réduisent à leur titre [58]) des documents résultats et des jugements de pertinence et en recalculant la qualité de la recherche sur cette base. Les résultats sont présentés dans le tableau 6.5. Nous remarquons alors une augmentation de 1% de la *R-PREC* pour la RI classique par *BM25* et de 4% de la *R-PREC* pour le modèle TDintT, ce qui laisse penser qu'une bonne prise en compte de la taille des documents doit bénéficier assez nettement au modèle TDintT.

	R-PREC	MAP
$BM25$	0.31	0.32
$TDintT$	0.22	0.23

Table 6.5 – BM25 vs. $TDintT$ sans documents courts

58. Même s'il y a d'autres documents courts qui ont une partie textuelle mais minimale.

6.3.3.3 Impact du relâchement rapide des valeurs d'activation avec la distance

Parmi les phénomènes qu'on arrive à observer avec notre modèle, on note que les valeurs d'activation des nœuds diminuent rapidement quand on s'éloigne des points de départ de la propagation. En réalité, ce phénomène réduit l'impact de la co-occurrence qui sert surtout à renforcer les valeurs d'activation de nœuds déjà activés et très peu à découvrir ou activer de nouveaux nœuds pertinents dans le graphe.

Considérons l'exemple de la requête `REQ7` [59] (voir AnnexeC). La *R-PREC* et la *MAP* obtenues sont stables pour les deux modèles mais on note qu'il y a deux documents pertinents qui n'ont pas été retournés ni par le modèle vectoriel ni par notre modèle. Ces deux documents sont `87203237`(lactose(7), milk(7)) et `87104937`(lactose(4), milk(3), malabsorpt(4)).

L'analyse de la propagation d'activation à partir de cette requête montre que ces deux documents pertinents qui n'ont pas été retrouvés par *BM25* sont atteints par la propagation dès la troisième itération du modèle TDintT, du fait de la co-occurrence de nouveaux termes pertinents découverts dans les premiers documents activés à la première itération. Ces termes sont notamment "lactos", "milk", "treat", "intol", "isol", "enzym", etc. Pour autant, ces deux documents pertinents ne sont pas retournés en fin de la propagation parce que leurs valeurs d'activation sont restées trop faibles.

On note donc que la co-occurrence permet de découvrir de nouveaux documents pertinents sur le graphe, mais que les valeurs d'activation diminuent trop vite avec la distance sur le graphe pour tirer un plein bénéfice de ce phénomène. Au final, la co-occurrence ne sert qu'à renforcer les valeurs d'activation des documents déjà activés lors la première itération.

6.3.3.4 Impact de la propagation des erreurs

Des erreurs peuvent se produire au cours de la propagation d'activation du fait des problèmes d'ambiguïtés. Il faut donc s'interroger pour savoir si l'erreur se propage et quel est son impact.

Reprenons l'exemple de la requête `REQ7`. On note une erreur de propagation En raison de la co-occurrence avec des notions traitées dans ces documents, et notamment avec les termes "lymphona" et "hodgkin", d'autres documents non pertinents sont susceptibles d'immerger par propagation à partir de la troisième itération. On remarque cependant que l'erreur ne se propage pas suffisamment pour retourner d'autres documents non pertinents dans les premiers rangs. Pour ce cas d'ambiguïté, le relâchement rapide des valeurs d'activation est bénéfique : il permet de limiter l'impact de l'erreur.

Notre système ne se comporte pas différemment d'un système de recherche classique face à tel problème d'ambiguïté. Les trois documents cités gardent les premiers rangs à la fin de la propagation, car l'erreur s'est produite dès la première itération, tout comme la RI classique par *BM25*.

La figure 6.3 présente une autre erreur de propagation qui se produit pour la requête `REQ9`. Cette erreur est due au document court `87253613` qui est indûment activé lors de la première itération car il contient le terme de la requête "isoimmune". Ce document active lors de la deuxième itération des termes comme "intravenous", "immunoglobulin" , "thrombocytopenia", dont les valeurs d'activation sont importantes à l'issue de la troisième itération. La co-occurrence entre ces termes et les termes proches de la requête (comme "woman", "rheus", "pregnanc") permet de retrouver les documents `87049380`, `87182723` et `87166815` qui ne sont pas pertinents mais qui

59. *REQ*7 : "Une jeune femme avec une insuffisance en Lactase".

relèvent d'une thématique connexe à la requête et au sujet du document 87253613 [60]. On note cependant que ces 3 documents non pertinents ne sont pas retournés par notre système à la fin de la propagation, ce qui est également dû au relâchement rapide des valeurs d'activation sur le graphe.

Au final, il semble que l'affaiblissment des valeurs d'activation bloque en partie la propagation des erreurs. Il faudra donc trouver un compromis entre ces deux phénomènes.

6.3.4 Conclusions

L'analyse de la propagation d'activation sur le modèle en graphe pondéré TDintT, sans sémantique, montre certains phénomènes intéressants et certaines lacunes dans notre modèle.

Les documents courts, sans partie textuelle, ont un impact fort sur les résultats : dès lors qu'ils contiennent au moins un terme de la requête, ils sont privilégiés par rapport à tous les autres documents pertinents. Ce problème ne vient pas des documents courts en tant que tels mais du fait que la taille des documents impacte trop fortement la recherche dans notre modèle. Il faudra sans doute réviser le poids des arcs et la formule de normalisation de notre modèle. Amit Singhal *et al.* se sont intéressés à ce problème des documents courts autour de la fonction Cosinus du modèle vectoriel. Ils ont proposé une normalisation permettant de réduire l'impact des documents courts lorsque leur nombre devient important. Cette normalisation dite *à pivot* [Singhal et al., 1996] se traduit par une fonction à deux paramètres, *pivot* et *slope*. Nous n'en tenons pas compte dans notre modèle mais cela pourrait être une piste intéressante à étudier.

Nous avons aussi noté que les valeurs d'activation diminuent rapidement avec la distance de propagation dans le graphe. Le réglage utilisé pour les expériences ci-dessus permet de renforcer les valeurs d'activation de nœuds déjà activés mais pas d'en activer de nouveaux. C'est un inconvénient quand de nouveaux documents sont découverts au fil de la propagation mais abandonnés car insuffisamment activés mais c'est un avantage, à l'inverse, quand cela permet de limiter la propagation d'erreurs, c'est-à-dire la propagation à partir de noeuds indûment activés. Il faudra étudier plus ce phénomène pour déterminer le bon compromis entre relâchement des valeurs d'activation et risque de propagation d'erreur.

Dans la suite, nous essayons d'intégrer une couche sémantique, à notre modèle, sous la forme de classes sémantiques notamment, pour analyser son impact sur la RIS.

6.4 Expériences sur le modèle Terme/Document/Concept avec annotation manuelle

L'objectif de cette section est de comparer la RI classique représentée par le modèle TDintT et la RI sémantique, en intégrant une couche sémantique dans le graphe pondéré. Dans un premier temps, nous ne prenons en compte que l'annotation manuelle fournie dans la compagne TREC-9 mais différentes modélisations sont possibles qui correspondent à autant de traductions différentes du réseau sémantico-documentaire en graphes pondérés (TDCintT$_{Manu}$, TDCintTC$_{Manu}$, TDCintC$_{Manu}$) et nous considérons alors différentes formes d'interrogation.

6.4.1 Interrogation par les termes

Nous allons dans cette partie considérer le modèle TDCintT$_{Manu}$ où le graphe pondéré est composé comme suit :

60. La requête REQ14 traite de cette thématique connexe "thrombocytopenia in pregnancy").

— les nœuds du graphe sont les documents, les termes et les concepts ($N = N_d \uplus N_t \uplus N_c$) ;
— les arcs correspondent aux relations d'occurrence et d'annotation ($R = R_{occ} \uplus R_{ann}$) ;
— les valeurs des arcs sont calculées comme suit :
  — pour les relations R_{occ} entre un terme t et un document d : $w_2 = w(t,d) = w(d,t) = \frac{tf(t,d)}{MaxTf(d)}$,
  — pour les relations R_{ann} entre un document et un concept : 1 si la relation existe et 0 sinon ;

et où l'interrogation se fait par les termes, c'est-à-dire par les unités lexicales normalisées extraites des requêtes : si t est un terme de la requête, la valeur d'activation initiale du nœud correspondant dans le graphe est $a_0(t) = 1$; sinon elle est nulle.

Les résultats de la propagation d'activation des 63 requêtes sur le graphe modèle TDCintT$_{Manu}$ sont présentés dans le tableau 6.4.

	R-PREC	MAP
$TDCintT_{Manu}$	0.19	0.19
$TDintT$	0.18	0.18

FIGURE 6.4 – Résultats de la propagation d'activation sur un graphe pondéré TDCintT$_{Manu}$

L'objectif de cette expérimentation est de comparer les résultats de la propagation d'activation sans sémantique sur le graphe TD, avec la propagation d'activation sur un graphe TDC, en interrogeant les deux modèles par les termes de la requête. Il s'agit d'apprécier l'apport de la sémantique quand on ne possède que des liens de catégorisation (des liens concept/document et des liens document/concept). La comparaison porte sur différents niveaux :
— la qualité de la recherche : *MAP*, *R-PREC* ;
— l'identification des concepts annotant la requête à travers la co-occurrence des concepts,
— la résolution des problèmes d'ambiguïtés, des erreurs de propagation, etc., s'il y en a, *via* les concepts de catégorisation.

On constate une modeste amélioration au niveau de la *MAP* et *R-PREC* pour toutes les requêtes (1%) : On a une amélioration de la *R-PREC* pour 4 requêtes et une amélioration de 1% de la MAP pour toutes les requêtes, ce qui signifie que le classement de certains documents pertinents s'est légèrement amélioré et qu'il n'y a pas de dégradation de la qualité de la recherche avec l'ajout de la sémantique.

Nous essayons d'expliquer pourquoi on n'a qu'une faible amélioration (1%) dans ce qui suit.

6.4.1.1 Impact de la co-occurrence

Le modèle $TDCintT$ met en place deux types de co-occurrences : une co-occurrence de termes et une co-occurrence de concepts. C'est le phénomène principal qui explique l'amélioration de la recherche avec le modèle TDCintT$_{Manu}$ même si elle est modeste.

Prenons l'exemple de la requête `REQ9` (voir section 6.3.3.1) où on observe une amélioration de la *R-PREC*. Cette requête est annotée par les concepts suivant de *MeSH* : `#adult, #women, #pregnancy, #first_pregnancy_trimester et #rh_isoimmunization`. Pour cette requête, le modèle TDCintT$_{manu}$ donne :
— une amélioration de la *R-PREC* qui passe de 0.5 avec le modèle TDintT à 0.8 avec le modèle TDCintT ;

— une amélioration des classements de tous les documents pertinents retrouvés, ce qui explique l'amélioration au niveau de la *MAP*.

De fait, on observe que les documents pertinents retrouvés sont tous activés dès la première itération du processus de propagation d'activation, qu'à la deuxième itération, les documents activés activent tous leurs termes et tous les concepts qui les annotent et qu'à la troisième itération, des phénomènes de co-occurrence permettent de :

— retrouver les concepts communs aux documents activés et à la requête : `#adult, #women #pregnancy, #rh_isoimmunization`, ainsi que d'autres concepts en accord avec la requête comme `#rh-hr_blood-group_system, #blood, #isoantibody`, etc. ;

— retrouver des termes co-occurrents avec ceux des documents activés, qui sont également pertinents pour la requête.

Les concepts et les termes activés par co-occurrence renforcent les valeurs d'activation des documents pertinents retrouvés à la première itération, dont les rangs s'améliorent à ce stade. On a par conséquent une amélioration de la *MAP* et de la *R-PREC*. Cette amélioration reste légère car l'impact de la sémantique vient un peu tard dans la propagation, du fait du relâchement rapide des valeurs d'activation avec la distance.

Considérons maintenant la requête `REQ31`[61] (voir AnnexeC). Cette requête est annotée par les concepts suivants de *MeSH* : `#adult, #women, #vein, #anticoagulant`, et `#warfarin`. Les résultats du modèle TDCintT donnent :

— une légère amélioration au niveau de la *MAP* qui passe de 0.14 avec le modèle TDintT, à 0.17 avec le modèle TDCintT ;

— une *R-PREC* stable à 0.10 pour les deux modèles (il n'y a pas de nouveau document dans les 19 premiers documents pertinents) ;

— deux documents pertinents de plus avec le modèle TDCintT qu'avec le modèle TDintT, mais ceux-ci sont insuffisamment activés pour être bien classés : il s'agit des documents `87155706` (rang 340) et `87225357` (rang 397).

Ces résultats s'expliquent principalement par les phénomènes de co-occurrence :

Co-occurrence entre termes : à la deuxième itération, la co-occurrence permet de sélectionner des termes proches de la requête : "necrosi", "induc", "heparin", "venou", "skin", "gangren", "thrombosi", "lupu", "warfarin" (synonyme de "coumadin"), "circul", "heart", etc. ; ces termes ayant des valeurs d'activation assez importantes, ils renforcent à la troisième itération les valeurs d'activation des documents pertinents activés à la première itération ; c'est ce que le modèle TDintT permet de faire d'ailleurs ;

Co-occurrence entre concepts : à la deuxième itération, la co-occurrence des concepts permet d'activer des concepts pouvant annoter les requêtes, comme `#human, #warfarin, #heparin, #coagulant, #blood_coagulation_factor, #prothrombine`[62], etc. ; ces concepts ayant déjà des valeurs d'activation assez importantes, ils :

— renforcent les valeurs d'activation des documents où ils co-occurrent : par exemple, le document `87302652` qui a été activé à la première itération et avait le rang 386, mais dont la valeur d'activation est renforcée à la troisième itération par les termes co-occurrents (notamment "warfarin" qui est synonyme de "coumadin") et les concepts co-occurrents (#human, #warfarin, #prothrombin) et qui avance ainsi au rang 169 ;

— activent de nouveaux documents pertinents là où leurs co-occurrents apparaissent : par exemple, les deux documents `87155706`(`#human, #warfarin, #prothrombin`), et

61. "Une jeune femme avec une Thrombose Veineuse Profonde (DVT), mise sous anticoagulant Coumadi".

62. prothrombine: examen du taux de coagulation du sang

`87225357`(`#prothrombin, #adult, #warfarin, #human`, etc.) ;
— résolvent implicitement la synonymie ; les concepts pertinents activés (ex. `#warfarin`) activent à la troisième itération de nouveaux documents qui activent à leur tour les termes qui y sont contenus (ex. le terme "warfarin") et qui peuvent être des termes proches des termes de la requête (ex. "warfarin" est synonyme de "coumadin").

6.4.1.2 Propagation de l'erreur

Considérons la requête `REQ7` étudiée précédemment et qui est annotée par les concepts `#women, #lactase,#lactose_intolerance`.

Pour vérifier si l'erreur constatée lors de la propagation d'activation du modèle TDinT se trouve corrigée par l'ajout des concepts de catégorisation, il suffit de vérifier si les documents qui contiennent l'abréviation "WF" sont encore bien classés à la fin de la recherche. On observe que ces documents `87130618`, `87170573`, `87226978` occupent toujours les rangs 1, 2, 3, avec le modèle TDCintT : les documents activés à la première itération restent toujours les mieux classés à la fin de la recherche.

Les deux documents pertinents qui n'ont pas été classés par le modèle TDintT ne sont toujours pas classés par le modèle TDCintT, bien qu'ils apparaissent à la troisième itération, par la co-occurrence des termes et des concepts. Ceci est encore une fois dû au relâchement rapide des valeurs d'activation avec la distance.

Si on filtre sur les concepts à la fin de la recherche, on trouve dans l'ordre des valeurs d'activation les plus importantes : `#human,#adult,#child, #t-lymphocyte, #animation, #middle_aged, #adolescent, #infant,..., #milk,#health,#infection, #breath_test, #lymphoma,#hydrogen_peroxide, #hydrogenation`, etc. On ne retrouve donc pas les concepts qui annotent la requête, ni d'ailleurs des concepts proches, hormis `#milk`. Cela signifie que les trois documents non pertinents ont eux aussi orienté l'activation des concepts et induit tout le graphe sémantique en erreur. L'apport de la sémantique n'est donc pas manifeste.

6.4.1.3 Conclusion

En étudiant la propagation d'activation avec le modèle TDCintT, on constate que l'ajout des classes sémantiques n'a pas beaucoup amélioré la qualité de la recherche. La co-occurrence des termes et des concepts qui annotent les documents contribue à l'amélioration des classements de certains documents pertinents et à la découverte de nouveaux documents pertinents mais les documents retrouvés à la première itération sont surclassés au cours de la propagation. Si des documents pertinents sont découverts à la deuxième itération, par la co-occurrence, ils restent mal classés du fait du relâchement rapide des valeurs d'activation à chaque itération. Par ailleurs, si des documents non pertinents sont activés à la première itération, ils gardent d'assez grandes valeurs d'activation à la fin de la recherche sur le graphe : la sémantique freine la propagation des erreurs mais n'en corrige pas la source.

Nous essayons dans ce qui suit d'interroger le même graphe avec cette fois des concepts de catégorisation, pour résoudre les problèmes d'ambiguïté et accélérer la prise en compte de la sémantique dès la première itération.

6.4.2 Interrogation par les concepts

Avec le modèle TDCintC$_{Manu}$, l'objectif est de mesurer l'apport de la co-occurrence entre termes. En effet, à partir des nœuds concepts de la requête, on s'attend à retrouver en premier lieu les documents annotés par ces concepts, puis tous les termes contenus dans ces documents,

mais les termes communs aux documents activés à la deuxième itération (co-occurrents) doivent avoir une meilleure valeur d'activation et ainsi permettre soit d'augmenter la valeur d'activation des documents retrouvés, soit de découvrir de nouveaux documents où ces termes co-occurrent.

Les résultats du modèle TDCintC sont donnés dans le tableau 6.5. Ils montrent :

	R-PREC	MAP
$TDCintC_{Manu}$	0.11	0.12
$TDCintT_{Manu}$	0.19	0.19

FIGURE 6.5 – Résultats de la propagation d'activation sur un graphe pondéré TDCintC$_{Manu}$

— une baisse globale de la *R-PREC* et de la *MAP* par rapport au modèle TDCintT ;
— une amélioration de la *R-PREC* par rapport au modèle TDCintT pour 11 requêtes ;
— un résultat stable pour 25 requêtes comparativement au modèle TDCintT.

Ces résultats s'expliquent principalement par deux raisons :

Problème de couverture : le fait d'interroger le graphe par les concepts uniquement ne permet pas d'interroger avec des termes orphelins qui n'ont pas de concepts associés dans *MeSH* ;

Prise en compte de l'aspect distributionnel : le fait d'interroger le graphe par les concepts de catégorisation ne permet pas de tenir compte de la fréquence d'occurrence des termes pertinents que tardivement ; par conséquent, tous les documents contenant un concept donné de la requête ont en première itération la même pertinence par rapport à cette requête, alors que le poids des différents concepts dans la requête et dans les documents peut être différent.

6.4.2.1 Problème de couverture

Considérons la requête `REQ57` [63] (voir AnnexeC) annotée par `#anemia, #hypochromic_anemia` et `#diagnosis`.

On peut facilement mettre en évidence un problème de couverture sémantique pour cette requête : certains termes de la requête, comme “iron deficiency“ ou “iron“, n'ont pas de correspondant sémantique, ce qui fait que la R-PREC de cette requête baisse de 0,5 avec le modèle TDCintT à 0 avec le modèle TDCintC, parce que certains documents pertinents pour la requête contiennent le terme “iron deficiency” (ex. les documents `87225360`(iron-deficiency(4)) et `87209636`(iron deficiency(4)).

Pourtant, on voit que dès la deuxième itération la propagation d'activation réussit par co-occurrence des termes à reconstituer les éléments de la requête : on retrouve alors dans l'ordre des nœuds termes à la fin de la recherche : “patient“, “anemia“, “cell“, “diseas”, “iron”, “ ‘aplast”, “infect”, “anaemia”, “defici”, etc.

6.4.2.2 Problème de calcul distributionnel

Prenons la requête `REQ35` [64] (voir Annexe C), qui est annotée par les concepts `#adult, #women` et `#bulimia`. On constate pour cette requête que la *R-PREC* chute de 0.78 avec le modèle TDCintT à 0.26 avec le modèle TDCintC, bien que tous les éléments clés de la requête possèdent des annotations.

63. “Une carence en fer, une anémie“.
64. “Une femme de 26 ans qui souffre de la boulimie. Évaluation des complications et gestion de la boulimie“.

A la première itération, les documents pertinents sont classés de 54 à 101, alors qu'ils sont tous annotés notamment par le concept `#bulimia` et que ces documents sont classés du rang 3 au rang 21 avec le modèle TDCintT.

A la deuxième itération, la co-occurrence des termes activés par les documents permet de classer les nœuds termes correspondant aux termes de la requête dans l'ordre suivant : "woman", "nervoa", "aneroxia", "bulimia", "eat", etc. ainsi que de nouveaux concepts proches des concepts de la requête comme `#human, #aneroxia_nervos, #aneroxia`, etc.

A la fin de la propagation, le classement des documents pertinents s'améliore par la prise en compte de l'aspect distributionnel, après activation des termes de la requête ; on obseve une nette amélioration, avec un nouveau document retrouvé au rang 137 (il s'agit d'un document pertinent mais qui n'est pas annoté par `#bulimia` au départ) mais cette amélioration des rangs reste insuffisante : l'aspect distributionnel intervient un peu tard dans le processus de propagation d'activation.

6.4.3 Interrogation par les termes et les concepts

Pour "sémantiser" davantage le graphe et résoudre d'emblée les erreurs de propagation, nous proposons d'interroger le graphe conjointement par les termes et par les concepts de catégorisation.

L'objectif est de profiter dès le départ de l'apport sémantique des classes sémantiques (concepts de catégorisation) et de l'aspect distributionnel de la RI classique. Cette interrogation doit permettre de corriger les erreurs d'identification des concepts de la requête par la co-occurrence, car les concepts de la requête sont spécifiés d'emblée lors de l'interrogation. Elle doit également favoriser la résolution des problèmes de couverture sémantique, car l'interrogation se fait par tous les termes de la requête.

Les résultats du modèle TDCintTC$_{Manu}$ présentés dans le tableau 6.6 montrent :

	R-PREC	MAP
$TDCintTC_{Manu}$	0.21	0.20
$TDCintT_{Manu}$	0.19	0.19

FIGURE 6.6 – Résultats de la propagation d'activation sur un graphe pondéré TDCintTC$_{Manu}$

- une amélioration de 2% de la *R-PREC* avec le modèle TDCintTC$_{Manu}$, par rapport au modèle TDCintT$_{Manu}$;
- des résultats de *R-PREC* et *MAP* stables pour 34 requêtes par rapport au modèle TDCintT. (voir figure 6.7) ;
- une amélioration de la qualité de la recherche pour 16 requêtes (voir figure 6.8) ;
- une dégradation de la qualité de la recherche pour 13 requêtes (voir figure 6.9).

Nous essayons de montrer sur quelques exemples, pourquoi la qualité de la recherche s'améliore dans certains cas et est dégradée dans d'autres.

6.4.3.1 Apport de la co-occurrence et de l'interrogation double

Prenons l'exemple de la requête `REQ48` [65] (voir Annexe C) pour laquelle on constate :

65. "Un personne de 24 ans avec le Virus de l'Immunodéficience Humaine (HIV), bilan de santé à propos de la démence due à la SIDA".

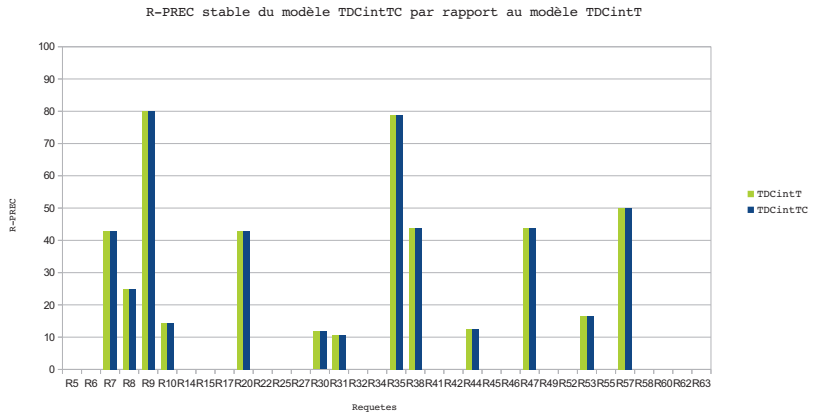

FIGURE 6.7 – Requêtes ayant une R-PREC stable entre les modèle TDCintTC et TDCintT

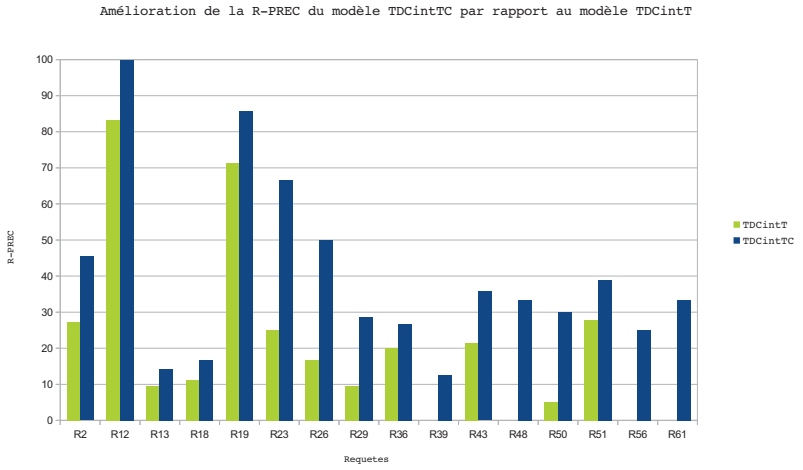

FIGURE 6.8 – Requêtes avec une R-PREC améliorée avec modèle TDCintTC par rapport au modèle TDCintT

— une amélioration de la *R-PREC* qui passe de 0.0 avec le modèle TDCintT à 0.33 avec le modèle TDCintTC ; trois documents pertinents remontent dans le classement et se retrouvent parmi les 9 premiers documents pertinents retournés ;
— une amélioration de la *MAP* qui passe de 0.07 à 0.17 avec le modèle TDCintTC.

En analysant le mécanisme de la propagation, on observe qu'à la première itération, les 3 documents qui contribuent à l'augmentation de la *R-PREC* sont déjà classés parmi les 9

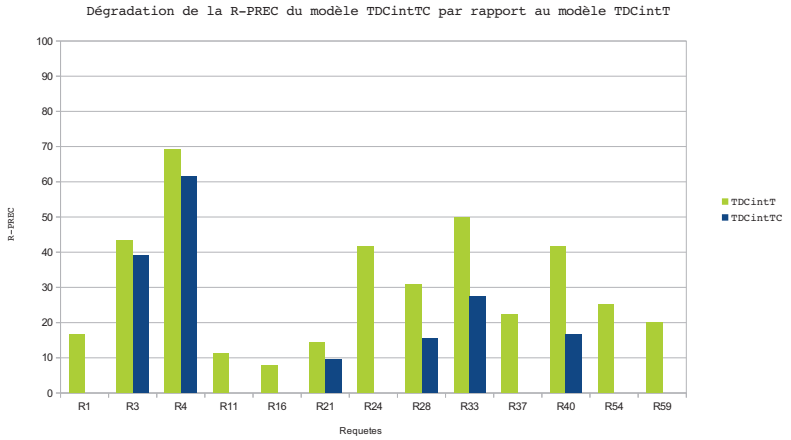

FIGURE 6.9 – Requêtes avec R-PRECs dégradées du modèle TDCintTC par rapport au modèle TDCintT

premiers documents retournés du fait de la complémentarité entre l'interrogation par termes et par concepts. Néanmoins, comme ces documents contiennent les concepts de la requête, leur classement s'améliore avec l'interrogation jointe par termes et concepts. Tous les documents pertinents sont retrouvés à la première itération.

A la fin de la propagation, les rangs des documents pertinents se retrouvent améliorés, car la co-occurrence a renforcé les valeurs d'activation des documents pertinents activés.

Au final, la plupart des documents pertinents voient leurs rangs améliorés par rapport au modèle TDCintT.

6.4.3.2 Impact de la taille des documents et du relâchement des valeurs d'activation

Avec la requête `REQ33`[66] (voir Annexe C), nous constatons :

— une dégradation de la *R-PREC* qui passe de 0.5 avec TDCintT à 0.27 avec le modèle TDCintTC ;

— une dégradation de la *MAP* qui passe de 0.34 avec TDCintT à 0.13 avec le modèle TDCintTC ;

— deux nouveaux documents pertinents retrouvés avec le modèle TDCinTC en plus de ceux qui sont retournés pas le modèle TDCintT.

Quand on analyse le processus de propagation en détail, on trouve dans les premiers rangs des documents, notamment des documents courts, qui contiennent les termes et concepts de la requête mais qui ne figurent pas dans les jugements de pertinence. Le fait d'interroger avec les concepts en plus des termes a fait remonter dans le classement ces documents qui passent un peu inaperçus autrement. A notre avis, il faudrait réviser les jugements de pertinence et considérer ces documents comme effectivement pertinents.

66. "Une femme de 65 ans avec une masse mammaire".

Même si les deux nouveaux documents retrouvés par le modèle TDCintTC restent mal classés (les documents `87139861` et `87318560` ont respectivement les rangs 553 et 567), il faut noter qu'ils sont retrouvés à la première itération et que leurs rangs s'améliorent ensuite avec la co-occurrence. Avec le modèle TDCintT en revanche, ils ne sont retrouvés qu'à la deuxième itération et ne sont pas classés.

Si on filtre sur les nœuds concepts pertinents à la fin de la propagation, on trouve les mêmes concepts et dans le même ordre avec les deux modèles TDCintT et TDCintTC : même sans interrogation par les concepts, la co-occurrence suffit à détecter les concepts associés à la requête.

6.4.3.3 Impact de la propagation d'erreur

Reprenons la requête `REQ7` déjà analysée plus haut (voir section 6.3.3.3). Nous avions détecté avec le modèle TDinT une erreur de propagation et qui n'était pas corrigée avec le modèle TDCinT. Quand on passe au modèle TDCintTC, on observe une *R-PREC* stable à 0.42, une petite dégradation de la *MAP* de 0.22 à 0.18 et un nombre constant de documents pertinents retournés.

Pour voir si l'erreur constatée lors de la propagation d'activation a été corrigée par l'interrogation avec des concepts de catégorisation, il faut regarder les rangs des documents qui contiennent l'abréviation "WF" . On constate que ces documents `87130618`, `87170573`, `87226978` occupent toujours les rangs 1, 2, 3, avec le modèle TDCintTC, signe que l'interrogation conjointe par termes et concepts n'a pas pu corriger cette erreur liée à l'ambiguïté de cette abréviation.

Cependant, si on filtre sur les nœuds termes à la fin de la recherche faite avec le modèle TDCintTC, on trouve dans l'ordre `#women, #lactase, #lactose_intolerance, #human, #physician, #adult, #health, #child, #case_report, #hospital, #pregnancy`, etc. alors qu'avec le modèle TDCinT on avait trouvé l'ordre suivant : `#human, #adult, #child, #t-lymphocyte, #animation, #middle_aged, #adolescent, #infant`, etc. On constate ainsi que les concepts de la requête (`#women, #lactase, #lactose_intolerance`) occupent les premiers rangs quand on les utilise pour interroger le graphe avec le modèle TDCintTC, alors qu'ils sont dilués dans les résultats du modèle TDCintT.

6.4.4 Conclusion

Il est difficile de faire la synthèse de ces expériences sur l'annotation manuelle. L'interrogation du graphe pondéré par les termes et les concepts améliore les résultats pour quelques requêtes mais on observe aussi des cas de dégradation qui sont dus pour certains à des lacunes dans les jugements de pertinence, et pour d'autres à des erreurs d'annotations. On remarque aussi que le fait d'interroger par les termes permet au modèle de tenir compte dès la première itération de l'aspect distributionnel, alors que l'interrogation par les concepts permet d'appuyer et d'orienter la sémantique de la propagation dès le départ. Il est enfin intéressant de noter que les concepts co-occurrents découverts par le modèle TDCintT correspondent dans la majorité des requêtes aux concepts qui annotent les requêtes.

Dans ce qui suit, nous refaisons les mêmes expériences, mais avec une annotation automatique, faite à l'aide de l'annotateur médical MetaMap.

6.5 Expériences avec l'annotation automatique par MetaMap

Cette section reprend certaines des expériences précédentes mais en procédant à l'annotation automatique de la collection *Ohsumed87* par l'annotateur médical MetaMap. Nous cherchons

en effet à remédier aux problèmes de couverture sémantique de l'annotation manuelle. Ceux-ci sont liés au fait que l'ontologie exploitée pour l'annotation manuelle n'est pas l'ontologie *MeSH* complète. Par ailleurs, l'annotateur médical MetaMap identifie, pour chaque concept reconnu dans le texte, le terme auquel il fait référence, ce qui permet de modéliser des liens terminologiques (termes-concepts) dans le réseau sémantico-documentaire, et donc de produire le modèle DTCintT$_{MM}$.

Après avoir décrit l'annotation sémantique obtenue par MetaMap, nous présentons les expériences réalisées avec cette annotation automatique.

6.5.1 Description de l'annotation sémantique

Nous exploitons l'annotateur MetaMap dans sa version 2016, téléchargée sur le site de la *National Library of Medicine*. Nous l'utilisons sous sa version locale, en l'interrogeant en ligne de commande. La commande utilisée est la suivante :

```
metamap -R MSH -y -XMLf document_entree document_sortie
```

tel que :

- l'option `-R MSH` permet de spécifier la ressource sémantique exploitée (dans notre cas : *MeSH* (`MSH`) ;
- l'option `-y` permet d'exploiter la désambiguïsation par défaut qui est implémentée dans MetaMap et qui consiste à sélectionner les concepts ayant les plus grands poids ;
- l'option `-XMLf` permet de sélectionner le format de sortie XML ; c'est le format le plus simple à parcourir et il mentionne explicitement le terme associé à chaque concept dans le document ;
- `document_entree` correspond au document à annoter ;
- `document_sortie` correspond au document résultant de l'annotation (dans notre cas, il est au format XML).

L'annotateur MetaMap a besoin de 8G de mémoire virtuelle (RAM) au minimum. L'annotation des 54 710 documents de la collection *Ohsumed87*, est de ce fait assez lourde. Elle nous a pris environ 1 mois.

Un échantillon de l'annotation en XML est donné dans la figure 6.10 : on voit que le concept `HIV infections` du *MeSH* a été identifié dans le document ; il figure dans la balise <CandidateMatched>. Ce concept a comme Identifiant Unique (CUI) `C0019693`, et comme label préféré "HIV infections". Le poids affecté à cette annotation est 853/1000, ce qui est assez élevé. Ce concept fait référence dans le texte à une séquence de 4 mots qui figurent dans la balise <MatchedWords> ("human", "immunodeficiency", "virus" et "infection") et qui forment le terme "human immunodeficiency virus infection". Ce dernier doit être normalisé avant d'être ajouté au réseau sémantico-documentaire.

Un parcours des documents annotés permet d'extraire les concepts associés à chaque document et les termes qui sont annotés. Ces informations sont stockées dans un format d'annotation allégé [67] qui sert de format d'entrée pour la construction du graphe. La figure 6.11 donne un exemple de ce format allégé : elle présente toutes les annotations du document `87312108` y compris le concept `HIV infections` et le terme "human immunodeficiency virus infection" auquel il fait référence dans le texte.

67. Il s'agit du même format imposé aux annotations données en entrée à la plate-forme *Terrier SIR*

```
<PhraseLength>58</PhraseLength>
<Candidates Total="11" Excluded="3" Pruned="0" Remaining="8" />
<Mappings Count="1">
 <Mapping>
  <MappingScore>-853</MappingScore>
  <MappingCandidates Total="1">
   <Candidate>
    <CandidateScore>-853</CandidateScore>
    <CandidateCUI>C0019693</CandidateCUI>
    <CandidateMatched>HIV Infections</CandidateMatched>
    <CandidatePreferred>HIV Infections</CandidatePreferred>
    <MatchedWords Count="4">
     <MatchedWord>human</MatchedWord>
     <MatchedWord>immunodeficiency</MatchedWord>
     <MatchedWord>virus</MatchedWord>
     <MatchedWord>infection</MatchedWord>
    </MatchedWords>
    <SemTypes Count="1">
     <SemType>dsyn</SemType>
    </SemTypes>
```

FIGURE 6.10 – Échantillon d'annotation *MetaMap* au format XML

```
<Document id="87312108">
  <Concept term=" screening">Screening</Concept>
  <Concept term=" human immunodeficiency virus">HIV</Concept>
  <Concept term=" public health">Public Health</Concept>
  <Concept term=" mandatory screening">Mandatory Testing</Concept>
  <Concept term=" program effectiveness">Program Effectiveness</Concept>
  <Concept term=" human immunodeficiency virus infection">HIV Infections</Concept>
  <Concept term=" united states">United States</Concept>
  <Concept term=" epidemiology">epidemiology</Concept>
  <Concept term=" human immunodeficiency virus">HIV</Concept>
  <Concept term=" antibodies">Antibodies</Concept>
  <Concept term=" human immunodeficiency virus">HIV</Concept>
  <Concept term=" logistic">Logistics</Concept>
  <Concept term=" economic">economics</Concept>
  <Concept term=" universal">Universal</Concept>
  <Concept term=" screening">Screening</Concept>
  <Concept term=" united states">United States</Concept>
  <Concept term=" human immunodeficiency virus 1">HIV-1</Concept>
  <Concept term=" individual">Persons</Concept>
  <Concept term=" cost">cost</Concept>
  <Concept term=" individual">Persons</Concept>
  <Concept term=" education">Teaching</Concept>
  <Concept term=" counseling">Counseling</Concept>
  <Concept term=" individual">Persons</Concept>
  <Concept term=" human immunodeficiency virus infection">HIV Infections</Concept>
  <Concept term=" mandatory screening">Mandatory Testing</Concept>
  <Concept term=" population">Population Group</Concept>
  <Concept term=" prevalence">Prevalence</Concept>
  <Concept term=" infection">Communicable Diseases</Concept>
  <Concept term=" use">utilization</Concept>
  <Concept term=" resources">Resources</Concept>
</Document>
<Document id="87312119">
  <Concept term=" genetic markers">Genetic Markers</Concept>
  <Concept term=" pre symptomatic disease">Asymptomatic Diseases</Concept>
```

FIGURE 6.11 – Annotation par MetaMap formatée du document 87312108

6.5.2 Modèle Terme/Document/Concept

Nous considérons dans cette partie le modèle TDC_{MM} qui s'appuie sur le même graphe pondéré que TDCintT_{Manu}, sauf que les concepts extraits *via* l'annotation automatique par MetaMap et ajoutés au graphe pondéré (seuls les concepts annotant les documents sont ajoutés).

Notre but, ici est de comparer l'annotation manuelle à l'annotation automatique de MetaMap avant d'utiliser cette dernière pour d'autres modèles de réseau sémantico-documentaire, notamment le modèle DTC (voir section 6.5.3) et le modèle DTCD (voir section 6.5.4).

Nous commençons par interroger le modèle TDC_{MM} d'abord par les termes (sous-section 6.5.2.1), puis conjointement par les termes et les concepts (sous-section 6.5.2.2).

6.5.2.1 Interrogation par les termes

Le graphe pondéré du modèle TDC_{MM} est interrogé au départ par les termes qui sont des unités lexicales normalisées des requêtes. Tous les termes t qui figurent dans la requête possèdent des valeurs d'activation initiales $a_0(t) = 1$ Le modèle TDCintT_{MM} est au départ comparé au modèle TDCintT_{Manu}, où l'annotation est faite manuellement à partir d'une sous-partie de *MeSH*.

Les résultats sont donnés dans le tableau 6.6 et montrent que les deux modèles reposant

Table 6.6 – Comparaison des résultats obtenus pour le modèle TDC avec une annotation manuelle (Manu) ou automatique (MM) et en interrogeant par les termes (TDCintT).

	R-PREC	MAP
TDCintT_{MM}	0,18	0,19
TDCintT_{Manu}	0,19	0,19

respectivement sur une annotation automatique et manuelle donnent des résultats très proches. On note une dégradation de 1% de la *R-PREC* pour le modèle TDCintT_{MM}. L'analyse des résultats requête par requête montre qu'ils sont quasi-stables pour les deux types d'annotations, à l'exception de la requête `REQ9`[68] (voir AnnexeC)., dont la *R-PREC* passe de 0,8 avec le modèle TDCintT_{Manu} à 0,3 avec le modèle $TDCintT_{MM}$.

Cette dégradation s'explique par des erreurs dans l'annotation automatique des documents pertinents de la requête `REQ9`. Ces erreurs sont dues à des ambiguïtés dans l'ontologie *MeSH* qui n'ont pas été résolues avec la désambiguïsation par défaut de Metamap. En effet, le terme de la requête "rh" (abréviation de "rhesus") a été annoté à tort dans quelques documents pertinents par le concept `#Macaca mulatta`, dont les labels dans *MeSH* sont "Macaca mulatta" (label préféré), "Rhesus Monkey", "Rhesus Macaque" et "Monkey, Rhesus". Cette annotation a engendré un changement dans le classement, qui a concerné les documents pertinents suivants : `87186146`(du rang 7 à 13) et `87158184` (du rang 6 à 14). D'autres documents pertinents ont également été annotés à tort par le concept `#Rhodium`, dont l'abréviation est aussi "rh". Cette annotation a concerné les documents pertinents suivants : `87272399` (du rang 8 à 12), `87210428` (du rang 5 à 11), `87077394`(du rang 4 à 10), et `87097496` (du rang 29 à 37). Ces documents sont de surcroît des documents courts pour la plupart d'entre eux, ce qui renforce la dégradation.

Cette faute d'annotation témoigne entre autres, de l'ambiguïté du terme "rh" dans *MeSH*. L'annotation manuelle est par conséquent plus efficace sur les termes ambigus.

Ces résultats sont par contre encourageants, car l'annotation par MetaMap ne semble pas beaucoup dégrader la qualité de la recherche par rapport à l'annotation manuelle, d'où on peut tester d'autres modes d'interrogation et d'autres modèles à partir de ces annotations.

68. *REQ*9 : "une femme enceinte de 3 mois, avec une rh isoimmunisation"

6.5.2.2 Interrogation par les termes et les concepts

On peut également interroger le graphe TDC$_{MM}$ par les termes et les concepts. Les résultats figurent dans le tableau 6.7, et montrent une nette amélioration par rapport au modèle

	R-PREC	MAP
TDCintTC$_{MM}$	0,30	0,30
TDCintT$_{MM}$	0,18	0,19
TDCintTC$_{Manu}$	0,21	0,20

TABLE 6.7 – Comparaison des résultats obtenus pour le modèle TDC avec une annotation manuelle (Manu) ou automatique (MM) et en interrogeant par les termes (TDCintT) ou conjointement par les termes et les concepts (TDCintTC).

TDCintT$_{MM}$ mais également par rapport au même modèle interrogé de même par les termes et concepts mais là où les annotations sont manuelles (le modèle TDCintTC$_{Manu}$).

Comparaison entre les modèles TDCintTC$_{MM}$ et TDCintT$_{MM}$

La comparaison des modèles TDCintTC$_{MM}$ et TDCintT$_{MM}$ montre une nette amélioration (12%) de la *R-PREC*. Dans le détail, on remarque que :

- cette amélioration concerne 29 requêtes, comme le montre le diagramme de la figure 6.12 ; elle vient principalement de ce qu'en interrogeant le graphe TDC conjointement par les termes et les concepts on profite dès la première itération des calculs distributionnels, mais aussi de la sémantique, pour corriger certaines ambiguïtés ;
  Prenons l'exemple de la requête `REQ58`[69](voir annexe C) dont la *R-PREC* passe de 0 à 1. Le terme de la requête "scheurmann" est un terme absent dans la collection documentaire, car il se trouve dans le texte sous sa variante lexicale "Scheu**e**rmann". Il s'agit d'un problème de *term mismatch*. Comme MetaMap a permis d'annoter et les documents et la requête avec le concept `#scheuermann's_disease`, l'interrogation double permet de résoudre ce problème : l'interrogation par les concepts (et les termes) et l'utilisation des liens d'annotation permettent ainsi de résoudre le *term mismatch*, comme nous avions vu l'interrogation par les termes et l'exploitation des liens terminologiques (voir chapitre5, sous-section 5.4.3) pouvaient le faire.
- la *R-PREC* est stable pou 28 requêtes (voir digramme 6.13) ;
- la *R-PREC* diminue légèrement pour 5 requêtes (voir diagramme 6.14), sauf pour la requête `REQ7`[70] analysée plus haut (voir section 6.3.3.3) où la diminution est importante. Une erreur dans l'annotation automatique de la requête `REQ7` est à l'origine de la dégradation de la *R-PREC*, car cette requête a été annotée à tort par le concept `#Rats, Inbred WF`. Ceci a renforcé l'erreur produite par le terme ambiguë "WF" , d'où la dégradation.

Comparaison entre TDCintTC$_{MM}$ et TDCintTC$_{Manu}$

Le modèle TDCintTC, qui profite de l'annotation automatique de MetaMap, semble plus efficace que le même modèle avec l'annotation manuelle (section 6.4). On remarque une amélioration de 9% de la *R-PREC*. L'examen des résultats requête par requête montre :

- une amélioration pour 27 requêtes (voir le diagramme 6.15) qui s'explique principalement par :

69. `REQ58:` une femme de 26 ans avec un dos thoracique, la maladie de Scheurmann et son traitement
70. *REQ*7 : "Une jeune femme avec une insuffisance en Lactase".

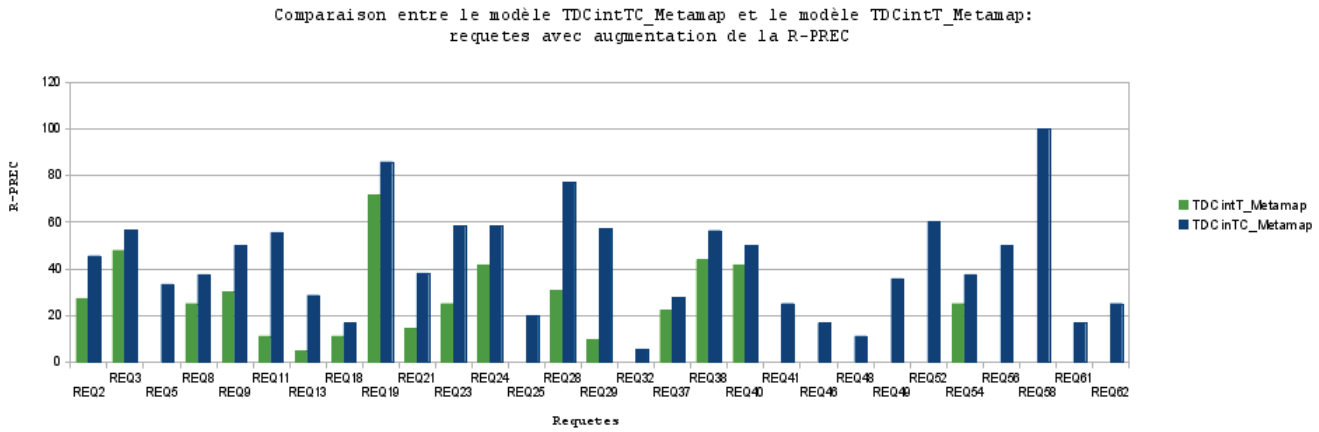

FIGURE 6.12 – Diagramme de comparaison entre les deux modèles TDCintTC$_{MM}$ et TDCintT$_{MM}$: augmentation de la R-PREC de quelques requêtes

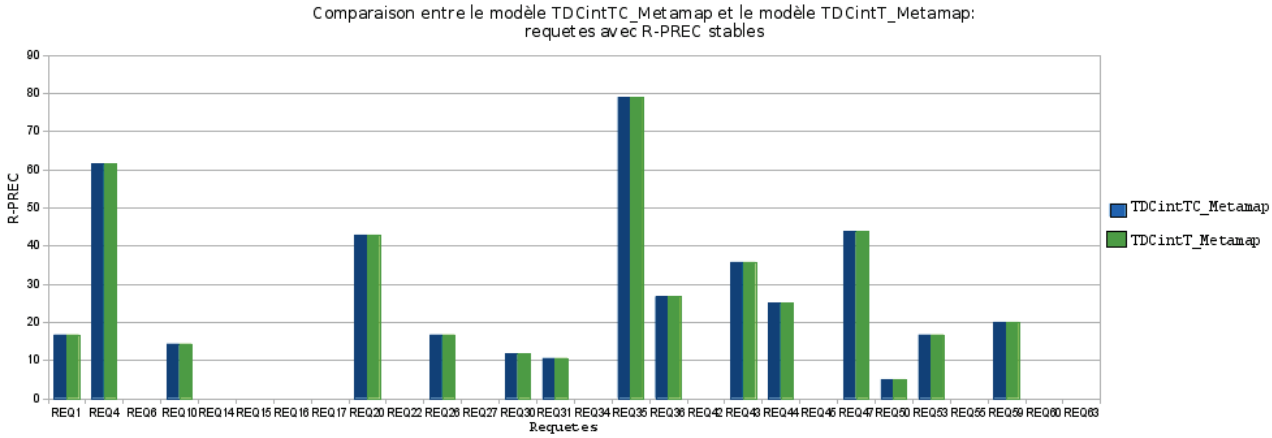

FIGURE 6.13 – Diagramme de comparaison entre les deux modèles TDCintTC$_{MM}$ et TDCintT$_{MM}$: requêtes à R-PREC stables

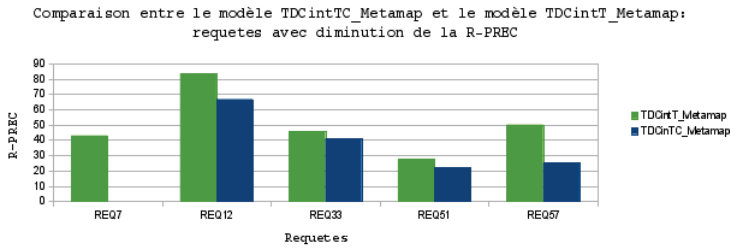

FIGURE 6.14 – Diagramme de comparaison entre les deux modèles TDCintTC$_{MM}$ et TDCintT$_{MM}$: diminution de la R-PREC de quelques requêtes

1. la résolution des problèmes de couverture par la ressource *MeSH* complète (utilisée dans MetaMap) : soit l'exemple de la requête `REQ58` [71] où l'annotation manuelle n'a

71. `REQ58`: une femme de 26 ans avec un dos thoracique, la maladie de Scheurmann et son traitement

pas permis d'annoter par le concept `#scheuermann's_disease` que donne l'annotation par MetaMap ;

2. l'exactitude et la précision de l'annotation automatique : si on prend l'exemple de la requête `REQ52` [72] (voir annexe C), on a une amélioration de la *R-PREC* qui passe de 0 à 0.6 ; cette requête est annotée manuellement par les concepts `#middle_aged, #lung, #lung_disease,#lung_absces` et `#drainage`, c'est-à-dire par des concepts proches de `#lung_absces` (`#lung, #lung_disease`) qui sont ici plus généraux ; avec Metamap, l'annotation de cette requête se limite aux concepts `#Lung Abscess, #Surgery` et `#Drainage`, ce qui minimise le bruit que peut introduire les concepts généraux proches.

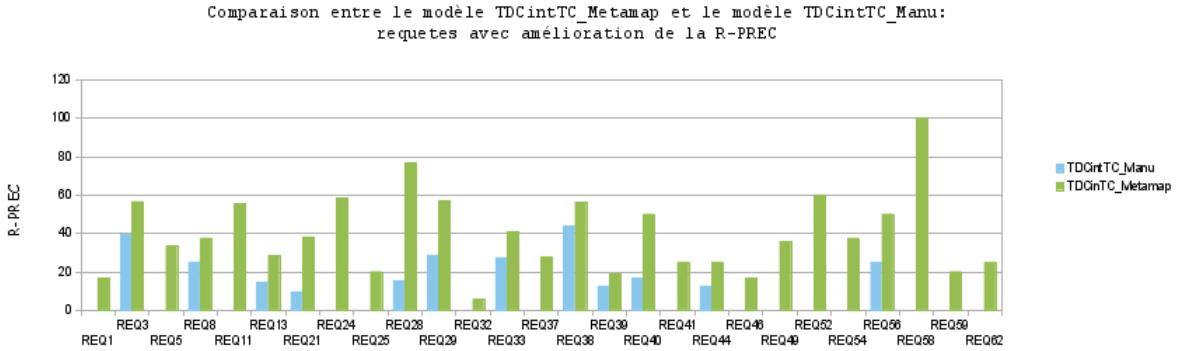

FIGURE 6.15 – Diagramme de comparaison entre les deux modèles TDCintTC$_{MM}$ et TDCintTC$_{Manu}$: amélioration de la R-PREC de quelques requêtes

— 25 requêtes tel que la *R-PREC* est stable (voir diagramme de la figure 6.16).

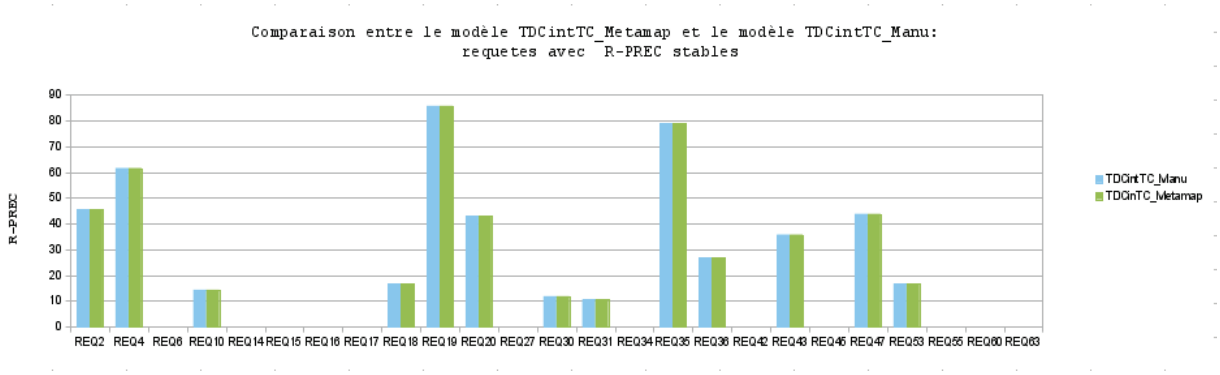

FIGURE 6.16 – Diagramme de comparaison entre les deux modèles TDCintTC$_{MM}$ et TDCintTC$_{Manu}$: R-PREC stable pour quelques requêtes

— une dégradation de la *R-PREC* au niveau de 10 requêtes (voir diagramme de la figure 6.17) : ces dégradations sont dans la majorité des cas dues à des fautes dans l'annotation automatique par MetaMap à cause de quelques problèmes d'ambiguïté. On cite par exemple le cas de la requête `REQ7` [73] analysée précédemment, tel que l'annotation du terme ambigu "WF" par MetaMap par le concept `#Rats, Inbred WF` a induit la pro-

72. une personne de 60 ans avec abcès pulmonaire, Chirurgie vs. drainage percutané pour abcès pulmonaire
73. *REQ7* : "Une jeune femme avec une insuffisance en Lactase".

pagation plus en erreur. On prend également le cas de la requête `REQ9`[74], là où la requête est annotée par les concepts `#Rh Isoimmunization` (annotataion du terme "rh") et `#Gravidity`[75](annotation du terme "pregnant"), alors que les documents pertinents ne le sont pas tous. En effet, comme le terme "rh" est annoté dans les documents pertinents par les concepts suivants : `#Macaca mulatta` et `#Rhodium`, et que le terme "pregnant" est annoté plutôt par le concept `#Pregnancy`, ceci a engendré une dégradation de la *R-PREC*. Cette requête témoigne donc de l'ambiguïté des termes de la requête "rh" et "pregnant" dans *MeSH* car il sont des labels respectivement des concepts `#Pregnancy`, `#Gravidity` et des concepts `#Rh Isoimmunization`, `#Macaca mulatta`, `#Rhodium`.
Par ailleurs, on est conscient que l'utilisation des liens ontologiques (hiérarchiques ou rôles) aurait peut-être permis de rapprocher les deux concepts `#Pregnancy` et `#Gravidity`, ou d'éloigner les concepts `#Rh Isoimmunization`, `#Macaca mulatta` et `#Rhodium`, et par conséquent évité la dégradation de la *R-PREC*.

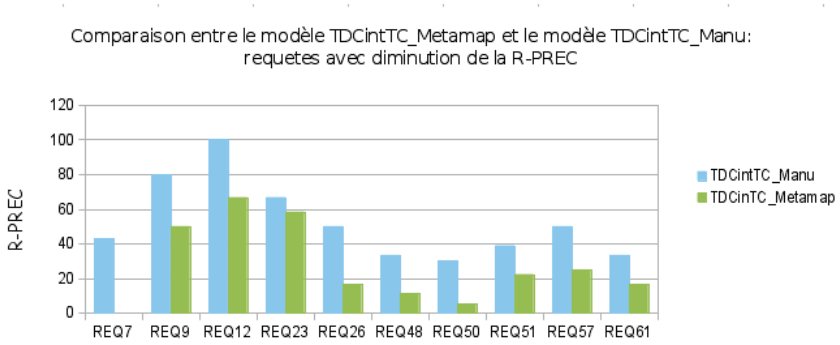

FIGURE 6.17 – Diagramme de comparaison entre les deux modèles TDCintTC_{MM} et TDCintTC_{Manu} : dégradation de la R-PREC pour quelques requêtes

6.5.2.3 conclusion

L'évaluation de la propagation par le modèle TDC_{MM} a permis de montrer que l'annotation automatique par Metamap est précise et améliore la qualité de la recherche, notamment quand on interroge conjointement par les termes et les concepts (modèle TDCintTC_{MM}).

Les expériences réalisées avec ce modèle ont également permis de se comparer à l'annotation manuelle de la section 6.4. On a alors conclu que l'annotation automatique a permis de résoudre certains problèmes de couverture, car l'ontologie *MeSH* exploitée pour MetaMap est complète, ce qui n'est pas le cas pour l'annotation manuelle. Elle a de même permis de résoudre certains problèmes de *term mismatch* entre le vocabulaire de l'utilisateur et la collection. On conclut également que le fait d'annoter que par le concept le plus spécifique (généralement un concept composé), et uniquement avec le concept spécifique (Metamap n'annote pas avec les concepts proches : père, fils, etc), a nettement amélioré les performances du système et minimisé le bruit.

Cependant ces expériences nous ont également permis de détecter certaines erreurs dans l'annotation automatique, mais qui sont généralement dues à des problèmes d'ambiguïté existants dans le vocabulaire (tel que l'annotation des termes "WF" et "rh") dont le désambiguïsateur de MetaMap n'a pas pu résoudre ou dans la ressource (le terme "pregnant" est label des deux concepts `#pregnancy` et `#Gravidity`).

74. *REQ*9 : "une femme enceinte de 3 mois, avec une rh isoimmunisation"
75. le nombre de fois où une femme est enceinte

De manière générale l'impact de l'annotation automatique et des liens de catégorisation sur la qualité de la recherche se trouve positif, et le modèle TDCintTC$_{MM}$ donne des résultats comparables même à *BM25* (voir section 6.3.1).

L'annotation automatique par MetaMap va nous servir dans la suite (section 6.5.3) pour tester de nouvelles traductions du graphe sémantico-documentaire, à savoir le modèle DTC$_{MM}$, puisque l'annotateur MetaMap fournit les termes auxquels les concepts font référence dans les documents et nous permet donc de modéliser les liens terminologiques.

6.5.3 Modèle Document/Terme/Concept

Dans cette partie, nous considérons le modèle DTC$_{MM}$ tel que le graphe pondéré est définit comme suit :

- les nœuds du graphe sont les documents, les termes et les concepts ($N = N_d \uplus N_t \uplus N_c$) ;
- les arcs représentent les relations d'occurrence et les relations terminologiques ($R = R_{occ} \uplus R_{ter}$) ;
- les valeurs des arcs des relations d'occurrence R_{occ}, pour un terme t et un document d sont $w_2 = w(t,d) = w(d,t) = \frac{tf(t,d)}{MaxTf(d)}$, tandis que les valeurs des arcs des relations R_{ter} entre les termes et les concepts sont fixées à 1 quand elles existent et à 0 sinon ;
- l'interrogation est soit par les termes (normalisés) de chacune des requêtes, soit par les termes et les concepts, tel que leurs valeurs d'activation initiales a_0 sont fixées à 1.

Les relations terminologiques R_{ter} sont extraites des annotations de MetaMap. Comme les termes qui font référence à un concept donné dans un document ne sont pas nécessairement des termes simples (`TS`),et peuvent être des termes composés(`TC`), on procède comme suit :

- les termes composés contenus dans les annotations des documents et ne figurant pas dans le graphe pondéré sont normalisés et ajoutés au graphe lors de sa construction ;
- les termes composés figurants dans les annotations des requêtes sont également normalisés et ajoutés à la requête et éventuellement au graphe ;

Ceci permet de garantir que tous les concepts annotant les documents sont connectés au graphe à travers les termes alors qu'il s'agit souvent de termes composés qui ne figurent pas dans l'indexation initiale fournie par Terrier IR et qui, autrement, n'apparaîtraient donc pas dans le graphe pondéré.

Nous appliquons tout d'abord la propagation d'activation sur le modèle sans sémantique, mais en ajoutant aux nœuds termes du graphe des termes composés extraits des annotations de MetaMap. On note ce modèle par TDintT$_{TSTC}$. Cette expérience doit nous permettre dans la suite de distinguer l'apport de l'ajout des termes composés et celui de la sémantique sur le graphe pondéré : il s'agit notamment de comparer le modèle DTC et le modèle DTCD que nous introduisons plus loin.

6.5.3.1 Modèle TD et exploitation des nœuds termes composés

	R-PREC	MAP
$TDintT_{TSTC}$	0,21	0,21
$TDintT$	0,18	0,18

TABLE 6.8 – Comparaison entre les modèles TDintT$_{TSTC}$ et TDintT

Les résultats du modèle TDintT$_{TSTC}$ sont présentés dans le tableau 6.8, et montrent que l'utilisation conjointement des `TS` et `TC` dans le graphe pondéré améliore nettement le résultat de

la propagation d'activation : on gagne 3% pour la *MAP* et la *R-PREC*. Ceci confirme encore une fois que l'exploitation des nœuds termes composés en plus des nœuds termes simples impacte la qualité des résultats, du fait qu'elle impacte le mode interrogatoire (on interroge par les `TS` et les `TC`), mais aussi les co-occurrences sur le graphe pondéré.

Nous allons alors désormais considérer le graphe pondéré avec et les `TS` et `TC` pour le modèle DTC_{MM}. Le but de ce modèle est bien évidemment de voir l'impact de l'exploitation des relations terminologiques sur la qualité de la recherche, en ce qui concerne la résolution de la synonymie, des problèmes de *Term mismatch*, etc., dans la lignée des conclusions que nous avions tirées pour le chapitre 5 mais en testant cette fois le modèle à grande échelle avec *Ohsumed87* et en l'interrogeant à la fois par les termes et par les termes et les concepts.

A noter que, dans la suite, aucune comparaison n'est possible avec le modèle TDC_{MM} ou avec le modèle classique pondéré par *Okapi-BM25*, étant donné que ces derniers ne disposent pas de la couche terminologique complète (avec les termes simples et composés), ce qui modifie l'ensemble des nœuds termes du graphe.

6.5.3.2 Modèle DTC et interrogation par termes

	R-PREC	MAP
$DTCintT_{MM}$	0,22	0,22
$DTCintTC_{MM}$	0,21	0,22
$TDintT_{TSTC}$	0,21	0,21

TABLE 6.9 – Comparaison entre le modèle DTCintT_{MM}, le modèle DTCintTC_{MM} et le modèle TDintT_{TSTC}

Le tableau 6.9 montre que si on compare les résultats des deux modèles TDintT_{TSTC} et DTCintT_{MM}, on constate des résultats très proches. On note globalement une petite amélioration de 1% en *MAP* et en *R-PREC* et l'analyse requête par requête montre que les résultats sont quasi-stables pour toutes les requêtes, avec pas ou de faibles améliorations.

Cette amélioration insignifiante est due, à notre avis, aux phénomènes de synonymie et de co-occurrence, mais les problèmes que nous avons identifiés dans la section 6.4 empêchent d'avoir une amélioration de plus grande ampleur.

On note un enrichissement intéressant de la composante lexicale de l'ontologie, qui s'est mis en place à partir des annotations sémantiques par MetaMap. Cet enrichissement aurait pu avoir plus d'impact sur la recherche, si on aurait résolu et les problèmes jadis cités : relâchement rapide des valeurs d'activation, la normalisation des valeurs des arcs entre termes et documents qui impose à *priori* que les documents courts soient privilégiés dans la recherche. On cite à titre d'exemple les concepts suivants :

— `#Lung cancer` (`REQ51`) : qui possède à présent plusieurs labels (synonymes) dans le graphe en plus de "lung cancer" et qui sont : "pulmonary malignant neoplasm", "cancer of the lung", "pulmonary malign", "malignant lung tumor", "lung primary canc", "pulmonary malignant tumor", "pulmonary canc", "cancer of lung", etc.

— `#Fever` (`REQ6`, `REQ32`, `REQ43`, `REQ49`) : qui à présent possède également plusieurs labels (synonymes) dans le graphe en plus de "fever" et qui sont : "highly temperatur", "body temperature elev", "increased temperatur", "pyrexia", "elevated core temperatur", "body temperature increas", "elevated temperatur", "high body temperatur", "hypertherm", "febril", etc.

D'autre part, le fait d'enrichir la couche terminologique par les termes composés augmente la taille du vocabulaire, ce qui peut engendrer un peu de bruit. En effet, l'enrichissement cité du volet terminologique des concepts de l'ontologie peut être considéré comme négatif, dans le sens où un nœud concept qui a beaucoup de labels peut être considéré comme un nœud très connecté au graphe et donc ayant un sens trop large. Pour résoudre ce point, on pourrait distinguer les labels "préférés" et les autres en attribuant des poids différents aux liens terminologiques. Nous n'explorons pas cette piste ici.

Comme nous avions noté, dans les premières expérimentations sur le modèle DTC, que le fait de passer par le concept pour profiter de la synonymie entre deux termes constitue un handicap (cela augmente la distance et conduit au relâchement de la valeur d'activation), nous avons aussi essayé d'interroger le modèle directement par les termes et par les concepts pour permettre aux labels synonymes des concepts de la requête d'être activés dès la première itération. Les résultats de l'expérimentation DTCintTC$_{MM}$ sont également présentés dans le tableau 6.9, et ne montrent pas d'amélioration à cause des mêmes phénomènes cités, et aux erreurs d'annotation détectées lors de l'annotation automatique, qui se renforcent quand on interroge par les mauvais concepts.

6.5.3.3 Conclusion

On ne peut pas conclure à ce stade que l'exploitation des liens terminologiques pour résoudre les problèmes d'ambiguïté améliore la qualité de la recherche pour un cadre de test large. Cependant, on est conscient que des comparaisons entre le modèle DTC$_{MM}$ et le modèle vectoriel pondéré par *BM25* et le modèle TDC$_{MM}$ (en fixant le mode d'interrogation) doivent être faites après avoir intégré les termes composés. En effet la stagnation de la qualité de la recherche provient de l'ajout des nœuds termes composés de l'apport des liens d'annotation dasn le cadre large de la collection *Oshumed87*, car ces premiers permettent de lier directement les classes sémantiques aux documents et de minimiser la distance sur le graphe.

Pour confirmer ou rejeter cette intuition, nous n'allons pas conduire à nouveau ces expérimentations, cependant nous allons plutôt expérimenter avec le modèle DTCD$_{MM}$, qui introduit ainsi les deux types de relations, terminologiques et d'annotation, avec une couche terminologique introduisant conjointement les termes simples et composés. Ce modèle nous permettra de se comparer au modèle DTC$_{MM}$, et de savoir l'apport des liens d'annotation.

6.5.4 Modèle Document/Terme/Concept/Document

Nous allons dans cette partie considérer le modèle DTCD$_{MM}$ tel que le graphe pondéré est composé comme suit :

- les nœuds du graphe $N = N_d \uplus N_t \uplus N_c$;
- les arcs du graphe présentent les relations suivantes $R = R_{occ} \uplus R_{ter} \uplus R_{ann}$;
- les valeurs des arcs des relations R_{occ}, pour un terme t et un document$d$, $w_2 = w(t,d) = w(d,t) = \frac{tf(t,d)}{MaxTf(d)}$, les valeurs des arcs des relations R_{ter} entre un terme et un concept, sont fixés à 1 si elles existent, et les valeurs des arcs des relations d'annotation R_{ann} sont également fixés à 1 ;
- l'interrogation est soit par les termes des requêtes normalisés, soit par les termes et les concepts, tel que leur valeurs d'activation initiales a_0 sont fixés à 1 ;

Le but de cette d'expérimentation avec le modèle DTCD$_{MM}$ est de voir l'impact de l'exploitation de toutes les relations disponibles possibles (R_{occ}, R_{ter} et R_{ann}), sur la qualité de la recherche, et l'apport de l'ajout des liens d'annotation au modèle DTC$_{MM}$, mais également de voir l'impact sur la taille du graphe pondéré, le nombre d'itérations de la propagation d'acti-

vation, et les temps d'exécution, à savoir le temps de construction du graphe et le temps de la propagation d'activation sur ce graphe.

6.5.4.1 Qualité de la recherche

Nous avons d'abord interrogé le modèle DTCD avec les termes seuls, puis avec les termes et les concepts conjointement. Le tableau 6.10 présente ces résultats en termes de *MAP* et *R-PREC*. Différentes comparaisons sont possibles à ce stade.

	R-PREC	MAP
DTCDintTC$_{MM}$	**0.26**	**0.28**
DTCDintT$_{MM}$	0.23	0.24
DTCintTC$_{MM}$	0.21	0.22
DTCintT$_{MM}$	0.22	0.22
TDintT$_{TSTC}$	0.21	0.21

TABLE 6.10 – Résultats obtenus en interrogeant par les termes (DTCDintT$_{MM}$) et par les termes et concepts conjointement (DTCDintTC$_{MM}$) le modèle DTCD construit à partir de l'annotation MetaMap et comparaison avec les modèles DTCintT$_{MM}$ et TDintT$_{TSTC}$.

La confrontation des deux modèles DTCDintTC$_{MM}$ et DTCintTC$_{MM}$ par exemple, laisse penser que l'ajout des liens d'annotation font gagner 5% de *R-PREC* et 6% de *MAP*. Ce gain vient de ce que les liens d'annotation présentent un lien direct entre un concept et un document, ce qui évite le relâchement des valeurs d'activation avec la distance (dans le modèle DTC, cette distance est égale à 2 car on passe par la couche terminologique), en plus d'autres facteurs comme la co-occurrence des concepts, qui intervient très tôt dans la propagation (deuxième itération) quand on dispose de ces liens concepts-documents.

On note également que l'interrogation double permet de profiter d'emblée du potentiel distributionnel et sémantique du modèle (comparaison des modèles DTCDintTC$_{MM}$ et DTCDintT$_{MM}$).

Il semble donc que toute connaissance ajoutée au graphe, au moment de sa construction ou lors de son interrogation, a un impact positif sur les résultats de la propagation d'activation. Ces améliorations ne sont certes pas très significatives mais on peut en espérer de meilleures si on arrive à corriger certains points comme le relâchement rapide des poids avec la distance ou la normalisation des arcs R_{occ} qui privilégie actuellement trop les documents courts.

Il faudrait enfin comparer notre modèle au modèle vectoriel classique par *BM25* en y introduisant les termes composés. Nous ne faisons pas ici car cela nécessiterait certains réglages de la plateforme *Terrier SIR*, à savoir l'utilisation d'extracteur de termes tel que *Yatea* [Aubin and Hamon, 2006],et peut-être le passage d'un *stemming* à une lemmatisation du texte pour la normalisation des termes.

6.5.4.2 Tailles des graphes et temps d'exécution

Le modèle sémantico-documentaire mis en place est un modèle simple. Nous n'avons pas introduit beaucoup de paramétrages et la propagation d'activation n'est pour l'instant contrainte que par le fait qu'un nœud ne peut propager son activation qu'une seule fois. Nous n'avons en effet pas mis en place des contraintes de chemin, de seuil, etc.

Il faut donc analyser le processus de propagation d'activation et vérifier qu'il ne faut pas un trop grand nombre d'itérations avant d'observer une stabilisation.

Modèle	Taille du graphe		Temps d'exécution		Iit.
	Nœuds	Relations	Constr.	PA	
TDintT	106 184	5 029 046	2h 25min	1jr 38min	6
TDintT$_{TSTC}$	139 583	5 738 204	5h 39min	1jr 6h 5min	6
DTCintT$_{metamap}$	165 554	5 829 516	9h 54min	1jr 6h 32min	8
DTCintTC$_{metamap}$	165 554	5 829 516	9h 54min	1jr 6h 30min	8
DTCDintT$_{metamap}$	165 554	7 772 340	11h 17min	1jr 17h 15min	6
DTCDintTC$_{metamap}$	165 554	7 772 340	11h 17min	1jr 17h 34min	6

TABLE 6.11 – Analyse du processus de propagation d'activation : taille du graphe construit à partir des différents modèles, temps d'exécution pour le construction du graphe (Constr.) et la propagation d'activation (PA) ainsi que nombre d'itérations requises pour la propagation d'activation (It.).

Le tableau 6.11 montre que le nombre d'itérations n'a pas dépassé les 8 itérations sur une collection de plus de 57 000 documents. Il montre également qu'ajouter des liens d'annotations au modèle *DTC* pour donner le modèle *DTCD* n'augmente pas le nombre d'itérations : bien au contraire, le graphe arrive à se stabiliser après un plus petit nombre d'itérations.

Le tableau 6.11 donne également une idée sur le nombre de nœuds et d'arcs qui composent chacun des modèles ainsi que des temps de construction des graphes pondérés et de propagation d'activation (PA) sur le graphe. Des optimisations sont donc à prévoir notamment sur le temps de propagation : : nous n'avons pas cherché jusqu'ici à faire des optimisations dans nos expérimentations et nous n'utilisons aucune contrainte alors que certaines pourraient aider à restreindre la propagation et à réduire le temps de recherche.

6.6 Bilan

Dans le premier chapitre expérimental (chapitre 5), notre but était d'explorer, sur un petit jeu de test, les potentialités du modèle proposé en termes de reproduction de la RI classique, de sémantique latente permise par le phénomène de co-occurrence et de résolution des problèmes classiques de RI(S) (synonymie, *term mismatch*, défaut de couverture, etc.) par l'exploitation d'une sémantique explicite.

Dans ce chapitre, l'évaluation du modèle sémantico-documentaire sur une collection de grande taille comme *Oshumed87* a permis en premier lieu de montrer que la capacité de notre modèle à passer l'échelle et sa robustesse. Ceci concerne autant la structure de représentation interne, grâce à la plate-forme de gestion des graphes *JUNG*, que l'algorithme de recherche ou les mécanismes de calcul et de contrôle de la propagation d'activation.

En deuxième lieu, l'expérimentation du modèle avec une collection de 57 000 documents a permis de de tester la qualité de la recherche sur un grand volume de données, avec différentes modélisations du réseau sémantico-documentaire. Ce chapitre présente donc de nombreuses expériences. Les résultats ne semblent pas spectaculaires mais nous avons cherché à analyser la manière dont notre méthode se comporte en faisant varier systématiquement certains paramètres :

— le type des connaissances prises en compte dans le réseau sémantico-documentaire : les documents, les termes et les relations d'occurrence évidemment mais aussi les concepts et les relations terminologiques ou d'annotation ;

— leur couverture : la couverture terminologique qui varie selon que l'on prend en compte les termes complexes ou seulement les termes simples et la couverture sémantique qui

dépend du type d'annotation exploitée, l'annotation automatique étant nécessairement plus systématique et plus couvrante qu'une annotation manuelle ;
— le mode d'interrogation qui peut reposer seulement sur les termes ou prendre aussi en compte les concepts.

Différentes modélisations en graphe pondéré et différents modes d'interrogation ont été testés, comme le résume le tableau 6.2. Nous avons procédé à différentes catégories d'expériences : les expériences de RI sans sémantique, avec sémantique et une annotation manuelle ou avec sémantique et une annotation automatique au regard de *MeSH*.

A l'issue des premières expériences, une comparaison entre le modèle vectoriel classique pondéré par *Okapi-BM25* et le modèle TD a été réalisée :
— les documents courts sont privilégiés au cours de la propagation d'activation sur le graphe pondéré : ceci provient du fait que la taille des documents est mal prise en compte dans notre formule de normalisation des poids des arcs des relations d'occurrences ;
— les valeurs d'activation diminuent trop vite avec la distance sur le graphe, ce qui atténue l'impact du phénomène de co-occurrence et restreint son rôle dans le renforcement des valeurs d'activation des nœuds activés précédemment, sans permettre d'en découvrir de nouveaux (les nouveaux nœuds pertinents découverts au fil de la propagation sont alors abandonnés à la fin de la recherche).

Les expérimentations menées avec l'exploitation de la sémantique tournent autour des modèles TDC, DTC et DTCD, interrogés par les termes ou conjointement par les termes et concepts, et ont permis de tirer les conclusions suivantes :
— la co-occurrence des termes et des concepts permettent d'améliorer le classement de la majorité des documents pertinents par rapport au modèle TD, à travers le renforcement de leurs valeurs d'activation ;
— l'interrogation conjointe par les termes et les concepts du graphe pondéré TDC, donne de meilleurs résultats que *BM25* quand l'annotation automatique est exploitée : le fait d'interroger par les termes permet de pallier les problèmes de couverture et de profiter dès le départ des calculs distributionnels ; le fait d'interroger par les concepts permet d'orienter la sémantique de la recherche tôt dans la recherche et de résoudre les problèmes d'ambiguïté, et de *term mismatch*, d'où une amélioration de la *R-PREC* au niveau de quelques requêtes ; cette expérimentation plaide aussi en faveur de l'exploitation des liens d'annotation pour la RI ;
— le modèle DTC_{MM} n'améliore pas beaucoup la qualité de la recherche par rapport au modèle TD_{TSTC}, en dehors de petites améliorations des classements des documents pertinents ; sa comparaison au modèle DTCD_{MM} prouve l'importance de l'exploitation des liens d'annotation par rapport aux liens terminologiques pour la RI ; la distance entre un concept et un document se trouve minimisée par l'exploitation des liens d'annotation ce qui réduit l'effet du relâchement des valeurs d'activation avec la distance ;

Les résultats de ces expérimentations peuvent sembler mitigés, mais il faut souligner que nous avons travaillé avec un algorithme de propagation générique. Nous n'avons modifié ni le mécanisme de contrôle de la propagation, ni les poids associés aux arcs, ni les valeurs d'activation initiales. Ces différents facteurs mériteraient d'être étudiés en détail mais il était difficile de faire varier tous les paramètres en même temps.

Pour résumer, aussi simple soit-il, le mécanisme de propagation que nous avons proposé permet de bien comprendre l'apport de l'approche par propagation d'activation pour la recherche d'information sémantique :
— le modèle de réseau documentaire que nous avons proposé peut intégrer différents types d'informations : nous aurions pu aller plus loin et prendre en compte par exemple des

relations intertextuelles (entre documents) ;
— notre approche permet de traiter de larges collections de documents et de grandes ressources sémantiques même si elle doit être optimisée pour être utilisée en pratique et pour passer réellement l'échelle ;
— nous intégrons dans une même approche les calculs distributionnels qui font la force des modèles de RI traditionnels et des calculs sémantiques qui permettent d'en dépasser les limites (prise en compte de phénomènes de co-occurrence, résolution du *term mismatch*, désambiguïsation) ;
— notre modèle permet de tester différents scénarios de RI(S), différents modes d'interrogation, différents moyens d'accès à l'information, et tout ceci sans changer de système et sans avoir à reconstruire l'espace de recherche.

La comparaison des résultats obtenus pour différents modèles et, dans certains cas, l'analyse détaillée des différentes requêtes montrent cependant que l'enrichissement du modèle n'apporte au mieux qu'un bénéfice marginal par rapport aux approches traditionnelles de RI mais nous avons identifié certains phénomènes qui limitent cet impact : il s'agit principalement de la vitesse avec laquelle les valeurs d'activation s'atténue quand on parcourt le graphe à cause des poids associés aux arcs dans le réseau sémantico-documentaire. Jusqu'ici, nous n'avons utilisé que des paramétrages simples mais il est clair qu'il faut aussi conduire des expériences pour déterminer la meilleure manière de paramétrer le graphe et le mécanisme de propagation qui s'y applique. D'autre part, l'exploitation de plus de sémantique sur le réseau sémantico-documentaire, notamment la structure de l'ontologie, en plus d'un phénomène aussi intéressant que l'est la co-occurrence, nous font espérer de grandes marges de bénéfice.

Chapitre 7

Conclusion et perspectives

Sommaire

7.1 Bilan des propositions

Cette thèse propose un modèle de RI sémantique, qui donne une représentation unifiée, cohérente et homogène d'un modèle documentaire et d'un modèle sémantique. Il intègre l'ensemble des informations disponibles, qu'elles soient symboliques ou numériques. Il préserve la nature des analyses et raisonnements que l'on fait traditionnellement sur les modèles documentaires et sémantiques, notamment les analyses distributionnelles largement éprouvées en RI et l'exploitation des connaissances sémantiques explicites ou implicites.

7.1.1 Modélisation

Notre modèle de RI sémantique repose sur une représentation en graphe et sur le mécanisme de propagation d'activation.

La modélisation en graphe a l'avantage de représenter de manière unifiée la collection documentaire, le besoin en information de l'utilisateur et une ressource sémantique (phase d'indexation). Quant au mécanisme de propagation d'activation, il permet de mettre en correspondance les requêtes et la base documentaire et sémantique : il consiste à propager l'information de pertinence de proche en proche sur le graphe en partant des éléments de la requête (phase de recherche).

Nous avons proposé de modéliser la collection documentaire et les ressources associées sous la forme d'un réseau sémantico-documentaire qui est composé de différents types de nœuds (documents, termes et concepts) et des relations possibles entre ces nœuds (relations d'occurrence, d'intertextualité, d'annotation, etc.). Ces différents nœuds et relations permettent d'intégrer dans une représentation unique les trois niveaux : terminologique, documentaire et sémantique. Ce réseau sémantico-documentaire se traduit sous la forme de graphe pondéré, où seuls les nœuds et

relations disponibles sont représentés et où les poids des arcs expriment la force des relations qu'ils encodent.

Ce graphe pondéré permet d'exprimer des propriétés classiquement utilisées en RI : la fréquence des termes dans les documents (poids des relations d'occurrence), le nombre de nœuds documents en relation avec un nœud terme, ou inversement le nombre d'arcs reliant un document à la couche terminologique. De même, le graphe pondéré prend en compte des propriétés sémantiques : le fait qu'un nœud terme est en relation avec plusieurs nœuds concepts peut traduire son ambiguïté, les nœuds termes connectés à un nœud concept forment le volet terminologique de ce concept (ensemble des labels), etc.

L'ensemble de ces informations distributionnelles et sémantiques encodées dans le graphe pondéré est pris en compte dans le mécanisme de propagation d'activation lors de l'appariement requêtes/documents : elles interviennent dans les calculs de pondération des nœuds. Les noeuds portent en effet des *valeurs d'activation* qui sont gérées par une fonction d'activation. La propagation d'activation est un processus itératif qui, à partir d'un ensemble de nœuds initialement activés (nœuds de la requête), diffuse l'information de pertinence de proche en proche sur le graphe pondéré. La fonction d'activation qui se charge du calcul des valeurs d'activation des nœuds activés, dépend de plusieurs facteurs, notamment :

La structure du graphe indique pour un nœud quels sont ses prédécesseurs et leurs degrés. La valeur d'activation d'un nœud est :

- proportionnelle à la somme des valeurs d'activation des nœuds prédécesseurs à l'itération précédente, ce qui nous rappelle les fonctions de correspondance du modèle vectoriel (produit scalaire, fonction cosinus, etc.) ;
- proportionnelle aux valeurs des arcs entre les nœuds, ce qui permet de tenir compte des fréquences terminologiques notamment ;
- inversement proportionnelle aux degrés de ces nœuds prédécesseurs, ce qui intègre les paramètres de fréquence documentaire, de taille de document, ou encore de d'ambiguïté conceptuelle.

Par analogie au modèle vectoriel classique, ces valeurs d'activation présentent les pondérations des termes, les scores des documents, etc.

L'état du graphe et des noeuds qui le composent est également important. Seuls les nœuds actifs à l'étape k du processus de propagation sont pris en compte dans les calculs d'activation de l'étape $k+1$. Cela permet que l'algorithme de propagation d'activation soit déterministe et de garantir sa terminaison en évitant les boucles infinies. Le mécanisme de contrôle de la propagation mis en place ne permet à un nœud de ne s'activer et de ne propager sa pertinence qu'une seule fois, avant d'atteindre son état final désactivé. En revanche, la valeur d'activation d'un nœud désactivé peut continuer à augmenter sous l'influence de la propagation de ses voisins.

Au final, le résultat de la propagation d'activation est la distribution des valeurs d'activation sur les nœuds du graphe. Un filtrage simple par type de nœud permet de retrouver, pour une requête utilisateur, l'ensemble des documents pertinents triés selon leurs valeurs d'activation. On peut également sélectionner différents types de nœuds, pour avoir en plus des documents les nœuds concepts ou les nœuds termes pertinents par rapport à la requête, ces éléments pouvant parfois servir à justifier la pertinence des documents retournés. Il faut noter que, pour ce type de modèle, le *ranking* des résultats est directement intégré à la phase de recherche (la propagation d'activation), qu'il dépend de la requête et qu'il repose sur les propriétés distributionnelles, sémantiques et documentaires tout à la fois.

7.1.2 Fonctionnalités du modèle

La modélisation en graphe proposée, le processus de propagation d'activation et le mécanisme de *ranking* des résultats permettent de reproduire la RI classique mais aussi de dépasser certaines de ses limites liées à l'ambiguïté de la langue.

C'est en effet un modèle de RI sémantique dont nous avons pu mettre en évidence certaines fonctionnalités sémantiques.

Rôle des classes sémantiques : A travers les concepts de la couche sémantique et les liens terminologiques qu'ils entretiennent avec le vocabulaire de la collection, des classes sémantiques sont modélisées dans le graphe pondéré ; elles permettent de :
- tenir compte de la synonymie (deux termes synonymes sont deux labels d'un même concept ou classe sémantique), ce qui a un impact positif sur la précision ;
- pallier au problème de *term mismatch* entre le vocabulaire de la requête et de la collection : on peut répondre à des requêtes comportant des termes hors vocabulaire en passant par le volet terminologique de l'ontologie, ce qui améliore notamment le rappel.

Rôle de la co-occurrence : La prise en compte de la co-occurrence entre les termes du vocabulaire fait partie des phénomènes intéressants que nous avons pu observer sur le graphe ; elle révèle une sémantique latente dans le graphe pondéré qui permet de désambiguïser le vocabulaire et de diminuer le bruit, mais également d'améliorer le rappel en activant de nouveaux documents contenant des termes co-occurrents.

Résolution de la couverture sémantique : Dans le cas des approches conceptuelles de RIS, certains problèmes de couverture peuvent surgir, quand des"termes orphelins" figurent dans la requête mais ne sont associés à aucun concept ; on peut dans ce cas limiter l'interrogation aux seuls termes sans changer de système ni de stratégie de pondération ou de recherche.

Rôle des concepts de catégorisation : La prise en compte des relations d'annotation sur le graphe pondéré permet d'exploiter la co-occurrence entre concepts et d'améliorer la précision, ou encore de découvrir de nouveaux documents pertinents et ainsi d'améliorer le rappel.

Notre modèle est en outre très polyvalent. Le graphe pondéré peut être interrogé de différentes manières (par les termes, les documents, les concepts ou même conjointement par les termes et les concepts, etc.), ce qui permet de dérouler plusieurs scénarios de RI(S) sans changer de système, de modèle ou d'indexation : recherche par mots-clefs, par concept, par l'exemple, etc. Selon le ou les type(s) de nœuds qui sont sélectionnés, on peut aussi passer d'un type de recherche à un autre : RI, RIS conceptuelle, recherche de données, etc.

7.1.3 Implémentation et validation expérimentale

Un travail d'implémentation et de test a été effectué pour mesurer la faisabilité et l'intérêt de l'approche de RIS proposée.

Nous avons enrichi la plate-forme de RI *Terrier IR* avec la plate-forme de gestion des graphes *JUNG* et proposé une implémentation modulaire. Notre système comporte deux modules :

Le module de construction du graphe pondéré est en charge de la construction du graphe, avec l'ensemble de ses nœuds, de ses arcs et la modélisation des propriétés associées aux couples de nœuds et/ou d'arcs. Un processus d'annotation sémantique permet de lier, à partir de l'ontologie exploitée, la couche documentaire et la couche sémantique

du graphe pondéré. Le graphe est construit à l'aide la plate-forme *JUNG* et des tests d'unicité sont établis sur les différents éléments du graphe.

Le module de propagation d'activation sur le graphe pondéré permet l'interrogation et la recherche sur le graphe. Le processus débute quand l'utilisateur interroge la collection documentaire et trois sous-processus sont déclenchés pour chaque requête utilisateur : (i) la recharge en mémoire du graphe construit, (ii) le pré-traitement de la requête qui permet d'identifier les unités correspondant à des nœuds du graphe et (iii) la propagation d'activation sur le graphe, qui débute par l'ancrage de la requête dans le graphe et l'initialisation des valeurs d'activation des nœuds de la requête.

Pour valider expérimentalement notre modèle, son implémentation et ses fonctionnalités sémantiques, nous avons monté différentes expériences sur deux collections documentaires. Le premier jeu de test est de taille réduite et a été utilisé à des fins exploratoires. Il nous a permis d'observer en détail le déroulement de l'algorithme de propagation d'activation et de nous assurer de sa viabilité face aux modèles de RI classiques comme le modèle vectoriel. A travers l'analyse de 6 requêtes, nous avons pu étudier les fonctionnalités offertes par notre modèle et l'impact de certains choix sur la qualité de la recherche. Le deuxième jeu de test a permis de tester le passage à l'échelle de notre modèle, différentes traductions du graphe pondéré [76] exploitant plus ou moins de connaissances sémantiques et différents modes d'interrogation. L'analyse des résultats obtenus montre l'impact de certains choix, notamment les types de connaissances mis en œuvre et le mode d'annotation (manuelle *vs.* automatique) utilisée.

Les expérimentations réalisées et l'analyse détaillée de certaines requêtes exemples montrent l'intérêt de l'approche proposée :

1. L'exploitation du modèle de RI sans sémantique (le modèle terme/document) montre que les choix de modélisation dans le graphe pondéré ont un impact direct sur la qualité de la recherche :
   - l'ajout des nœuds termes composés au graphe pondéré et à la requête améliore nettement la qualité de la recherche ;
   - la normalisation des valeurs des arcs termes/documents affecte beaucoup la qualité de la recherche, par rapport à un scénario sans normalisation ;
   - le modèle proposé permet de reproduire la RI classique : il tient compte de certaines propriétés documentaires et de la distribution des termes dans le document.
2. Différentes traductions du modèle sémantico-documentaire en graphes pondérés avec sémantique qui ont été testées : les graphes comportent tous des documents, des termes et des concepts mais, selon les cas, les concepts sont liés aux termes (DTC), aux documents (TDC), ou aux termes et documents à la fois (DTCD). Ces expériences montrent que le modèle proposé assure de surcroît des fonctionnalités sémantiques qui varient selon la nature des connaissances qui sont intégrées dans le graphe pondéré :
   - l'ajout de connaissances dans le graphe pondéré n'induit aucune dégradation de précision ou rappel par rapport au modèle sans sémantique (TD) et peut avoir au contraire un impact positif sur la qualité de la recherche :
     - les expériences sur le modèle DTC mettent en relief le rôle des classes sémantiques dans la résolution des problèmes de synonymie et de *term mismatch*, ainsi que le rôle de la co-occurrence entre les termes du vocabulaire ; une amélioration de la qualité de la recherche a été constatée ;

76. Modèles terme/document (TD), terme/document/concept (DTC), terme/document/concept/terme (DTCD), terme/document/concept (TDC), etc.

  - dans le modèle TDC où les liens entre l'ontologie et le vocabulaire de la collection manquent, on voit l'apport des concepts de catégorisation en RI ; ils permettent de "sémantiser" le graphe et de profiter de la co-occurrence entre concepts, ce qui se traduit par un gain en précision ;
  - les erreurs d'ambiguïté qui peuvent surgir au début de la propagation ne se propagent pas sur le graphe pondéré et n'affectent pas plus la qualité de la recherche que dans un modèle de RI classique ;
- le choix du mode d'interrogation a aussi un impact sur les résultats :
  - l'interrogation double par les termes et les concepts donne dans la majorité des scénarios de meilleurs résultats et c'est le modèle TDCintTC qui repose sur l'annotation automatique et l'ontologie MeSH complète qui a donné les meilleurs résultats en terme de *MAP* et *R-PREC* ;
  - l'interrogation par les seuls termes ne permet pas de tenir compte suffisamment tôt de l'aspect sémantique ;
  - l'interrogation par les seuls concepts est limitée par la couverture de la ressource sémantique et ne permet pas de profiter dès le début de la propagation de la distribution des termes de la requête dans les documents.

Les expérimentations réalisées montrent cependant certaines limites de la modélisation proposée qui peuvent expliquer la dégradation des résultats lors du passage à l'échelle par rapport au modèle vectoriel par $BM25$:

- La formule de normalisation des valeurs des arcs terme/document que nous utilisons ne fait pas un bon compromis entre la taille du document et sa pertinence par rapport à la requête utilisateur. Elle privilégie les documents courts dans la recherche et cela a un impact important dans les expériences sur la collection *Ohsumed87* qui contient beaucoup de documents sans partie textuelle (elle a été critiquée pour cela).
- Le relâchement rapide des valeurs d'activation donne un rôle important à la distance dans le graphe. Nous avons remarqué, lors de l'étude de certaines requêtes, que notre modèle fait apparaître des documents pertinents à la deuxième itération de propagation, principalement grâce aux phénomènes de co-occurrences, sans pour autant que ces documents obtiennent une valeur d'activation suffisamment haute pour figurer dans les R premiers documents retournés.

Au total et comparativement avec l'état de l'art des modèles de RI par propagation d'activation, notre approche repose sur un modèle simple, qui fait appel à peu de contraintes ou de paramétrages, et qui autorise la construction automatique du graphe pondéré. Ce modèle générique peut de surcroît exploiter différents types de ressources sémantiques.

Nous avons pu tester le mécanisme de propagation d'activation sur une collection de taille significative et l'étudier sur quelques exemples de requête. La recherche manque cependant d'efficacité mais nous avons mis l'accent sur la qualité de la recherche plutôt que sur les temps d'accès ou de réponse et il faut noter qu'aucune optimisation de l'algorithme n'a été réalisée à ce stade.

7.2 Perspectives

Par une étude théorique et expérimentale, nous avons montré l'intérêt de mettre en place un modèle dédié à la RIS mais beaucoup reste à faire, tant dans la formalisation du modèle proposé que sur le volet expérimental.

Concernant la formalisation et modélisation, il faut encore :

- formaliser les fonctionnalités sémantiques comme la co-occurrence, la synonymie, le *term*

mismatch, etc. que nous avons observées expérimentalement : nous avons travaillé de manière exploratoire et expérimentale et la théorie du modèle reste à construire ;
— étudier la manière dont la structure des ontologies (liens hiérarchiques et rôles) peut être exploitée dans la propagation d'activation, ce qui permettra de comparer notre approche avec les méthodes de RIS de l'état de l'art qui exploitent le voisinage sémantique : nous voudrions vérifier si la distance entre deux concepts sur le graphe pondéré permet de rendre compte de leur similarité sémantique ;
— tenir compte de la sémantique des liens, notamment les liens ontologiques ;
— étudier d'autres mécanismes de contrôle de l'algorithme de propagation d'activation ou d'autres contraintes de propagation (contrainte de chemin, de seuil, de connectivité, etc. qui sont citées de la littérature) avec l'objectif notamment de diminuer le bruit, en évitant la propagation dans tout le graphe.

D'un point de vue expérimental, nous voudrions à court terme :
— résoudre les problèmes rencontrés lors du passage à l'échelle, notamment en :
  — améliorant la normalisation des valeurs des arcs terme/document de manière à minimiser l'impact des documents courts ; cette question a été souvent abordée dans la littérature et nous nous intéressons notamment à la proposition de normalisation dite "à pivot" de [Singhal et al., 1996] (cette normalisation se traduit par une fonction à deux paramètres : *pivot* et *slope*) ;
  — travaillant sur le relâchement des valeurs d'activation et les facteurs de la fonction de propagation qui en sont responsables (le degré, la normalisation des valeurs des arcs terme/document, etc.) ;
— tester d'autres traductions du modèle sémantico-documentaire ainsi que d'autres modes de recherche (recherche par l'exemple, recherche de données, etc.) en jouant sur le mode interrogatoire et le type des nœuds filtrés en sortie ;
— enrichir le modèle en ajoutant d'autres types de liens comme les liens d'intertextualité ou les liens directs de synonymie entre les termes, etc.

A ces objectifs de modélisation et expérimentaux, s'ajoutent des questions que nous nous sommes posées en cours de route, qui nous ont préoccupées et que nous nous contentons de lister ici :
— Faut-il ou pas distinguer différents types de degré selon le type de nœuds dans la formule de propagation ? On peut en effet distinguer les degrés terminologiques, sémantiques, documentaires, etc.
— Comment peut-on contrôler l'activation des nœuds si l'on relâche la contrainte dure que nous avons introduite (un nœne peut s'activer qu'une seule fois) ?
— Faut-il différencier la fonction d'activation en fonction du type de nœud ?
— Quel est l'impact du passage par la classe sémantique (concept) et de la distance dans la résolution de la synonymie ?

A long terme, on pourra envisager d'enrichir le modèle proposé en :
— intégrant plusieurs ressources sémantiques dans le graphe sémantico-documentaire tout en garantissant la cohérence des analyses ;
— réduisant les temps de construction et de recherche sur le graphe pondéré ;
— proposant, à titre de trace et de justification pour l'utilisateur, une visualisation graphique de la propagation d'activation et de l'état du graphe ou d'une de ses sous-parties à chaque itération ;
— apprenant la structure et les poids du graphe pondéré à partir de données, de trace d'interrogation ou de retour des utilisateurs.

Bibliographie

[Agirre and Rigau, 1996] Agirre, E. and Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 16–22, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Amardeilh, 2008] Amardeilh, F. (2008). *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. PhD thesis, Paris.

[Aronson, 2001] Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. *Proc AMIA Symp*, pages 17–21.

[Aronson and Lang, 2010] Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap : historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3) :229–236.

[Aubin and Hamon, 2006] Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer.

[Aussenac-Gilles, 2008] Aussenac-Gilles, N. (2008). Le web sémantique, quel renouvellement pour la recherche d'information ? In Boughanem, M. and Savoy, J., editors, *Recherche d'information : état des lieux et perspectives*, Recherche d'information et Web, pages 97–132. Hermès, http ://www.editions-hermes.fr/.

[Baader et al., 2003] Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors (2003). *The Description Logic Handbook : Theory, Implementation, and Applications*. Cambridge University Press, New York, NY, USA.

[Badra et al., 2008] Badra, F., Bendaoud, R., Bentebibel, R., Champin, P.-A., Cojan, J., Cordier, A., Després, S., Jean-Daubias, S., Lieber, J., Meilender, T., Mille, A., Nauer, E., Napoli, A., and Toussaint, Y. (2008). TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In Schaaf, M., editor, *9th European Conference on Case-Based Reasoning - ECCBR 2008, Workshop Proceedings*, pages 219–228, Trier, Germany.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Bannour and Zargayouna, 2012] Bannour, I. and Zargayouna, H. (2012). Une plate-forme open-source de recherche d'information sémantique. In *COnférence en Recherche d'Information et Applications (CORIA)*, pages 167–178.

[Bannour et al., 2016a] Bannour, I., Zargayouna, H., and Nazarenko, A. (2016a). Modèle unifié pour la recherche d'information sémantique. In *27es Journées Francophones d'Ingénierie des Connaissances*, Montpellier, France.

[Bannour et al., 2016b] Bannour, I., Zargayouna, H., and Nazarenko, A. (2016b). Propagation d'activation dans les graphes pour la recherche d'information sémantique. In *Actes de l'atelier RISE (Recherche d'Information Sémantique)*.

[Baziz, 2005] Baziz, M. (2005). *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. PhD thesis.

[Baziz et al., 2003] Baziz, M., Aussenac-Gilles, N., and Boughanem, M. (2003). Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques . *Revue des Sciences et Technologies de l'Information (RSTI) série ISI*, 8(4/2003) :113–136.

[Beitzel et al., 2009] Beitzel, S. M., Jensen, E. C., and Frieder, O. (2009). Average r-precision. In *Encyclopedia of Database Systems*, page 195.

[Berners-Lee and Lassila, 2001] Berners-Lee, T. and Lassila, J. H. O. (2001). The semantic web. *Scientific American*.

[Bhagdev et al., 2008] Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., and Petrelli, D. (2008). Hybrid search : effectively combining keywords and semantic searches. In *Proceedings of the 5th European semantic web conference on The semantic web : research and applications*, ESWC'08, pages 554–568, Berlin, Heidelberg. Springer-Verlag.

[Bhatia, 1998] Bhatia, V. K. (1998). Intertextuality in legal discourse. *JALT Journal Online*, (1).

[Bhogal et al., 2007] Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information Processing and Management*, 43(4) :866 – 886.

[Bikakis et al., 2010] Bikakis, N., Giannopoulos, G., Dalamagas, T., and Sellis, T. (2010). Integrating keywords and semantics on document annotation and search. In *Proceedings of the 2010 international conference on On the move to meaningful internet systems : Part II*, OTM'10, pages 921–938, Berlin, Heidelberg. Springer-Verlag.

[Blanco and Lioma, 2012] Blanco, R. and Lioma, C. (2012). Graph-based term weighting for information retrieval. *Inf. Retr.*, 15(1) :54–92.

[Boubekeur and Azzoug, 2013] Boubekeur, F. and Azzoug, W. (2013). Concept-based indexing in text information retrieval. *CoRR*, abs/1303.1703.

[Boughanem et al., 2010] Boughanem, M., Mallak, I., and Prade, H. (2010). A new factor for computing the relevance of a document to a query. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–6.

[Bougouin et al., 2013] Bougouin, A., Boudin, F., and Daille, B. (2013). Topicrank : Graph-based topic ranking for keyphrase extraction. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 543–551.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy o f a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web (WWW7)*, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

[Brouard, 2013] Brouard, C. (2013). Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation. In *CORIA*, pages 1–10, Neuchâtel, France.

[Buckley, 1996] Buckley, C. (1996). New retrieval approaches using smart : Trec 4. pages 25–48.

[Castells et al., 2007] Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19 :261–272.

[Chevallet, 2009] Chevallet, J. P. (2009). *Ressources endogènes et exogènes pour une indexation conceptuelle intermédia*. Université de Grenoble.

[Claveau, 2012] Claveau, V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF. In *TALN - Traitement Automatique des Langues Naturelles*, pages –, Grenoble, France.

[Cleverdon et al., 1966] Cleverdon, C., Mills, J., Keen, M., and Project, A. C. R. (1966). *Factors Determining the Performance of Indexing Systems*. Number vol. 1,ptie. 1 in Factors Determining the Performance of Indexing Systems. College of Aeronautics.

[Cohen and Kjeldsen, 1987] Cohen, P. R. and Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manage.*, 23(4) :255–268.

[Collins and Loftus, 1975] Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6) :407 – 428.

[Collins-Thompson and Callan, 2005] Collins-Thompson, K. and Callan, J. (2005). Query expansion using random walk models. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 704–711, New York, NY, USA. ACM.

[Corby et al., 2006] Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., and Gandon, F. (2006). Searching the semantic web : Approximate query processing based on ontologies. *IEEE Intelligent Systems*, 21 :20–27.

[Crestani, 1997] Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6) :453–482.

[Crestani, 2000] Crestani, F. (2000). Exploiting the similarity of non-matching terms at retrieval time. *Information Retrieval*, 2(1) :27–47.

[Croft et al., 1989] Croft, W., Lucia, T., Cringean, J., and Willett, P. (1989). Retrieving documents by plausible inference : An experimental study. *Information Processing and Management*, 25(6) :599 – 614.

[Croft, 1986] Croft, W. B. (1986). User-specified domain knowledge for document retrieval. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '86, pages 201–206, New York, NY, USA. ACM.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6) :391–407.

[Déjean et al., 2010] Déjean, S., Baccini, A., Kompaoré, D., and Mothe, J. (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. *Revue des Sciences et Technologies de l'Information - Série TSI : Technique et Science Informatiques*, 29(3) :289–308.

[Despres et al., 2009] Despres, S., Zargayouna, H., and Bentebibel (2009). Quelles connaissances pour se mettre à taaable ? pages 151, 168.

[Dinh and Tamine, 2010] Dinh, D. and Tamine, L. (2010). Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients. *Conférence francophone en Recherche d'Information et Applications, CORIA 2010*, pages 325–336.

[d'Amato et al., 2012] d'Amato, C., Fanizzi, N., Fazzinga, B., Gottlob, G., and Lukasiewicz, T. (2012). Ontology-based semantic search on the web and its combination with the power of inductive reasoning. *Annals of Mathematics and Artificial Intelligence*, 65(2-3) :83–121.

[Fang and Zhai, 2005] Fang, H. and Zhai, C. (2005). An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 480–487, New York, NY, USA. ACM.

[Fang and Zhai, 2006] Fang, H. and Zhai, C. (2006). Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 115–122, New York, NY, USA. ACM.

[Fernandez et al., 2011] Fernandez, M., Cantador, I., Lopez, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced information retrieval : an ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, 9(4).

[Gonzalo et al., 1998] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarrán, J. M. (1998). Indexing with wordnet synsets can improve text retrieval. *CoRR*, cmp-lg/9808002.

[Guarino et al., 1999] Guarino, N., Masolo, C., and Vetere, G. (1999). Ontoseek : Using large linguistic ontologies for accessing on-line yellow pages and product catalogs. *National Research Council, LADSEBCNR.*

[Guha et al., 2003] Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 700–709, New York, NY, USA. ACM.

[Hamadan et al., 2012] Hamadan, H., Albitar, S., Bellot, P., Espinasse, B., and Fournier, S. (2012). Lsis at trec 2012 medical track – experiments with conceptualization, a dfr model and a semantic measure. In *The Twenty-First Text REtrieval Conference (TREC 2012) Notebook*, volume Special Publication, page 12 p., Gaithersburg (USA).

[Hersh et al., 1994] Hersh, W., Buckley, C., Leone, T., and Hickam, D. (1994). Ohsumed : An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192–201. Springer.

[Huang et al., 2012] Huang, L., Milne, D., Frank, E., and Witten, I. H. (2012). Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology.*

[Hussein et al., 2007] Hussein, T., Westheide, D., and Ziegler, J. (2007). Context-adaptation based on ontologies and spreading activation. In Hinneburg, A., editor, *LWA*, pages 361–366. Martin-Luther-University Halle-Wittenberg.

[Jacquemin and Zweigenbaum, 2000] Jacquemin, C. and Zweigenbaum, P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. *In Le document Multimédia en Sciences du Traitement de l'Information, CÉPADUÈS-Éditions, Toulouse.edition ACM press*, pages 71–110.

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 :11–21.

[Kang and Lee, 2005] Kang, B.-Y. and Lee, S.-J. (2005). Document indexing : a concept-based approach to term weight estimation. *Information Processing et Management*, 41(5) :1065 – 1080.

[Khan, 2000] Khan, L. R. (2000). *Ontology-based Information Selection.* PhD thesis, Faculty of the Graduate School, University of Southern California.

[Kiryakov et al., 2004] Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semant.*, 2(1) :49–79.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5) :604–632.

[Koopman et al., 2012] Koopman, B., Zuccon, G., Nguyen, A., Vickers, D., Butt, L., and Bruza, P. D. (2012). Exploiting snomed ct concepts and relationships for clinical information retrieval : Australian e-health research centre and queensland university of technology at the trec 2012 medical track. In *21st Text REtrieval Conference (TREC 2012)*, pages 1–8, Gaithersburg, Md. National Institute of Standards and Technology - NIST.

[Krovetz and Croft, 1992] Krovetz, R. and Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10(2) :115–141.

[Lei et al., 2006] Lei, Y., Uren, V., and Motta, E. (2006). Semsearch : A search engine for the semantic web. In *Proc. 5th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks, Lect. Notes in Comp. Sci., Springer, Podebrady, Czech Republic*, pages 238–245. Springer-Verlag.

[Lu and Keefer, 1995] Lu, X. A. and Keefer, R. B. (1995). Query expansion/reduction and its impact on retrieval effectiveness. In *Proceedings of the Third Text Retrieval Conference TREC-3*, pages 231–239.

[Maisonnasse, 2008] Maisonnasse, L. (2008). *Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale. (Vocabulary supports for precision oriented information retrieval systems : application to graphs for medical information retrieval).* PhD thesis, Joseph Fourier University, Grenoble, France.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, USA.

[Mihalcea and Moldovan, 2000] Mihalcea, R. and Moldovan, D. (2000). Semantic indexing using wordnet senses. In *Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval : Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11*, RANLPIR '00, pages 35–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Mihalcea and Tarau, 2004] Mihalcea, R. and Tarau, P. (2004). TextRank : Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 404–411.

[Miller, 1995] Miller, G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM*, 38(11) :39–41.

[Mimouni, 2015] Mimouni, N. (2015). *Querying a semantic network of documents : intertextuality in legal information access.* Theses, Doctorat de l'Université Paris 13 - Sorbonne Paris Cité.

[Mimouni et al., 2014] Mimouni, N., Nazarenko, A., Paul, È., and Salotti, S. (2014). Towards graph-based and semantic search in legal information access systems. In *Legal Knowledge and Information Systems - JURIX 2014 : The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014*, pages 163–168.

[Preece, 1981] Preece, S. (1981). *A Spreading Activation Network Model for Information Retrieval.* University of Illinois at Urbana-Champaign.

[Prud'hommeaux, 2008] Prud'hommeaux, E. (2008). Sparql query language for rdf.

[Quillian, 1968] Quillian, M. R. (1968). Semantic memory. In Minsky, M., editor, *Semantic information processing.* MIT Press, Cambridge.

[Radhouani, 2008] Radhouani, S. (2008). *Un modèle de recherche d'information orienté précision fondé sur les dimensions de domaine.* PhD thesis, Co-tutelle Université Joseph Fourier Grenoble, Université de Genève (Suisse).

[Rehder et al., 1998] Rehder, B., Littman, M. L., Dumais, S., and Landauer, T. K. (1998). Automatic 3-language cross-language information retrieval with latent semantic indexing. In *In The Sixth Text Retrieval Conference Notebook Papers (TREC6), 103–110. National Institute of Standards and Technology Special Publication*, pages 6–233.

[Réhel, 2011] Réhel, S. (2011). *Catégorisation Automatique de Textes Et Cooccurrence de Mots.* Editions universitaires europeennes EUE.

[Robertson et al., 1996] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1996). Okapi at trec-3. pages 109–126.

[Robertson and Jones, 1976] Robertson, S. E. and Jones, S. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3) :129–146.

[Robertson et al., 1998] Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at TREC-7 : automatic ad hoc, filtering, VLC and interactive. In *Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998*, pages 199–210.

[Rocha et al., 2004] Rocha, C., Schwabe, D., and Aragao, M. P. (2004). A hybrid approach for searching in the semantic web. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 374–383, New York, NY, USA. ACM.

[Sabetghadam, 2014] Sabetghadam, S. (2014). *Which One to Choose : Random Walks or Spreading Activation ?*, pages 112–119. Springer International Publishing, Cham.

[Salton, 1968a] Salton, G. (1968a). *Automatic Information Organization and Retrieval.* McGraw Hill Text.

[Salton, 1968b] Salton, G. (1968b). *Automatic Information Organization and Retrieval.* McGraw Hill Text.

[Salton, 1983] Salton, G. (1983). Some research problems in automatic information retrieval. In *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '83, pages 252–263, New York, NY, USA. ACM.

[Salton and Buckley, 1988a] Salton, G. and Buckley, C. (1988a). On the use of spreading activation methods in automatic information. In *SIGIR '88 : Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–160, New York, NY, USA. ACM Press.

[Salton and Buckley, 1988b] Salton, G. and Buckley, C. (1988b). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5) :513–523.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11) :613–620.

[Sanderson, 1994] Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 142–151, New York, NY, USA. Springer-Verlag New York, Inc.

[Savoy, 1992] Savoy, J. (1992). Bayesian inference networks and spreading activation in hypertext systems. *Information Processing and Management*, 28(3) :389 – 406.

[Schumacher et al., 2008] Schumacher, K., Sintek, M., and Sauermann, L. (2008). Combining fact and document retrieval with spreading activation for semantic desktop search. In Bechhofer, S., Hauswirth, M., Hoffmann, J., and Koubarakis, M., editors, *The Semantic Web : Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*, pages 569–583. Springer Berlin Heidelberg.

[Schütze and Pedersen, 1995] Schütze, H. and Pedersen, J. (1995). Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

[Shoval, 1981] Shoval, P. (1981). Expert/consultation system for a retrieval data-base with semantic network of concepts. In *Proceedings of the 4th Annual International ACM SIGIR Conference on Information Storage and Retrieval : Theoretical Issues in Information Retrieval*, SIGIR '81, pages 145–149, New York, NY, USA. ACM.

[Singhal et al., 1996] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 21–29, New York, NY, USA. ACM.

[Sirin et al., 2007] Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet : A practical owl-dl reasoner. *Web Semant.*, 5(2) :51–53.

[Sowa, 1984] Sowa, J. F. (1984). *Conceptual Structures. Information Processing in Mind and Machine.* The system programming series. Addison Wesley Publishing Company, Reading, MA.

[Spritzer, 2000] Spritzer, F. (2000). *Principles of Random Walk.* Springer.

[Stetina et al., 1998] Stetina, J., Kurohashi, S., and Nagao, M. (1998). General word sense disambiguation method based on a full sentential context. In *IN USAGE OF WORDNET IN NATURAL LANGUAGE PROCESSING, PROCEEDINGS OF COLING-ACL WORKSHOP.*

[Torsten et al., 2008] Torsten, S., , Ulf, L., and Jörg, H. (2008). *Word Sense Disambiguation in Biomedical Applications : A Machine Learning Approach.* Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration. IGI Global, 2009.

[Voorhees, 1999] Voorhees, E. M. (1999). The trec-8 question answering track report. In *In Proceedings of TREC-8*, pages 77–82.

[Wang et al., 2011] Wang, H., Tran, T., Liu, C., and Fu, L. (2011). Lightweight integration of ir and db for scalable hybrid search with integrated ranking support. *Web Semant.*, 9(4) :490–503.

[Xu and Croft, 2000] Xu, J. and Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1) :79–112.

[Zargayouna, 2005] Zargayouna, H. (2005). *Indexation sémantique de documents XML.* PhD thesis, Université Paris-Sud.

[Zargayouna et al., 2015] Zargayouna, H., Roussey, C., and Chevallet, J.-P. (2015). Recherche d'information sémantique : état des lieux. *Traitement Automatique des Langues*, 56(3).

[Zhao, 2012] Zhao, L. (2012). Modeling and solving term mismatch for full-text retrieval. *SIGIR Forum*, 46(2) :117–118.

Annexe A

Algorithme de propagation d'activation

Algorithm 1: Algorithme de propagation d'activation sur le graphe pondéré

```
1  Data: Graph
   Result: Graph
2  Integer i = 0;
3  HashMap ActifsNodes= new HashMap <MyNode,Double>;
4  repeat
5      // Recherche des nœuds actifs qui vont se propager
6      Collection Nodes = Graph.getVertices();
7      Iterator itOfNodes = nodes.iterator();
8      while (itOfNodes.hasNext()) do
9          currentNode= itOfNodes.next();
10         if (currentNode.getWeight() !=0) && (currentNode.getState()=="inactif") then
11             ActifsNodes.put(currentNode,currentNode.getWeight())
12         end
13     end
14     // Propagation des nœuds actifs
15     ItActifsNodes = ActifsNodes.keySet().iterator();
16     while (ItActifsNodes.hasNext()) do
17         currentActifNode = ItActifsNodes.next();
18         currentActifNode.setState("actif");
19         neighbors = getOutNeighbors(currentActifNode,Graph);
20         ItNeighbors = neighbors.iterator();
21         // Parcours des nœuds voisins de currentActifNode atteints
22         while (ItNeighbors.hasNext()) do
23             neighbor = ItNeighbors.next();
24             link = getLinkBetweenTwoNodes(graph, currentActifNode, neighbor);
25             double linkWeight = 1;
26             if (link != null) then
27                 linkWeight = link.getWeight();
28             end
29             currentActifNodeDegree = neighbors.size();
30             // Mise à jour la valeur d'activation du nœud neighbor activé par
                  currentActifNode
31             neighbor.weight +=
                NodesToFire.get(currentActifNode)*linkWeight/currentActifNodeDegree);
32         end
33         currentActifNode.setState("desactivated") ;
34     end
35     i=i+1 ; // Étape de propagation suivante
36 until (ActifsNodes.isEmpty()) // La liste des nœuds actifs est vide
37 ;
```

Annexe B

Jugements de pertinence

Requêtes	Jugements de pertinence
REQ1	Doc 355 : *Clear Spinach Soup (Korean Malgun Sigumchi Kuk)* Termes pertinents : soup(1), scallion(3)[1], Korean(1)
	Doc 479 : *Drunken leeks* Termes pertinents : leek(6)
	Doc 775 : *Ken's Hot-And-Sour Soup* Termes pertinents : soup(3), scallion(2), chinese(2)
	Doc 797 : *Leeks with Celery* Termes pertinents : leek(5), sauce(4)
	Doc 1363 : *Thai Spiced Mussel Soup with Leeks And Carrot Spaghetti* Termes pertinents : leek(3), soup(1), Thai(2), onion(2)
	Doc 731 : *Indonesian Soy Sauce* Termes pertinents : Asian(1), Indonesian(2), onion(2), sauce(4)
REQ2	Doc 1322 : *Summer Fruit Bowl* Termes pertinents : fruit(3), fruit bowl(1)fruit bowl[2], Lime juice(2)[3]
	Doc 585 : *Frozen Fruit Salad #1* Termes pertinents : fruit(3), salad(2), fruit Salad(1), lemon juice(1)
	Doc 904 : *Molded Waldorf Salad* Termes pertinents : fruit(2), salad(3), lemon(1), fruit salad(1)
	Doc 388 : *Cranberry Fruit Salad* Termes pertinents : fruit(1), salad(2), fruit Salad(1), lemon juice(2), orange juice(2)
	Doc 544 : *Festive Fruit bowl* Termes pertinents : Fruit bowl(1), fruit(4), salad(2), lemon(3)
	Doc 1467 : *Winter Fruit Bowl* Termes pertinents : fruit(3), fruit bowl(1), lemon(2), juice(2)

1. Scallion : oignon
2. Fruit bowl=Fruit salad
3. Lime juice=Lemon juice=Orange juice
4. kiwano=horned melon=jelly melon=melano=African Horned Cucumber
5. watermelon=pastèque(kiwano et watermelon deux fils de melon dans l'ontologie)
6. plum=prune
7. souffle=pizza
8. jambalaya=prépaation de riz et de saussissons

REQ3	Doc 1490 : *Salad with Edamame and Creamy Horned Melon Dressing* Termes pertinents : salad(4), horned melon(3)[4] , jelly melon(1)[4]
	Doc 1491 : *Kalahari melano Salad* Termes pertinents : melano(1)[4], salad(4), horned melon(2)melano
	Doc 1492 : *African Horned Cucumber Salad* Termes pertinents : salad(2), African Horned Cucumber(2)[4],melano(2)[4]
	Doc 1493 : *Chicken BLT Salad with Creamy Avocado and Horned Melon Dressing* Termes pertinents : salad(3), Horned Melon(2)[4], melano(1)[4]
	Doc 576 : *Fresh Watermelon salad* Termes pertinents : Watermelon(3)[5], salad(2)
REQ4	Doc 58 : *Baby Plum Cake* Termes pertinents : cake(1), plum(3)[6]
	Doc 141 : *Blueberry Cake Roll* Termes pertinents : cake(6), blueberry(2)
	Doc 144 : *Blueberry Crumbcake Squares* Termes pertinents : crumbcake(1), blueberry(2), cake(3)
	Doc 1055 : *Plum and Wheatgerm Muffins* Termes pertinents : muffin(2), plum(4)
	Doc 1323 : *Summer Fruit Pizza* Termes pertinents : plum(1), Peach(1), biscuit(5), shortcake(1)
	Doc 715 : *Hot Fruit Compote* Termes pertinents : plum(2), Peach(2)

REQ5	Doc 722 : *Idiot Bread Pizza* Termes pertinents : pizza(4), sausage(2))
	Doc 1051 : *Pizza Bread - Breadmaker* Termes pertinents : pizza(2), sausage(2)
	Doc 1053 : *Pizza Quiche* Termes pertinents : pizza(2), sausage(4)
	Doc 842 : *Pizza Quiche* Termes pertinents : sausage(5), souffle(2)[7]
REQ6	Doc 181 : *Brown Rice Jambalaya* Termes pertinents : rice(8), sausage(4),jambalaya(1)[8]
	Doc 200 : *Cajun Jambalaya* Termes pertinents : rice(4), sausage(2), jambalaya(1)
	Doc 714 : *Hot and Spicy Jambalaya* Termes pertinents : rice(5), sausage(2), jambalaya(1)
	Doc 749 : *Jambalaya #3* Termes pertinents : rice(2), sausage(2), jambalaya(1)
	Doc 1495 : *Skillet Sausage and Rice* Termes pertinents : rice (6), sausage(6)
	Doc 1496 : *Sausage fried rice* Termes pertinents : rice(6), sausage(4)

Annexe C

Quelques requêtes corpus Ohsumed 87

REQ9

```
<top>
<num>9</num>
<title>29 yo female 3 months pregnant</title>
<desc>Rh isoimmunization, review topics</desc>
</top>
```

Cette requête possède 10 documents pertinents dans les jugments de pertinence et qui sont les documents suivants : 87090179,87097505, 87186146, 87158184, 87283585, 87272399, 87210428, 87274428, 87077394 et 87097496.

REQ7

```
<top>
<num>7</num>
<title>young wf with lactase deficiency</title>
<desc>lactase deficiency therapy options</desc>
</top>
```

Cette requête possède 7 doucments pertinents dans les jugements de pertinence et qui sont les suivants : 87203237, 87103870, 87153184, 87104937, 87253647, 87124599, 87267477.

REQ31

```
<top>
<num>31</num>
<title>24 y. o. w. f. s/p DVT currently on coumadin</title>
<desc>course of anticoagulation with coumadin</desc>
</top>
```

Cette requête possède 19 documents pertinents dans les jugements de pertinence et qui sont les suivants : 87201761, 87058532, 87170012, 87302652, 87302655, 87181741, 87155706, 87096562, 87140264, 87141596, 87155637, 87079248, 87223174, 87318250, 87254703, 87245272, 87225357, 87138393 et 87063007

REQ57

```
<top>
<num>57</num>
<title>anemia</title>
<desc>iron deficiency anemia, which test is best</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#anemia</concept>
<concept>http://org.snu.bike/MeSH#hypochromic_anemia</concept>
<concept>http://org.snu.bike/MeSH#diagnosis</concept>
</annotation>
</top>
```

Cette requête possède 4 documents pertinents dans les jugements de pertinence et qui sont les suivants : 87225360, 87197122, 87209636, 87209897.

REQ35

```
 <top>
<num>35</num>
<title>26 y o female with bulimia</title>
<desc>evaluation for complications and management of bulimia</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#adult</concept>
<concept>http://org.snu.bike/MeSH#women</concept>
<concept>http://org.snu.bike/MeSH#bulimia</concept>
</annotation>
</top>
```

Cette requête possède 19 documents pertinents dans les jugements de pertinence et qui sont les suivants : 87125380, 87323663, 87111399, 87186290, 87239755, 87058486, 87050653, 87286014, 87155695, 87182112, 87309369, 87309370, 87240776, 87094393, 87154031, 87101884, 87325426, 87073356, 87143665

REQ48

```
<top>
<num>48</num>
<title>24 y o with HIV</title>
<desc>aids dementia, workup</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#adult</concept>
<concept>http://org.snu.bike/MeSH#hiv</concept>
<concept>http://org.snu.bike/MeSH#acquired_
immunodeficiency_syndrome</concept>
<concept>http://org.snu.bike/MeSH#dementia</concept>
<concept>http://org.snu.bike/MeSH#immunologic_
deficiency_syndrome</concept>
</annotation>
</top>
```

Cette requête possède 9 documents pertinents dans les jugements de pertinence et qui sont les suivants :

REQ33

```
 <top>
<num>33</num>
<title>65 yo female with a breast mass</title>
<desc>diagnostic and therapeutic work up of breast mass</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#aged</concept>
<concept>http://org.snu.bike/MeSH#women</concept>
<concept>http://org.snu.bike/MeSH#breast</concept>
<concept>http://org.snu.bike/MeSH#breast_disease</concept>
<concept>http://org.snu.bike/MeSH#breast_neoplasm</concept>
</annotation>
</top>
```

REQ10

```
<top>
<num>10</num>
<title>endocarditis</title>
<desc>endocarditis, duration of antimicrobial therapy</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#endocarditis</concept>
<concept>http://org.snu.bike/MeSH#bacterial_endocarditis</concept>
<concept>http://org.snu.bike/MeSH#drug_therapy</concept>
```

```
</annotation>
</top>
```

Cette requête possède 7 documents pertinents dans les jugements de pertinence et qui sont les suivants : 87110886, 87215782, 87069371, 87183439, 87183754, 87269476, 87139867

REQ58

```
<top>
<num>58</num>
<title>26 yo woman with mid-thoracic back pain</title>
<desc>scheurmann's disease, treatment</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#adult</concept>
<concept>http://org.snu.bike/MeSH#women</concept>
<concept>http://org.snu.bike/MeSH#thoracic_injury</concept>
<concept>http://org.snu.bike/MeSH#thoracic_vertebrae</concept>
<concept>http://org.snu.bike/MeSH#back_pain</concept>
</annotation>
</top>
```

REQ52

```
 <top>
<num>52</num>
<title>60 year old with lung abscess</title>
<desc>surgery vs. percutaneous drainage for lung abscess</desc>
<annotation>
<concept>http://org.snu.bike/MeSH#middle_aged</concept>
<concept>http://org.snu.bike/MeSH#lung</concept>
<concept>http://org.snu.bike/MeSH#lung_disease</concept>
<concept>http://org.snu.bike/MeSH#lung_absces</concept>
<concept>http://org.snu.bike/MeSH#drainage</concept>
</annotation>
</top>
```

Résumé

La recherche d'information sémantique (RIS), cherche à proposer des modèles qui permettent de s'appuyer, au delà des calculs statistiques, sur la signification et la sémantique des mots du vocabulaire, afin de mieux caractériser les documents pertinents au regard du besoin de l'utilisateur et de les retrouver. Le but est ainsi de dépasser les approches classiques purement statistiques (de « sac de mots »), fondées sur des appariements de chaînes de caractères sur la base des fréquences des mots et de l'analyse de leurs distributions dans le texte.

Pour ce faire, les approches existantes de RIS, à travers l'exploitation de ressources sémantiques externes (thésaurus ou ontologies), procèdent en injectant des connaissances dans les modèles classiques de RI de manière à désambiguïser le vocabulaire ou à enrichir la représentation des documents et des requêtes. Il s'agit le plus souvent d'adaptations de ces modèles, on passe alors à une approche « sac de concepts » qui permet de prendre en compte la sémantique notamment la synonymie. Les ressources sémantiques, ainsi exploitées, sont « aplaties », les calculs se cantonnent, généralement, à des calculs de similarité sémantique.

Afin de permettre une meilleure exploitation de la sémantique en RI, nous mettons en place un nouveau modèle, qui permet d'unifier de manière cohérente et homogène les informations numériques (distributionnelles) et symboliques (sémantiques) sans sacrifier la puissance des analyses. Le réseau sémantico-documentaire ainsi modélisé est traduit en *graphe pondéré*. Le mécanisme d'appariement est assuré par une propagation d'activation dans le graphe. Ce nouveau modèle permet à la fois de répondre à des requêtes exprimées sous forme de mots clés, de concepts ou même de documents exemples.

L'algorithme de propagation a le mérite de préserver les caractéristiques largement éprouvées des modèles classiques de recherche d'information tout en permettant une meilleure prise en compte des modèles sémantiques et de leurs richesse. Selon que l'on introduit ou pas de la sémantique dans ce graphe, ce modèle permet de reproduire une RI classique ou d'assurer en sus certaines fonctionnalités sémantiques. La co-occurrence dans le graphe permet alors de révéler une sémantique implicite qui améliore la précision en résolvant certaines ambiguïtés sémantiques. L'exploitation explicite des concepts ainsi que des liens du graphe, permettent la résolution des problèmes de synonymie, de *term mismatch* et de couverture sémantique. Ces fonctionnalités sémantiques, ainsi que le passage à l'échelle du modèle présenté, sont validés expérimentalement sur un corpus dans le domaine médical.

Mots-clés: Recherche d'information sémantique, Modèle, Ontologies, Réseau sémantico-documentaire, Graphe pondéré, Propagation d'activation

Abstract

Semantic information retrieval (SIR) aims to propose models that allow us to rely, beyond statistical calculations, on the meaning and semantics of the words of the vocabulary, in order to better represent relevant documents with respect to user's needs, and better retrieve them.

The aim is therefore to overcome the classical purely statistical (« bag of wordsé») approaches, based on strings' matching and the analysis of the frequencies of the words and their distributions in the text.

To do this, existing SIR approaches, through the exploitation of external semantic resources (thesauri, ontologies, etc.), proceed by injecting knowledge into the classical IR models (such as the vector space model) in order to disambiguate the vocabulary or to enrich the representation of documents and queries.

These are usually adaptations of the classical IR models. We go so to a « bag of concepts » approach which allows us to take account of synonymy. The semantic resources thus exploited are « flattened », the calculations are generally confined to calculations of semantic similarities.

In order to better exploit the semantics in RI, we propose a new model, which allows to unify in a coherent and homogeneous way the numerical (distributional) and symbolic (semantic) information without sacrificing the power of the analyzes of the one for the other. The semantic-documentary network thus modeled is translated into a *weighted graph*. The matching mechanism is provided by a *Spreading activation* mechanism in the graph. This new model allows to respond to queries expressed in the form of key words, concepts or even examples of documents. The propagation algorithm has the merit of preserving the well-tested characteristics of classical information retrieval models while allowing a better consideration of semantic models and their richness.

Depending on whether semantics is introduced in the graph or not, this model makes it possible to reproduce a classical IR or provides, in addition, some semantic functionalities. The co-occurrence in the graph then makes it possible to reveal an implicit semantics which improves the precision by solving some semantic ambiguities. The explicit exploitation of the concepts as well as the links of the graph allow the resolution of the problems of synonymy, term mismatch, semantic coverage, etc. These semantic features, as well as the scaling up of the model presented, are validated experimentally on a corpus in the medical field.

Keywords: Semantic information retrieval, Model, Ontologies, Graph, Spreading activation