Modèle de recherche d'information sémantique en graphe : interrogation par propagation d'activation

Ines Bannour, Haïfa Zargayouna, Adeline Nazarenko

LABORATOIRE D'INFORMATIQUE DE PARIS NORD (LIPN, UMR 7030)
Université Paris 13 – Sorbonne Paris Cité & CNRS
Email: prenom.nom@lipn.univ-paris13.fr

Résumé : La recherche d'information sémantique (RIS) cherche à dépasser les approches classiques purement statistiques en injectant des connaissances de manière à désambiguïser le vocabulaire ou enrichir la représentation des documents et des requêtes. Néanmoins, les modèles sémantiques, censés aller au-delà des mots et raisonner à un niveau conceptuel, sont en fait eux-mêmes « aplatis » : on ne représente plus les documents comme des « sacs de mots » mais comme des « sacs de concepts ».

Nous proposons de modéliser les données sémantiques et documentaires sous la forme de graphe pondéré et l'interrogation 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 recherche d'information tout en permettant une représentation adéquate des modèles sémantiques. 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 que l'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 un corpus dans le domaine médical (Ohsumed87) qui permet de vérifier le passage à l'échelle de notre modèle et de faire une première analyse qualitative et quantitative des performances de l'algorithme de propagation.

Mots-clés : graphe, ontologie, annotation sémantique, interrogation, propagation d'activation

1 Introduction

La recherche d'information (RI) consiste à retrouver des documents classés par ordre de pertinence qui répondent à une requête utilisateur généralement exprimée sous le formes de mots-clés. Les modèles de recherche d'information définissent une représentation des documents et des requêtes ainsi qu'une fonction de correspondance qui permet de calculer des similarités entre documents et requêtes et de classer les résultats. Quel que soit le modèle de recherche d'information et le paradigme sous-jacent (géométrique, probabiliste ou logique), les calculs sont purement numériques et reposent essentiellement sur la fréquence des mots et l'analyse de leur distribution. La recherche d'information sémantique (RIS) cherche à dépasser cette approche formelle, en injectant des connaissances de manière à désambiguïser le vocabulaire ou enrichir la représentation des documents et des requêtes.

Les travaux en recherche d'information sémantique sont nombreux et consistent pour la plupart à adapter les modèles classiques de RI (Zargayouna *et al.*, 2015) mais on observe que les modèles sémantiques, qui sont sensés permettre d'aller au-delà des mots et raisonner à un niveau conceptuel, sont en fait eux mêmes aplatis et on ne fait que passer d'une représentation des documents en « sacs de mots » à une représentation en « sacs de concepts ». Les calculs de correspondance intègrent les calculs de similarité sémantique entre concepts mais les liens sémantiques entre les mots ne sont pas explicités et exploités, alors même que c'est l'un des intérêts principaux des modèles sémantiques. La RIS semble atteindre un plateau, en dépit de l'intérêt toujours croissant qu'elle suscite et de l'influence des nouvelles technologies séman-

tiques du web de données et on peut se demander si l'approche consistant à adapter les modèles documentaires classiques par injection de sémantique n'a pas atteint ses limites.

Nous cherchons 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. Nous présentons dans ce travail, un nouveau modèle pour la recherche d'information sémantique qui repose sur une intégration du modèle documentaire et sémantique dans une représentation en graphe, la fonction de correspondance consistant à appliquer une fonction de propagation dans ce graphe.

Nous présentons dans un premier temps un état de l'art sur les travaux qui exploitent les ressources sémantiques en RI et nous dressons le bilan des approches sémantiques à base de graphes (section 2). Dans la section 3, nous détaillons notre modèle et expliquons le mécanisme de propagation dans les graphes que nous instantions. La section 4 présente les expériences faites sur la collection médicale *Ohsumed*.

2 État de l'art

2.1 Approches de RIS

L'exploitation des ressources sémantiques externes a pour but de pallier aux problèmes des modèles classiques de RI, comme le modèle vectoriel ou le modèle probabiliste.

S'attaquer aux problèmes des modèles classiques de type « sac de mots », revient 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 système de recherche d'information de renvoyer un document contenant non pas les mots de la requête mais des mots sémantiquement proches et ainsi de réduire le silence, ou, à l'inverse, d'exclure des documents ambigus et de réduire le bruit.

Le but de l'enrichissement de la représentation des documents est d'aboutir à une représentation fidèle des documents (et des requêtes) qui soit plus riche et moins ambiguë qu'une représentation classique par mots. Étant donné l'ambiguïté des ressources sémantiques les plus utilisées en RIS – notamment le thésaurus WordNet et les thésauri médicaux, UMLS et MeSH – leur utilisation en RIS repose souvent sur un processus préalable de désambiguïsation (voir entre autres les travaux de Stetina *et al.* (1998); Guarino *et al.* (1999); Khan (2000); Baziz *et al.* (2003)). Les étapes d'une indexation sémantique conceptuelle, qui tend à enrichir la représentation des documents, sont (i) l'annotation sémantique qui assure le lien entre les unités du texte et les unités sémantiques de la ressource et (ii) le choix des unités d'index et leur calcul de pondération. Ces unités d'index peuvent être des termes, des concepts ou encore une combinaison de termes et de concepts pour pallier le défaut de couverture sémantique de la ressource (Dinh & Tamine, 2010; Hamadan *et al.*, 2012).

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, 1968). Plusieurs méthodes d'expansion de requêtes existent dans l'état de l'art : Bhogal *et al.* (2007) passent en revue celles qui exploitent des ontologies.

Zargayouna *et al.* (2015) dressent un état de l'art de ces modèles classiques adaptés pour la RIS. Un bilan est difficile à dresser, l'évaluation d'un système de RIS reste complexe car 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.

2.2 Approches à base de graphes

Les travaux en Web Sémantique (WS) adoptent une modélisation en graphe de connaissances. Les documents sont représentés comme des instances, l'accès à ces documents se fait *via* les annotations sémantiques. Cela pose le problème de couverture des ressources, puisque toute information non annotée est perdue. De plus, les interrogations étant exactes, les réponses aux requêtes ne sont pas classées.

Les travaux récents en Web Sémantique proposent plus ou moins d'intégrer des moteurs classiques à des moteurs d'interrogation du WS (Zhang *et al.*, 2007; Fernández *et al.*, 2011; Wang *et al.*, 2011). Narula & Jain (2014) présentent un aperçu de quelques systèmes. Dans les approches de ce type, les documents retrouvés viennent en appui à la réponse factuelle apportée par le système, ils ne constituent pas le cœur de la réponse du système. Ces modèles d'accès sur le WS sont des modèles de recherche de données qui ont été adaptés pour la recherche de documents mais où le document est une information accessoire donnant la provenance des connaissances.

La RI a cependant intégré la notion de graphe bien avant l'avènement du WS. 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 années 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 (Brin & Page, 1998).

Nous proposons un modèle dédié à la RIS, qui représente les documents sous la forme de graphes intégrant des caractéristiques du modèle documentaire et du modèle sémantique. Ce modèle de RI permet de représenter de manière homogène des informations de natures différentes et de les intégrer dans un même calcul de pertinence. Nous présentons dans ce qui suit notre modèle en graphe, l'algorithme de propagation d'activation et les fonctionnalités sémantiques que ce type de raisonnement permet d'assurer.

3 Modèle en graphe et propagation d'activation

Notre modèle de RIS permet de représenter les données textuelles et leurs propriétés statistiques (fréquences d'occurrences, etc.) ainsi que les connaissances sémantiques issues d'ontologies dans un même *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.

3.1 Modélisation en graphe

Nous proposons de représenter la base documentaire et le modèle sémantique qui lui est associé sous la forme d'un réseau sémantico-documentaire. Cette structure permet d'introduire différents types de nœuds et différents types de relations selon ce qu'on souhaite représenter.

Le réseau *sémantico-documentaire* comporte de ce fait trois types de nœuds : nœuds documents (N_d), nœuds termes (N_t) et nœuds concepts (N_c). Ces nœuds sont liés les uns aux autres par cinq types de relations qui peuvent porter certaines propriétés numériques ou symboliques : les relations d'occurrence (R_{occ}), d'intertextualité (R_{int}), terminologiques (R_{ter}), lexicales (R_{lex}) d'annotation (R_{ann}) et ontologiques (R_{ont}) (Bannour *et al.*, 2016).

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, ainsi qu'en introduisant les éléments nécessaires au calcul distributionnel. Ces paramétrages sont dépendants des calculs à effectuer sur le graphe. Nous reviendrons sur cet aspect dans la section suivante.

Nous proposons de représenter ce type de réseau sémantique sous la forme d'un *graphe pondéré*, $G = \langle N, R \subseteq N \times \mathcal{R} \times N \rangle$, qui est constitué d'un ensemble de nœuds ($N = N_d \uplus N_t \uplus N_c$) et d'arcs orientés et pondérés ($R = R_{occ} \uplus R_{int} \uplus R_{ter} \uplus R_{lex} \uplus R_{ann} \uplus R_{ont}$).

Les arcs traduisent les relations qu'entretiennent les différents nœuds. Les pondérations peuvent être binaires ou calculées (relations booléennes ou numériques) pour prendre en compte la force des liens exprimés par les relations. 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, mais l'interprétation de ces valeurs dépend des types des nœuds 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 lexicale indique si deux termes sont variantes l'un de l'autre et éventuellement le degré de confiance qu'on a en cette relation ;

le poids d'une relation d'annotation indique si un concept est associé à un document ou non ; ce peut être une valeur booléenne ou une valeur de confiance fournie par l'outil de catégorisation ;

le poids d'une relation d'intertextualité indique si deux documents sont reliés (valeur booléenne) et éventuellement la force de cette relation (valeur numérique) [1] ;

1. 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.

le poids d'une relation ontologique indique s'il y a 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. On peut ainsi choisir par exemple de privilégier les liens de spécialisation dans la recherche ou au contraire de bloquer les parcours inverses, en leur attribuant des poids nuls ou négatifs, etc. [2].

Les arcs peuvent être orientés ou non, selon le type du lien :

les relations terminologiques, lexicales, d'occurrences et d'annotations peuvent être parcourues dans les deux sens et ne sont pas orientées ;

les relations d'intertextualité peuvent être orientées ou non (par ex. les relations `est_cité_par` et `cite` n'ont pas nécessairement le même poids [3]) selon le type de la relation considérée et les choix de modélisation ;

les relations ontologiques sont généralement considérées comme orientées quand il s'agit de relations hiérarchiques mais certaines relations peuvent aussi être considérées comme symétriques, si c'est explicité dans le modèle sémantique.

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, des documents, des concepts, une combinaison de différents types de nœuds, etc. De même, les réponses attendues peuvent être de différents types. Selon l'application, un filtrage est effectué sur l'ensemble des nœuds retournés pour sélectionner le/les type(s) de nœud(s) qu'on recherche. 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 sur ce graphe pondéré, consiste à appliquer une méthode de *propagation d'activation*, qui part des nœuds de la requête et active de proche en proche les nœuds voisins dans le graphe.

3.2 Appariement par propagation d'activation

La propagation d'activation (PA) est un processus qui permet de propager une information de proche en proche sur un graphe. Elle reprend 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 de sorte que l'activation en mémoire d'une unité active aussi par association les unités voisines.

Dans le cadre de l'accès à l'information dans le WS, la propagation d'activation sur les graphes de connaissances RDF a fait l'objet de plusieurs travaux qui proposent une hybridation de la RI et du WS avec une interrogation par mots-clés (Rocha *et al.*, 2004; Jiang & Tan, 2006; Schumacher *et al.*, 2008). Ces travaux nécessitent le plus souvent d'avoir des ontologies avec

2. La similarité entre concepts dépend plutôt de la distance dans l'ontologie par exemple et n'est donc pas une propriété intrinsèque.

3. 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.

une composante textuelles riche afin d'éviter le silence et de garantir une amélioration de la recherche par la prise en compte de la sémantique.

Crestani (1997) a formalisé le problème de la propagation d'activation pour la RI et Brouard (2013) a prouvé mathématiquement la correspondance entre son modèle de PA et le modèle vectoriel.

Nous détaillons dans ce qui suit le mécanisme de propagation d'activation que nous appliquons au graphe pondéré.

La propagation d'activation consiste en un ensemble d'*étapes de propagation* et un ou plusieurs *mécanismes de contrôle* de la propagation. A chaque étape, les *valeurs d'activation* associées aux nœuds du graphe sont calculées. 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 nœuds sources et de recalculer les *valeurs d'activation* des nœuds cibles.

A une étape de propagation n, l'*activation* s'applique 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). 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. Comme les graphes peuvent contenir des cycles, il n'est pas garanti que le processus de propagation itératif se stabilise et il faut introduire un mécanisme de contrôle.

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, comme celles qui sont présentées par Crestani (1997) : contrainte de distance, de chemin, *fan-out* et de seuil. Le mécanisme de contrôle fait partie intégrante de l'algorithme de propagation d'activation et repose sur *la modification de l'état des nœuds du graphe*.

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. 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. Un nœud désactivé ne peut plus être réactivé mais sa valeur d'activation peut continuer à croître sous l'influence de ses voisins. La terminaison de l'algorithme est assurée : il s'arrête quand l'ensemble des nœuds *actifs* est vide, c'est-à-dire au plus tard au bout de $|N|$ itérations, où N est le nombre de nœuds du graphe.

A chaque étape de propagation, de nouvelles valeurs d'activation a_k sont calculées sur la base des valeurs d'activation a_{k-1} et en fonction de la structure du graphe. Soient un nœud i et $a_{k-1}(x)$ la valeur d'activation d'un noeud x à l'issue de l'itération $k-1$. La valeur d'activation de i à l'itération k est définie par l'équation 1 suivante :

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

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

Les valeurs d'activation dépendent de *la structure du graphe* et de *l'état des nœuds du graphe*. 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. Ceci rend la fonction de mise à jour *synchrone* [4].

3.3 Fonctionnalités sémantiques

L'interrogation du graphe pondéré par différents points d'entrée, permet de :

— poser des requêtes en utilisant des termes qui n'existent pas dans le vocabulaire de la collection, mais peuvent appartenir au volet terminologique de la ressource sémantique, ce qui répond au problème de *term mismatch* décrit par Crestani (2000) ;

— 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 ;

— résoudre le défaut de couverture de la ressource en s'appuyant sur le vocabulaire de la collection.

Au cours de la propagation, certaines autres fonctionnalités sémantiques peuvent être assurées par notre modèle, *via* l'exploitation de la sémantique implicite ou explicite sur le graphe pondéré :

— la sémantique implicite se manifeste par le phénomène de *co-occurrence* des nœuds du graphe : la prendre en compte permet d'améliorer la précision, le rappel et le classement des résultats, par son impact sur les valeurs d'activation des nœuds, et grâce à la désambiguïsation sémantique du vocabulaire ou de la ressource ;

— la sémantique explicite se manifeste *via* la prise en compte des classes sémantiques (concepts de la couche sémantique), qui permet de traiter la *synonymie*, même en l'absence de ressource dédiée.

4 Expériences

Le but de notre expérimentation est double : tester le passage à l'échelle du modèle en vérifiant qu'on peut le déployer sur une grande collection et étudier les différentes traductions possibles du graphe pondéré ainsi que les différents modes d'interrogation qui en découlent.

La collection *Ohsumed* (Hersh *et al.*, 1994) est une sous collection de MedLine, qui est utilisée pour la tâche de filtrage de TREC-9. Elle rassemble 348 566 références de MedLine, extraites de 270 journaux médicaux sur une période de 5 ans (1987 à 1991) [5]. Nous utilisons dans la suite la partie de la collection datée de 1987, *Ohsumed87* : elle est composée de 54 710

4. C'est à dire indépendante de l'ordre dans lequel les valeurs d'activation des nœuds sont calculées.

5. En général, ces références comportent un titre et un résumé, mais certaines d'entre elles peuvent ne comporter qu'un titre.

documents et la taille moyenne des documents est de 74 56 termes. Nous disposons, toujours dans le cadre de TREC-9, de 63 requêtes avec leurs jugements de pertinence, 3 à 22 documents jugés pertinents par requête.

Des annotations manuelles ont été associées à la collection *Ohsumed* dans le cadre de la tâche de TREC-9, au regard d'une sous-partie du thésaurus MeSH qui a été formalisée en OWL et qui comporte 40 247 concepts (termes MeSH) avec des liens taxonomiques. Nous disposons également d'annotations manuelles des requêtes au regard de MeSH [6].

Nous avons réalisé une annotation automatique du corpus et des requêtes avec l'outil d'annotation METAMAP (Aronson, 2001) qui été conçu pour associer des concepts du méta-thésaurus UMLS à des documents, *a priori* médicaux. Pour expérimenter le passage à l'échelle de notre modèle, nous avons enrichi la plateforme sources ouvertes TERRIER SIR (Bannour & Zargayouna, 2012) avec la plate-forme de gestion et d'analyse des graphes JUNG [7].

Nous avons conduit plusieurs expériences comparatives. Nous avons commencé par distinguer deux modèles de graphe :

— le modèle terme/document (TD), un graphe minimal sans sémantique, avec comme nœuds des termes et des documents, et comme arcs, des liens d'occurrence (termes-documents) pondérés par la fréquence normalisée du terme dans le document [8] ;

— le modèle terme/document/concept (TDC), avec en sus des classes sémantiques et des liens d'annotation (liens documents-concepts avec une valeur de 1) entre les documents et les concepts qui les annotent ; ce modèle a lui-même deux instantiations selon qu'on prend en compte les liens d'annotation manuelle ($TDCman$) ou les liens d'annotation calculés automatiquement ($TDCauto$) [9].

Les résultats sont présentés en termes de R-Précision (R-PREC) qui est la précision après que R documents ont été retrouvés, où R est le nombre de documents pertinents pour la requête considérée. Le nombre d'itérations n'a pas dépassé les 8 itérations par requêtes mais lel temps de traitement des requêtes reste long : nous n'avons pas cherché à ce stade à optimiser l'algorithme de propagation mais à nous assurer de sa robustesse et des fonctionnalités sémantiques qu'il offre.

4.1 Impact de la sémantique

Il s'agit tout d'abord de comparer la RI sans sémantique représentée par le modèle considéré comme la *baseline* et la RI sémantique, en intégrant dans le graphe pondéré une couche sémantique formée des liens terminologiques (termes-documents) et des liens d'annotations (documents-concepts) provenant de l'annotation manuelle ou automatique. Les deux modèles, TD et TDC, sont interrogés par les termes.

Les résultats montrent une différence négligeable (on passe de 18% à 19%, soit un point de pourcentage d'amélioration) mais l'analyse détaillée des requêtes montre des éléments intéressants.

---

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

7. JUNG (*Java Universal Network/Graph Framework*) accessible à `http://jung.sourceforge.net/`

8. Pour un terme t et un document d : $w(t,d) = w(d,t) = \frac{tf(t,d)}{MaxTf(d)}$

9. Par défaut, ce sont les résultats de l'annotation manuelle qui sont rapportés.

La *co-occurrence des termes et des concepts* apporte une légère amélioration dans l'ordre des documents pertinents retournés pour certaines requêtes. Prenons l'exemple de la requête `REQ9` : « 29 yo female 3 months pregnant. Rh isoimmunization, review topics ». Nous observons une amélioration notable de la R-PREC au niveau de cette requête qui passe de 50% avec le modèle TD à 80% avec le modèle TDC : la co-occurrence des concepts, à la deuxième itération, a permis de renforcer les valeurs d'activation des documents pertinents activés à la première itération. Ces concepts co-occurrents correspondent aux concepts annotant la requête (`#adult, #women #pregnancy`, `#rh_isoimmunization,`) et à d'autres concepts en accord avec la requête comme `#rh-hr_blood-group_system`, `#blood`, `#isoantibody`, etc.

Il y a de *nouveaux documents pertinents découverts au cours de la propagation mais mal classés* en fin de propagation. En fait, on remarque que les documents retrouvés à la première itération restent en haut du classement au cours de la propagation. Ceci montre que les valeurs d'activation s'affaiblissent trop rapidement quand on parcourt le graphe et c'est un point sur lequel le paramétrage du modèle demanderait à être amélioré.

4.1.1 Impact de la densité du graphe

Nous comparons également les modèles $TDCman_{TC}$ et $TDCauto_{TC}$ pour analyser la propagation en fonction de la densité du graphe, les liens étant beaucoup plus nombreux dans le graphe construit à partir de l'annotation automatique. L'interrogation s'est faite par termes et concepts ($_{TC}$).

Le modèle $TDCauto_{TC}$, qui profite de l'annotation automatique de METAMAP, semble plus efficace que le même modèle avec l'annotation manuelle ($TDCman_{TC}$) : 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 de la figure 1) qui s'explique principalement par :

— *la résolution des problèmes de couverture* avec la ressource *MeSH* complète utilisée pour l'annotation automatique : pour la requête `REQ58` (« 26 yo woman with mid-thoracic back pain. scheurmann's disease, treatment. ») par exemple, l'annotation manuelle ne permet pas d'annoter le concept `#scheuermann's_disease` que donne l'annotation par METAMAP ;

— *l'exactitude et la précision de l'annotation automatique* : dans la requête `REQ52` (« 60 year old with lung abscess. surgery vs. percutaneous drainage for lung abscess »), on a une amélioration de la *R-PREC* qui passe de 0 à 60% ; 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`) mais plus généraux ; avec METAMAP, l'annotation de cette requête se limite aux concepts `#Lung Abscess, #Surgery` et `#Drainage`, ce qui minimise le bruit que peuvent introduire des concepts plus généraux.

On note également une dégradation de la R-PREC au niveau de 10 requêtes (voir diagramme de la figure 2). Ces dégradations sont majoritairement dues à des erreurs d'annotation faites par METAMAP pour des problèmes d'ambiguïté. Citons par exemple le cas de la requête `REQ7` (« young wf with lactase deficiency. lactase deficiency therapy options ») où l'annotation du terme ambigu « wf » par METAMAP avec le concept `#Rats, Inbred WF` induit la propagation en erreur.

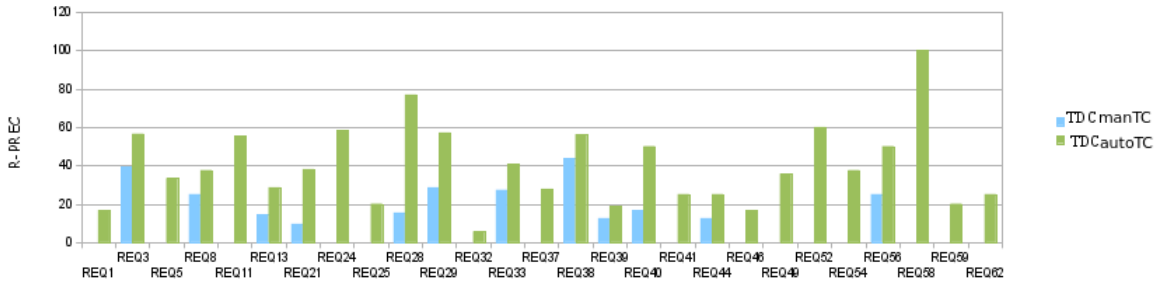

FIGURE 1 – Diagramme de comparaison entre les deux modèles $TDCman_{TC}$ et $TDCauto_{TC}$: amélioration de la R-PREC pour quelques requêtes

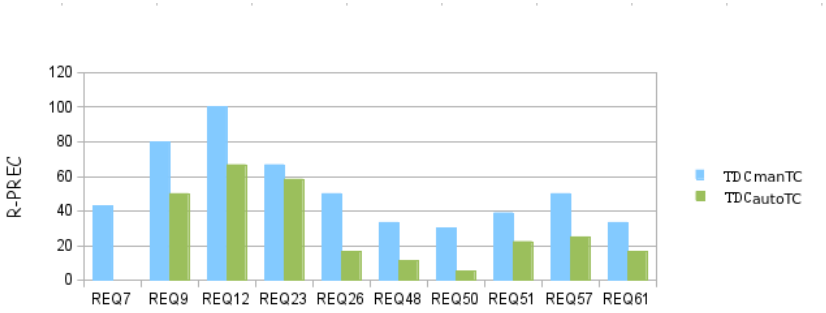

FIGURE 2 – Diagramme de comparaison entre les deux modèles $TDCman_{TC}$ et $TDCauto_{TC}$: dégradation de la R-PREC pour quelques requêtes

L'annotation automatique permet de résoudre certains problèmes de couverture, car l'ontologie MeSH exploitée par MetaMap est complète, ce qui n'est pas le cas pour l'annotation manuelle mais aussi certains problèmes d'ambiguïté et de *term mismatch* entre le vocabulaire de l'utilisateur et la collection. On note également qu'annoter par les concepts les plus spécifiques et uniquement avec ceux-là améliore nettement les performances du système en minimisant le bruit.

5 Conclusion et perspectives

Nous avons présenté dans ce travail un modèle dédié à la recherche d'information sémantique qui donne une vision unifiée des modèles documentaire et sémantique. 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.

Nous avons testé ce modèle sur un corpus médical et montré l'impact des nœuds et des liens pris en compte. Nous avons ainsi pu mesurer l'apport des classes sémantiques ainsi que des liens entre documents et concepts qu'ils soient construits manuellement ou automatiquement.

Ces expériences ont permis de faire une analyse à la fois quantitative et qualitative des résultats : elles montre les fonctionnalités sémantiques qu'apporte le modèle proposé et suggère des pistes d'amélioration. Ainsi, nous avons observé l'impact des problèmes de couverture des ressources. La prise en compte de la co-occurrence termes-concepts permet d'en atténuer l'effet. Nous avons aussi vérifié que nous arrivons à retrouver des documents même si les termes de la requête ne font pas partie du vocabulaire des documents (*term mismatch*).

Certaines erreurs d'annotation introduisent cependant des problèmes d'ambiguïté et dégradent les résultats pour quelques requêtes. Ces erreurs pourraient être corrigées avec l'ajout des liens entre concepts. Cette piste reste à explorer car elle nécessiterait de revoir le mécanisme de contrôle de la propagation qui provoque le relâchement trop rapide des valeurs d'activation avec la distance. Un autre mécanisme de contrôle permettrait de réduire le nombre de nœuds visités et agirait ainsi sur les temps de propagation. D'autres expériences sont à mener pour étudier en détail d'autres paramétrages, concernant par exemple les poids des différents arcs du graphe, mais il faut les conduire de manière systématique pour ne faire varier qu'un paramètre à la fois et analyser son impact sur les résultats globaux. Ces expériences nous permettraient à terme de proposer un calibrage automatique de ces paramètres.

Remerciements

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

Références

ARONSON A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. *Proc AMIA Symp*, p. 17–21.

BANNOUR I. & ZARGAYOUNA H. (2012). Une plate-forme open-source de recherche d'information sémantique. In *COnférence en Recherche d'Information et Applications (CORIA)*, p. 167–178.

BANNOUR I., ZARGAYOUNA H. & NAZARENKO A. (2016). Modèle unifié pour la recherche d'information sémantique. In *27es Journées Francophones d'Ingénierie des Connaissances*, Montpellier, France.

BAZIZ M., AUSSENAC-GILLES N. & 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.

BHOGAL J., MACFARLANE A. & SMITH P. (2007). A review of ontology based query expansion. *Information Processing and Management*, **43**(4), 866 – 886.

BRIN S. & 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)*, p. 107–117, Amsterdam, The Netherlands, The Netherlands : Elsevier Science Publishers B. V.

BROUARD C. (2013). Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation. In *CORIA*, p. 1–10, Neuchâtel, France.

CRESTANI F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, **11**(6), 453–482.

CRESTANI F. (2000). Exploiting the similarity of non-matching terms at retrieval time. *Information Retrieval*, **2**(1), 27–47.

DINH D. & 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*, p. 325–336.

FERNÁNDEZ M., CANTADOR I., LÓPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : an ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, **9**(4), 434–452.

GUARINO N., MASOLO C. & VETERE G. (1999). Ontoseek : Using large linguistic ontologies for accessing on-line yellow pages and product catalogs. *National Research Council, LADSEBCNR*.

HAMADAN H., ALBITAR S., BELLOT P., ESPINASSE B. & 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, p. 12 p., Gaithersburg (USA).

HERSH W., BUCKLEY C., LEONE T. & HICKAM D. (1994). Ohsumed : An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, p. 192–201 : Springer.

JIANG X. & TAN A.-H. (2006). Ontosearch : A full-text search engine for the semantic web. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, p. 1325–1330 : AAAI Press.

KHAN L. R. (2000). *Ontology-based Information Selection*. PhD thesis, Faculty of the Graduate School, University of Southern California.

MIMOUNI N., NAZARENKO A., PAUL È. & SALOTTI S. (2014). Towards graph-based and semantic search in legal information access systems. In *Legal Knowledge and Information Systems - JURIX*, volume 271, p. 163–168.

NARULA G. S. & JAIN V. (2014). Information retrieval (IR) through semantic web (SW) : an overview. *CoRR*, **abs/1403.7162**.

QUILLIAN M. R. (1968). Semantic memory. In *Semantic information processing*. Cambridge : MIT Press.

ROCHA C., SCHWABE D. & 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, p. 374–383, New York, NY, USA : ACM.

SALTON G. (1968). *Automatic Information Organization and Retrieval*. McGraw Hill Text.

SCHUMACHER K., SINTEK M. & SAUERMANN L. (2008). Combining fact and document retrieval with spreading activation for semantic desktop search. In *The Semantic Web : Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*, p. 569–583. Springer Berlin Heidelberg.

STETINA J., KUROHASHI S. & NAGAO M. (1998). General word sense disambiguation method based on a full sentential context. In *Proceedings of COLING-ACL workshop, Usage OF Wordnet in natural langage processing*.

WANG H., TRAN T., LIU C. & FU L. (2011). Lightweight integration of ir and db for scalable hybrid search with integrated ranking support. *Web Semant.*, **9**(4), 490–503.

ZARGAYOUNA H., ROUSSEY C. & CHEVALLET J.-P. (2015). Recherche d'information sémantique : état des lieux. *Traitement Automatique des Langues*, **56**(3), 49–73.

ZHANG L., LIU Q., ZHANG J., WANG H., PAN Y. & YU Y. (2007). Semplore : An ir approach to scalable hybrid query of semantic web data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, volume 4825 of *LNCS*, p. 645–658, Berlin, Heidelberg : Springer Verlag.