Modèle unifié pour la recherche d'information sémantique

Ines Bannour, Haifa Zargayouna, Adeline Nazarenko

Modèle unifié pour la recherche d'information sémantique

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 : 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). Les modèles documentaires sont généralement exploités en recherche d'information pour la représentation des documents et des requêtes et ils autorisent des calculs de pertinence numériques fondés sur la répartition des unités d'indexation dans la collection de documents.

Le modèle sémantique définit les unités sémantiques (concepts, instances de concepts, etc.) qui peuvent être reliées par des relations (relations hiérarchiques ou rôles). En recherche d'information, ils permettent d'aller au-delà des mots et de raisonner au niveau des concepts ou instances de concepts.

Nous proposons d'unifier les modèles documentaire et sémantique dans un unique réseau sémantico-documentaire pour représenter l'ensemble des propriétés, qu'elles soient numériques ou symboliques. La propagation d'activation est utilisée pour propager l'information de pertinence de proche en proche sur le graphe, depuis les éléments de la requête utilisateur.

Dans cet article, nous présentons notre modèle et les requêtes qu'il permet de prendre en compte. Nous présentons également des résultats sur des expérimentations préliminaires effectuées sur un banc de test de l'état de l'art. Nous obtenons des performances comparables à l'état de l'art pour des modèles simples, ce qui fait espérer des marges de progrès avec l'introduction de la sémantique.

Mots-clés : Documents, ontologie, annotation sémantique, graphe, propagation d'activation, recherche d'information

1 Introduction

L'avènement du Web, des moteurs de recherche et du Web Sémantique (WS) a décuplé l'information disponible et les moyens d'accéder à cette information. Les modèles sous-jacents sont cependant hétérogènes : les moteurs de recherche reposent essentiellement sur la fréquence des mots et l'analyse de leurs distributions dans les documents ; la recherche d'information sémantique (RIS) exploite à l'inverse des connaissances sémantiques généralement consignées dans des ressources comme les ontologies ou les thesaurus Zargayouna *et al.* (2015).

Des travaux récents (Castells *et al.*, 2007; Bhagdev *et al.*, 2008; Fernández *et al.*, 2011) proposent de combiner différents espaces d'indexation qui pour exploiter au mieux les modèles sémantiques tout en gardant une représentation classique du modèle documentaire tel que le modèle vectoriel défini par Salton *et al.* (1975).

Le défi aujourd'hui consiste à proposer un modèle unifié qui permette à l'utilisateur d'avoir accès à l'ensemble de ces fonctionnalités dans un unique système d'accès à l'information. Il doit pouvoir interroger une base documentaire à l'aide de mots-clefs ou de concepts mais aussi retrouver des documents similaires à un texte source voire les concepts associés à un ensemble de termes.

Dans ce travail nous proposons d'exploiter à la fois les connaissances sémantiques des ontologies du WS et les caractéristiques distributionnelles largement éprouvées en RI. Nous intégrons pour cela 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.

La suite de cet article est organisée comme suit : la section 2 présente le modèle proposé, la section 3 présente les travaux en propagation d'activation pour la RI, la section 4 donne les premiers résultats obtenus.

2 Modèle

Un modèle de Recherche d'Information propose une manière unifiée de représenter les requêtes et les documents ainsi qu'une *fonction de correspondance* qui associe des scores aux couples requête-document permettant ainsi de trier les documents en fonction de la requête.

Notre approche repose sur un modèle de graphe qui permet de représenter dans un modèle unique la base documentaire, avec notamment les relations termes-documents, et le réseau sémantique qui comporte par exemple une structure de concepts et des associations entre termes et concepts. La correspondance entre les requêtes et les documents du graphe est calculée par un mécanisme de propagation d'activation sur ce graphe.

2.1 Réseau sémantico-documentaire

Nous proposons de représenter le modèle 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 noeuds et différents types de relations selon ce qu'on souhaite représenter.

Nous proposons de prendre en compte trois types de noeuds : les *noeuds documents* représentent tous les documents de la collection documentaire ; les *noeuds termes* représentent le vocabulaire de la collection documentaire et les *noeuds concepts* représentent les concepts et instances de l'ontologie associée à la collection documentaire.

Ces noeuds sont reliés par 5 types de relations qui peuvent porter des propriétés :

— les *relations d'occurrence* 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 ;

— les *relations d'intertextualité* 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 [1] ;

— les *relations terminologiques* 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 termes « préférés » ;

— les *relations d'annotation* sont des relations entre documents et concepts associant des concepts ou des catégories comme méta-données à des documents et qui sont souvent

1. 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.

issues d'un travail d'annotation sémantique des documents ;
— les *relations ontologiques* sont des relations entre concepts (ou concepts et instances) qui peuvent représenter aussi bien les rôles que les liens hiérarchiques.

2.2 Graphe pondéré

Ce réseau sémantico-documentaire peut être représenté sous la forme d'un graphe pondéré. Ce graphe $G = \langle N, R \subseteq N \times \mathbb{R} \times N \rangle$ est constitué d'un ensemble de noeuds ($N = N_d \uplus N_t \uplus N_c$) et d'arcs qui sont orientés et pondérés ($R = R_{occ} \uplus R_{int} \uplus R_{ter} \uplus R_{ann} \uplus R_{ont}$).

Chacune de ces relations peut porter un poids qui reflète son importance dans le graphe : le poids d'une relation d'occurrence peut représenter la fréquence d'un terme dans un document ; le poids d'une relation terminologique peut permettre de distinguer le label « préféré » d'un concept par rapport aux autres termes qui lui sont associés ; etc. Nous n'entrons pas ici dans le détail du calcul de ces poids, considérant que différents paramétrages sont possibles, depuis un graphe booléen (sans poids) à un graphe entièrement pondéré, qu'ils reflètent différents choix de modélisation mais qu'ils sont tous compatibles avec le modèle à base de graphe que nous proposons.

Le graphe pondéré peut être interrogé de plusieurs manières 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 : $Q = \{t_1, t_2, \ldots, C_1, C_2, \ldots, D_1, D_2, \ldots\}$. Les réponses attendues peuvent également être de différents types (documents, termes, concepts). Le modèle unifié à base de graphe 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.

2.3 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 noeuds du graphe : au départ les noeuds qui correspondent à la requête ont des valeurs d'activation positives et les autres noeuds sont neutres ; le processus de propagation est ainsi déclenché ; quand celui-ci s'arrête, les valeurs d'activation obtenues sur les noeuds du graphe [2] déterminent l'ordre de pertinence des noeuds de ce graphe au regard de la requête initiale.

La propagation à proprement parler consiste à 1) sélectionner les noeuds à activer parmi les noeuds dont la valeur d'activité est non nulle et qui n'ont pas encore été activés, puis à 2) propager l'activité de ces noeuds à leurs voisins et les désactiver, ce processus étant itéré jusqu'à ce que plus aucun noeud ne puisse être sélectionné.

On comprend que la propagation s'applique aux valeurs d'activation qui sont mises à jour par « contagion » sur les voisins mais qu'elle est aussi contrôlée par l'état des noeuds, lequel évolue au cours du processus, un noeud étant tour à tour *inactif*, *activé* et *désactivé*.

La valeur d'activation du noeud i à la k^{ime} itération est calculée de la manière suivante :

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

2. Les noeuds peuvent être filtrés si on veut restreindre le type de résultat.

Elle dépend de la structure du graphe, à savoir les prédécesseurs du noeud i ($j \in pred(i)$) et le degré de ces noeuds (deg(j)) mais aussi de l'état des noeuds du graphe, seuls les actifs à l'itération $(k-1)$ étant pris en compte ($j \in actif(k-1)$). On note que la fonction d'activation pour un noeud est croissante : un noeud désactivé ne peut plus être réactivé mais sa valeur d'activation peut continuer à croître sous l'effet de ses voisins. La propagation d'activation s'arrête quand il n'y a plus de noeuds actifs : il ne reste que des noeuds déjà désactivés ou des noeuds inactifs qui ne peuvent être atteint par la propagation d'activation.

3 État de l'art

Les approches à base de graphes ont pris beaucoup d'importance en recherche d'information où les méthodes des marches aléatoires sont répandues depuis PageRank (Brin & Page, 1998). Même si le modèle mathématique de la propagation d'activation est moins solidement fondé que celui des algorithmes à base de marches aléatoires, il a une complexité moindre et s'adapte mieux au cadre d'une fonction de correspondance, qui est dirigée par la requête.

L'application de la propagation d'activation en recherche d'information n'est pas récente (Preece, 1981; Cohen & Kjeldsen, 1987; Croft *et al.*, 1988; Salton & Buckley, 1988; Savoy, 1992). Crestani (1997) présente un état de l'art sur les travaux en recherche d'information qui ont proposé l'utilisation de la propagation d'activation dans des réseaux associatifs ou des réseaux sémantiques.

Plusieurs travaux en WS proposent de peupler la base de connaissances avec des documents mais la représentation des documents dans une base de connaissances ne permet pas de mettre en place une recherche documentaire. Les bases de connaissances sont utiles pour une recherche précise avec une vision orientée données qui nécessite de connaître la structure de la base. Les graphes de connaissances, tels que les graphes RDF ou les graphes conceptuels ne sont pas adaptés à la RI car ils modélisent essentiellement des données symboliques. En effet, il est difficile de mettre en place des calculs distributionnels car ils ne permettent pas de prendre en compte des informations telles que les fréquences et le nombre d'occurrences.

Le modèle de propagation d'activation que nous proposons permet de reproduire la RI classique et d'exploiter des informations ontologiques difficiles à mettre en place dans une base de connaissances.

4 Expérimentations

Nous avons enrichi la plateforme Terrier RIS proposée par Bannour & Zargayouna (2012) par la plateforme d'analyse de graphes JUNG (*Java Universal Network/Graph Framework*) [3]. La plateforme JUNG permet de modéliser différents types de graphes (orienté, non orienté, etc.). Elle implémente également de nombreux algorithmes sur les graphes.

Les premières expérimentations ont porté sur le corpus de recettes de cuisine exploité par Bannour & Zargayouna (2012). Le corpus est composé de 1 489 recettes de cuisines et de 4 requêtes avec leurs jugements de pertinence [4].

3. http ://jung.sourceforge.net/

4. Une mise à jour des jugements de pertinence a été effectuée. Exemple, la requête 1 : Cook an Asian soup with leek (6 documents pertinents).

La *baseline* consiste en une implémentation classique du modèle vectoriel (Salton *et al.*, 1975) avec la formule de pondération TF-IDF (*Term Frequency-Inverse Document Frequency*).

Nous avons voulu, dans un premier temps, nous assurer que notre modèle permet de simuler un comportement classique de RI dans le cas où le modèle représente uniquement les termes, les documents et les liens entre eux. Le poids des liens terme-document et document-terme est calculé par :

$$w(t, doc) = \frac{tf(t, d)}{maxTF(doc)}$$

tel que $maxTF(doc)$ présente la fréquence maximale d'un terme dans le document d.

Nous présentons les résultats en termes de MAP et R-PREC : la MAP (*Mean Average Precision*) est la moyenne de la précision obtenue après chaque document pertinent retourné ; la R-PREC (*R-Precision*) est la précision après R documents pertinents retournés, R étant le nombre de documents pertinents. Les résultats de notre approche (GraphTerm) sont meilleurs que ceux de la base line (TF.IDF) : le tableau 4 fait apparaître une augmentation de près de 10% et, pour la requête 1 par exemple, une nette amélioration en termes de MAP et R-précision.

Méthode	MAP				R-PREC			
	R1	R2	R3	R4	R1	R2	R3	R4
TF.IDF	0.25	0.6	0.15	0.47	0.33	0.66	0.0	0.5
Total	0.36				0.37			
GraphTerm	0.5	0.78	0.12	0.5	0.66	0.66	0.0	0.5
Total	0.47				0.46			

FIGURE 1 – Tableau récapitulatif des résultats

Ces résultats, même s'ils sont préliminaires, montre que notre modèle donne des résultats similaires à ceux qu'on peut obtenir avec un modèle éprouvé, en recherche d'information classique. Brouard (2013) a présenté des résultats équivalents mais avec un modèle qui exploite d'une manière différenciée la couche documents de la couche termes en proposant deux formules de propagation différentes.

5 Conclusion

Nous avons proposé d'unifier les modèles documentaire et sémantique dans un unique réseau sémantico-documentaire qui permet de représenter l'ensemble des propriétés du modèle documentaire et du modèle sémantique. Ce réseau est représenté sous forme de graphe pondéré. L'intérêt de cette représentation est de combiner des propriétés numériques et symboliques. Nous proposons d'appliquer une propagation d'activation qui permet de propager l'information de pertinence de proche en proche sur le graphe. Les premières expérimentations montrent que nous obtenons de bonnes performances par rapport à l'état de l'art pour des modèles simples, ce qui fait espérer des marges de progrès avec l'introduction de la sémantique. Des expérimentations à plus grande échelle sont en cours. Ces expérimentations vont permettre de calibrer les formules de propagation, de mettre en place les heuristiques de recherche nécessaires, et d'étudier l'impact des structures de connaissances sur la propagation.

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

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.

BHAGDEV R., CHAPMAN S., CIRAVEGNA F., LANFRANCHI V. & PETRELLI D. (2008). Hybrid search : Effectively combining keywords and semantic searches. In *Proceedings of the 5th European Semantic Web Conference, ESWC*, p. 554–568.

BRIN S. & PAGE L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web (WWW7)*, p. 107–117.

BROUARD C. (2013). Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation. In *CORIA*, p. 217–226.

CASTELLS P., FERNANDEZ M. & VALLET D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Know. and Data Eng.*, **19**(2), 261–272.

COHEN P. R. & KJELDSEN R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manage.*, **23**(4), 255–268.

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

CROFT W. B., LUCIA T. J. & COHEN P. R. (1988). Retrieving documents by plausible inference : A priliminary study. In *Proceedings of the 11th Annual International ACM SIGIR*, SIGIR '88, p. 481–494.

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.

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.

PREECE S. (1981). *A Spreading Activation Network Model for Information Retrieval*. University of Illinois at Urbana-Champaign.

SALTON G. & BUCKLEY C. (1988). On the use of spreading activation methods in automatic information. In *Proceedings of the 11th Annual International ACM SIGIR*, SIGIR '88, p. 147–160.

SALTON G., WONG A. & YANG C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, **18**(11), 613–620.

SAVOY J. (1992). Bayesian inference networks and spreading activation in hypertext systems. *Information Processing Management*, **28**(3), 389 – 406.

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