Propagation d'activation dans les graphes 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. *Cet article présente un modèle de recherche d'information sémantique qui regroupe dans une même représentation le modèle sémantique et le modèle documentaire. La représentation en graphe que nous proposons permet des interrogations riches par termes, concepts ou même par documents. Nous appliquons sur cette représentation « unifiée » une méthode de propagation d'activation qui permet, à partir d'une requête, de retourner un ensemble de résultats ordonnés par pertinence, en propageant l'information de pertinence de proche en proche sur le graphe, depuis les éléments de la requête utilisateur.*

ABSTRACT. *This paper presents a model of semantic information retrieval which brings together in one representation the semantic and the documents' model. We propose a graph representation that allows rich queries with terms, concepts or even documents. We apply to this "unified" representation a method of spreading activation which is striggered by the elements of the user query and which enables to return a set of ordered relevant results.*

MOTS-CLÉS : *RIS, Ontologies, Graphe, Propagation d'activation*

KEYWORDS: *SIR, Ontologies, Graph, Spreading activation*

1. Introduction

L'avènement du Web, des moteurs de recherche et du Web Sémantique (Berners-Lee et Lassila, 2001) ont 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.

La plupart des modèles de RIS existants sont des adaptations des modèles classiques de RI où l'exploitation de la ressource sémantique se réduit à une fonction de similarité intégrée dans la fonction de correspondance (Zargayouna *et al.*, 2015).

Des travaux récents (Castells *et al.*, 2007 ; Fernández *et al.*, 2011) proposent de combiner différents espaces d'indexation qui permettent à la fois d'exploiter au mieux les modèles sémantiques et de garder une représentation classique du modèle documentaire tels que le modèle vectoriel défini par Salton *et al.* (1975).

Le défi aujourd'hui consiste à proposer un modèle dédié à la RIS qui permette d'exploiter le modèle sémantique et le modèle documentaire d'une manière « unifiée ».

Nous proposons dans cet article une représentation en graphe qui permet d'intégrer dans une même représentation différents niveaux : le niveau conceptuel, le niveau terminologique et le niveau documentaire. Le modèle sémantique pris en compte peut être un thesaurus, un réseau terminologique ou une ontologie. Nous utilisons dans cet article le mot ontologie d'une manière générique qui regroupe ces différents types de ressources.

La construction du graphe se fait à l'aide d'un processus d'indexation et d'annotation sémantique qui permet d'établir des liens entre concepts, termes et documents. L'intérêt d'une telle représentation est de permettre de répondre à des requêtes plus complexes que juste des mots-clés. La mise en correspondance consiste en une méthode de propagation d'activation dans le graphe qui permet de raisonner de proche en proche et de rapprocher des documents, des termes et des concepts à un niveau sémantique.

L'application de la propagation d'activation en recherche d'information n'est pas récente (Preece, 1981 ; Cohen et Kjeldsen, 1987 ; Croft *et al.*, 1988 ; Salton et 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.

Plus récemment, Brouard (2013) montre qu'une propagation d'activation très simple dans un réseau associatif modélisant une couche documents et une couche termes donne des résultats équivalents au modèle vectoriel.

Dans la suite de cet article, nous décrivons la représentation en graphe, l'interrogation du graphe et le mécanisme de propagation proposé.

2. Représentation en graphe

Un modèle de recherche d'information concerne à la fois la représentation des documents et des requêtes (généralement dans le même formalisme) et une fonction de correspondance qui permet de rapprocher les deux représentations et d'attribuer un score aux documents.

Nous proposons une représentation en graphe qui permet d'intégrer dans une même représentation, le niveau conceptuel, terminologique et documentaire (voir figure1).

Le graphe est constitué d'un ensemble de nœuds et d'arcs valués et orientés $\langle N, E \subseteq N \times \mathbb{R} \times N \rangle$[1].

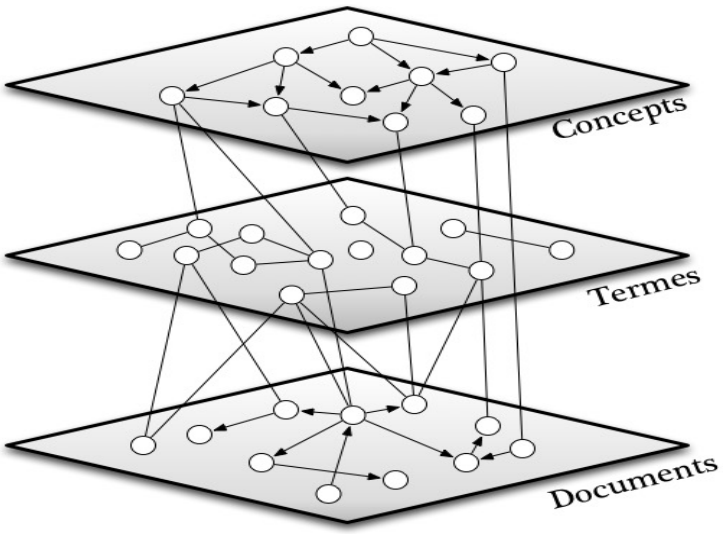

Figure 1. *Représentation en graphe des trois niveaux*

Les nœuds représentent trois types d'entités ($N = N_d \uplus N_t \uplus N_c$) :

– les *nœuds documents (N_d)* représentent les documents de la collection ;

– les *nœuds termes (N_t)* représentent le vocabulaire de la collection ainsi que les termes de l'ontologie ;

1. Dans la figure1, les arcs orientés dans les deux sens sont présentés par des arcs simples.

– les *nœuds concepts (N_c)* représentent les concepts de l'ontologie.

Les arcs traduisent les liens entre les différents nœuds. Ces arcs peuvent être pondérés, le poids d'un arc exprimant la force du lien qu'il traduit. Nous notons $w(i, j)$ le poids de l'arc allant du nœud i vers le nœud j. Les liens entre les nœuds sont :

– concept↔concept ($N_c \times \mathbb{R} \times N_c$) : ces arcs traduisent les relations ontologiques entre concepts ; il peut s'agir de relations hiérarchiques ou de rôles et le poids peut servir à distinguer la force des différents types de relations ontologiques ;

– concept↔terme ($(N_c \times \mathbb{R} \times N_t) \cup (N_t \times \mathbb{R} \times N_c)$) : ces arcs traduisent les relations entre concepts et termes qui peuvent apparaître dans le volet lexical d'une ontologie ; le poids peut éventuellement servir à distinguer le label « préféré » d'un concept (*preflabel*) par rapport aux autres termes qui lui sont associés (altlabel) ;

– concept↔document ($(N_c \times \mathbb{R} \times N_d) \cup (N_d \times \mathbb{R} \times N_c)$) : ces arcs représentent les liens de catégorisation qui peuvent être calculés lors de la phase d'annotation.

– terme↔terme ($N_t \times \mathbb{R} \times N_t$) : ces arcs représentent les liens terminologiques entre termes (ex. synonymie) ;

– terme↔document ou document-terme ($(N_t \times \mathbb{R} \times N_d) \cup (N_d \times \mathbb{R} \times N_t)$) : ces arcs représentent l'apparition d'un terme dans un document et leur poids peut naturellement traduire la fréquence d'occurrence du terme dans le document auquel il est relié ;

– document↔document ($N_d \times \mathbb{R} \times N_d$) : ces arcs représentent des relations « intertextuelles » entre documents (Mimouni *et al.*, 2014), notamment les liens de citation qui sont les plus fréquents.

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.

3. Interrogation du graphe

La représentation unifiée de toutes ces informations sous la forme d'un graphe 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 : par les termes, les concepts ou les documents. Les nœuds pertinents renvoyés peuvent être des documents comme cela est classique mais aussi des termes, des concepts ou même une combinaison de plusieurs types de nœuds.

Il est ainsi possible de :

– poser des requêtes par des termes qui n'existent pas dans le vocabulaire de la collection, ce qui répond au problème de *term mismatch* décrit dans (Crestani, 2000) ;

– poser des requêtes par des concepts de structuration, qui ne possèdent pas forcément une dénotation lexicale dans le vocabulaire, mais qui servent à la catégorisation

du domaine et donnent un accès direct aux documents ;

– poser des requêtes par l'exemple, en soumettant comme requête un document et en recherchant les documents similaires.

Un processus de propagation d'activation permet d'intégrer à la recherche un mécanisme d'enrichissement de la requête qui prend en compte des termes synonymes, des concepts ou des documents reliés.

En effet, la fonction de correspondance que nous proposons repose un mécanisme de propagation dans le graphe : l'activation part des nœuds de la requête et se propage de proche en proche sur les nœuds voisins dans le graphe de sorte que les valeurs d'activation des nœuds du graphes traduisent, en fin de propagation, la pertinence des nœuds du graphe au regard de la requête.

4. Propagation d'activation dans le graphe

La propagation d'activation est un processus qui permet de propager une information de proche en proche sur un graphe. Cette information se représente par une valeur d'activation associée aux nœuds du graphe.

De manière générale, un algorithme de propagation d'activation dans un graphe, tel que décrit dans (Crestani, 1997), se compose d'une séquence d'étapes de propagation élémentaires et d'une condition d'arrêt. Il permet d'associer des valeurs numériques aux nœuds du graphe, les *valeurs d'activation*. Le graphe de départ est un graphe neutre, où toutes les valeurs d'activation sont nulles. L'initialisation du processus se fait par l'activation des nœuds qui figurent dans la requête. A chaque étape de propagation, le calcul de la valeur d'activation d'un nœud dépend de sa valeur précédente et des valeurs transmises par ses voisins lorsqu'ils s'activent. Le processus itératif se termine quand la condition d'arrêt est vérifiée. Le résultat est une distribution de valeurs d'activation sur l'ensemble des nœuds du graphe.

Nous détaillons les éléments clefs de cet algorithme de propagation dans ce qui suit.

4.1. *Valeurs d'activation des nœuds*

Étant donné $G = \langle N, E \rangle$, a une fonction d'affectation de valeurs d'activation sur le graphe G tel que $a : N \rightarrow \mathbb{R}^+$, nous notons $a(i)$ la valeur du nœud i dans l'affectation a.

Un algorithme de propagation peut être défini comme une suite finie de fonctions d'affectation $\pi = a_0 \dots a_k \dots a_n$ où a_0 représente l'affectation issue de la requête et a_n l'affectation vérifiant la condition d'arrêt.

L'activation initiale des nœuds (a_0) est fixée *a priori* avant que le processus de propagation d'activation ne soit déclenché. Ainsi, pour une tâche de RI dont la requête est

un ensemble de termes, les nœuds activés au départ peuvent correspondre aux termes de la requête, et leurs valeurs d'activation initiales dépendre de leur fréquence d'occurrence dans la requête. Le résultat de la requête est un sous-ensemble N_d de nœuds du graphe G déterminé en fonction de a_n et éventuellement restreint à un certain type de nœuds (documents, termes et/ou concepts).

4.2. *Étape de propagation*

Une étape de propagation se décompose elle-même en deux opérations : la sélection des nœuds à activer et la propagation de leur l'activation à leur voisins, qui se traduit par le calcul d'une nouvelle affectation a_k. Nous revenons dans la section 4.3 sur les règles que nous proposons pour la sélection des nœuds. Il suffit ici de considérer qu'à chaque étape de propagation, tous les nœuds actifs se déclenchent.

Le calcul d'une nouvelle affectation a_k dépend de l'affectation précédente a_{k-1} et de la structure du graphe. La fonction de propagation peut être également dépendante du type de nœud mais, pour simplifier la présentation, nous considérons ici une fonction de propagation identique pour tous les nœuds, qu'ils soient concepts, termes ou documents.

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 se définie comme suit :

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

Les suites d'activation des nœuds sont croissantes. La valeur d'activation d'un nœud est augmentée à l'itération k de la somme des valeurs d'activation obtenues à l'itération précédente $(k-1)$ pour ses nœuds prédécesseurs $(pred(i))$ à condition qu'ils soient actifs, c'est-à-dire sélectionnés pour l'étape $k-1$ $(actif(k-1))$. Cette somme est pondérée par le poids des arcs reliant les prédécesseurs au nœud $(w(j,i))$. Elle est également atténuée par le degré des nœuds prédécesseurs $deg(j)$. Ce degré peut intégrer d'une manière globale tous les arcs comme il peut tenir compte des types de nœuds reliés. Il est par exemple possible de calculer le degré d'ambiguïté d'un nœud terme en prenant en compte le nombre de nœuds concepts auxquels il est relié. Entre un nœud terme et des nœuds documents, le degré permet de rendre compte du *document frequency*, le nombre de documents dans lesquels le terme apparaît.

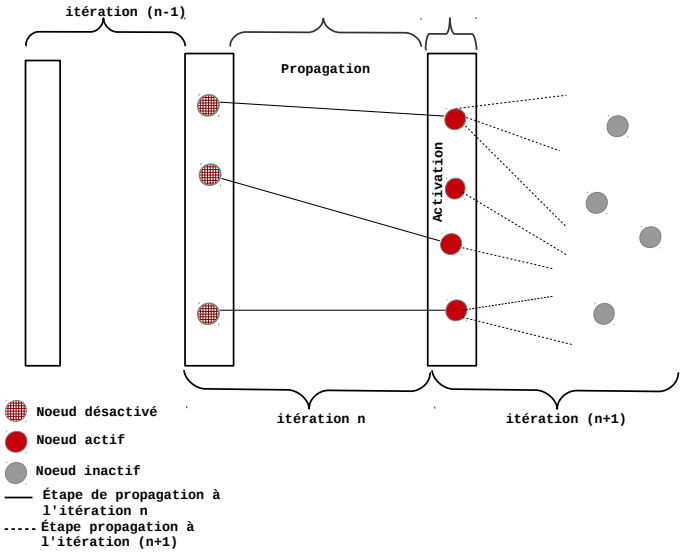

Figure 2. *Processus de propagation d'activation*

4.3. *Condition d'arrêt*

Dans le cas où le graphe contient au moins un cycle, l'itération du processus de propagation en partant d'une affectation initiale pourrait se poursuivre à l'infini en l'absence de condition d'arrêt. Il existe plusieurs façons de terminer un algorithme de propagation : on peut borner *a priori* les étapes de propagation, jouer sur la sélection des nœuds actifs ou garantir que les valeurs se stabilisent à un moment donné.

Dans ce travail nous utilisons la condition d'arrêt proposée par Rocha *et al.* (2004). Elle repose sur le contrôle de la sélection des nœuds. Tous les nœuds ayant une valeur d'activation non nulle peuvent-être sélectionnés mais un nœud ne peut être sélectionné qu'une seule fois. Ce mécanisme est contrôlé par une variable d'état associée à chaque nœud. Un nœud change d'état au cours du processus de propagation : il est *inactif* avant d'être atteint par « l'onde » de propagation, *actif* quand il est sélectionné pour se déclencher et activer ses voisins puis et *désactivé* une fois qu'il a été déclenché.

La propagation d'activation s'arrête donc quand il n'y a plus de nœuds actifs : il n'y a plus que des nœuds déjà désactivés ou des nœuds inactifs qui ne peuvent être atteint par la propagation d'activation.

4.4. *Analyse*

Le calcul d'activation dépend de la structure du graphe ($pred(i)$). Dans notre approche, tous les nœuds du graphe sont susceptibles d'être activés mais qu'un nœud ne peut se déclencher qu'une seule fois. Il s'ensuit que le calcul de la propagation est linéaire par rapport à la taille du graphe : $O(|N| * deg_{max})$, où $|N|$ est le nombre de nœuds du graphe et deg_{max} son degré maximal, c'est-à-dire le degré du nœud ayant le plus de successeurs. Le nombre d'étapes de propagation dépend de la forme du graphe : il est borné par le rayon du graphe ayant pour centre les nœuds de la requête.

5. Conclusion

Nous avons proposé une représentation unifiée en graphe du modèle sémantique et documentaire. Cette représentation permet d'exprimer des requêtes plus riches que les simples requêtes par mots-clés. Nous avons présenté une méthode de propagation d'activation dans le graphe qui permet de trouver des rapprochements entre requêtes et résultats en tenant compte des proximités sémantiques.

Les premières expérimentations ont porté sur le corpus de recettes de cuisine exploité par (Bannour et Zargayouna, 2012) et les premiers résultats sont encourageants. Des expérimentations à plus grande échelle sont en cours et vont permettre de calibrer les formules de propagation, de mettre en place les heuristiques de recherche nécessaires, d'ajuster la condition d'arrêt et d'étudier le passage à l'échelle d'un tel algorithme en utilisant la plate-forme de modélisation de graphes JUNG[2].

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.

6. Bibliographie

Bannour I., Zargayouna H., « Une plate-forme open-source de recherche d'information sémantique », *COnférence en Recherche d'Information et Applications (CORIA)*, p. 167-178, 2012.

2. Java Universal Network/Graph Framework http ://jung.sourceforge.net/

Berners-Lee T., Lassila J. H. O., « The Semantic Web », *Scientific American*, 2001.

Brouard C., « Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation », *CORIA*, Neuchâtel, France, p. 1-10, April, 2013.

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

Cohen P. R., Kjeldsen R., « Information Retrieval by Constrained Spreading Activation in Semantic Networks », *Inf. Process. Manage.*, vol. 23, n° 4, p. 255-268, July, 1987.

Crestani F., « Application of Spreading Activation Techniques in Information Retrieval », *Artificial Intelligence Review*, vol. 11, n° 6, p. 453-482, 1997.

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

Croft W. B., Lucia T. J., Cohen P. R., « Retrieving Documents by Plausible Inference : A Priliminary Study », *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, ACM, New York, NY, USA, p. 481-494, 1988.

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

Mimouni N., Nazarenko A., Paul È., Salotti S., « Towards Graph-based and Semantic Search in Legal Information Access Systems », *Legal Knowledge and Information Systems - JURIX 2014, Krakow, Poland*, vol. 271 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, p. 163-168, 2014.

Preece S., *A Spreading Activation Network Model for Information Retrieval*, University of Illinois at Urbana-Champaign, 1981.

Rocha C., Schwabe D., Aragao M. P., « A Hybrid Approach for Searching in the Semantic Web », *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, ACM, New York, NY, USA, p. 374-383, 2004.

Salton G., Buckley C., « On the Use of Spreading Activation Methods in Automatic Information », *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, ACM, p. 147-160, 1988.

Salton G., Wong A., Yang C. S., « A Vector Space Model for Automatic Indexing », *Commun. ACM*, vol. 18, n° 11, p. 613-620, November, 1975.

Savoy J., « Bayesian inference networks and spreading activation in hypertext systems », *Information Processing Management*, vol. 28, n° 3, p. 389 - 406, 1992.

Zargayouna H., Roussey C., Chevallet J. P., « Recherche d'information sémantique : état des lieux », *TAL (Traitement Automatique des Langues)*, vol. 56, n° 3, p. 49-73, 2015.