Un outil de navigation dans un espace sémantique

Yann Vigile Hoareau*, Murat Ahat**
David Medernach*** Marc Bui****

*Université Paris 8, vigilehoareau@gmail.com
**Ecole Pratique Des Hautes Etudes, murat.ahat@etu.epeh.sorbonne.fr
***Université Paris 8, david.medernach@gmail.com
****Ecole Pratique Des Hautes Etudes, marc.bui@ephe.sorbonne.fr

1 Introduction

L'outil que nous allons présenter à pour objectif d'exploiter les propriétés des espaces sémantiques et des méthodes de visualisations de graphes pour proposer un moteur de recherche disposant des fonctionnalités suivantes :

1. recherche à partir de mot-clés
2. recherche à partir de documents entiers
3. recherche dite "sémantique" qui n'est plus basé sur la croisement des mots-clés de la requête avec les mots qui apparaissent dans les documents, mais sur le croisement du sens des mots qui compose la requête avec le sens des mots qui composent les documents. À titre d'illustration, contrairement à la recherche classique par mot-clé, la recherche "sémantique" serait capable de retrouver le document "les troubles à Bangkok" à partir de la requête "les émeutes en Thailand".
4. visualisation globale d'une grande collection de documents (plusieurs milliers de documents) pour permettre à l'utilisateur d'appréhender "la structure" de la collection de documents. Si l'on prend l'exemple d'un centre de recherche scientifique donné, on pourrait saisir instantanément les thèmes de recherches les plus populaires ainsi que les documents les plus centraux.
5. visualisation locale de la similarité sémantique d'un nombre réduit de documents (plusieurs dizaines) afin de dépasser la simple liste ordonnée de documents pour offrir à l'utilisateur une représentation visuelle exprimant intuitivement les relations de proximités sémantiques parmi les résultats retournés à partir de la requête.

Notre outil s'appuie sur un modèle d'espace sémantique appelé Random Indexing (Kanerva et al., 2000) pour représenter les connaissances. Notre méthode de visualisation est une alternative à la méthode classique de visualisation d'espace sémantique par la méthode de *multi-dimensionnal scaling*. Elle consiste à représenter l'espace vectoriel obtenue à partir de RI sous la forme d'un graphe qui garde les propriétés de l'espace sémantique.

2 La recherche sémantique au moyen d'espace sémantique

Notre moteur de recherche sémantique est basé sur un modèle d'espace sémantique appelé Random Indexing (Kanerva et al., 2000). Les méthodes de représentation vectorielle de la sémantique des mots qui servent à construire des espaces sémantiques relèvent d'une famille de modèles qui représentent la similitude sémantique entre les mots en fonction de l'environnement textuel dans lequel ces mots apparaissent.

Divers traitements mathématiques et statistiques, permettant d'extraire la signification des concepts, peuvent être appliqués à la matrice des fréquences de co-occurrence des termes. Par exemple LSA (Landauer et Dumais, 1997) emploie une méthode générale de décomposition linéaire de matrice de fréquence d'occurence : la décomposition de valeur singulière (SVD). Le modèle RI n'est pas basé sur des méthodes de réduction matricielle mais sur des méthodes de projections aléatoires. La première étape consiste à générer un vecteur aléatoire composé de centaines de 0 et d'une dizaine de -1 et de $+1$, de dimension N, avec $N > 1000$, pour chaque document de la collection de tel sorte à ce que chaque vecteur soit différent des autres. Ces vecteurs aléatoires sont appelés vecteurs *index*. La seconde étape consiste à initialiser une matrice nulle de dimension N. La troisième étape est un processus itératif qualifié d'accumulation. Chaque fois qu'un mot m apparait dans un document d, par une opération de somme algébrique, le vecteur $\vec{d}$ est accumulé au vecteur $vecm$ correspondant au mot m. Si l'on considère les vecteurs *index* comme des empreintes uniques, au terme du processus d'accumulation, les mots qui sont apparut dans les mêmes documents ont accumulé les mêmes empreintes. Les vecteurs correspondant à ces mots seront considéré comme similaire. Le modèle RI a montré des performances supérieures à LSA (Kanerva et al., 2000). Le modèle RI est bien moins couteux que LSA d'un point de vue computationnel. À titre d'illustration, le modèle RI permet d'indexer plus de 9 millions de résumé sur du corpus Medline sur un ordinateur portable de consommation courante. L'implémentation du modèle RI que nous avons utilisé est SemanticVector. La librairie permet de réaliser des requêtes à partir de mots ou de documents.

3 Visualisation d'un espace sémantique et d'une requête

La visualisation de l'espace sémantique pose le problème de réduction de la dimentionnalité qui est généralement résolu au moyen de la méthode du *multi-dimensional scaling*. Notre outil propose une méthode de visualisation de l'espace sémantique à partir d'un graphe. Bien que la visualisation de connaissances sémantiques sous la forme de graphes soit très populaires pour les ontologies (Wong et al., 2006), la proposition qui consiste à représenter un espace sémantique de type LSA ou Random Indexing par un graphe demeure à notre connaissance inédite. La propriété du graphe que nous proposons est de représenter la similarité sémantique entre les documents. Le procédé consiste à calculer la distance euclidienne pondérée entre chaque document de l'espace sémantique afin de construite une matrice de connexité. Cette matrice de connexité correspond alors à une représentation de l'espace sémantique sous la forme d'un graphe à N noeud et N^2 arcs. Ce graphe est représenté visuellement au moyen de la librairie Prefuse. Notre méthode n'est pas limitée par le nombre de documents car elle offre la possibilité de "zoomer" pour spécifier la recherche dans un domaine précis, ou au contraire prendre connaissances des différences entre les thématiques présentes. Notre prototype permet de visualiser la structure sémantique d'une grande collection de documents de façon instanta-

née. La fonction *Derived Force* ou *Radial Graph* (Yee et al., 2001), respectivement représenté sur la partie (a) et b de la figure 1.

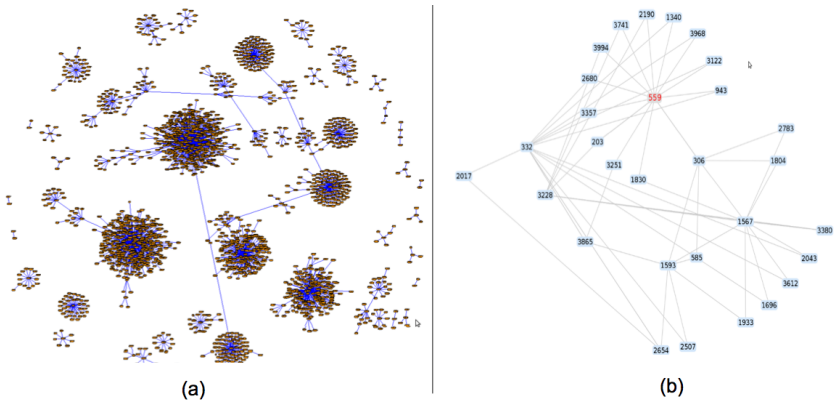

FIG. 1 – *Représentation visuelle d'une la collection de 4400 résumés des rapport de recherche de l'INRIA avec la fonction Derived Force (a) et représentation des résultats d'une requête avec la fonction Radial Graph (b) pour 30 documents.*

Références

Kanerva, P., J. Kristoferson, et A. Holst (2000). Random Indexing of Text Samples for Latent Semantic Analysis. In L. Gleitman et A. Josh (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah. Lawrence Erlbaum Associates.

Landauer, T. K. et S. T. Dumais (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review 104*(2), 211–240.

Yee, K.-P., D. Fisher, R. Dhamija, et M. Hearst (2001). Animated exploration of dynamic graphs with radial layout. In *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, Washington, DC, USA, pp. 43–. IEEE Computer Society.

Pak Chung Wong, Chin, G., Foote, H., Mackey, P. and Thomas, J. (2006). Have Green – A Visual Analytics Framework for Large Semantic Graphs In *Proceedings of the 2006 IEEE Symposium On Visual Analytics Science And Technology*, pp. 67–74.

Summary

We propose a software for semantic space visualization. It is ,unlike classical visulization based on multi-dimensional scaling, based on the graph structure derived from a semantic space. This provides another method of navigation as well as finding information in the semantic space.