

Sujet de stage de M2R (module fouille de données - UMINR 306) :

## Ajout de connaissances syntaxiques avec Sygmart pour améliorer LSA

**Encadrement :** Mathieu Roche et Jacques Chauché

Équipe TAL, LIRMM, UMR 5506, Université Montpellier 2

Bureau 2.113, tel : 04 67 41 85 11

*mathieu.roche@lirmm.fr et jacques.chauche@lirmm.fr*

### 1. Contexte

Pour regrouper les termes appartenant à un même concept, nous allons nous appuyer sur la méthode automatique appelée Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997; Landauer *et al.*, 1998). LSA est une méthode statistique appliquée à des corpus de grande dimension consistant à regrouper les termes apparaissant dans le même contexte. Cette méthode qui s'appuie sur l'hypothèse « harrissienne », est fondée sur le fait que des mots qui apparaissent dans le même contexte sont sémantiquement proches. Le corpus est représenté sous forme matricielle. Les lignes représentent les mots et les colonnes représentent les différents contextes choisis (un document, un paragraphe, une phrase, etc.). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le cosinus de l'angle entre les deux vecteurs.

#### *Caractéristiques théoriques de LSA*

La théorie sur laquelle s'appuie LSA est la décomposition en valeurs singulières (SVD). Une matrice  $A = [a_{ij}]$  où  $a_{ij}$  est la fréquence d'apparition du mot  $i$  dans le contexte  $j$ , se décompose en un produit de trois matrices  $USV^T$ .  $U$  et  $V$  sont des matrices orthogonales et  $S$  une matrice diagonale. La figure 1 représente le schéma bien connu d'une telle décomposition où  $r$  représente le rang de la matrice  $A$ .

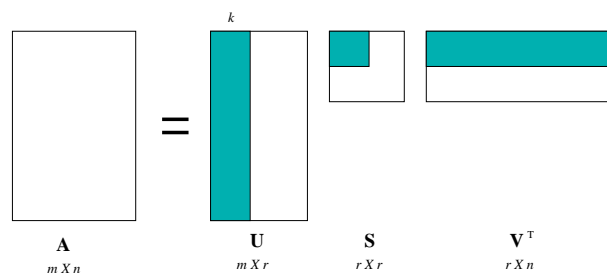


Figure 1: Décomposition en valeurs singulières. La matrice  $A$  représente le corpus d'origine de  $m$  lignes (mots du corpus) et  $n$  colonnes (contextes).

Soit  $S_k$  où  $k < r$  la matrice produite en enlevant de  $S$  les  $r - k$  colonnes qui ont les plus petites valeurs singulières. Soit  $U_k$  et  $V_k$  les matrices obtenues en enlevant les colonnes correspondantes des matrices  $U$  et  $V$ . La matrice  $U_k S_k V_k^T$  peut alors être considérée comme une version compressée de la matrice originale  $A$ .

## 2. Objectifs du stage

### 2.1. État de l'art

(Bestgen, 2004) précise que la taille des contextes (documents) est primordiale pour obtenir une qualité des résultats satisfaisante. Cette affirmation confirme les travaux de (Rehder *et al.*, 1998) qui ont effectué des expérimentations pour estimer la taille minimale d'un contexte afin d'obtenir des résultats intéressants avec LSA. Ces expérimentations ont consisté à découper les documents d'un corpus correspondant à des essais d'étudiants en documents de 10 mots, 20 mots, et ceci jusqu'à 200 mots. Les expérimentations ont montré que si les contextes (documents) possèdent moins de 60 mots alors la méthode LSA se révèle décevante.

Afin d'améliorer la performance de LSA, (Wiemer-Hastings, 2000) propose de transformer les phrases en structures syntaxiques. Pour ce faire, une segmentation syntaxique des phrases en trois groupes de mots est effectuée :

- syntagmes nominaux représentant les sujets,
- verbes en prenant en compte les adverbes et les syntagmes adverbaux,
- syntagmes nominaux représentant les objets.

Ainsi, chaque phrase est représentée sous la forme (« verbe » « sujet » « objet »). Lorsqu'il y a deux objets (« objet1 » et « objet2 ») affectés à un même verbe, la phrase sera représentée sous la forme (« verbe » « sujet » « objet1 ») et (« verbe » « sujet » « objet2 »), de même dans le cas de la présence de deux sujets associés à un seul verbe.

Initialement, LSA ne prend pas en compte un certain nombre de mots (« stops words ») tels que « if », « because », « have », etc. Contrairement à la version originale de LSA, (Wiemer-Hastings, 2000) prend en compte de tels mots et peut les utiliser pour construire les structures de certaines phrases. Par exemple la phrase *if the new motherboard uses the same type of RAM* sera représentée sous la forme (« if uses » « the new motherboard » « the same type of RAM »). En ajoutant l'ensemble de ces connaissances syntaxiques à la méthode LSA, les performances sont améliorées. En effet, la méthode développée par (Wiemer-Hastings, 2000) est efficace à partir de contextes ayant une longueur de 16 mots. Ceci est plus intéressant que la longueur de 60 mots décrite dans (Rehder *et al.*, 1998).

### 2.2. But à atteindre et méthode à appliquer

Afin d'effectuer des regroupements de termes (Roche, 2004), il est nécessaire d'obtenir des similarités entre les termes de bonne qualité deux à deux. Ainsi, dans les expérimentations que nous allons décrire, nous nous intéressons aux couples de termes trouvés automatiquement par LSA. Nous allons nous appuyer sur un corpus des Ressources Humaines (décrit ci-dessous) pour lequel plus de 1800 termes présents ont été associés manuellement à un concept par un expert du domaine. Ainsi, il est aisé de vérifier si les termes des couples obtenus avec le système LSA appartiennent à un même concept.

Pour les expérimentations à mener, nous avons à notre disposition un corpus des Ressources Humaines écrit en français<sup>1</sup>. Ce corpus issu du domaine des Ressources Humaines (société PerformanSe) correspondant à des commentaires de tests de psychologie de 378 individus. Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique avec l'utilisation de tournures souvent littéraires.

Différentes expérimentations pour regrouper les termes en utilisant LSA ont déjà été menées avec ce corpus (Roche & Kodratoff, 2003). Les résultats obtenus se sont révélés décevants. Ceci s'explique par la taille des contextes utilisés (phrases) dans notre approche. Nous proposons donc d'apporter des connaissances syntaxiques à LSA pour améliorer les résultats obtenus.

La méthode de (Wiemer-Hastings, 2000) a été mise en œuvre sur des textes écrits en anglais. Le premier objectif du stage sera de mettre en place cette méthode sur des corpus en français. Pour cela, le stagiaire devra utiliser l'analyseur syntaxique Sygmart (Chauché, 2005). La deuxième étape consiste à enrichir les relations syntaxiques de (Wiemer-Hastings, 2000). Par exemple, les objets peuvent être décomposés en plusieurs relations syntaxiques. Ce découpage plus fin des phrases pourrait sensiblement améliorer la qualité des résultats obtenus avec LSA.

## Références

- BESTGEN Y. (2004). Analyse sémantique latente et segmentation automatique de textes. In *Proceedings of JADT'04 (Journées Internationales d'Analyse Statistique des Données Textuelles)*, volume 1, p. 171–181.
- CHAUCHÉ J. (2005). Application des vecteurs sémantique à la fouille de texte. In *Dans les actes de la conférence "Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, volume 2, p. 113–124.
- LANDAUER T. & DUMAIS S. (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.
- LANDAUER T. K., FOLTZ P. W. & LAHAM D. (1998). Introduction to latent semantic analysis. In *Discourse Processes*, volume 25, p. 259–284.
- REHDER B., SCHREINER M., WOLFE M., LAHAM D., LANDAUER T. & KINTSCH W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. In *Discourse Processes*, volume 25, p. 337–354.
- ROCHE M. (2004). *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. PhD thesis, Université de Paris 11.
- ROCHE M. & KODRATOFF Y. (2003). Utilisation de LSA comme première étape pour la classification des termes d'un corpus spécialisé. In *Actes (CD-ROM) de la conférence MAJECSTIC'03 (MANifestation des JEunes Chercheurs dans le domaine STIC)*.
- WIEMER-HASTINGS P. (2000). Adding syntactic information to LSA. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, p. 989–993.

---

<sup>1</sup>Fragment du corpus disponible à l'adresse : <http://www.lirmm.fr/~mroche/Recherche/corpusPsy.html>