Fadoua Ataa Allah, Siham Boulaknadel,
Abderrahim El qadi, Driss Aboutajdine
GSCM-LRIT (Groupe Signaux, Communications et multimédia-Laboratoire de Recherche d'Information et Télécommunication)
B.P 1014 ; Faculté des Sciences de Rabat ; Université Mohamed V-Agdal
{fadoua_01, siham_06}@ yahoo.fr ; elqadi_a@ yahoo.com ; aboutaj@ fsr.ac.ma

L'apport des termes génériques sur la recherche d'information en langue Arabe

Résumé.

L'augmentation significative des informations en langue arabe au format électronique s'est accompagnée d'une prise de conscience de l'importance de développer des nouveaux moyens informatiques plus efficaces pour la recherche d'information. Dans ce but, nous proposons d'utiliser une technique d'exclusion des termes génériques pour améliorer la performance du modèle vectoriel.

Mots-Clés : Termes génériques, Langue Arabe, Analyse Sémantique Latente, Modèle Vectoriel.

1. Introduction

Les volumes astronomiques des données électroniques, la diversité et l'hétérogénéité des sources, durant les dernières décades, nécessitent une mise à niveau de la philosophie des traitements de ces données. Dans cet objectif, plusieurs techniques ont été adaptées pour améliorer la performance des systèmes de recherche d'information (SRI), basés sur le modèle vectoriel (Salton, 1983). Malgré le succès apporté par ce modèle ; celui ci soufre d'un inconvénient majeur résidant dans l'hypothèse d'indépendance faite sur les termes d'indexation où chaque terme constitue une dimension de l'espace vectoriel, sans considération d'éventuelles relations entre termes ou l'importance du terme vis à vis des thèmes des documents traités.

Pour remédier à ce problème, des techniques basées sur l'aspect sémantique ont été intégrées dans ce modèle.

En parallèle, vu que les données en langue arabe commencent à représenter une importante portion des données électroniques publiées, des études spécialisées dans la recherche d'information en cette langue sont menées ; plus particulièrement celles basées sur des traitements linguistiques : la racinisation (Al-Kharashi, 1991), la pseudo-racinisation (Larkey et al., 2002), N-Gram (Mustafa et al., 2004) et les Syntagmes-Nominaux (Ataa Allah et al., 2006). Tandis que le seul modèle introduisant la sémantique et utilisé pour la langue arabe est l'analyse sémantique latente (LSA) (Boulaknadel et al., 2005), nous proposons dans cet article d'introduire une autre technique basée sur l'exclusion des termes génériques (Hua Yan, 2005).

2. Exclusion des termes génériques

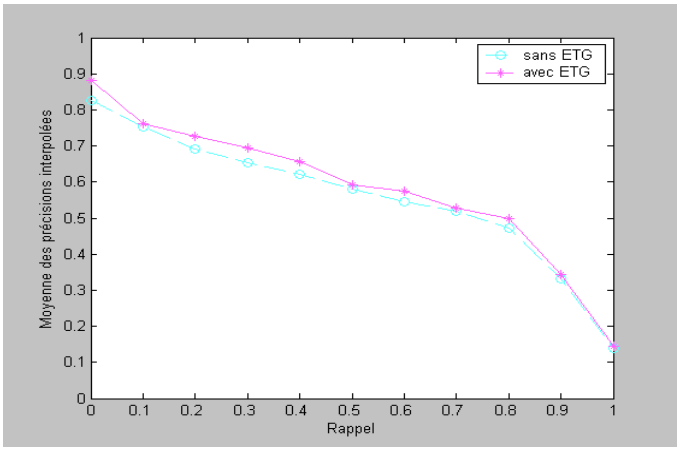
En général, les modèles de recherche d'information ne prennent pas en considération lors de l'indexation la différence entre les termes spécifiques à un domaine donné et entre les termes qui peuvent être utilisés pour tous les thèmes, appelés termes Génériques. Alors que la dominance de ces derniers dans une collection de documents induit une dégradation de la performance d'un système de recherche d'information, particulièrement pour un système basé sur l'analyse sémantique latente où les concepts résultants de la décomposition en valeurs singulières peuvent être perturbés par la fréquence prépondérante de ces termes.

Pour l'extraction des termes génériques dans une collection, un algorithme de classification basé sur l'algorithme des k-means sphérique est utilisé nécessitant la spécification du nombre des termes à extraire et un nombre de classes lors de l'initialisation de l'algorithme.

3. Résultats

Les travaux réalisés, au cours de cette étude, ont pour objectif d'évaluer l'impact de l'exclusion des termes génériques sur la performance d'un système de recherche d'information basé sur le modèle LSA en utilisant un corpus en langue arabe spécialisé dans le domaine de l'environnement.

La Figure 1 montre l'apport de l'exclusion de ces termes sur notre SRI, où nous avons constaté une amélioration de 2%. Alors que les courbes de la Figure 2 montrent l'influence du choix du nombre des classes sur la performance du système. sur la performance du système



ETG : exclusion des termes génériques

Figure 1. L'apport de l'exclusion des termes génériques

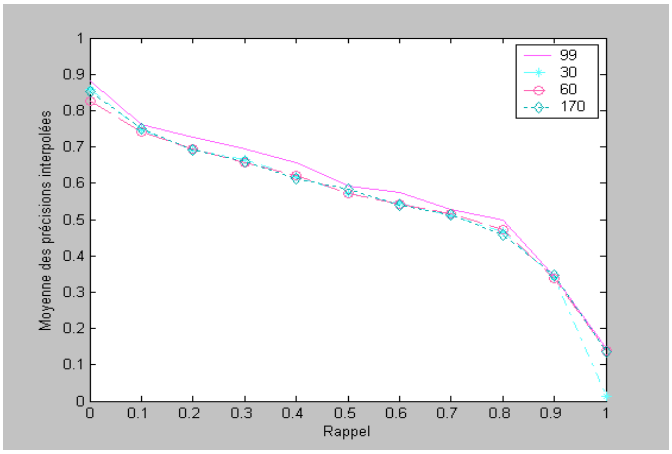


Figure 2. L'influence du choix du nombre des classes

Plusieurs tests réalisés, ont permis de conclure que l'initialisation du nombre des classes par la dimension réduite du modèle LSA correspondante au corpus utilisé, et qui égale aussi au nombre des concepts résultants, donne de meilleurs résultats qu'une initialisation aléatoire ; car vu la structure de l'algorithme utilisé, une telle initialisation permet la réduction de l'influence des termes génériques sur chaque concept du modèle LSA.

4. Conclusion

Nous avons abouti dans ce travail à des résultats similaire à l'étude originale qui a été appliquée pour un corpus en langue anglaise, nous avons aussi pu améliorer l'efficacité de l'algorithme lorsque nous avons initialisé le nombre des classes par la dimension réduite du modèle LSA correspondante au corpus ; mais il reste de voir comment le nombre des termes génériques à exclure peut être déterminé.

Références

1. Al Kharashi I., *Microcomputer based Arabic Information Retrieval System, Comparing words, stems, and roots as index terms*, Ph.D. dissertation, Illinois Institute of Technology, University Microfilm, Ann Arbor, MI, 1991.

2. Ataa Allah F., Boulaknadel S., El Qadi A. et Aboutajdine D., «Arabic Information Retrieval System Based on Noun Phrases », *ICTTA'06*, Damas, Syrie, Avril 2006.

3. Hua Y., *Techniques for Improved Lsi Text Retrieval,* Doctoral dissertation, Wayne State University, 2005.

4. Larkey L. S., Ballesteros L. and Connell M., «Improving Stemming for Arabic Information Retrieval : Light Stemming and Cooccurrence Analysis», *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*, Tampere, Finland, August 2002, p. 275-282. , 1983.

5. Boulaknadel S., Ataa Allah F., «Recherche d'information en langue arabe : influence des paramètres linguistiques et de pondération de LSA», *RÉCITAL 2005*, Dourdan, 6-10 juin 2005.

6. Mustafa S. H., Al-Radaideh Q. A., «Using N-grams for Arabic text searching», *In JASIST*, September 2004, Vol 55 n.11, p.1002-1007.

7. Salton G.and. McGill M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill,