

Expansion de la requête basée sur WordNet Arabe

Mohammed El Amine ABDERRAHIM
Université Abou Bekr Belkaid Tlemcen, Algérie

Résumé

L'amélioration de la qualité d'un Système de Recherche d'Information (SRI) passe nécessairement par la réduction de la distance entre la pertinence utilisateur et la pertinence système. Les solutions proposées tournent autour du raffinement de la fonction de correspondance et du processus de reformulation de la requête en utilisant la technique d'expansion de la requête (query expansion) ou la technique de la réinjection de la pertinence (relevance feedback). Ces techniques utilisent le plus souvent des ressources linguistiques externes comme les ontologies, les thésaurus, les bases terminologiques génériques ou spécifiques à un domaine.

Dans cet article, nous proposons d'évaluer l'apport réel de l'utilisation des ontologies pour l'expansion de la requête dans un SRI ad-hoc Arabe. Pour ce faire, nous avons effectué une expérimentation basée essentiellement sur un corpus Arabe de taille moyenne et une ontologie générique de la langue arabe « WordNet Arabe ».

Nous montrerons par ailleurs, contrairement à ce que nous avons estimé au départ, que l'expansion de la requête Arabe n'apporte pas un gain considérable en précision.

Mots Clés : TALN Arabe, Recherche d'Information Arabe, expansion de la requête, WordNet Arabe.

Keywords: Arabic NLP, Arabic Information Retrieval, query expansion, Arabic WordNet.

1. Introduction

Un Système de Recherche d'Information (SRI) repose sur les trois fonctions suivantes : stocker, organiser (indexer) et rechercher des données (en réponse à des requêtes utilisateurs). Il fait appel à trois types de connaissances :

- les connaissances sur les documents : ils regroupent les informations sur le contenu et le contenant ;
- les connaissances sur les utilisateurs ;
- et les connaissances sur le domaine d'application : ils permettent d'organiser les différents termes utilisés, on retrouve par exemple les dictionnaires, les thesaurus...

Dans le cadre de la recherche d'information pour les textes en langue arabe, la récupération de mots clé est jugée insuffisante, car les termes utilisés dans la requête peuvent présenter par rapport aux documents de la base, des différences sur plusieurs plans, par exemple :

- des variations morphologiques comme dans « مدرسة » et « خيل », « خيل » et « خيول » ;
- des variations lexicales (on utilise pour le même sens des mots différents) comme dans le cas dans « فرس » et « خيل » ;
- des variations sémantiques comme dans le cas de « الحجر : أنثى الخيل » et « الحجر : مرادف الصخر ».

L'utilisation des ontologies peut constituer une solution pour résoudre le problème des variations lexicales et sémantiques, ce qui a pour conséquence l'amélioration des résultats de la recherche. Par ailleurs l'utilisation d'un analyseur morphologique peut suffire pour résoudre le problème des variations morphologiques.

Les ontologies peuvent être utilisées à différents niveaux dans un SRI. Elles peuvent participer au processus d'indexation des documents et requêtes, nous parlons alors d'indexation sémantique ou conceptuelle. Elles peuvent également contribuer à faire l'appariement entre les documents et la requête. Enfin les ontologies peuvent aider à la formulation du besoin de l'utilisateur qui peut être formulé sous forme de requête le plus souvent en langage libre (langage proche du langage naturel). Il faut noter toutefois que la formulation de la requête est un processus très important car de sa qualité dépend la qualité des documents restitués par le SRI.

Dans ce qui suit nous allons décrire différentes utilisations d'une ontologie dans un SRI Arabe.

2. Indexation conceptuelle pour les textes Arabes

Dans ce contexte et dans le cadre des SRI pour les textes Arabes, nous avons procédé à l'évaluation des performances de l'indexation conceptuelle. A cet effet l'ontologie lexicale WordNet arabe a été utilisée. La figure 1 montre l'architecture proposée.

Les expérimentations réalisées sur un corpus de texte Arabe nous ont permis de mesurer l'apport de cette approche de reformulation de requête dans un SRI Arabe.

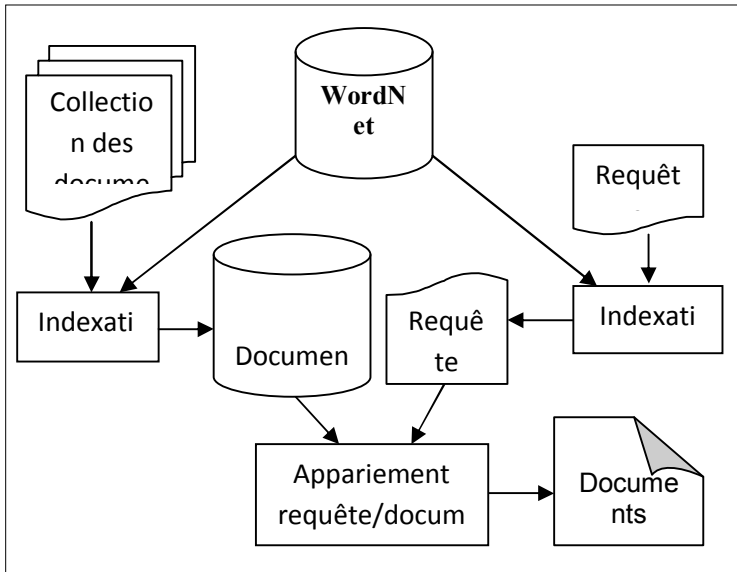


Figure 1 Architecture du SRI basée sur une indexation conceptuelle des documents et des requêtes

Nous avons aussi examiné l'approche de reformulation de la requête dans un SRI pour les textes arabe. Dans cette optique, il existe plusieurs approches, en effet, nous distinguons :

- La reformulation par l'utilisation d'une représentation du domaine de recherche.
- La reformulation par l'utilisation des relations sémantiques de bases terminologiques.
- La reformulation par l'utilisation d'un espace d'information structuré et construit automatiquement.

L'interrogation se fait par navigation (query by navigation).

- La reformulation par l'utilisation des points de vue. Les points de vue représentent des besoins élémentaires en information par exemple : causalité, définition, citation, thème,...

L'examen de ces approches, permet de dégager trois grandes démarches pour la reformulation de la requête (voir figure 2) :

- L'utilisation des ressources externes : consiste à utiliser les ressources externes comme les ontologies ou les thésaurus pour trouver des termes similaires à la requête initiale.
- L'analyse globale : cette approche consiste à analyser tout l'ensemble des documents de la collection pour extraire les termes pertinents à ajouter à la requête initiale. Deux techniques existent: le thésaurus de similarité (similarity thesaurus) et le thésaurus statistique (statistical thesaurus).
- L'analyse locale : les documents retournés en réponse à une requête sont analysés pour extraire des termes pertinents qui serviront à étendre la requête. Deux techniques sont alors proposées dans la littérature :
 - La classification locale (local clustering) : consiste à construire une matrice d'association qui quantifie les relations de corrélation entre les termes issus de l'ensemble des documents retournés en réponse à la requête initiale. Selon

la méthode de construction des relations de corrélation on distingue trois types de clusters : association clusters, metric clusters et scalar clusters. Nous développons dans la suite de cet article cette technique et nous implémentons le premier type de cluster (association clusters).

- L'analyse du contexte local : consiste à utiliser les concepts à la place de mot-clés pour représenter les documents.

Les architectures que nous avons proposées dans cet axe de recherche sont articulées autour de l'évaluation de l'apport réel de ces approches dans un SRI arabe afin de déterminer la meilleure approche à intégrer dans un SRI arabe. En effet plusieurs architectures ont été étudiées:

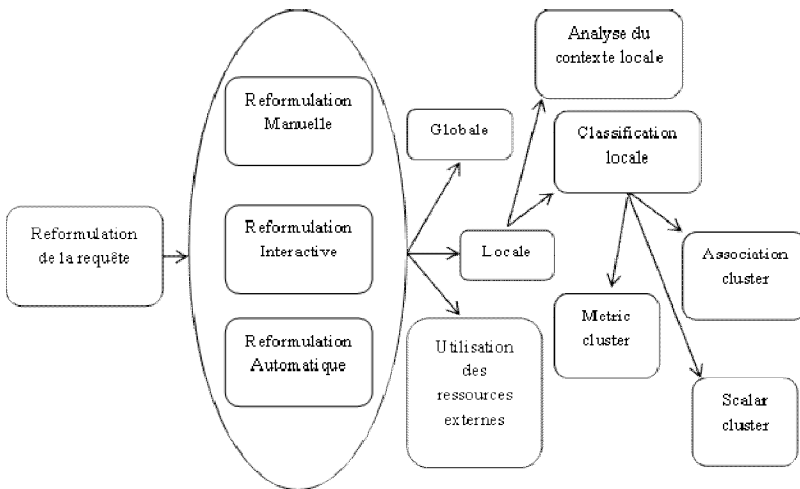


Figure 2 *Les approches pour la reformulation de la requête*

3. Reformulation de la requête par utilisation d'une ressource externe avec le moteur de recherche Google (voir figure 3).

L'évaluation de l'apport réel de l'enrichissement de la requête arabe dans le cas de l'architecture étudiée de la figure 3 est une tâche très délicate et demande beaucoup d'investigations, c'est pour cette raison que nous avons orienté notre évaluation vers l'utilisation d'un corpus fixe avec un moteur de recherche sous la forme d'une API (Lucene).

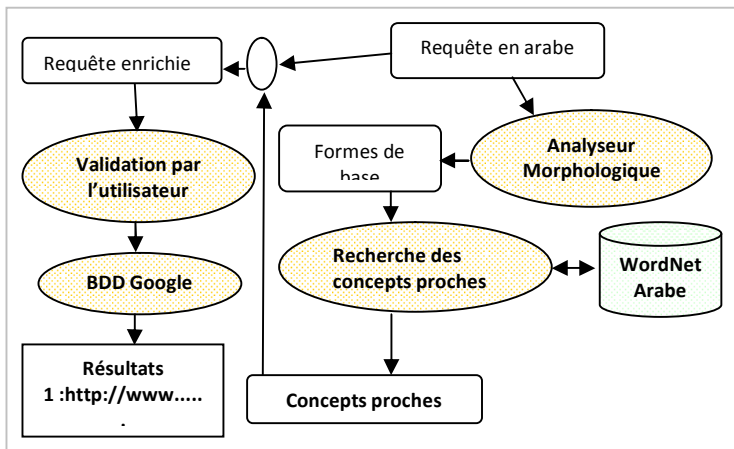


Figure 3 Architecture de l'interface de recherche pour la reformulation des requêtes (utilisation d'une ressource externe avec le moteur de recherche Google)

4. Reformulation de la requête par utilisation d'une ressource externe avec l'API Lucene (voir figure 4).

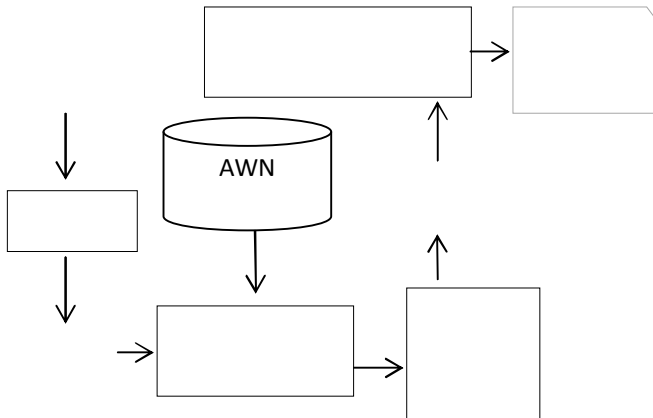


Figure 4 *Architecture de l'interface de recherche pour la reformulation des requêtes (utilisation d'une ressource externe et l'API Lucene)*

Les résultats obtenus dans le cadre de cette architecture (figure 4) nous ont permis d'une part de confirmer que cette technique de reformulation améliore considérablement le rappel, d'autre part de mesurer l'apport (6%) d'une telle approche dans l'amélioration des performances globales d'un SRI Arabe.

5. Conclusion

Le processus de recherche d'information se compose de trois parties : construire la requête, construire la réponse, évaluer la réponse. La qualité de la réponse dépend largement de la qualité de la requête construite, ainsi, une requête clairement formulée est beaucoup plus complexe que sa réponse. L'idée de cet article est d'exploiter une ressource lexicale (ontologie) dans un SRI pour les textes Arabes. Pour tester cette approche nous avons utilisé plusieurs stratégies. Les résultats obtenus de l'expérimentation des différentes approches pour la recherche d'information, nous ont permis d'une part, de confirmer l'hypothèse de départ, à savoir, l'amélioration des performances du SRI Arabe. D'autre part, d'ouvrir la voie pour tester d'autres techniques avec les mêmes données de ces expérimentations pour déterminer la meilleure approche afin de l'intégrer dans un SRI Arabe.

Bibliographie

- [1] Abderrahim M. A., Abderrahim Med Alaeddine (2010) *Using Arabic WordNet for query expansion in information retrieval system* ; IEEE, The Third International Conference on Web and Information Technologies, 16-19 June, 2010, Marrakech – Morocco.
- [2] Abderrahim M. A., Abderrahim Med Alaeddine (2012) *Réinjection Automatique de la pertinence pour la Recherche d'Informations dans les textes Arabes* ; IEEE, 4th International Conference on Arabic Language Processing, May 2–3, 2012, Rabat, Morocco; pp 77-81.
- [3] Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999) *Modern Information Retrieval*. Addison-Wesley, New York City, NY, ACM Press.

- [4] Bassam Hammo, Azzam Sleit , Mahmoud El-Haj (2007) *Effectiveness of Query Expansion in Searching the Holy Quran. Colloque internationale Traitement automatique de la langue Arabe, CITALA'07*, 18-19 juin.
- [5] Bodo Billerbeck (2005) *Efficient Query Expansion*. PHD Thesis, RMIT University, Melbourne, Australia.
- [6] Claudio Carpineto, Giovanni Romano (2012) *A Survey of Automatic Query Expansion in Information Retrieval* ; Computing Surveys (CSUR), Volume 44 Issue 1; January.
- [7] Efthimis N. Efthimiadis ((1996) *Query Expansion* ; Williams, Martha E., ed. Annual Review of Information Systems and Technology (ARIST), v31, pp 121-187, 1996.
- [8] Farag, A., Andreas, N. (2008). *AraSearch: Improving Arabic text retrieval via detection of word form variations*. SIIE'2008, Hammamet – Tunisie, 14-16 Février.
- [9] Hany M. Harb ; Khaled M. Fouad ; Nagdy M. Nagdy (2011) *Semantic Retrieval Approach for Web Documents*; International Journal of Advanced Computer Sciences and Applications, Vol. 2, No. 9.
- [10] Hlaoua L. (2007) *Reformulation de Requêtes par Réinjection de pertinence dans les Documents Semi-Structurés*. PHD Thesis, université Paul Sabatier.
<http://nrrc.mitre.org/NRRC/publications.htm>
- [11] IJsbrand Jan Aalbersberg, (1992) *Incremental relevance feedback*, Proceeding SIGIR '92 Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 11 – 22, ACM New York, NY, USA.
- [12] Jinxi Xu, W. Bruce Croft (2000) *Improving the effectiveness of information retrieval with local context analysis*. Transactions on Information Systems (TOIS), Volume 18 Issue 1, ACM, January.
- [13] Jinxi Xu, Alexander Fraser , Ralph Weischedel (2002) *Empirical Studies in Strategies for Arabic Retrieval*.

- [14] Kanaan, G., Al-Shalabi, R., Abu-Alrub, M., and Rawashdeh, M. (2005) *Relevance Feedback: Experimenting with a Simple Arabic Information Retrieval System with Evaluation*. International Journal of Applied Science & Computations, Vol. 12, No.2, USA.
- [15] Kyung Soon Lee W. Bruce Croft James Allan (2008) *A Cluster-Based Resampling Method for Pseudo-Relevance Feedback*. SIGIR'08, July 20-24, Singapore. pp. 235-242 ACM.
- [16] Marie-France BRUANDE, Jean-Pierre CHEVALLET (2003) *Assistance intelligente à la recherche d'information*. Lavoisier ; pp. 99-129.
- [17] Musa, A. (2006). Arabic WordNet and Arabic NLP. JETALA 5-7 June, Rabat.
- [18] Sabri, E., William, B., Piek, V., David, F., Adam, P., Christiane, F. (2006). *Arabic WordNet and the Challenges of Arabic*. <http://www.mt-archive.info/BCS-2006-Elkateb.pdf>.
- [19] Salton, G., and C. BUCKLEY (1990) *Improving Retrieval Performance by Relevance Feedback*. Journal of the American Society for Information Science, 41(4), 288-97.
- [20] Soraya, Zaidi, M-T. Laskri (2007) « **تمديد طلب البحث على شبكة الانترنت باللغة العربية** », Barmajiat (CSLA): Les applications logicielles en arabe: Pas vers le e-gouvernement, 9-10 décembre Alger.
- [21] William, B., Sabri, E., Horacio, R., Musa, A., Piek, V., Adam, P., Christiane, F. (2006). *Introducing the Arabic WordNet Project*. <http://www.globalWordNet.org/AWN/meetings/GWApaper.pdf>