

**Titre:** Classification des méthodes d'extraction d'informations à partir des documents arabes

**Auteurs:** Abd El Salam AL HAJJAR&§, Mohammad HAJJAR&, Khaldoun ZREIK§

**Affiliation:** & Institut Universitaire de Technologie, Université Libanaise, Liban  
§ Laboratoire Paragraphe, Université de Paris 8 - Vincennes- Saint-Denis, France

**Mots Clés:** Classification, Extraction d'information, Langue arabe, N-gram, Stemmer.

## **1. Objectif**

La langue arabe est utilisée par plus de 330 millions arabophones, mais il n'y a que 26 millions parmi eux qui utilisent l'Internet, d'après les estimations de 2006. Les sites arabes sont de plus en plus nombreux, ils sont 50 000 selon les mêmes estimations [8, 17]. Cependant, la performance de la recherche d'information en langue arabe reste très problématique à cause des problèmes structurels de la langue (polysémie, formes dérivés irrégulières, infléchie...). Pour remédier à ces problèmes, plusieurs méthodes ont été proposées. L'objectif de cet article est de proposer une première classification des méthodes d'extraction d'informations à partir des documents arabes. Dans cet article, nous nous limitons aux méthodes traitant de l'extraction du texte.

## **2. Problématique**

Pour rechercher un mot dans un dictionnaire arabe, il faut d'abord extraire la racine de ce mot puis en suite rechercher cette racine dans le dictionnaire. Ceci est dû au fait que le vocabulaire de la langue arabe est essentiellement construit à partir de la dérivation des racines. Les racines sont des mots formés de trois à cinq lettres consonnes. La langue arabe possède cinq à sept milles racines, 85 % de ces racines sont trilatérales. La dérivation de mots se fait en ajoutant à la racine des affixes (préfixe, infixé, ou suffixe) selon plusieurs modèles qui sont aux alentours de 120 [4]. Par exemple, si on prend la racine (كتب), les mots (مكتوب، كاتبة، مستكتب، كاتب) sont dérivés respectivement à partir de cette racine selon les modèles (فاعل، مستفعل، فاعلة، مفعول).

Pour extraire de l'information d'un document arabe, les méthodes associées doivent d'abords répondre à la question suivante: "Comment peut-on trouver la racine du mot à rechercher?".

## **3. Résultats**

L'étude bibliographique que nous avons réalisée nous a permis de recenser plusieurs méthodes traitant de ce sujet. Ces méthodes peuvent être classées en deux grandes catégories. La première catégorie appelée "Stemmer" exige des connaissances approfondies de la langue. La deuxième appelée "N-gram" est basée sur des approches statistiques pour extraire les informations indépendamment de la complexité de la langue. Par ailleurs, certaines méthodes utilisent à la fois les deux approches "Stemmer" et "N-gram".

### **3.1. Catégorie Stemmer**

Le Stemmer est un processus automatique utilisé pour réduire les mots de différentes formes morphologiques à la racine (Stem) commune pour améliorer la capacité du système d'extraction. Pour ce faire, plusieurs approches sont proposées. La première est basée sur un dictionnaire. En générale, ces méthodes donnent des bons résultats surtout que le dictionnaire contient tous les mots connus avec leurs inflexions. Par contre, les résultats sont modestes quand il s'agit des mots absents du dictionnaire ou étranges. L'autre approche est basée sur l'application d'un ensemble de règles prédéfinies. Cette classe présente l'avantage de ne pas exiger un dictionnaire, tâche non évidente surtout pour la langue arabe. Par contre, les points faibles de ces méthodes résident dans le fait que la construction de ces







On enlève les diacritiques qui représente des voyelles, on normalise les différentes formes de l'hamza à la forme (أ), on enlève les préfixes de longueur trois et deux, on enlève le connecteur "و" s'il précède un mot, on normalise أ, إ, إ par ا, et on retourne le stem s'il est plus petit ou égale à 3.

Ensuite, l'extraction de la racine se fait selon la longueur du stem. Si la longueur est 4, on extrait le stem pertinent et on le retourner en enlevant les préfixe et les suffixes de longueur 1 (S1, P1). Si la Longueur est 5, on extrait le stem de longueur 3 (modèles PR53), si aucun de ce modèles ne correspondent pas, alors on enlève les préfixes et les suffixes pour avoir un stem de longueur 3, si le mot reste de longueur 5, il faut utiliser le modèle PR54 pour déterminer s'il contient un stem de longueur 4 et ainsi de suite.

Les auteurs ont testé leur méthode sur la collection "Arabic Trec 2001" qui contient 383,872 documents. Ils ont aussi comparé leur méthode avec Khoja Stemmer [3] et Light Stemmer [13, 14], selon eux, leur méthode a donné une précision meilleure.

### **3.2. Catégorie N-gram**

La deuxième catégorie est basée sur des approches statistiques: N-gram. Le N-gram permet de trouver si deux mots sont semblables ou non sémantiquement à partir des structures de caractères de ces mots. Deux mots sont considérés semblables s'ils ont en commun plusieurs sous chaînes de N caractères, ceci est fait en calculant un coefficient sur ces deux mots. Les avantages de N-gram sont qu'il n'exige sur la langue ni une connaissance préalable, ni l'établissement des règles prédéfinies et ni la construction d'une base de données du vocabulaire. Cette catégorie donne un bon résultat dans plusieurs langages, même en langue arabe (3-gram et 4-gram ). Dans cette catégorie on peut trouver les classes suivantes:

#### **3.2.1. N-gram basé sur le calcul du coefficient de ressemblance**

W. Adamson George et J. Boreham [1] ont développé la première technique de la classification automatique basée sur la structure des mots. Le coefficient de la ressemblance est calculé à partir du nombre de digrammes (2-gram) assortis dans des paires des sous-caractères. Un échantillon des mots d'une base de données chimique a été choisi. Cette base contient certains stems dérivés des noms des éléments chimiques. Chaque groupe est caractérisé par une racine et les mots dérivés.

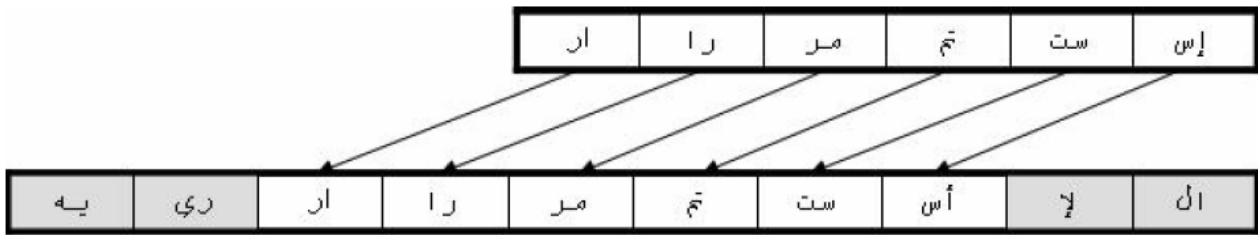
H. Suleiman Mustafa [9] a testé la performance de deux techniques N-gram sur la recherche des racines arabes, N-gram contigu et N-gram hybride. Les deux techniques ont été testées en utilisant trois expériences qui impliquent différents niveaux de stemmer du mot, un corpus textuel contient environ 25 mille mots (avec une dimension totale d'environ 160KB), et un ensemble de 100 mots de requête textuelle. Les résultats de l'approche hybride ont montré l'amélioration de la performance considérable sur l'approche contiguë.

F. Ahmed et A. Nürnberger [18] ont présenté le modèle du N-gram qui peut être utilisé pour calculer la ressemblance entre deux chaînes de caractère en comptant le nombre des N-grams semblables qu'ils partagent. Le coefficient de ressemblance par l'équation:

$$\delta_n(a, b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}$$

Où  $\alpha$  et  $\beta$  sont les ensembles de N\_gram.

Par exemple: prenons les 2 mots: استمرار استمرار



M. Sinane et al. [19] proposent une approche qui utilise N-gram basée sur le mot et les caractères. Quatre types de base ont été explorés, parfois séparément et parfois en association: Mot, Racine lexicale, Racine, N-gram. En général, les N-grams basées sur les stems sont plus meilleurs que celles basées sur les mots, car les N-grams basées sur les mots peuvent être des préfixes et des suffixes ce qui aboutit à des erreurs de similarité entre le document et la requête.

Pour expérimenter leur méthode, ils ont construit un corpus constitué des 2667 documents issus du Journal officiel de l'année 2002. Ces documents sont classifiés et chacun d'eux appartient à un ou plusieurs classes prédéfinies. En général on a trois classifications: Classification administrative, Classification juridique, Classification thématique. En utilisant ces documents classifiés, on va les segmenter en essayant de différents paramètres pour la méthode N-gram (basée sur mot, stem ou caractère). Ensuite, on va essayer de chercher les mots candidats pour chaque document et comparer les résultats obtenus par des paramètres différents. Le terme peut être un mot, 3-gram, 4-gram ou 5-gram. On a relevé toutes les diacritiques. Et annuler les mots parasites. Pour segmenter les 2667 documents qui forment le corpus, en utilisant la méthode N-gram, j'ai fait un programme (en utilisant le langage VB.net) qui utilise N-gram avec 3,4 et 5 caractères et donne les résultats dans une table (top 50).

Les résultats de cette expérience ont montré que l'utilisation de la méthode de N-gram dans la recherche d'information arabe est plus efficace que la méthode "Keyword matching", mais reste insuffisante. Il faut donc considérer un certain niveau linguistique. Les raisons de cette insuffisance sont en générale la spécificité de la langue arabe: grand nombre de synonymes, directives....Par exemple: Le mot "الحرب" a plusieurs autres synonymes comme "المعركة", "العدوان"...

### 3.2.2. N-gram basé sur la technique statistique de la fréquence

L. Khreisat [16] a étudié la technique Statistique de la Fréquence du N-gram pour classer des documents arabes. La technique emploie une mesure de la dissemblance appelée "Distance Manhattan", et une mesure de ressemblance appelée "Dicemeasure". Un corpus des documents du texte arabes a été collecté des journaux arabes en ligne. 40% du corpus a été utilisé comme classes de formations (training classes) et l'en rester 60% pour classification. Tout documents (training et document à classifier) a traversé une phase de la normalisation qui enlève signes de ponctuation, les mots parasites, les diacritiques, et les non lettres. Pour les documents de la formation, le profil de fréquence N-gram (N=3) (les trigrammes pour le mot المودعين est: عين، دع، ودع، دعي، عين) a été produit pour chaque document et été sauvé dans les dossiers texte. Ce profil a été produit pour chaque document utilisé pour classification, et été comparé avec le profil de toutes les classes de formation. Cette comparaison se fait en calculant Distance Manhattan et Dicemeasure.

Un document appartient à une telle catégorie, s'il a une plus petite Distance Manhattan, et une plus grande DiceMeasure, entre ce document et une classe de cette catégorie. Pour Comparer la performance de la technique du trigramme qui utilise Distance Manhattan et DiceMeasure, les valeurs du rappel et de la précision ont été calculées. Le meilleur résultat pour la méthode du trigramme qui utilise le Distance Manhattan a une valeur du rappel de 0.88, et le plus mauvais a une valeur du rappel

de 0.409. DiceMeasure donne des meilleurs résultats de la classification en comparant avec Distance Manhattan.

### **3.3. Catégorie: Stemmer et N-gram**

Chacune des deux approches ont des avantages et des désavantages, tant que l'approche Stemmer dépend de la langue, et sa complexité morphologique, et ne donne pas toujours la meilleure performance... et l'approche statistique N-gram est indépendant de la langue mais a des inconvénients au niveau de synonymes. Pour cela, il y a des auteurs essayent de fusionner les deux approche, pour avoir une bonne méthode.

A. N. De Roeck et W. Al-Fares [5] ont présenté une méthode pour les mots arabes qui partagent la même racine. Pour implémenter cette méthode "Clustering Algorithm", il faut passer dans deux étapes, pour cela, cette algorithme est nommée "Two-Stage". Les mots sont soumis à Light Stemmer pour enlever les affixes. La deuxième étape est basée sur l'algorithme d'Adamson avec quelques modifications. Chaque bi-gram est assigné à un poids (0.25 pour bi-gram qui contiennent des lettres faibles, 0.5 pour bi-gram qui contiennent la lettre non-faible, 1 pour tous les autres bi-gram).

## **4. Perspective**

Finalement, cette étude nous a permis de réaliser une première classification sur les méthodes qui permettent d'extraire de l'information à partir de documents arabes. La prochaine étape sera la réalisation d'une étude comparative détaillée des méthodes existantes (testes, avantages, inconvénients, utilisation, ...) afin d'en utiliser la meilleure ou pourquoi pas en proposer une nouvelle.

## **5. Référence:**

1. W. Adamson George, J. Boreham "The use of an association measure based on character structure to identify semantically related pairs of words and document titles", *Information Storage and Retrieval*, Vol. 10, pp 253-260, 1974.
2. R. Al-Shalabi, M. Evens "A Computational Morphology System for Arabic", *Proceedings of COLING-ACL*, New Brunswick, NJ, 1998.
3. S. Khoja, R. Garside "Stemming Arabic text", Computing Department, Lancaster University, Lancaster, [www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps](http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps), 1999.
4. I. Al Kharashi "A Web Search Engine for Indexing, Searching and Publishing Arabic Bibliographic Databases", 1999.
5. A. N. De Roeck, W. Al-Fares "A morphologically sensitive clustering algorithm for identifying Arabic roots". In *Proceedings ACL-2000*. Hong Kong, 2000.
6. B. Hammo, H. Abu-Salem, S. Lytinen, M. Evens "A Question Answering System to Support the Arabic Language". *Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages Philadelphia, Pennsylvania Pages: 1 – 11*. 2002
7. A. Chen, F. Gey "Building an Arabic stemmer for information retrieval". *TREC-11 conference* 2002.
8. Zeinab Ghosn, Les sites Internet gouvernementaux au Moyen-Orient 2003, The Arab Advisors Group, [www.arabadvisors.com/Pressers/presser-230101.htm](http://www.arabadvisors.com/Pressers/presser-230101.htm), 2003.
9. H. Suleiman Mustafa "Character contiguity in N-gram based word matching: the case for Arabic text searching". *Information Processing and Management*.41 (4), 819-827, 2004.
10. N. Thabet "Stemming the Qur'an" *WORKSHOP ON Computational Approaches to Arabic Script-based Languages*, University of Geneva, Geneva, Switzerland, August 28, 2004.
11. G. Kanaan, R. Al-Shalabi, J. Jaarn, M. Al-Kabi, A. Hasnah "A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots", 2004.

12. K. Taghva, R. Elkoury, J. Coombs "Arabic Stemming without a root dictionary". International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I pp. 152-157, 2005
13. H. Al Ameen, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, S. Al Muhairi "Arabic Light Stemmer: A new Enhanced Approach", The Second International Conference on Innovations in Information Technology (IIT'05), 2005.
14. L. Larkey, L. Ballesteros, M. Connell "Light Stemming for Arabic IR" Arabic Computational Morphology: Knowledge-based and Empirical Methods, A.Soudi, A. van en Bosch, and Neumann, G., Editors. Kluwer/Springer's series on Text, Speech, and Language Technology, 2005.
15. Y. Kadri, J. Nie "Effective Stemming for Arabic Information Retrieval" in proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.
16. L. Khreisat "Arabic Text Classification Using N-gram Frequency Statistics A Comparative Study". The 2006 International Conference on Data Mining Part of the 2006 World Congress in Computer Sciences DMIN 2006: 78-82, 2006.
17. Censure de l'internet dans les pays arabes, Tribune des Droits Humains - Genève 2006 - [www.humanrights-geneva.info](http://www.humanrights-geneva.info), 2006.
18. F. Ahmed, A. Nürnberger, "N-grams Conflation Approach for Arabic", ACM SIGIR Conference, Amsterdam, 27 July 2007.
19. M. Sinane, M. Rammal, K. Zreik "Arabic documents classification using N-gram", Conference ICHSL6, Toulouse, 2008.