REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Hadj Lakhdar – Batna
Faculté des Sciences
Département d'Informatique

MÉMOIRE
Présenté par
Tahar DILEKH

Pour obtenir le grade de
Magister
Spécialité : **S**ystème d'**I**nformation et de **C**onnaissance (**SIC**)

Implémentation d'un outil d'indexation et de recherche des textes en arabe

---

Soutenue publiquement le 28 /09 /2011 devant le jury formé de :

Pr. Mohammed BENMOHAMMED	Professeur	**Président**	Université de Constantine
Dr. Abdelmadjid ZIDANI	M.C.	**Rapporteur**	Université de Batna
Dr. Ali BEHLOUL	M.C.	**Co-Rapporteur**	Université de Batna
Pr. Azzeddine BILAMI	Professeur	**Examinateur**	Université de Batna
Dr. Brahim BELATTAR	M.C.	**Examinateur**	Université de Batna

ABSTRACT

Arabic, one of the six official languages of the United Nations, is the mother tongue of more than 300 million people[1]. Becoming a center of research and commercial development, the importance of the domain of information retrieval (IR) is due to the essential need for such tools to people in the Net era. The number of Internet users in 2002 was about 4.4 million which represents about 1.5 % of the population of the Arab world[2]. Few search engines are available for the growing number of Arabic Internet users; however many effort are being deployed,

Arabic is a highly inflected language and has a complex morphological structure. The information retrieval (IR) on Arabic texts requires the basic form of the word (root or lemma); therefore stemming process is necessary. This process can be defined as a process of removing all affixes (prefixes, infixes and/or suffixes) from words in order to bring them to their roots or lemmas.

Morphological complexity of Arabic language makes particularly developing software for natural languages process difficult. In Semitic languages like Arabic, the majority of names, adjectives and verbs are derived from few thousand roots by adding new letters, for example, the words مكتبة (library), كتاب (book) كتب (pounds) كتب (he wrote), and نكتب (writing), the root كتب [Wig98].

Every natural language has its own characteristics and features. Thus, it is difficult to follow the same pattern of stemming and apply the same techniques for all languages. A technique of stemming could be relevant to a language, but could not be effectively applied to other languages. There are several techniques used for word stemming; these include technical dictionaries, morphological analysis, deletion of affixes, statistics, and translation.

In this work, we proposed a hybrid method that incorporates three different techniques so as the Arabic stemming process resolve problems in connection with each mentioned technique.

These three techniques are: removal of affix given by Kadri [Kad08], dictionaries, and morphological analysis

These techniques require some adjustments to be relevant for use. Each technique is fitted and adapted individually to solve practical problems related to itself.

Therefore, the main contribution of this experiment was to demonstrate the effectiveness of the *hybrid method* compared to other methods, and the choice of removing the suffix before prefix during the operation of Arabic stemming process. For example

[1] http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm. Accédé le 12/12/2008.
[2] http://sonomabusiness.com/archives/2002-09-column-levini.html. Accédé le 05/12/2008.

Word	Root	Suffix – Préfixe	Prefix - Suffix
ألامهم	ألم	ألام	مها
Their pain	pain	Pains	error

Key Words: Information retrieval, lemmatization, Arabic language.

RESUME

L'arabe, une des six langues officielles des Nations Unies, est la langue maternelle de plus de 300 millions de personnes[1]. Le domaine recherche d'information (RI) arabe, devenu un centre de la recherche et du développement commercial est du à la nécessité essentielle de tels outils pour des personnes dans l'ère électronique. Le nombre d'internautes arabophones en 2002 était environ 4.4 millions, environ 1.5% de la population du monde arabe[2]. Mais, de l'autre côté de la réalité, peu de moteurs de recherche sont mis à la disposition des utilisateurs arabophones, bien que les efforts soient en marche pour servir le nombre croissant d'utilisateurs.

L'Arabe est une langue fortement flexionnelle qui a une structure morphologique complexe. La recherche d'information sur le texte arabe exige la forme de base du mot (racine ou lemme) pour être la plus pertinente, donc le processus de lemmatisation est nécessaire. La lemmatisation peut être définie comme un processus qui consiste à retirer tous les affixes (préfixes, infixes, ou/et suffixes) des mots pour ramener ces derniers à leurs lemmes ou racines.

La complexité morphologique de la langue arabe rend particulièrement difficile le développement des applications pour le traitement en langue naturelle. Dans les langues sémitiques comme l'arabe, la plupart des lemmes de nom, d'adjectif, et de verbe sont dérivés de quelques mille racines par l'insertion de nouvelles lettres, par exemple, les mots مكتبة (bibliothèque), كتاب (livre), كُتُبٌ (livres), كَتَبَ (il a écrit), et نَكْتُبُ (nous écrivons), de la racine كتب [Wig98].

Chaque langue naturelle a ses propres caractéristiques et dispositifs. Ainsi, il est difficile de suivre la même configuration de lemmatisation et d'appliquer les mêmes techniques pour toutes les langues. Une technique de lemmatisation pourrait être pertinente à une langue, mais ne peut pas être effectivement appliquée à d'autres langues. Il existe plusieurs techniques utilisées pour la lemmatisation des mots. Celles-ci incluent, des techniques de dictionnaires, d'analyse morphologique, de suppression des affixes, de statistiques, et de traduction.

Dans ce travail, nous avons proposé une méthode hybride qui incorpore trois techniques différentes pour que la lemmatisation arabe résolve les problèmes liés à chaque technique précédente.

Ces trois techniques sont: suppression d'affixe proposée par Kadri [Kad08], dictionnaires, et analyse morphologique.

[1] http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm. Accédé le 12/12/2008.
[2] http://sonomabusiness.com/archives/2002-09-column-levini.html. Accédé le 05/12/2008.

Ces techniques ont besoin d'une certaine adaptation pour être pertinentes pour l'utilisation. Chaque technique est adaptée individuellement pour résoudre les problèmes pratiques liés à elle-même.

La contribution principale de ce travail concerne la démonstration de l'efficacité de la *méthode hybride* comparée aux autres méthodes, et le choix de l'enlèvement des suffixes avant les préfixes pendant l'opération de lemmatisation Arabe. Par exemple

Mot	Racine	Suffixe – Préfixe	Préfixe - Suffixe
ألامهم	ألم	ألام	مها
Leurs douleurs	douleur	douleurs	Erreur

Mots clés : Recherche d'information, lemmatisation, langue arabe.

TABLE DES MATIERES