cide
Colloque International sur le Document Electronique

Session Recherche d'information

L'accès Multilingue à l'information scientifique et technologique : limitations des moteurs de recherche en langue Arabe

Majed Sanan, Mahmoud Rammal et Khaldoun Zreik
Article

Mots-clés :

Matching ; Moteur de recherche arabe ; Recherche d'information ; Précision ; Rappel

Résumé

L'Internet demeure la source essentielle d'accès à l'information scientifique et technique.

Dans certaines langues, telles que l'arabe, les moyens déployés pour la recherche d'information ne semblent avoir les mêmes performances que dans d'autres langues. Cette carence est probablement due à l'introduction tardive de l'Internet dans le monde scientifique et technique arabisant d'une part et d'autre part aux avancements dans le développement de traitement numérique de langue arabe.
Cet article vise à identifier et à expliquer les limitations et les problèmes de la recherche d'information, en langue arabe, lors de l'usage de trois moteurs de recherche demeurant « standards » basés sur le principe de comparaison des mots clés le "keyword matching" : Google, Yahoo, et Idrisi1 . Nous avons effectué une série d'expériences sur des documents juridiques arabes extraits du Journal officiel libanais. Nous avons adopté les techniques de calcul des taux de rappel et de précision comme critères de comparaisons afin d'identifier les limitations de cette méthode.
Cette étude soutenue par une expérimentation sur corpus réelle représente la problématique d'un projet de recherche entre trois établissements de recherche en France et au Liban.

Table des matières

Texte intégral

Introduction

L'Internet demeure la source essentielle d'accès à l'information scientifique et technique. Dans certaines langues, telles que l'arabe, les moyens déployés pour la recherche d'information ne semblent avoir les mêmes performances que dans d'autres langues. Cette carence est probablement due à l'introduction tardive de l'Internet dans le monde scientifique et technique arabisant d'une part et d'autre part aux avancements dans le développement de traitement numérique de langue arabe.

La langue Arabe standard moderne est la langue officielle utilisée dans les pays arabes. Elle relève certains défis, qui lui sont spécifiques, dans la recherche d'information, pour les raisons suivantes :

- Les variations orthographiques sont très répandues en arabe [1]; certaines combinaisons des lettres peuvent être écrites de différentes manières. Par exemple, parfois dans les « glyphes » combinant HAMZA2 ou le MADDA^2 avec la lettre ALEF. La présence de l'une ou de l'autre de ces accentuations peut introduire une sorte d'ambiguïté. Pour cette raison on a tendance, lors du traitement de la langue arabe, à omettre expressément ces accentuations.

XU, FRASER, WEISCHEDEL, 2002

- De même, par exemple, changer la lettre YEH (ي) en ALEF MAKSURA (ى) à la fin d'un mot est très commun. C'est d'autant plus perturbant que, les formes des deux lettres sont très semblables.

- C'est tout comme la confusion entre « résumé » en français et « resume » en anglais. Dès lors que le mot prévu et le mot écrit sont les mots valides, il est impossible de corriger les épellations sans faire appel au contexte et l'utiliser.

- L'arabe a une morphologie très compliquée [2]. En faite, la forme des mots arabes peut avoir 4 catégories d'affixes : les antéfixes, les préfixes, les suffixes et les postfixes. Ainsi un mot arabe peut avoir une forme plus compliquée s'il y a présence de tous ces affixes attachés à sa forme standard. On peut les catégoriser selon leur rôle syntaxique. Les antéfixes sont généralement des prépositions. Les préfixes représentés par une seule lettre indiquent la personne de la conjugaison des verbes au présent. Les suffixes sont les terminaisons de conjugaison des verbes et les signes du pluriel et du féminin pour les noms. Enfin, les postfixes représentent des pronoms.

KADRI, NIE, 2004

- Les pluriels cassés sont communs. Les pluriels cassés souvent ne ressemblent pas à la forme singulière, ils n'obéissent pas à des règles morphologiques normales, et ne sont pas manipulés par des stemmers existants.

XU, FRASER, WEISCHEDEL, 2002

- Les mots arabes sont assez ambigus, ce qui est dû au système tri-littéral de racine (root). En arabe, un mot est généralement dérivé d'une racine, composée le plus souvent de trois lettres. Dans certaines dérivations, une ou plusieurs lettres de racine peuvent être

lâchées, ce qui peut amplifier sensiblement l'ambiguïté entre les mots arabes.

- L'omission presque systématique des voyelles courtes dans la rédaction de textes arabes écrits [2].

- Les synonymes sont répandus, car la variété dans l'expression est souvent appréciée en tant qu'élément d'un bon modèle d'écriture par les rédacteurs ou les auteurs (littéraires, politiques ou scientifiques) arabes (Noamany, 2001).

Dans ce contexte, notre projet de recherche consiste à optimiser l'exploit des ressources d'information en langue arabe. Nous travaillons sur une base de données textuelle rassemblant des textes (en arabe) extraits de différentes sources administratives dont les journaux officiels. Cette base de données est mise à disposition des juristes et des politologues pour consultations en ligne3.

Dans un premier temps, nous nous intéressons au traitement et à la recherche d'information textuelle pour lesquels nous proposons une approche structurelle, indépendante de ressources linguistiques.

Dans ce papier nous rappelons brièvement la problématique de recherche d'information en générale avant de focaliser notre contribution sur les spécificités et les défis posés par les ressources en langue arabe. Par la suite nous présentons notre approche que nous illustrons via 3 moteurs de recherche.

La recherche d'information en langue arabe : un processus standard dans sa forme

HARMANANI, KEYROUZ, RAHEEL, 2006

La recherche d'information en langue arabe est un processus standard dans sa forme. Il est déclenché par une demande spécifique exprimée sous forme d'une « requête ». Ce processus consiste à vérifier l'existence de l'information requise et son adresse (localisation) en cas d'une réponse positive. Les requêtes fournies par les utilisateurs se composent généralement d'un ensemble de mots clés et des opérateurs booléens très simplifiés ; le système répond en localisant, à l'aide d'une procédure de comparaisons, les documents satisfaisant le plus ces combinaisons de mots. Ce procédé est fortement influencé par l'approche d'indexation adoptée (tout texte, ontologie, web sémantique, …) ainsi que par les spécificités de la langue de document indexé (segmentation, analyse lexicale, etc.) [3].

Limitations des critères « standard » d'évaluation de

l'efficacité de la recherche

Le nombre de sites culturels et « scientifiques » est en forte croissance. L'Internet est désormais un instrument indispensable dans le champ opératoire des producteurs et des consommateurs de l'information culturelle, scientifique et technique. Par conséquent la question du degré d'efficacité de la recherche dans les documents en langue arabe, doit être posée pour ne pas dire « reposer » (c'est-à-dire prendre en compte la spécificité de la langue).

LAVRENKO, 2005

Les résultats assortis d'une approche de recherche d'information « standard » en langue arabe peuvent être imprécis et inconsistant en comparaison avec la base de données, d'où une importance accrue est accordée pour mesurer l'efficacité de la recherche d'information [4]. La qualité des moyens de restitution de documents en langue arabe (via les moteurs de recherche) n'a pas été examinéeencore.

Lapremière mesure que nous avons retenue dans le cadre de notre approche méthodologique consistait à identifier les particularités (différences) de la langue arabe pouvant affecter l'efficacité de la performance de la recherche et de la restitution de documents. Dans un premier temps nous avons identifié une explication simplifiée basée sur trois caractéristiques structurelles de la langue arabe :

- Le préfixe : Pour l'article défini et quelques formes plurielles. Par exemple pour les articles définis on a : "ال،كال،فال..." et ces préfixes sont collés aux radicaux des mots.

Charlotte WIEN, 1997

- L'infixe : Pour quelques formes plurielles comme "بيوت" (maison –beit بيت , maisons –boyout بيوت) dans ce cas, on ajoute une lettre au milieu du radical (و) pour former le pluriel d'un mot.

- Le suffixe : Pour quelques pronoms et pour des formes plurielles comme "هما،هم،وا.." .

Par la suite nous nous sommes intéressés à étudier la possibilité de mesurer l'effet de ces caractéristiques. Nous avons adopté une approche comparant l'efficacité de récupération de l'arabe. Pour mettre en place cette approche comparative nous avons considéré les mesures de rappel et de précision [5] utilisées et validées depuis les années 60 pour évaluer l'efficacité de recherche d'information.

Précision : La précision (pertinence) est la proportion de documents pertinents restitués.

$$\text{Précision} = |\text{pertinents} \cap \text{restitués}| \div |\text{restitués}| = P(\text{pertinents} \mid \text{restitués})$$

Précision = $(\frac{a}{a+b}).100\%$ [4]

Où *a* représente le nombre des documents pertinents restitués, et *b* les documents restitués et qui sont jugés non pertinents.

Rappel : Le rappel (complétude) est la proportion de documents pertinents restitués par le moteur de recherche par rapport à l'ensemble des documents pertinents existants sur l'espace de recherche

$$\text{Rappel} = |\text{pertinents} \cap \text{restitués}| \div |\text{pertinent}| = P(\text{restitués} \mid \text{pertinents})$$

Rappel = $(\frac{a}{a+c}).100\%$ [4]

Où *a* représente le nombre de documents restitués et pertinents et *c* celui des documents non restitués mais pertinents.

Mesures par un nombre unique: LAVRENKO, 2005

> Nous pouvons également employer des mesures qui utilisent un nombre unique pour l'efficacité, comme suit:
>
> F1 = 2PR / (P+R) … où F1 est une moyenne harmonique de la précision et du rappel [4].

Dans le cadre de notre projet de recherche la notion de pertinence a été limitée au niveau du sujet : « un document est pertinent par rapport à une requête si le document porte le même sujet que celui de la requête, autrement le même ensemble des mots clés et des opérateurs ».

Défis de la langue arabe

Différentes approches technologiques et méthodologiques ont été proposées pour adapter (au contexte de la langue arabe) le calcul des mesures de rappel et précision [6] afin d'optimiser la qualité de la recherche d'information. Ahmed ABDELALI, 2004

Les conférences TREC sur la recherche de textes (Text REtrieval Conferences) en 2001 et 2002, puis le forum CLEF (Cross-Language Evaluation Forum) en 2002 ont contribué à montrer les accomplissements de différents groupes de recherche dans le secteur, et ont permis une évaluation concrète de différents systèmes ayant y participés. La table 1 récapitule

les techniques et les approches employées par les participants dans TREC 2001. [7]

Team	Arabic Terms Indexed				Query Language	Translation Resources Used			
	Word	Stem	Root	n-gram		MT	Lexicon	Corpus	Translit
BBN		X			A,E	X	X	X	
Hummingbird		X			A				
IIT	X	X	X		A,E	X	X		
JHU-APL	X			X	A,E,F	X			
NMSU	X	X			A,E		X		
Queens	X			X	A,E	X			
UC Berkeley		X			A,E	X	X		
U Maryland	X	X	X	X	A,E	X			X
U Mass	X	X			A,E	X	X		
U Sheffield	X				A,E,F	X			

Table 1. Techniques utilisées dans TREC 2002.

A: Arabe, E: Anglais, F: Français.

Ahmed ABDELALI, 2004

Les équipes participantes à ces colloques ont abordé différentes alternatives d'indexation de termes, de langages de requêtes, de traduction (cross-languages) et des sources de connaissance. Chaque équipe participante a adopté une technique de « sac de termes » basée sur des statistiques d'indexation au sujet de l'occurrence des termes dans chaque document. Une grande variété de techniques spécifiques ont été employées, y compris des modèles de langue, les modèles cachés de Markov, les modèles de l'espace de vecteur et les réseaux d'inférence.

Quatre types de base pour l'indexation de termes ont été explorés, parfois séparément et parfois en association :

- Mot: Un lexème isolé et unique qui représente une certaine signification.
- Racine lexicale(Stem): un morphème ou un groupe de morphèmes concaténés qui peuvent accepter un affixe.
- Racine(Root): un morphème unique qui constitue la signification basique d'un mot.
- N-gramme: Le texte est décomposé en N-gramme, i.e., les sous chaînes de longueur N, qui souvent consistent en un ensemble adjacents de caractères d'un texte. Les diagrammes contiennent deux caractères et les trigrammes trois.

Caractéristiques de la langue arabe

Ahmed ABDELALI, 2004

La langue arabe est une langue flexionnelle, ce n'est pas une langue analytique [1]. La dérivation en arabe est basée sur les modèles morphologiques et le verbe joue un rôle flexionnel plus important que dans d'autres langues. En outre, les mots arabes sont constitués

des racines représentant des connecteurs lexicologiques et sémantiques. L'arabe offre la possibilité de combiner des particules et des pronoms apposés aux mots. En d'autres termes, l'arabe permet beaucoup de liberté dans la commande des mots dans une phrase.

Ainsi, la syntaxe de la phrase peut changer selon les mécanismes transformationnels comme une extraposition, affrontement et omission, ou selon le remplacement syntactique tel qu'un nom d'agent au lieu d'un verbe.

La langue arabe est distinguée par sa sensibilité élevée au contexte sur plusieurs dimensions. Au niveau d'écriture, la forme de la lettre dépend de la lettre qui la précède et de celle qui la suit. Au niveau syntactique, les différentes relations synthétiques de concordance telles que « case-ending », « matching », connexion, association et pronominalisation représentent différents exemples de sensitivité syntactique.

Ahmed ABDELALI, 2004

Le caractère de sensibilité au contexte est non seulement limité aux lettres, aux mots, et aux phrases. Des phrases arabes sont incluses et normalement reliées par les particules de copulatives, exceptionnelles et adversatives. Pour cette raison, il est plus difficile d'identifier la fin d'une phrase arabe qu'est le cas dans d'autres langues. En outre, le « shadda » dans la langue arabe représente un accent plus élevé sur le caractère (dans d'autres langues le shadda est représenté en doublant le caractère en écrivant). Alors nous pouvons avoir deux mots : un avec le « shadda » (l'équivalent de la répétition de lettre en français : tt, mm, …) et un autre le même que le premier mais sans « shadda » ; ces deux mots peuvent avoir de sens différents.

Il faut noter que l'écriture explicite de « shadda » et de la majorité des accentuations voyelles en arabe est de moins en moins fréquente surtout dans les domaines scientifiques et techniques.

Ambiguïté

Ahmed ABDELALI, 2004

L'ambiguïté des langues est l'une des questions la plus complexe que les moteurs de recherche confrontent. Le taux de l'ambiguïté dans la recherche d'information en langue arabe est considéré parmi les plus élevés des langues pratiquées [1]. Ce point rend difficile d'adopter des moteurs de recherche dédiés à la langue arabe que les moteurs de recherche soient en langue arabe « non native » ou bien indigène. Cependant, plusieurs approches et solutions ont été mises en application et évaluées. Xu, Fraser et Weishedel (2001) ont examiné deux techniques pour manipuler une requête. Dans un premier temps ils traitent les stems -sûrs : le mot sera réduit à un stem si et seulement si le mot a un seul stem (forme fléchie) possible. En second lieu, ils traitent tout stem : considérer tous les stems probables (toutes les formes fléchies d'un mot) et tous les stems obtenus seront équitablement probables. Si un mot avait n stems possibles, chaque stem obtient la probabilité 1/n.

Comme nous l'avons signalé plus haut, le processus de « stemming » ne peut échapper à

l'ambiguïté de la langue. Certes l'ampleur de ce problème peut être réduit en adoptant des solutions complémentaires de type analyseur de syntaxe qui peut déterminer et éliminer certaines des stems qui ne s'adapteraient pas dans la structure de la phrase (procédé de désambiguïsation de sens de mot) ou en adoptant des mesures statistiques qui peuvent être tirées de l'analyse de corpus pour élire des sens basés sur la fréquence ou la Co-occurrence.

Stratégies de recherche en arabe

Ahmed ABDELALI, 2004

Introduction

Fournir à l'utilisateur le mécanisme clair de commande et une réponse rapide comprenant les documents pertinents sont les objectifs principaux des systèmes de recherche d'information connus de plus en plus sous la forme de moteurs de recherche.

Les travaux de recherche et de développement (R&D) sur le texte arabe ont toujours un long chemin à parcourir. Bien que le milieu universitaire ait fait des accomplissements significatifs, la structure morphologique complexe de la langue arabe pose des défis ; des techniques doivent s'avérer pour rendre la recherche d'information efficace pour la langue arabe (Abdelali, Cowie et Soliman, 2004). Les systèmes existants de recherche et de restitution de textes arabes pourraient être classifiés en deux groupes [1] :

- Systèmes basés sur une approche plein texte (full form based approach) : C'est le cas de la plupart des moteurs commerciaux utilisés dont le moteur web ayna (www.ayna.com) ainsi que d'autres moteurs multilingues et Unicode tels que www.alltheweb.com ou www.google.com.

Ahmed ABDELALI, 2004

- Systèmes basés sur la morphologie (morphological based approach): Les efforts qui ont été faits dans le milieu universitaire pour étudier des systèmes plus sophistiqués ont permis d'avoir une idée sur la prochaine génération des moteurs de recherche arabes. Des expérimentations ont été effectuées sur des systèmes utilisant différentes approches tenant compte de la morphologie lors de la recherche des formes fléchies (méthode de racine, méthode de stem, méthode de stem léger) [7, 8].

Ces expérimentations ont montré que, généralement, en utilisant des « stemmers » on améliore la mesure du rappel ainsi que la précision. Les expériences de Larkey, Connell en 2002 ont prouvé que la performance du « stemmer » léger est meilleur que le « stemmer » régulier.

Tandis que chacune de ces méthodes est proposée comme une solution alternative pour la recherche et la restitution des textes arabes, aucune d'elles ne peut prétendre fournir la solution optimale. Par exemple, les méthodes basées sur le mot ou le « stem » sont efficaces

pour fournir un résultat assez focalisé néanmoins elles peuvent omettre des textes pertinents. La méthode de racine, permet une recherche plus exhaustive proposant tous les textes concernés (matchés) mais ces résultats peuvent contenir de textes non pertinents. Ainsi le besoin d'une méthode plus efficace pour la recherche d'information pertinente en langue arabe demeure à l'ordre du jour.

Normalisation et mise en correspondance

XU, FRASER, WEISCHEDEL, 2002

L'orthographe arabe est fortement variable. Un type plus problématique de variation d'épellation est que certains glyphes combinant ALEF avec HAMZA ( ) ou MADDA ( ) sont parfois écrits comme ALEF plat ( ), probablement en raison de leur similitude apparente. Souvent, le mot prévu et le mot écrit réellement sont des mots valides.

Nous avons exploré deux techniques pour étudier ce problème.

1) la technique de normalisation, où nous remplaçons par exemple toutes les occurrences de l'ALEFs diacritique par l'ALEF plat.

XU, FRASER, WEISCHEDEL, 2001

2) la technique de mise en correspondance (mapping), où nous mettons en correspondance un mot portant l'ALEF plat avec un ensemble de mots pouvant potentiellement être écrits comme ce mot en changeant l'ALEF diacritique en ALEF plat. En cette absence des données de formation, nous supposerons que tous les mots dans l'ensemble sont également probables.

Les deux techniques ont des avantages et des inconvénients. La technique de normalisation est simple, mais elle augmente l'ambiguïté. La technique de mise en correspondance (mapping) bien qu'elle n'augmente pas l'ambiguïté mais elle est plus complexe.

« Stemming » : Recherche de forme fléchie en arabe

L'arabe a une morphologie complexe. La plupart des mots arabes (à l'exception de quelques noms propres et mots empruntés à d'autres langues) sont dérivés d'une racine. Une racine se compose habituellement de trois lettres. Nous pouvons considérer un mot comme dérivé en appliquant d'abord un modèle à une racine pour produire une stem et en attachant ensuite des préfixes et des suffixes à la stem pour produire le mot [10]. Pour cette raison, un « stemmer » arabe peut être basé sur une racine ou sur une stem.

XU, FRASER, WEISCHEDEL, 2002

Notre approche

Corpus

UNDP

La Chambre de députés au Liban est formée de 128 membres (Majlis al-Nuwwab), élus pour une période d'activité de quatre ans par le suffrage universel. L'Assemblée nationale libanaise a la grande commande et influence législatives. Il joue un rôle crucial dans l'orientation de la vie publique, économique, politique, et sociale du pays. L'Assemblée diffère de beaucoup d'autres pays dans le fait qu'il n'y en a aucune maison supérieure à partager les processus législatifs. Pour cette raison, nous sommes intéressés par les documents du parlement libanais dans lesquels le Journal officiel libanais forme la partie principale.

Notre corpus est constitué des documents (numériques) libanais issus du Journal officiel de l'année 2002, qui se composent de 2667 documents.

Expérience

Dans notre approche on a choisit trois moteurs de recherche offrant des services de recherche d'information en arabe :

- 1. Idrisi
- 2. Google
- 3. Yahoo

Ces trois moteurs sont utilisés comme moyens de recherche d'information adoptant le principe de "Keyword matching" à partir de notre corpus.

Cette expérience consiste, dans un premier temps, à :

- choisir des mots-clés à l'aide des experts juridiques,

- lancer la recherche à partir de ces mots-clés en utilisant les trois moteurs de recherche cités ci-dessus

- appliquer les critères d'évaluation de l'efficacité de la recherche (cf. paragraphe 2.2).

Préalablement et avant de déclencher la recherche par les moteurs, nous avons identifié, manuellement avec l'aide des experts juridiques, les documents pertinents satisfaisant les critères (a+c) correspondant à chaque requête choisie par les experts, et on a obtenu les résultats suivants :

1- En effectuant une recherche à partir du mot-clé " حرب " (guerre) nous devons trouver 75 documents pertinents (que nous avons indiqués leurs titres manuellement).

2- En une recherche à partir du mot-clé " المجلس " (assemblé) nous devons trouver 28 documents pertinents.

Résultats

En utilisant les deux mots-clés " المجلس " ، " حرب ", on obtient les résultats suivants selon le moteur utilisé.

1 - Le moteur de recherche : Idrisi

Mot-clé	a	b	c	Rappel	Précision	F1
حرب	11	110	64	14.67	9.09	11.22
المجلس	11	2023	17	39.28	0.54	1.065
Moyen	11	1066.5	40.5	26.975	4.815	6.1425

Où :

a est le nombre de documents pertinents restitués,

b est le nombre de documents restitués et qui sont jugés non pertinents (par les experts)

c est le nombre de documents non restitués mais pertinents.

2 - Le moteur de recherche : Google

Key word	a	b	c	Rappel	Précision	F1
حرب	10	115	65	13.33	8	9.99
المجلس	12	2064	16	42.85	0.58	1.14
Moyen	11	1147	40.5	28.09	4.29	5.565

3 - Le moteur de recherche :Yahoo

Key word	a	b	c	Rappel	Précision	F1
حرب	13	105	62	17.33	9.07	11.91
المجلس	15	1983	13	53.57	0.75	1.48
Moyen	14	1044	37.5	35.45	4.91	6.695

Ensuite on calcule la moyenne des moyennes des trois moteurs de recherche cités ci-dessus, et on obtient le tableau suivant :

Moyen	a	b	c	Rappel	Précision	F1
Moyen 1	11	1066.5	40.5	26.975	4.815	6.1425
Moyen 2	11	1147	40.5	28.09	4.29	5.565
Moyen 3	14	1044	37.5	35.45	4.91	6.695
Moyen	12	1085.83	39.5	30.17	4.67	6.134

Discussion

Comme le montrent les tableaux des résultats (cf. ci-avant), l'efficacité de la recherche d'information arabe des moteurs de recherche qui utilisent la méthode de « keyword matching » est très problématique. La valeur moyenne finale F1 de toutes les valeurs moyennes est 6.695%), et cela en raison des défis linguistiques de la langue arabe (cf. paragraphe 3).

Par exemple en recherchant le mot-clé " حرب " (guerre) des moteurs ont restitué des documents non pertinents contenant le mot " حربتا ", (le mot recherché étant inclus dans le mot trouvé) qui représente une région au Liban, ou bien contenant le mot-clé" حرب " qui est un nom propre (le nom d'un député libanais), ce qui n'est pas conforme au mot-clé en requête.

En outre, d'autres documents contenant par exemple les mot-clé " عدوان "," معركة "," واقعة "qui sont des synonymes de" حرب " , ne sont pas restitués, pourtant ils sont

pertinents.

Il faut rappeler que cette expérimentation a porté sur un corpus limité mais riche en informations (politique, sociale, économique,…) avec quelques mots recherchés pour montrer la limite de la recherche, sans prendre en considération les techniques évoquées en paragraphe 4, des moteurs de recherche. Les résultats auraient été nettement plus surprenants (mais difficilement mesurables) sur une recherche ouverte sur l'espace informationnel de l'Internet.

Certes, ce problème d'efficacité n'est pas spécifique à la recherche d'information en langue arabe, mais il est plus accentué dans ce cadre car les travaux de R&D en cette matière sont nettement moins importants que dans d'autres langues.

Aussi ces expérimentations montrent la complexité des traitements, par les moteurs, lexical et syntaxique de la langue arabe sans évoquer les analyses sémantiques ou pragmatiques, …

En somme, il est indispensable de doter les moteurs de recherche de moyens leur permettant d'améliorer rapidement leur performance de recherche en langue arabe pour atteindre au moins des performances analogues aux d'autres langues de même importance.

Dans notre projet de recherche nous optons pour une approche de traitements structurels/distributionnels [13, 14, 15] qu'elle soit la plus indépendante de la complexité linguistique que possible. Pour cela nous proposons une démarche d'optimisation en deux temps : l'indexation de documents par les moteurs et l'optimisation de la requête de l'utilisateur en utilisant une approche plus appropriée qui serait dérivée de la méthode n-gramme [12].

Conclusion

La langue arabe est l'une des langues parmi les plus largement répandues dans le monde (C'est une des langues officielles des nations unies), pourtant il y a relativement peu d'études sur la restitution par des moteurs de recherche de documents pertinents en arabe.

La contribution principale de cet article porte sur les expérimentations effectuées sur des moteurs de recherche utilisant le "keyword matching" comme approche pour restituer des documents arabes. Ces expérimentations ont bien confirmé que les spécificités de la langue arabe [1] rendent la méthode de « keyword matching » insatisfaisante.

Dans ce cadre nous avons présenté deux techniques qui sont bien connues dans le domaine de la recherche d'information : la normalisation de l'orthographe et le "stemming". Des expériences en d'autres langues [10, 11] ont montré que ces techniques peuvent améliorer de manière significative la qualité de la recherche. Notre étude a révélé que ces propositions ne

sont pas appropriées à la langue arabe.

Notre projet de recherche fait acte de ces résultats et tient compte de l'urgence de mettre en place d'une part une démarche d'indexation et de reformulation des requêtes et d'autre part de la complexité de la langue arabe. C'est pour cette raison que nous avons retenu une approche structurelle / distributionnelle [13, 14, 15] pour indexer les documents arabes afin de rendre la recherche d'information arabe plus efficace.

Bibliographie

[1] Ahmed Abdelali. Improving Arabic Information Retrieval Using Local variations in Modern Standard Arabic, New Mexico Institute of Mining and Technology, 2004.

[2] Xu, A. Fraser, and R. Weischedel. 2002. Empirical studies in strategies for Arabic information retrieval. In SIGIR 2002, Tampere, Finland.

[3] Haidar M. Harmanani, Walid T. Keirouz, and Saeed Raheel. A Rule-Based Extensible Stemmer for Information Retrieval with Application to Arabic.*The International Arab Journal of Information Technology, Vol.3, No.3, July 2006*

[4] Victor Lavrenko. Center for intelligent Information Retrieval University of Massachusetts Amherst. *Hopkings IR workshop, 2005*

[5] G. Salton. "Another Look at Automatic Text Retrieval systems," *CACM,* 9 (7), pp. 648-656, 1986.

[6] Leah Larkey, Lisa Ballesteros, and Margaret Connell. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In *SIGIR 200*2, pages 269to 274, 2002.

[7] Dumais, S. T. (1994) Latent Semantic Indexing (LSI) and TREC-2. In: D. Harman (Ed.),The Second Text REtrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215 , pp. 105-116

[8] Leah S. Larkey and Margaret E. Connell. Arabic information retrieval at UMass in TREC-10. In *Proceedings of TREC 1*0, 2002.

[9] Shereen Khoja and Roger Garside. Stemming Arabic text. Computer Science Department, Lancaster University, Lancaster,UK, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps,1999.

[10] James Mayfield, Paul McNamee, Cash Costello, Christine Piatko, and Amit Banerjee, JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video, and Web Retrieval.

In E. Voorhees and D. Harman (eds.), *Proceedings of the Tenth TextREtrieval Conference (TREC 2001),* Gaithersburg,Maryland, July 2002.

[11] P. McNamee, "Knowledge -light Asian Language Text Retrieval at the NTCIR-3 Workshop," *WorkingNotes of the 3rd NTCIR Worksho*p, 2002.

[12] Darwish et al, 2001; Mayfield et al, 2001; Kwok et al, 2001.

[13] D.T. Nguyen, K. Zreik. « Multilingual Web Documents: the system Hyperling » *In ICCTA 2006: 2d International IEEE Conference on Information & Communication Technologies*. Damascus- Syria, April 24 - 28, 2006. Editions IEEE.

[14] K. Zreik, D. Nguyen. « Catégorisation de documents multilingue : le système Hyperling ». *In CiDE.8. 25-28 Mai 2005,* Mission Culturelle Française, Beyrouth-Liban – Ed. Europia, Paris.

[15] D. Nguyen, K. Zreik. "HYPERLING : Système de reconnaissance et de classification des hyperdocuments multilingues ». *In RIVF05, 21-24 FEVRIER 2005, UNIVERSITE DE CANTHO* - VIETNAM

Notes

1 Moteur de recherche arabe, de la compagnie Sakhr, http://www.sakhr.com.

2 Elément d'accentuation utilisé en langue arabe

3 Le site en Langue Arabe : www.legallaw.ul.edu.lb

Pour citer ce document

Majed Sanan, Mahmoud Rammal et Khaldoun Zreik, «L'accès Multilingue à l'information scientifique et technologique : limitations des moteurs de recherche en langue Arabe», *cide* [En ligne], CIDE 10, Session Recherche d'information, mis à jour le : 02/05/2012, URL : lodel.irevues.inist.fr/cide/index.php?id=136

Quelques mots à propos de : Majed Sanan

Sinane80@hotmail.comUniversité de Caen

Quelques mots à propos de : Mahmoud Rammal

mrammal@ul.edu.lbUniversité Libanaise

Quelques mots à propos de : Khaldoun Zreik

zreik@univ-paris8.frUniversité de Paris

Retour au sommaire
Article suivant
Plan du site

Votre recherche Rechercher avancée

cide

- CIDE 10
  - Session Infrastructure et mondialisation
  - Session Recherche d'information
  - Session Le document numérique scientifique à des fins pédagogiques
  - Session Publication numérique scientifiques
  - Session Collections numériques
  - Session Document culturel
  - Session Bibliométrie

Index

- Auteur

- Index de mots-clés
- Index by keyword

Syndication

- Documents

Partenaires