International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:4, 2013

# Query Reformulation Guided by External Resource for Information Retrieval

Mohammed El Amine Abderrahim

***Abstract*—**Reformulating the user query is a technique that aims to improve the performance of an Information Retrieval System (IRS) in terms of precision and recall. This paper tries to evaluate the technique of query reformulation guided by an external resource for Arabic texts. To do this, various precision and recall measures were conducted and two corpora with different external resources like Arabic WordNet (AWN) and the Arabic Dictionary (thesaurus) of Meaning (ADM) were used. Examination of the obtained results will allow us to measure the real contribution of this reformulation technique in improving the IRS performance.

***Keywords*—**Arabic NLP, Arabic Information Retrieval, Arabic WordNet, Query Expansion.

## I. Introduction

IN information retrieval, the query formulation is a very important process because one of documents restored by the IRS depends on its quality. Therefore shows greatly the importance of Query Reformulation (QR) [8].

In this respect, researchers work distinguishes two approaches for the QR. The first approach is called direct reformulation and consists of adding new terms to the initial query and is the core of our research; it is based on external resources such as ontologies and thesaurus or the relation of co-occurrences between terms, i.e., concept-based QR [11], [14]. Whereas, the second approach, is called indirect reformulation and consists of modifying the user query by taking into account a list of documents already considered as selected. This process is called relevance feedback once supervised and pseudo relevance feedback once automatic. It should be noted however, that all the terms added to the query come only from the documents' collection and not from an external resource [1], [7], [10], [17], [19], and [21].

According to [7], [10], [14], [15] and [18] the QR guided by ontology has a positive effect in information retrieval for English language. Along the same line, some related works to Arabic are as follows:

- The system of [2] and [3] uses a lexical resource (Arabic WordNet) and a morphological analyzer to reformulate (by expansion) the user query. The expanded query is sent to Google search engine. In this system, the evaluation of the real contribution of the Arabic query enrichment is not achieved because it is very delicate and requires consequently many investigations.
- The system of [4] uses the Yahoo API search engine and Arabic WordNet to evaluate the contribution of the query user enrichment. The experimentation of this technique shows an improvement of the global pertinence of the Arabic text IRS.
- The work in [16] shows that the manual reformulation by reweighting query terms can improve the performance (recall and precision) of the Arabic IRS. For its experiment, [16] uses a corpus of 242 documents and a set of nine (9) queries.
- The work [13] on expanding the query with terms from a thesaurus shows an improvement in the recall of the Arabic IRS. It should be noticed that the corpus applied is Koran.
- The work [22] shows that the use of a thesaurus improves significantly (18%) the performance of an Arabic IRS.
- The work [23] on expanding the query using ontology in the legal field and WordNet provides significant improvements in the IRS performance.
- The system of [5] assists the user in reformulating his query by adding forms morphologically similar to the original query. This operation is based on the calculation of n-grams similarity between words in the original query and those recorded in a lexicon. Being based on the similarity of strings, this approach does not solve the problem of semantic or lexical variations. For the operations of indexing and retrieval, [5] uses the services of Google search engine.
- The study made by [20] shows that the automatic QR based on a thesaurus can improve the performance of an Arabic IRS from 10% to 20%.

This paper aims to evaluate the real contribution of the QR guided by a lexical ontology in the context of Arabic IRS. It is necessary to notice that the process of query modification is totally independent of the IRS, in other words, the enrichment does not modify the internal representation of the documents and the query of the IRS. The aim of the enrichment process is the formulation of a richer and more precise query. So, a direct consequence is the output improvement of the IRS by sending back more relevant results.

The QR (see Fig. 1) consists of analyzing the query aiming at detecting the terms which correspond to concepts in the ontology. These terms are replaced by the proximate concepts by using the ontology semantic relations. It should be noted that the synonymy relation is tested in our work.

Mohammed El Amine Abderrahim is with the University ofTlemcen, Faculty of Technology, Laboratory of Arabic Natural Language Processing BP 230 chetouane, Tlemcen, Algeria (e-mail: medamineabd@yahoo.fr).

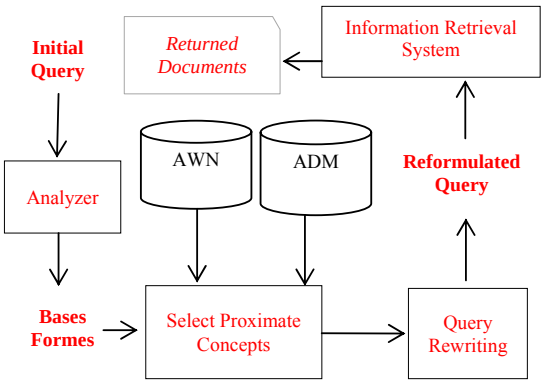


Fig. 1 The query reformulation

In the following description of experimentation and discussion of obtained results are displayed.

## II. Experiments' Description

Two different external resources for research similar concepts in the phase of QR are used in our experiment. The first resource is Arabic WordNet and is freely available for Arabic language (http://www.globalwordnet.org/AWN/AWNBrowser.html). It currently has 11,269 synsets and 23,481 words [6], [12], and [9]; the second resource is the ADM and is also freely available online (http://www.almaany.com). In its current version, it has 20,500 synonyms and 35,000 words. However, it should be noticed that ADM is more informative compared to ArabicWordNet.

Two different corpora are also used. The first is composed of more than 22,000 documents containing Arabic texts from different fields. The second, however, consists of 409 documents containing classical Arabic texts (Available at: ksucorpus.ksu.edu.sa). The main characteristics of each corpus can be found in Table I. It should be noted that 50 different queries were built for each corpus.

TABLE I
The Main Features of the Two Used Corpus

| | Corpus 1 | Corpus 2 |
|---|---|---|
| Number of text files | 22 428 | 409 |
| Fields | Health, sports, politics, sciences, religion, astronomy, nutrition, law, tales, family | Religion, linguistics, literature, science, sociology, biography |
| Size | 180 MB | 439 MB |
| Number of words | 17 000 000 | 50 602 412 |
| Number of terms | 193 736 | 350 388 |

To carry out our experimentation the freely available API Lucene search engine is used (http://lucene.apache.org/), were the whole of the documents of our two corpora is indexed and resulted in two indexes of approximately 37 MB and 72 MB respectively. Two strategies are planned for information retrieval:

- Simple Retrieval or without reformulation (SR): A set of 50 queries (see Table II) is used for each corpus.

TABLE II
Examples of Queries for Simple Retrieval

| Corpus 1 | Corpus 2 |
|---|---|
| Simple query (SR) | Simple query (SR) |
| تغذية صحية (Healthy nutrition) | ترقيقالصوت( Laminating sound) |
| فوائد البنوك ( Bank Interest) | تحالف ( Alliance) |

- Blind Retrieval (BR): A set of 50 queries deduced from the initial queries by a blind enrichment is used for each corpus. The set of the synonyms found in Arabic WordNet or ADM is added to the initial query (see Table III).

TABLE III
Examples of Queries for Blind Retrieval

| | Simple query (SR) | New enriched query (BR) From Arabic WordNet | New enriched query (BR) From ADM |
|---|---|---|---|
| Corpus 1 | تغذية صحية ( Healthy nutrition) | تغذية صحية ( Healthy nutrition)،تناول(Eat)،أكل(Meal) الطعام | مأكل (Plasmon) ، طعام (Food) ، وجبة (Meal) ، تغذية ( Healthy nutrition) صحية |
| | فوائد البنوك ( BANK (INTEREST | فوائد البنوك ( BANK (INTEREST)،عائدة (Profit)،صافي الدخل ( Net income)،صافي الربح (Net Profit)،دخل صافي (Net income)،فائدة (Interest)،كسب (Win)،مربح (Profitable)،ربح (Profit)،ربح صاف ( Net profit) | فوائد البنوك ( BANK (INTEREST)، جدوى (Feasibility)، رادة (Wills)، جداء (Multiplying)، فضل (Grace) نفع (Usefulness)، منفعة (Usefulness) |
| Corpus 2 | ترقيقالصوت(Laminating sound) | ترقيقالصوت ( Laminating sound)، أنقصَ (Reduce)، نَحُفَ (Skinny)، نَقصَ (Shortage)، قَلَّلَ (Reduce)، رَقَّ (Tenderize)، خَفَضَ (minimize)، ذعَنَ (Soft)، لانَ (supple)، سَوَّى (straighten)، إستِعبَاد (Enslavement)، عُبُودِيَّة (Enslavement)، رقّ (Enslavement) | ترقيقالصوت ( Laminating sound)، لطُفَ (Kindness) ، حَسَّنَ (improve)، رَفَقَ(Attach)، جَمَّلَ (embellish)، صنَّعَ (Manufacture) |
| | تحالف ( Alliance) | تحالف (Alliance)، إتَّحَاد فِدْرَالِي (Federation)، حلف (Treaty)، إتَّحَاد (Union)، كُنْفِدْرَالِيّ(Confederation)، مُحَالَفَة (Confederacy) | تحالف (Alliance) ، ائتلاف (Coalition) ، تآلف (Harmony) ، تكتل (coalesce) ، اتفاق (treaty)، توافق (harmonize)، وفاق(harmonize) ، تعاقد(Contract)، تعاهد (Compact)،اتفق (Agree)، تكلع (Gathering)، تجمع (Gathering) |

The retrieval results obtained by these various types of queries are saved in various files and various values of recalls and precisions of the system are calculated for each type of retrieval and query.

## III. Analysis and Discussion of the Results

### A. The Number of Documents Found

The most noticeable thing to mention, is that for all queries of the two corpora, the number of documents found after expansion is always higher than the number of documents found before, which implies an improvement of the IRS. Moreover, we found that for 47 queries (94%), the number of relevant documents found after expansion is greater than the number of relevant documents found before expansion. Therefore, we can say that the query expansion through Arabic WordNet or ADM, led to an improvement of the recall for an Arabic IRS.

### B. The Precision at Different Level of Documents

Fig. 2 presents the precision obtained at 5, 10, 20 and 100 documents (P@5, P@10, P@20, P@100) from different types of reformulation for corpus 1 and 2, and shows that the use of Arabic WordNet (or ADM) did not improve the precision at different levels of documents after the QR. Moreover, the reformulation made by Arabic WordNet is much better than ADM.

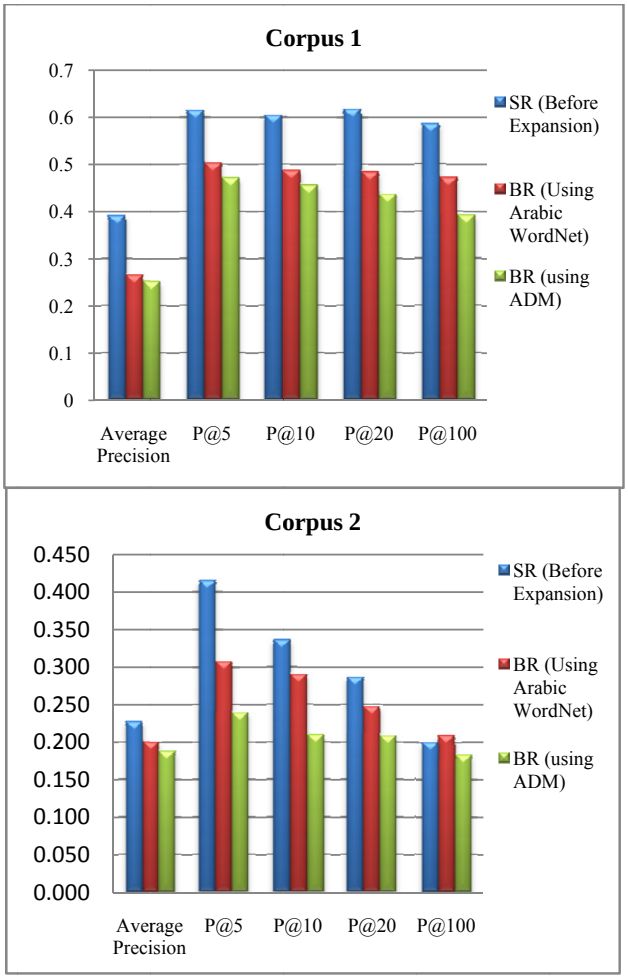


Fig. 2 Comparison of the reformulation results

### C. The Precision at Different Level of Recall

Table IV presents the precisions at 11 levels of recall for each retrieval type.

TABLE IV
The Precisions at 11 Levels of Recall (Depending on the Resource Used)

| Recall | Corpus 1 SR | Corpus 1 BR Using Arabic WordNet | Corpus 1 BR Using ADM | Corpus 2 SR | Corpus 2 BR Using Arabic WordNet | Corpus 2 BR Using ADM |
|---|---|---|---|---|---|---|
| 0 | 0.620 | 0.484 | 0.419 | 0.434 | 0.302 | 0.259 |
| 0.1 | 0.616 | 0.453 | 0.394 | 0.363 | 0.271 | 0.250 |
| 0.2 | 0.587 | 0.427 | 0.386 | 0.287 | 0.248 | 0.223 |
| 0.3 | 0.557 | 0.392 | 0.362 | 0.255 | 0.237 | 0.208 |
| 0.4 | 0.543 | 0.359 | 0.336 | 0.232 | 0.224 | 0.204 |
| 0.5 | 0.508 | 0.330 | 0.315 | 0.220 | 0.213 | 0.202 |
| 0.6 | 0.464 | 0.305 | 0.291 | 0.213 | 0.207 | 0.200 |
| 0.7 | 0.431 | 0.293 | 0.269 | 0.202 | 0.202 | 0.195 |
| 0.8 | 0.402 | 0.281 | 0.259 | 0.197 | 0.196 | 0.192 |
| 0.9 | 0.368 | 0.257 | 0.243 | 0.190 | 0.190 | 0.189 |
| 1 | 0.304 | 0.207 | 0.201 | 0.168 | 0.183 | 0.184 |

Fig. 3 shows the curves' recall / precision obtained from Table IV.

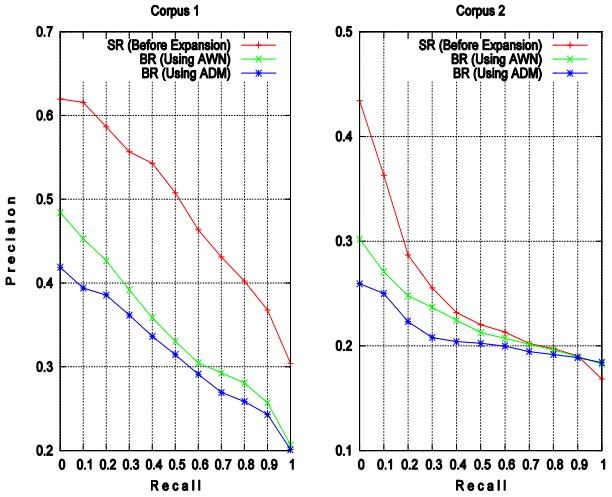


Fig. 3 Comparison between recall / precision curves

From Fig. 3 we can deduce that:

- The QR has not led to an overall improvement in the performance of the Arabic IRS.
- The use of Arabic WordNet is more beneficial than ADM.

To understand the effect of different types of retrieval on each query, various measures that are mainly based on the comparison of results before and after expansion have been established.

For a given query, three cases can arise and are as follows:

- Improvement (+): All precisions (at 11 points of recall) before are lower than those after. In other words, the curve (recall / precision) after is over before.
- No improvement (-): is the inverse of the previous case.

The curve (recall / precision) before is over after.

- No decision (X): for some precisions, there is an improvement but for others there is no improvement. In other words, there is an intersection of the two curves (recall / precision).

Table V shows for each query used in the experiment the indicator: Improvement (+), No improvement (-) or No decision (X).

TABLE V
THE INDICATOR: IMPROVEMENT (+), NO IMPROVEMENT (-) OR NO DECISION (X), OF THE DIFFERENT TYPES OF QUERY EXPANSION ACCORDING TO THE RESOURCE USED

| | Corpus 1 | | Corpus 2 | |
|---|---|---|---|---|
| Query N° | BR Using AWN | BR Using ADM | BR Using AWN | BR Using ADM |
| 1 | - | - | - | - |
| 2 | X | - | X | X |
| 3 | - | - | X | X |
| 4 | X | - | + | + |
| 5 | X | - | X | X |
| 6 | - | - | X | - |
| 7 | - | - | - | - |
| 8 | - | - | X | X |
| 9 | + | X | - | X |
| 10 | - | - | X | - |
| 11 | X | + | - | - |
| 12 | + | + | X | X |
| 13 | X | - | - | X |
| 14 | X | X | X | X |
| 15 | - | - | - | - |
| 16 | - | - | X | - |
| 17 | - | - | - | - |
| 18 | - | - | X | X |
| 19 | X | X | X | X |
| 20 | - | - | + | X |
| 21 | X | X | X | X |
| 22 | X | + | + | + |
| 23 | X | X | X | X |
| 24 | X | X | X | X |
| 25 | - | - | X | + |
| 26 | - | - | X | - |
| 27 | + | X | X | X |
| 28 | + | - | X | X |
| 29 | - | - | X | X |
| 30 | - | - | X | X |
| 31 | - | - | - | X |
| 32 | + | X | X | X |
| 33 | - | - | X | X |
| 34 | - | - | X | X |
| 35 | - | X | X | - |
| 36 | X | + | X | X |
| 37 | - | - | X | X |
| 38 | + | X | - | - |
| 39 | - | - | X | - |
| 40 | - | - | X | X |
| 41 | - | - | X | X |
| 42 | - | X | - | - |
| 43 | X | - | X | X |
| 44 | X | X | X | X |
| 45 | - | - | X | X |
| 46 | + | - | + | + |
| 47 | - | - | X | X |
| 48 | - | - | X | X |
| 49 | - | - | X | X |
| 50 | - | - | - | - |

From Table V we can deduce that:

**a)** For the first corpus, whatever the resource used in the reformulation, we note that there is:

- Improvement (+) in only one query (query number 12) (2%).
- No improvement (-) in 27 queries (54%).
- No decision (X) in 6 queries (12%).
- A set of 16 queries (32%) for which there is at least an improvement in 9 queries (18%). For the remaining 7 queries (14%), there is no improvement or indecision.

**b)** For the second corpus, whatever the resource used in the reformulation, one can that there is:

- Improvement (+) in 3 queries (queries number: 4, 22 and 46) (6%).
- No improvement (-) in 8 queries (16%).
- No decision (X) in 28 queries (56%).
- A set of 11 queries (22%) for which there is at least an improvement in 2 queries (4%). For the remaining 9 queries (18%), there is no improvement or indecision.

From the point of improvement view, the facts (a) and (b) suggest that:

- The reformulation by the use of an external resource can improve the performance of an Arabic IRS about 4%.
- There are queries that are hardly improvable with the QR.

In order to determine the best resource for the reformulation (Arabic WordNet or ADM) the number of queries for each retrieval type has been counted (see Table VI).

TABLE VI
THE NUMBER OF QUERIES SATISFYING THE CONDITIONS: IMPROVEMENT (+), NO IMPROVEMENT (-) OR NO DECISION (X), DEPENDING ON THE RESOURCE USED

| | Number of queries | | | |
|---|---|---|---|---|
| | Corpus 1 | | Corpus 2 | |
| | BR Using Arabic WordNet | BR Using ADM | BR Using Arabic WordNet | BR Using ADM |
| Improvement (+) | 7 (14%) | 4 (8%) | 4 (8%) | 4 (8%) |
| No improvement (-) | 29 (58%) | 34 (68%) | 11 (22%) | 14 (28%) |
| No decision (X) | 14 (28%) | 12 (24%) | 35 (70%) | 32 (64%) |

Results in Table VI can show that the reformulation by the use of Arabic WordNet has the best rate of improvement (14%). Furthermore, the analysis of the recall / precision curves (see Fig. 3) confirms this result.

Finally, we can conclude that the contribution of the use of an external resource in an Arabic IRS is about 4%. Moreover, it appears that Arabic WordNet is better than ADM.

## IV. CONCLUSION

In this article, one technique for QR in Arabic IRS is examined. This reformulation is based on an external resource,

to do this, two resources namely Arabic Wordnet and ADM with two different corpora were tested.

The results obtained, firstly, have confirmed that this technique improves significantly the recall reformulation, and allowed measure the contribution (4%) of this approach in improving the overall performance of the Arabic IRS. As for the queries' number that has really led to an improvement, comparison of results has showed that the use of Arabic WordNet is better than ADM. Moreover, in terms of recall / precision, the use of Arabic WordNet is more beneficial than ADM. As a conclusion Arabic WordNet offers a better rate of improvement in the Arabic IRS performance although ADM has the advantage of being more informative compared to Arabic WordNet.

As for perspectives it can be said that this study has paved the way to test and compare other methods of reformulation with the same data of this experiment in order to determine the most appropriate technique to be adopted for Arabic IRS.

## REFERENCES

[1] Aalbersberg, I.J.: Incremental Relevance Feedback. In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 11–22. ACM (1992).

[2] Abderrahim, M.E.A.: "vers la recherche d'information de contenus en arabe fondée sur l'enrichissement des requêtes. In: Proceedings of the 2nd Conférence Internationale, Systèmes d'Information et Intelligence Economique, SIIE2009 Hammamet – Tunisie, 12-14 Février, Proceedings IHE éditions, ISBN 9978-9973-868-21-3, pp. 598–607 (2009).

[3] Abderrahim, M.E.A.: Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. In: Proceedings of the 2nd ConférenceInternationalesurl'Informatiqueetses Applications (CIIA'09), Saida, Algeria, May 3-4, pp. 60–69 (2009).

[4] Abderrahim, M.E.A., A, M.A.: Using Arabic Wordnet for Query Expansion in Information Retrieval System. In: IEEE The Third International Conference on Web and Information Technologies 16-19 June Marrakech Morocco (2010).

[5] Ahmed, F., Nürnberger, A.: Arasearch: Improving Arabic Text Retrieval via Detection of Word Form Variations. In: SIIE 2008 Hammamet – TunisieFévrier 14-16, pp. 309–323 (2008).

[6] Alkhalifa, M.: Arabic Wordnet and Arabic NLP. In: Journées d'Etudes sur le Traitement Automatique de la Langue Arabe JETALA 5-7 Jun Rabat (2006).

[7] Baeza-Yates, R., Berthier, R.N.: Modern Information Retrieval. Addison-Wesley New York City NY ACM Press (1999).

[8] Baziz, M.: Indexation conceptuelle guidée par ontologie pour la recherche d'information. Ph.D. thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier (2005).

[9] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Fellbaum, C.: Introducing the Arabic WordNet Project. In: Proceedings of the Third International WordNet Conference, pp. 295–300 (2006).

[10] Carpineto, C., Romano, G.: A survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys (CSUR) 44(1), 1 (2012).

[11] Efthimiadis, E.N.: Query Expansion. Annual review of information science and technology 31, 121–187 (1996).

[12] Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., Alkhalifa, M.: Arabic WordNet and the Challenges of Arabic. In: Proceedings of Arabic NLP/MT Conference, London, UK. Citeseer (2006).

[13] Hammo, B., Sleit, A., El-Haj, M.: Effectiveness of Query Expansion in Searching the Holy Quran. In: colloque internationale Traitement automatique de la langue Arabe:, CITALA, vol. 7, pp. 18–19 (2007).

[14] Harb, H.M., Fouad, K.M., Nagdy, N.M.: Semantic Retrieval Approach for Web Documents. IJACSA) International Journal of Advanced Computer Science and Applications 2(9), 11–75 (2011).

[15] Hernandez, N.: Ontologies de domaine pour la modélisation du contexte en recherche d'information. Ph.D. thesis, Université Paul Sabatier-Toulouse III (2005).

[16] Kanaan, G., Al-Shalabi, R., Abu-Alrub, M., Rawashdeh, M.: Relevance Feedback: Experimenting with a Simple Arabic Information Retrieval System with Evaluation. International Journal of Applied Science and Computations Vol 12 No 2 USA (2005).

[17] Lee, K.S., Croft, W.B., Allan, J.: A Cluster-based Resampling Method for Pseudo-relevance Feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 235–242. ACM (2008).

[18] Nathalie, H., Gilles, H., Josiane, M., Bachelin, R.: Ri et ontologies – etat de l'art. Tech. rep., INP Toulouse, Université Paul Sabatier Toulouse III (2008).

[19] Salton, G., Buckley, C.: Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science 41(4), 288–97 (1990).

[20] Wedyan, M., Alhadidi, B., Alrabea, A.: The effect of using a thesaurus in arabic information retrieval system. International Journal of Computer Science Issues, IJCSI 9(1), 431–435 (2012).

[21] Xu, J., Croft, W.B.: Improving the Effectiveness of Information Retrieval with Local Context Analysis. ACM Transactions on Information Systems (TOIS) 18(1), 79–112 (2000).

[22] Xu, J., Fraser, A., Weischedel, R.: Empirical Studies in Strategies for Arabic Retrieval. In: Annual ACM Conference on Research and Development in Information Retrieval: Proceedings of the 25 th annual international ACM SIGIR conference on Research and development in information retrieval, vol. 11, pp. 269–274 (2002).

[23] Zaidi, S., Laskri, M.: Expansion de la requête Arabe sur le réseau internet. In: Barmajiat (CSLA): Les applications logicielles en arabe: Pas vers le e-gouvernement 9-10 Décembre Alger (2007).