# Performances of the Most Popular Search Engines in Arabic Language

A. Hajjar, M. Hajjar, G. Lebbos, K. Zreik, and M. El-Sayed

***Abstract*—The arabic language is different from Western languages especially at the morphology and spelling variations. Indeed, the performance of information retrieval systems in the arabic language is still problematic. For this reason, we are interested in studying the performance of search engines which is the most famous between 2006 and 2010, on a corpus of a thousand arabic documents. We found that morphological analysis is not taken in consideration in these engines. Morphological analysis of an arabic word is to identify its morphemes, its affixes, its model and its root.**

***Index Terms*—Search engine, information retrieval, arabic information extraction, Google desktop, corpus.**

## I. Introduction

A search engine is a communication software, making it possible to find a whole of the resources, which answer a request user [1]. These resources can be web pages, images, videos, files, etc. Which are represented by documents of different formats (HTML, JPEG, MPEG, PDF, etc.). The importance of this engine depends on the relevance of the overall result that can contain million web pages. Certain pages can be more relevant and accessible than others.

The performance of search engines varies with the language used, and depends on the nature and the complexity of the language, in which the request of research is formulated. The operation of an engine is mainly based on an automatic treatment of the natural language. These treatments differ from one language to another, and may depend on the particular characteristics of this language [2]. So it is easy to see the part played by the structure of a natural language in the access to the information in the document. The performance of search engines depends mainly on the effectiveness of the indexing methods and the information retrieval, which constitute the heart of these systems [3]-[6]. Most of the available engines which are primarily developed for the Western languages, such as English, are increasingly powerful in these languages. In addition, these performances are less, in the case of the Arabic language, probably because of specificities morphological and structural characteristics of Arabic compared to the Western languages [7]-[13]. Indeed, few studies have focused on studying its performance in the Arabic language. For these reasons, we are interested in studying the performance of these engines to be extracted the relevant information from the Arabic documents. With this intention, we choose the most famous  search engines, between 2006 and 2010, Google, Yahoo, Copernic, Bing, Ask, AOL Search and MSN / Live [14], to perform our experiments. Therefore, we present in this paper the search engines and their performance in Arabic language.

The following section presents the general architecture and the total function of a search engine. In Section III, we present the methodology and the corpus used to perform our experiments. Next, the results are given in Section IV. Finally, we finish by a conclusion.

## II. Search Engines

### A. General Principle

A search engine can provide a set of documents in response to a given query [1]. The entry of the engine is a query which can be only one word, a set of words or a phrase. The engine analyzes each word of the query and checks its index, while starting with the statistical analysis to find the documents containing exactly the word, or the phrase of the request. Then it tries to use the techniques of automatic processing of the natural language, to find a list of the most relevant documents. The result contains a short summary, containing the title and sometimes an outline of each document belonging to them. The search engines traverse all the visited pages of the web to feed their databases with copies of these documents. The search engines analyze then the contents of these documents, to determine the key words, as the titles, the headings, the contexts of the document, etc. The resulting data are stored in a database [15].

### B. Architecture of a Search Engine

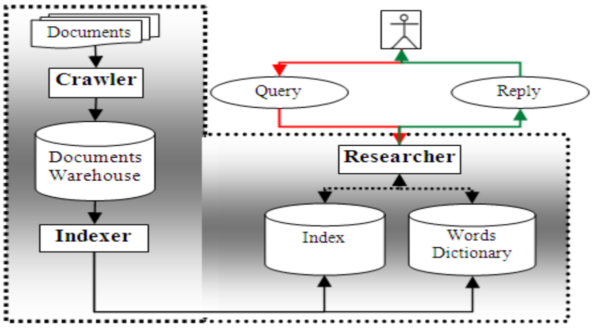


Fig. 1. General architecture of a search engine.

Fig. 1 presents the general architecture of a search engine. Its operation is structured around three main components: the crawler, the indexer and the searcher. Other components may be used, according to the complexity of the search engine.

Manuscript received June 7, 2013; revised August 15, 2013.

Abd El Salam Al Hajjar, Mohammad Hajjar, and George Lebbos are with the Lebanese University, University Institute of technology, Sidon, Lebanon (e-mail: abdsalamhajjar@hotmail.com).

Khaldoun Zreik is with Hypermedia Department, University Paris 8.

DOI: 10.7763/IJCTE.2014.V6.826

*1) Crawler*

The Crawler is a computer program, which regularly browses the web in an automatic way. The other terms used to designate the crawler are: Ants, Automatic indexing, Bots, Spiders, Robots, etc. Search engine use the crawler as a data provider. Then, the crawler is mainly used to create a copy of all pages visited and for storing them in a data warehouse. Other engine components are used to index these documents in order to propose future research fast offline [16].

*2) Indexer*

The indexer is an important component of search engine. Its work follows that of the Crawler. Indeed, the indexer must treat all the pages downloaded by the crawler, and analyzing them by breaking up them into several areas such as titles, headers, text, links, styles and the other portions of the page. In addition, it treats the pages by removing all HTML tags, stop words and words filtered, before storing in a database, called the index. In general, the database is divided into two parts: Part index and words dictionary. Most indexers treat only the internal factors of the page, but the most advanced is also the external factors, for example, the number links to another page. Indexing is then performed with this information and by using a suitable method of indexing. This method has a role is to analyze and create the relations between information stored in the database [17] However, the words of a language are formed according to specific rules and guidelines that differ from one language to another. The method of indexing has to cover the processing of all natural languages that are supported by the search engine.

*3) Researcher*

When a user poses a query containing one or more terms on a search engine, the researcher receives this request and consults the index makes it all the documents ordered by relevance. So the researcher is the intermediate component between the search engine and the database (index) already created and powered by indexer [15].

*4) Additional components*

The following additional components are often used [2]:

*a) Spellchecker*

This component corrects the spelling mistakes in the words of the query.

*b) Stemmer*

The Stemmer retrieves the Stem of each word in the query.

*c) Anti dictionary*

This feature erases all blank words a same time in the index and in the queries (as "the", "of", "to") which are disturb the research score by introducing a noise.

## C. Function

The function of a search engine consists of three main stages based on the preceding components. It begins with the exploration which collects the documents from web sites, using the Crawler component. Then, the indexing phase is to extract significant words of these documents and store them in an organized database as a large dictionary, with an index that makes it possible to find quickly in which part of the document is given a significant term. Non significant terms are called empty words. A weight value is associated in each significant term. This weight corresponds to the number of the word occurrence in a document. An algorithm is applied, to find in the corpus, the most relevant documents that match the query and present them in order of relevance. The simplest search engines are based on Boolean results to compare the words of a query with the words of the documents. But this method quickly reaches its limitations on large corpus. The most advanced search engines are interested in the weight of the words in documents, and use the method of semantic analysis (e.g. the word "war" is automatically associated with its similar words such as "weapon")[1], [2].

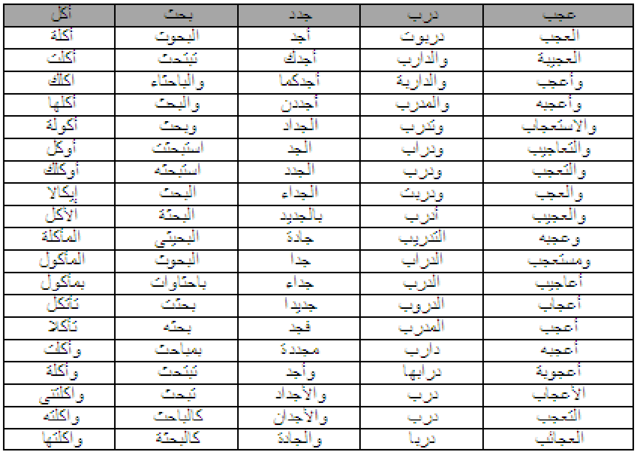
# III. Evaluation Methodology

The most search engines used between 2006 and 2010 are in the order: Google, Yahoo, Copernic, Bing, Ask, AOL Search and MSN / Live. These engines represent over 99% market [14]. To evaluate the performance of the search engines on Arabic documents, we chose the search engines that can run on a local computer. We were able to find local versions for all the search engines except the Bing. Thus, our study covers six search engines. In this section we will detail the corpus and the procedure which we used to perform our experiments.

## A. Corpus

The corpus which we built is a set of thousand Arab documents in several formats (TXT, PDF, HTML, XML, etc.). The core of these documents is a set of articles collected from sites of the daily Lebanese newspaper: Al-Jazeera, Al-Mustaqbal, Al-Nahar and Al-Safir [18]-[21]. The selected documents include one or more words in the Table I. The Table I contains the five roots <أكل, akl, he ate>, <بحث, bhhth, he searched> <جدد, jdd, he renewed>, <درب, drb, he coached> <عجب, ajb, he admired> and ten words derived from each. So this table has more than fifty words divided into five sets, where each set is formed by a root and its ten derived words. We then used these documents and modified them one by one in a way that a document contains, at most one word in each set. So each of these fifty documents contains one of the words in the Table I. We assume that a document which contains a word doesn't contain any other word of the same set.

TABLE I: List of Five Roots <أكل, AKL, he ate>, <بحث, BHHTH, he Searched> <جدد, JDD, he RENEWED>, <درب, DRB, he Coached> <عجب, AJB, He Admired> and Ten Words Derived from Each.



| أكل | بحث | جدد | درب | عجب |
|---|---|---|---|---|
| أكلة | البحوث | أجد | دربوت | العجب |
| أكلت | تبتحث | أجدك | والدارب | العجيبة |
| اكلك | والباحثاء | أجدكما | والدارية | وأعجب |
| أكلها | والبحث | أجدن | والمدرب | وأعجبه |
| أكولة | وبحث | الجداد | وتدرب | والاستعجاب |
| أوكل | استبحثت | الجد | ودراب | والتعاجيب |
| أوكلك | استبحثه | الجدد | ودرب | والتعجب |
| إيكالا | البحث | الجداء | ودربت | والعجب |
| الأكل | البحثة | بالجديد | أدرب | والعجيب |
| المأكلة | البحثي | جادة | التدريب | وعجبه |
| المأكول | البحوث | جدا | الدراب | ومستعجب |
| بمأكول | باحثاوات | جداء | الدرب | أعاجيب |
| تأكل | بحثت | جديدا | الدروب | أعجاب |
| تأكلا | بحثه | فجد | المدرب | أعجب |
| وأكلك | بمباحث | مجددة | دارب | أعجبه |
| وأكلة | تبتحث | وأجد | درابها | أعجوبة |
| واكلتي | تبحث | والأجداد | درب | الأعجاب |
| واكلته | كالباحث | والأجدان | درب | التعجب |
| واكلتها | كالبحثة | والجادة | دربا | العجائب |

These fifty documents constitute the core of our corpus. To reach the thousand documents of the corpus, we performed an automatic generation starting from this core. To do this, we take a document associated with a word Table I. We refer to this document as a source and keyword that contains as word origin. Then we build from the original word, twenty words derived by using a set of 7 affixes (4 prefixes and 3 affixes) (Table II). Then we take each word derived from this set and we construct a new document by copying the original document and replace only the original word by derived word. This process is repeated for all words in the Table I to produce the entire corpus. In this way, each document among the fifty of the core must be at the origin of the others twenty derived documents. Thus, a document may contain the word origin as it appears in the Table I or one of its affixes form. Take for example the word <مأكول, makwul, eatable> that derives from the root <أكل, akl, he ate> then we take one document which contains only <مأكول, makwul, eatable > from all that the root <أكل, akl, he ate>.A first derived document may contain <المأكول, al makwul, the eatable> the affix form of the word <مأكول, makwul, eatable > resulting from the addition of the affix (prefix) < ال, al> at the beginning of the word. A second derived document may contain <مأكولات, makwulat, food> the affix form of the word <مأكول, makwul, eatable> resulting from the addition of the affix (suffix) <ات, at> at the end of the word. Another derived document may contain <المأكولات, almakwulat, the food> the affix form of the word <مأكول, makwul, eatable> resulting from the addition of the affix (prefix) < ال, al> at the beginning and the affix (suffix) <ات, at> the end of the word. Every word derived from the word <مأكول, makwul, eatable> (Table II) is associated with a document [22].

### B. Procedure

Two preliminary steps are carried out. The first consists in making index the thousand documents of the corpus by each evaluated search engine. The second is the manual analysis performed for each document, which was described in the previous corpus section. This analysis is a reference which is used to validate and measure the results of queries of the following steps. The next step in this procedure is the definition of queries. In our case, we launched one hundred queries on each search engine. Each query consists of a single word. These words are those of the Table I and one form affix of each one of these words. Then these queries are asked, one after another, on search engines. The results of each query are compared to documents provided initially.

### C. Measures

To evaluate the results of each query, we used the classic measure, the precision and the recall, used in information retrieval. If proposing for a query Q, $S_{Found}$ is the number of the found documents and $S_{Relevant}$ is the number of the relevant documents, so these measures are:

Precision: For a query Q, the precision indicates the proportion of the relevant documents among the found documents (1).

$$P = \frac{|S_{Relevant} \cap S_{Found}|}{S_{Found}} \qquad (1)$$

Recall: For a query Q, Recall measures are the proportion of relevant documents to Q, which have been found (2).

$$R = \frac{|S_{Relevant} \cap S_{Found}|}{|S_{Relevant}|} \qquad (2)$$

## IV. Results

The results of hundreds of queries to the Google search engine are presented in Table III. The first column gives the application launched. These requests are all formed of single word. The second column contains the keywords of the documents found for each query. The third and fourth columns respectively show the precision and recall measure for each query. These results show that the Google search engine that can retrieve documents, contain exactly the query word and whatever the word is used. For this reason, the precision of the word used in Google search engine, it is 1 (one document found) or 0 (no documents found). Similarly, the recall of the word, it is 5% (one document found among twenty relevant) or 0% (no documents found). For example, if we have a query <أكل, akl, he ate> Google can retrieve only the documents containing the same form of the word, without changing any letter of the word. Thus, Google can retrieve the document containing the word <وأكل, wakl, and he ate» which is derived from the word <أكل, akl, he ate>, butdrow eht gniniatnoc tnemucod eht eveirter ton nac <يأكل , akl, he eats>. This is due, to the fact that the form of <أ, a> in the word <أكل, akl, he ate> is different from of <أ, a> in the word <يأكل , akl, he eat>. So it seems that the search engine gnizylana tuohtiw drow nettirw a fo mrof eht sredisnoc ti.These results are also confirmed in case of the word <عجب, åjb, he admire> is written with three letters <ع,- å>, <ج,-j-> and <ب, b->. In this case, taht stnemucod seveirter elgooG sdrow eht niatnoc <عجب, åjb, he admire>, <وعجب, wu åjb, and he admire> and <وأعجب, wua åjb, he impressed>, but it does not extract the documents that contain the words <يعجبهم, yi åjbhm, he impresses them> and <مستعجب, mst åjb, exclamatory> (Fig. 2). Indeed, the forms of the writing letters <ع, å> and <ب, b> in the word <يعجبهم, yi åjbhm, he impresses them> are : <ـعـ, - å> and <ـبـ, -b->, that are different from <عـ, - å> and <ـب, b-> that contained in the query of the word <عجب, åjb, he admire>. Similarly, in case of the request <مستعجب, mst åjb, exclamatory>, the form of the writing letter <ع, å> in the word <مستعجب, mst åjb, exclamatory> is <ـعـ, - å > which is different from that stated in the query word <عجب, åjb, he admire> that is <عـ, - å> (Fig. 2). In addition, the average precision of the Google search engine does not exceed 50% and the average recall is about 2%. From that result, the problems of the Google engine appear in its local version, in the extraction of information from Arabic documents. It seems that the specific treatments for the Arabic language, particularly the morphological analysis, are not included in this search engine.

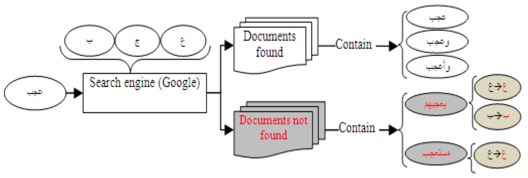


Fig. 2. Documents and extracts documents not retrieved by Google in the case of the query "عجب".

TABLE II: RESULTS OF HUNDRED QUERIES TO THE GOOGLE ENGINE.

| Query | Document contains | Precision | Recall | Query | Document contains | Precision | Recall | Query | Document contains | Precision | Recall | Query | Document contains | Precision | Recall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| أكله |  | 0 | 0 | تبتحث |  | 0 | 0 | جاده |  | 0 | 0 | الدروب | الدروب | 1 | 0.05 |
| أكل | أكل | 1 | 0.05 | البحثه |  | 0 | 0 | جاد | جاد | 1 | 0.05 | داربه |  | 0 | 0 |
| مستأكله |  | 0 | 0 | البحث | البحث | 1 | 0.05 | مجدده |  | 0 | 0 | دارب | دارب | 1 | 0.05 |
| مستأكل | مستأكل | 1 | 0.05 | الباحثه |  | 0 | 0 | مجدد | مجدد | 1 | 0.025 | متدربه |  | 0 | 0 |
| أكلته |  | 0 | 0 | الباحث | الباحث | 1 | 0.05 | الأجداده |  | 0 | 0 | متدرب | متدرب | 1 | 0.05 |
| أكلت | أكلت | 1 | 0.05 | البحوثه |  | 0 | 0 | الأجداد | الأجداد | 1 | 0.05 | عجبه |  | 0 | 0 |
| أكوله |  | 0 | 0 | البحوث | البحوث | 1 | 0.05 | الأجداته |  | 0 | 0 | عجب | عجب | 1 | 0.05 |
| أكول | أكول | 1 | 0.05 | بحثته |  | 0 | 0 | الأجدان | الأجدان | 1 | 0.05 | العجبه |  | 0 | 0 |
| أوكله |  | 0 | 0 | بحثت | بحثت | 1 | 0.05 | المجدده |  | 0 | 0 | العجب | العجب | 1 | 0.05 |
| أوكل | أوكل | 1 | 0.05 | تبحثه |  | 0 | 0 | المجدد | المجدد | 1 | 0.05 | العجيبه |  | 0 | 0 |
| الأكله |  | 0 | 0 | بحثه | بحثه | 1 | 0.05 | دربه |  | 0 | 0 | العجيب | العجيب | 1 | 0.05 |
| الأكل | الأكل | 1 | 0.05 | مباحثه |  | 0 | 0 | درب | درب | 1 | 0.05 | أعاجيبه |  | 0 | 0 |
| المأكله |  | 0 | 0 | مباحث | مباحث | 1 | 0.05 | الداربه |  | 0 | 0 | أعاجيب | أعاجيب | 1 | 0.05 |
| المأكل | المأكل | 1 | 0.05 | تبحثه |  | 0 | 0 | الدارب | الدارب | 1 | 0.05 | أعجابه |  | 0 | 0 |
| المأكوله |  | 0 | 0 | تبحث | تبحث | 1 | 0.05 | المدربه |  | 0 | 0 | أعجاب | أعجاب | 1 | 0.05 |
| المأكول | المأكول | 1 | 0.05 | جدده |  | 0 | 0 | المدرب | المدرب | 1 | 0.05 | أعجبه |  | 0 | 0 |
| مأكوله |  | 0 | 0 | جدد | جدد | 1 | 0.05 | تدربه |  | 0 | 0 | أعجب | أعجب | 1 | 0.05 |
| مأكول | مأكول | 1 | 0.05 | أجده |  | 0 | 0 | تدرب | تدرب | 1 | 0.05 | أعجوبه |  | 0 | 0 |
| تأكله |  | 0 | 0 | أجد | أجد | 1 | 0.025 | أدربه |  | 0 | 0 | أعجوب | أعجوب | 1 | 0.05 |
| تأكل | تأكل | 1 | 0.05 | الجداده |  | 0 | 0 | أدرب | أدرب | 1 | 0.05 | الأعجا |  | 0 | 0 |
| بحثه |  | 0 | 0 | الجداد | الجداد | 1 | 0.05 | التدريبه |  | 0 | 0 | الأعجا | الأعجاب | 1 | 0.05 |
| بحث | بحث | 1 | 0.025 | الجده |  | 0 | 0 | التدريب | التدريب | 1 | 0.05 | التعجبه |  | 0 | 0 |
| البحوثه |  | 0 | 0 | الجد | الجد | 1 | 0.05 | الدربه |  | 0 | 0 | التعجب | التعجب | 1 | 0.05 |
| البحوث | البحوث | 1 | 0.05 | الجدده |  | 0 | 0 | الدرب | الدرب | 1 | 0.05 | العجائبه |  | 0 | 0 |
| تبتحثه |  | 0 | 0 | الجدد | الجدد | 1 | 0.05 | الدروبه |  | 0 | 0 | العجائب | العجائب | 1 | 0.05 |
| All | - | 49% | 2.37% |  |  |  |  |  |  |  |  |  |  |  |  |

TABLE III: EXPERIMENTS RESULTS OF THE SEARCH ENGINES: WINDOWS SEARCH, X1, COPERNIC SEARCH AND AOL SEARCH.

| Query | Document | Precision | Recall | Query | Document | Precision | Recall | Query | Document | Precision | Recall |
|---|---|---|---|---|---|---|---|---|---|---|---|
| أكله | أكلهما | 1 | 0.2 | تبحثه | بحثههما | 1 | 0.2 | تدرب | تدرب, تدربكما, تدربهما, تدربكن | 1 | 0.05 |
| أكل | أكلكما, أكلهما, أكلكن, أكل | 1 | 0.1 | بحثه | بحثهكما, بحثههما, بحثهكن, بحثه | 1 | 0.05 | أدربه | أدربهما | 1 | 0.2 |
| مستأكله | مستأكلهما | 1 | 0.2 | مباحثه | مباحثهما | 1 | 0.2 | أدرب | أدرب, أدربكما, أدربهما, أدربكن | 1 | 0.05 |
| مستأكل | مستأكلكما, مستأكلهما, مستأكلكن, مستأكل | 1 | 0.05 | مباحث | مباحثكما, مباحثهما, مباحثكن, مباحث | 1 | 0.05 | التدريبه | التدريبهما | 1 | 0.2 |
| أكلته | أكلتهما | 1 | 0.2 | تبحثه | تبحثهما | 1 | 0.2 | التدريب | التدريب, التدريبكما, التدريبهما, التدريبكن | 1 | 0.05 |
| أكلت | أكلتكما, أكلتهما, أكلتكن, أكلت | 1 | 0.05 | تبحث | تبحثكما, تبحثهما, تبحثكن, تبحث | 1 | 0.05 | الدربه | الدربهما | 1 | 0.2 |
| أكوله | أكولهما | 1 | 0.2 | جدده | جددهما | 1 | 0.2 | الدرب | الدرب, الدربكما, الدربهما, الدربكن | 1 | 0.05 |
| أكول | أكولكما, أكولهما, أكولكن, أكول | 1 | 0.05 | جدد | جددكما, جددهما, جددكن, جدد | 1 | 0.05 | الدروبه | الدروبهما | 1 | 0.2 |
| أوكله | أوكلهما | 1 | 0.2 | أجده | أجدهما | 1 | 0.2 | الدروب | الدروب, الدروبكما, الدروبهما, الدروبكن | 1 | 0.05 |
| أوكل | أوكلكما, أوكلهما, أوكلكن, أوكل | 1 | 0.05 | أجد | أجدكما, أجدهما, أجدكن, أجد | 1 | 0.025 | داربه | داربهما | 1 | 0.2 |
| الأكله | الأكلهما | 1 | 0.2 | الجداده | الجدادهما | 1 | 0.2 | دارب | دارب, داربكما, داربهما, داربكن | 1 | 0.05 |
| الأكل | الأكلكما, الأكلهما, الأكلكن, الأكل | 1 | 0.05 | الجداد | الجدادكما, الجدادهما, الجدادكن, الجداد | 1 | 0.05 | متدربه | متدربهما | 1 | 0.2 |
| المأكله | المأكلهما | 1 | 0.2 | الجده | الجدهما | 1 | 0.2 | متدرب | متدرب, متدربكما, متدربهما, متدربكن | 1 | 0.05 |
| المأكل | المأكلكما, المأكلهما, المأكلكن, المأكل | 1 | 0.05 | الجد | الجدكما, الجدهما, الجدكن, الجد | 1 | 0.05 | عجبه | عجبهما | 1 | 0.2 |
| المأكوله | المأكولهما | 1 | 0.2 | الجدده | الجددهما | 1 | 0.2 | عجب | عجب, عجبكما, عجبهما, عجبكن | 1 | 0.05 |
| المأكول | المأكولكما, المأكولهما, المأكولكن, المأكول | 1 | 0.05 | الجدد | الجددكما, الجددهما, الجددكن, الجدد | 1 | 0.05 | العجبه | العجبهما | 1 | 0.2 |
| مأكوله | مأكولهما | 1 | 0.2 | جاده | جادهما | 1 | 0.2 | العجب | العجب, العجبكما, العجبهما, العجبكن | 1 | 0.05 |
| مأكول | مأكولكما, مأكولهما, مأكولكن, مأكول | 1 | 0.05 | جاد | جاد, جادكما, جادهما, جادكن | 1 | 0.05 | العجيبه | العجيبهما | 1 | 0.2 |
| تأكله | تأكلهما | 1 | 0.2 | مجدده | مجددهما | 1 | 0.1 | العجيب | العجيب, العجيبكما, العجيبهما, العجيبكن | 1 | 0.05 |
| تأكل | تأكلكما, تأكلهما, تأكلكن, تأكل | 1 | 0.05 | مجدد | مجدد, مجددكما, مجددهما, مجددكن | 1 | 0.025 | أعاجيبه | أعاجيبهما | 1 | 0.2 |
| بحثه | بحثهما | 1 | 0.2 | الأجداده | الأجدادهما | 1 | 0.2 | أعاجيب | أعاجيب, أعاجيبكما, أعاجيبهما, أعاجيبكن | 1 | 0.05 |
| بحث | بحث, بحثكما, بحثهما, بحثكن, بحث ,<br>بحثتكما, بحثتهما, بحثتكن | 1 | 0.025 | الأجداد | الأجداد, الأجدادكما, الأجدادهما, الأجدادكن | 1 | 0.05 | أعجابه | أعجابهما | 1 | 0.2 |
| البحوثه | البحوثهما | 1 | 0.2 | الأجداته | الأجداتهما | 1 | 0.2 | أعجاب | أعجاب, أعجابكما, أعجابهما, أعجابكن | 1 | 0.05 |
| البحوث | البحوثكما, البحوثهما, البحوثكن, البحوث | 1 | 0.05 | الأجدان | الأجدان, الأجداتكما, الأجداتهما, الأجداتكن | 1 | 0.05 | أعجبه | أعجبهما | 1 | 0.2 |
| تبتحثه | تبتحثهما | 1 | 0.2 | المجدده | المجددهما | 1 | 0.2 | أعجب | أعجب, أعجبكما, أعجبهما, أعجبكن | 1 | 0.05 |
| تبتحث | تبتحثكما, تبتحثهما, تبتحثكن, تبتحث | 1 | 0.05 | المجدد | المجدد, المجددكما, المجددهما, المجددكن | 1 | 0.05 | أعجوبه | أعجوبهما | 1 | 0.2 |
| البحثه | البحثهما | 1 | 0.2 | دربه | دربهما | 1 | 0.2 | أعجوب | أعجوب, أعجوبكما, أعجوبهما, أعجوبكن | 1 | 0.05 |
| البحث | البحثكما, البحثهما, البحثكن, البحث | 1 | 0.05 | درب | درب, دربكما, دربهما, دربكن | 1 | 0.05 | الأعجابه | الأعجابهما | 1 | 0.2 |
| الباحثه | الباحثهما | 1 | 0.2 | الداربه | الداربهما | 1 | 0.2 | الأعجاب | الأعجاب, الأعجابكما, الأعجابهما, الأعجابكن | 1 | 0.05 |
| الباحث | الباحثكما, الباحثهما, الباحثكن, الباحث | 1 | 0.05 | الدارب | الدارب, الداربكما, الداربهما, الداربكن | 1 | 0.05 | التعجبه | التعجبهما | 1 | 0.2 |
| البحوثه | البحوثهما | 1 | 0.2 | المدربه | المدربهما | 1 | 0.2 | التعجب | التعجب, التعجبكما, التعجبهما, التعجبكن | 1 | 0.05 |
| البحوث | البحوثكما, البحوثهما, البحوثكن, البحوث | 1 | 0.05 | المدرب | المدرب, المدربكما, المدربهما, المدربكن | 1 | 0.05 | العجائبه | العجائبهما | 1 | 0.2 |
| بحثته | بحثتهما | 1 | 0.2 | تدربه | تدربهما | 1 | 0.2 | العجائب | العجائب, العجائبكما, العجائبهما, العجائبكن | 1 | 0.05 |
| بحثت | بحثتكما, بحثتهما, بحثتكن, بحثت | 1 | 0.05 |  |  |  |  |  |  |  |  |
| All | - | 100% | 12.33% |  |  |  |  |  |  |  |  |

However, if we take the experiments that we performed by changing the query to add the words of the same group in the Table I to the keyword, we reach the value of 100% for both precision and recall. This is achieved through an application that sits between the user query and the Google search engine. This application takes the user's query and tries to find all the words in the same group that are relation to its keyword and add them to the query. We obtain the results of all the words in the same group having a relationship with the initial keyword in the query. These words will be added to the application and then launched into a single application and the documents associated with each of these words are given as a result (Fig. 3).

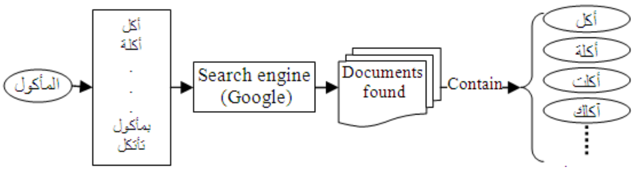


Fig. 3. The documents found by the engine Google in the case of the query "المأكول" and its dependent words.

Similarly, the Table III has the results of the engines: Windows Search, X1, Copernic Search, AOL Search in their local versions. This table has the same structure as the Table II. The first observation is that these four engines gave the same results for the hundred queries made in the conditions of our experiments. These requests are always formed a word. The second column of the Table III shows that these search engines can find in most cases, in addition to documents containing exactly the same word in the query, documents that contain a derived word or the same group. On the other hand, these engines are unable to find the words prefixed and this whatever the search term. The recall, by word, varies between 5% and 20% for a precision of 100%, because every queries word of the experimentation are postfixed words. In addition, the average of precision of these engines is 100% and the average recall is about 12%. These values could have been much worse if more of prefixed words are used in the queries of the experiment. In all the cases, these results are better than those given by Google. It also appears that morphological analysis is made, in part, included in this search engine. We also evaluated the engine "Ask Jeeves Desktop Search" in the same way as the previous engines. This engine is unable to answer any queries made under the conditions of our experiments and this whatever the search term.

## V. CONCLUSION

In this paper, we presented the performance of the search engines of the most widely used between 2006 and 2010, Google, Yahoo, Copernic, Ask, Aol Search and MSN / Live to extract relevant information from the Arabic documents. We chose a search engine that can run on a local computer. To achieve our experiments, we have built a corpus of one thousand Arabic documents that contain only words derived from roots <أكل, akl, he ate>, <بحث, bhhth, he searched> <جدد, jdd, he renewed>, <درب, drb, he coached> <عجب, ajb, he admired>. The results showed that the search engine Google, in its local version, can extract only those documents that contain exactly the query word. The search engines: Windows Search, X1, Copernic Search, AOL Search in their local versions gave the same results under the conditions of our experiments. These search engines can find in most cases, in addition to documents containing exactly the same word in the query, the documents that contain a derived word or same group. In all cases, these results are better than those given by Google. According to what precedes, the problems of these search engines on the level of the information extraction from Arabic documents appear clearly. It seems that the specific treatments for the Arabic language, particularly the morphological analysis, are not taken into account in these search engines.

## ACKNOWLEDGMENT

This work has been done as a part of the following projects: "Automatic information extraction form Arabic texts" supported by the Lebanese National Council for Scientific Research CNRSL, "Arabic speech synthesis from text, with natural prosody, using linguistic and semantic analysis" supported by the Lebanese-Syrian Scientific Research Cooperation Program, and “Extraction et fusion de l’information à partir du Web par agrégation des services et de l’analyse du contenu” supported by Lebanese university.

## REFERENCES

[1] I. Al Kharashi. (1999). A Web search engine for indexing, searching and publishing Arabic bibliographic databases. [Online]. Available: http://www.isoc.org /inet99/proceedings/posters/085/index.htm

[2] Wikipedia moteur de recherche (WWMR). (2010). [Online]. Available: http://fr.wikipedia. org/ wiki/ Moteur_de_recherche

[3] J. A. Kharashi and M. Evens, “Comparing words, stems, and roots as index terms in an Arabic Information Retrieval system,” *Journal of the American Society for Information Science*, pp. 548–560, 1994.

[4] L. Larkey, L. Ballesteros, and M. Connell, “Light stemming for Arabic IR, Arabic computational morphology: Knowledge-based and empirical methods,” in *Kluwer/Springer's Series on Text, Speech, and Language Technology*, A. Soudi, A. Van Bosch, and G. Neumann Eds. 2005.

[5] A. El-Halees, “Arabic text classification using maximum entropy,” *The Islamic University Journal (Series of Natural Studies and Engineering)*, pp. 157-167, 2007.

[6] A. M. Attia, “An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks,” *The Challenge of Arabic for NLP/MT*, pp. 48-67, 2007

[7] R. A. Shalabi, and N. Evens, “A Computational Morphology System for Arabic,” in *Proc. COLING-ACL*, New Brunswick, NJ, 1998, pp. 66-72.

[8] J. Dichy, and A. Farghaly, “Roots & patterns vs. stems plus grammar-lexis specifications: On what basis should a multilingual database centered on Arabic be built?" in *Proc.* MT Summit IX -- workshop: Machine Translation for Semitic Languages, New Orleans, USA, 2003.

[9] I. A. Sughaiyer and I. Al-Kharashi, “Arabic Morphological Analysis Techniques: A Comprehensive Survey,” *Journal of the American Society for Information Science and Technology*, pp. 189-213, 2004.

[10] I. Zitouni, J. Sorensen, X. Luo, and R. Florian, “The impact of morphological stemming on Arabic mention detection and coreference resolution,” in *Proc. the ACL Workshop on Computational Approaches to Semitic Languages*, 2005, pp. 63-70.

[11] N. Habash and O. Rambow, “MAGEAD: A morphological analyzer and generator for the Arabic dialects,” in *Proc. the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 2006, pp. 681-688.

[12] R. Sonbol, N. Ghneim, and M. S. Desouki, “Arabic Morphological Analysis: a New Approach,” *Information and Communication Technologies: From Theory to Applications, ICTTA*, 2008.

[13] R. Ryan, O. Rambow, N. Habash, M. Diab, and C. Rudin, “Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking,” Association for Computational Linguistics (ACL), Columbus, Ohio, 2008.

[14] SEO Search Consultants Directory. (2010). [Online]. Available: http://www. seoconsultants.com /search-engines/

[15] E. Paul *et al.*, “Inverted Index, dictionary of algorithms and data structures,” U.S. National Institute of Standards and Technology Oct 2006.

[16] Wikipedia WebCrawler (WWC). (2010). [Online]. Available: http://en.wikipedia.org/ wiki/Web_crawler

[17] Abacus Référencement (AR). (2010). [Online]. Available: http://www.abacus-referencement.com/lexique/indexeur.htm

[18] ALJAZEERA.NET. (2010). [Online]. Available: http://www.aljazeera.net

[19] Al-Mustaqbal. (2010). [Online]. Available: http:// website: www.almustaqbal.com

[20] AlSafir. (2010). [Online]. Available: http://www.assafir.com

[21] Al Nahar. (2010). [Online]. Available: http://www.annahar.com

[22] A. A. Hajjar, M. Hajjar, and K. Zreik, “Classification of Arabic Information Extraction methods,” in *Proc. 2nd International Conf. on Arabic Language Resources and Tools Cairo (Egypt)*, pp. 22–23, 2009.

**Abd El Salam Al Hajjar** was born in Lebanon. He works as an instructor at the Lebanese University, University Institute of technology, Sidon, Lebanon. He has a B.S. and Technical leader at the oger system company, Lebanon Branch. He has a B.A. in applied Mathematics, Computer Science from the Lebanese University – Faculty of Sciences, and Masters in Computer Science "Cooperation in sciences of information treatment" from the Lebanese university and Paul Sabatier University (IRIT France), and a Ph.D. in Computer Science from Paris8 University, France. His main research in the Arabic information extraction and processing.

**Mohammad Hajjar** is a professor at University Institute of Technology, Lebanese University, Lebanon. He received a Ph.D. in computer Science at Nantes University in France. His Interest domain concerns Arabic language processing, multimedia information research and data management in peer-to- peer systems.

**George Lebbos** was born in Lebanon. He has a Software Engineering and master’s degree from the CNAM (Conservatoire National des Arts et Métiers), Paris. He has thirteen years of experience in IT and software development and architecture; He is IT Manager at an insurance company and a teacher at the Lebanese University, University Institute of technology, Sidon, Lebanon; His research interests are especially in Root Extraction Methods and Semantic Methods in Arabic language and building a complete semantic corpus.

**Khaldoun Zreik** was a full professor at Hypermedia Department, University Paris 8 since September 1, 2006 and full professor at Computer Sciences Department, University of Caen, France from September 1, 1993 to August 31, 2006. He received a master degree in Informatics. Option: Artificial Intelligence (DEA Traitement Algorithmique de l'Information), University of Paris VI, Paris – France, on September 1985, and a Ph.D. degree at Ecole Nationale des Ponts et Chaussées, Paris - France on November 24, 1986. His main research areas are: hyperurban, communicative document design (C2D), computer and media arts (CMA), structure driven information extraction (Text Mining Based Approach), interactive learning support design (IHSL / Machine Learning Based Approach) and Arabic language processing.

**Mazen El-Sayed** was born in Lebanon. He works as an assistance professor and the head of Applied Bussiness Computer Department at the Lebanese University, University Institute of technology, Sidon, Lebanon. He has an engineer degree in computer science from the Lebanese University (LU), an M.S. in computer science from the Central School of Engineering (ECN), University of Nantes, France, and a Ph.D. in computer science from the Anger University, France.