Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.8, No.2, 2018

Development of Arabic Information Retrieval Systems in the 21st Century

Awatif Elmekawi
Assist. Professor of Library and Information Sciences, College of Arts, Imam Abdulrahman Bin Faisal University

Abstract
The present study deals with the development of Arabic Information Retrieval Systems starting from 2000, its vital role in the Text Retrieval Conference (TREC), and in the cross-language information retrieval track. It has overviewed the developments concerning the Holy Qur'an, Arabic language, terms relevant to Arabic information retrieval systems, and the characteristics of the Arabic language compared with other languages since the early 21st century. These developments include rich resources of up to date information so as to develop research in this area, modern developments in assessing and measuring Arabic information retrieval systems, relevant theses, and some research studies of contemporary universities on the use of TREC in Arabic information retrieval, and the researchers with no prior knowledge of Arabic language. The study ends with some studies of the Arab universities.
Keywords: Retrieval Systems, Arabic Information**,** Twenty- first century

Introduction

The Intellectual production in the field indicates that most information retrieval research focused on English language for more than 50 years. The lack of research in Information Retrieval in Arabic is partially due to the distinctive structure of the Arabic language, which differs from the other Latin-based languages, and to the small budgets allocated for Arabic Information Retrieval research.

Bamaflah (2000), Abu El-khair et al. (2013) and Qassem (1978) generally overviewed the beginning of interest in Arabic Information Retrieval, where the main challenge was how Arabic language is represented on the engines. This problem was solved within few years later. Ali (1988) referred to the study of Suwina'a (1994) on Arabic information linguistics and the study of Al-Kharashi (1994) on the use of word roots for indexing terms.

Ghoneim (2003) dealt with retrieval through the points of subject availability such as classification and thematic headings besides some contributions of Al-Atram (1990) and Al-Tayar (1998) and others.

Provided that the research on Arabic Information Retrieval (AIR) started from (1989) until the end of the twentieth century through small experiments and small testing groups, most of these researches focused on root extraction and comparing their effectiveness in indexing the Arabic text using roots and Stems ,or words. The twentieth century started with Text Retrieval Conference studies (TREC) which focused on Monolingual retrieval and Cross-Languages. The amount of research expanded in various foreign universities and some Arab universities. Thus, the focus of this study was on the twenty-first century

Problem and Questions of the Study

The problem of the study focuses on the attempt to identify the development of research and studies in the field of Arabic information retrieval, the characteristics of the Arabic language, and the extent of its entry into the field of Arabic information retrieval. The problem can be identified in the following questions:

- What are the characteristics of the Arabic language compared with other languages and the extent of adaptation of machines and search engines to suit them?
- What is the form of the words and their transformation from the roots of the language?
- What is the extent to which the use of the Internet in Arabic is growing and what are the sources of the new information?
- What is the extent of the recent developments in the evaluation and measurement of Arabic information retrieval systems?
- What are the most important theses dealing with the development of Arabic information retrieval systems from (2000) to (2014)?
- What are the research papers of the contemporary universities and their role in developing research tools?

Methodology and Tools of the Study

This study follows the analytical descriptive approach accompanied by the extrapolation of the intellectual production relevant to the Arabic Information Retrieval systems, benefiting from the search engines especially Google and the publications of the Arab intellectual production of Mohamed Fathi Abdel Hadi.

Terminology of the Study

- Morphology: is the branch of linguistics that studies word structure.
- Morpheme : is the smallest grammatical unit in function and meaning.
- Affixes: The parts added to the Root or Stem (*beginning*, *in*, or *end* of the word)
- Derivation : A branch of Morphology deals with word derivation in a dictionary or
- Lexicon from another word such as Keeper from keep, and Hopeless from Hope.
- Prefixes: The small units or letters that are placed before the Root or the Stem.
- Root: The part that can be analyzed into smaller parts.
- Stem: The part of the root that is combined to smaller ones (morphemes) at the beginning, inside, or at the end of the stem and it is the part of the word left after stripping the affixes.
- Suffixes: The small units added to the root or the stem.
- Inflection: It's the part of Morphology that is relevant to number, sound, or verb.
- Stemming: it is the word -reduction technique to its original grammatical roots.
- Some researchers (Abu El-Khair, 2003) see that stemming has four ways such as affix removal ,word segmentation ,table look up, or (N-grams) as follows:

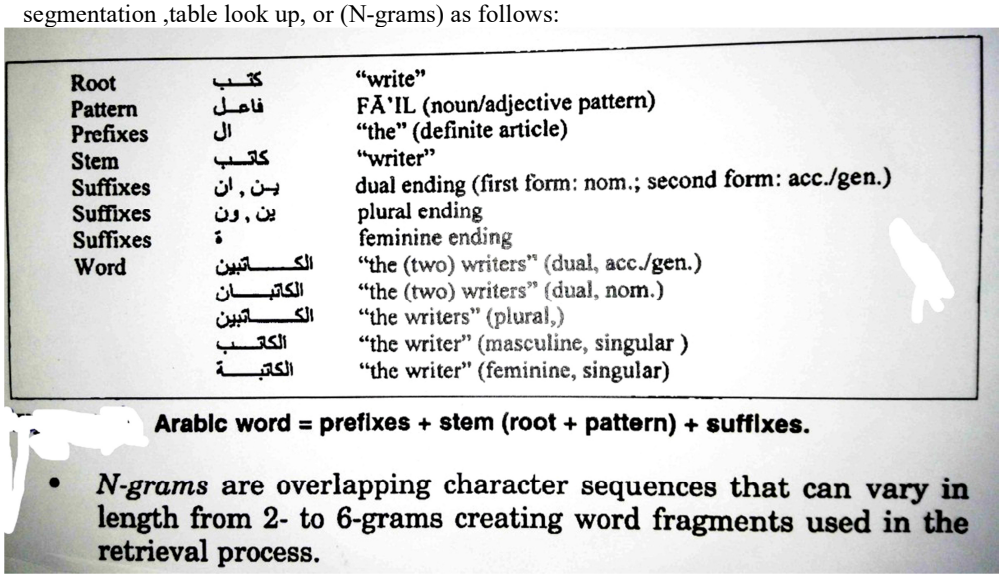


Figure 1

While others see that Arab Stemmers range between the Morphological deep analyzer to the Light Prefix-Suffix.

Some of the Arabic language characteristics and comparison with other languages

There are three forms of the Arabic language, Classical, Modern Standard and Colloquial. The first and second forms are considered as Classical language which is more rhetoric and is the language of the Koran. The modern standard Arabic language is derived from the classical language and most of the books printed in Arabic and newspapers are written in this form, while the colloquial (spoken) language is different from one Arab country to the other (Al-Dayel, 2013).

The Arabic language is characterized by writing from right to left and has eight diacritic marks which can totally change the meaning of the word relying on its place in the word. For example, the word "K**a**t**a**ba" means "he wrote" and "K**u**t**u**b" means books. The Arabic Alphabets consist of 28 letters, but each letter can have different forms and may reach four forms in writing: separate, at the beginning of the word, in the middle, or at the end, depending on the place of these letters in the word, and every letter in the Arabic letters is pronounced as a word. For example, letter (ا) is pronounced (Alif) and letter (ل) is pronounced (Lam), which means that the separate sound can't be used to refer to the Alphabets. Consequently, the abbreviations are not used in Arabic as in English (Moukdad, 2004).

The Arabic language possesses the potentials and the accurate precision and its sounds are musical weights (rhythms). It is the only living language in the world that has remained unchanged in its words, grammar, and its structures for about fifteen centuries, as well as its control of the rhetorical styles and the so many semantic meanings. Finally, most of its derivatives accept inflection to meet the needs of its users.

The Arabic Language and Information Retrieval

In spite of the continuous increase in Internet users who speak Arabic, the search engines in Arabic are still lagging behind.

The process of information retrieval is affected by language and how the search engines process the characteristics of this language (Moukdad, 2004). Arabic has about five million words derived from about 11.350 roots compared to English which has 1.3 million words, of which there are 400.000 keywords .Thus, Arabic has more words than English because Arabic contains hundreds of derivations from one root and there is a distinct characteristic of the Arabic language which is that words have multiple meanings, and this represents many challenges to the retrieval process. There is another problem relevant to the Arabic language on the online public access catalog in different regions in the Arab world which is that different words are used for the same thing. In some countries, the term "newspaper" is translated as a “Sahifa” and in the others as "Jarida". This results in an irrelevant response to the question. So, the hypothesis that the research would be more precise and more specific if it is based on meanings rather than words began to exist.

Though English occupies a central position in information retrieval research (the content is up to 67%) on the Internet, the online Arabic content is ranked 10th (content is about 0.01% on the Internet) and the content in other languages is about 32% on the same network (Nabil, 2003).

The lack of such Arab research in the field of information retrieval is due to the characteristics of the Arabic language, where it is usually said to be a derivative language, while the English language is an agglutinative language, as well as the lack of budgets and standard evaluation resources for the practitioners in the field of Arabic information retrieval. This study deals with the characteristics of Arabic language and research trends and their development especially in the 21st century with the introduction of TREC in research of the field in 2000.

The form of the words and their transformation from the roots of the language

Arabic is considered one of the five languages of the United Nations. It is the mother tongue spoken by about 300 million people, and it is the language of the Holy Quran, so, it is the second language of more than a billion and half Muslims around the world.

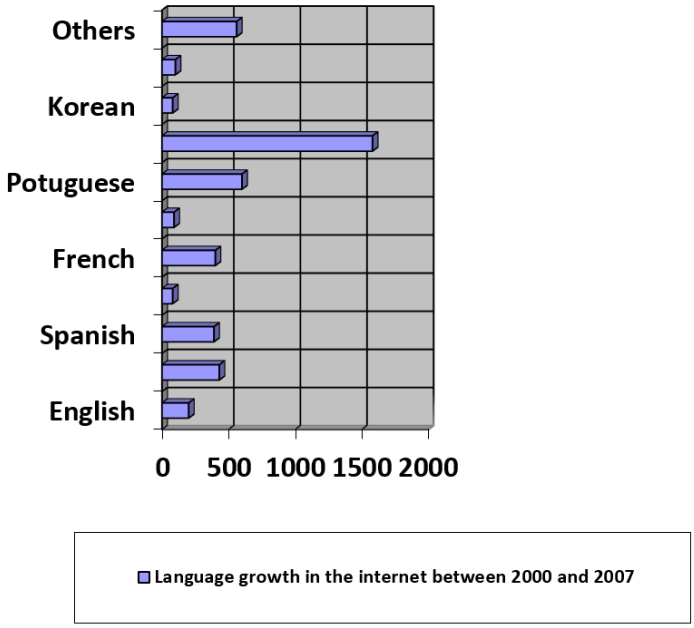
Retrieving information is the basic technology of the search engines on the web and the web has become a major source of information, dealing with billions of documents available for search that increase every day. According to April 2008 statistics, there are 165,000,000 fields of names on the Internet.

http://news.netcraft.com/archives/2008/04/14/April2008_web_server_survey.html

As for the internet users in the Middle East, their number increased by 920% from 2000 to 2007 while the number of people using the Internet in Arabic reached 46,359,140 in November 2007, with an increase of 1576% over their number in 2000.

http://www.internetworldstats.com.stats7.htm

The following graph shows this increase



Language growth on the internet between 2000 and 2007
Source: Internet World Stats [Miniwatts International, 2007], (Abdul Salam & Nwesri, 2008)

Table (1) Some derived forms from the root, Kataba (كتب)

Word form	Transliteration	Translation
كَتَبَ	Kataba	He wrote
يَكْتُبُ	Yaktubu	He writes
تَكْتُبُ	Taktubu	She writes
مَكْتُوبٌ	Maktūb	Written/Letter
كَاتِبٌ	Kātib	Writer
كَتَبَةٌ	Katabah	Writers
كُتَّاب	Kuttāb	Writers
كَاتَبَ	Kātaba	Correspond
تَكَاتَبَ	Takātaba	Write to
مَكْتَبٌ	Maktab	Office/Desk
مَكْتَبَةٌ	Maktabah	Library
مُكَيْتِبٌ	Mukaytib	Small office/Desk
كِتَابٌ	Kitāb	Book
كُتُبٌ	Kutub	Books
كِتَابَةٌ	Kitābah	Writing
كُتِبَ	Kutiba	Written
إِنْكَتَبَ	Inkataba	Became written
إِسْتَكْتَبَ	Istaktaba	Dictate
سَيَكْتُبُ	Sayaktubu	Will write

Abu El-Khair, 2006

Third: Sources of new information

TREC Text Retrieval Conference

This conference is considered the first evaluation initiative on large-scale in 2008 in the seventeenth year and is funded by the National Institute of Standards and Technology NIST in Maryland, U.S.A. to begin a new era in information retrieval research (Voorhees and Buckland, 2006) for the first time in information science, with the following proceedings:

- Encouraging research in information retrieval based on large test kits.
- Increasing communication of industry, universities and government agencies by opening the exchange in the research ideas.
- Accelerating the transfer of technology from research laboratories to commercial products concerning real problems.
- Increasing the possibility of obtaining the techniques used in industry and academic bodies including the evaluation methods.

Here, we should refer to the year (2001) when the (TREC), for the first time ,implemented a track for Arabic Retrieval and more precisely CLIR track to test using English and French questions concerning the Arabic documents and the Arabic retrieval questions.

The retrieval experiments were carried out with stems and roots, using automatic translation systems, and with words with their transliterations (Gey & Oand 2002).

Trek has recently organized his evaluation in 26 tracks .What follows are some of them:

- The QA track which requires systems that focuses on short answers to fact-based questions in different specializations.
- The most common track the researchers participated in was in 2005.It was in the field of genomics, which combines textual scientific documents with data and facts (Hersh, 2006). The conference

included the bio-informatics community as well as information retrieval specialists and finished in 2007.

- In the Robust retrieval track, new evaluation measures are used that focus on firm performance on all subjects rather than rewarding the systems that have provided a good average of Precision.
- The Blog Track, which began in 2006 to explore the behavior of information in large groups of computerized social data.

Some results for the new information resources (TREC)

Experiments Arabic language retrieval is still in their beginning, but researchers in Arabic information retrieval can benefit greatly from the English experience. The Arab researchers can replicate what has been done on the foreign side in many experiments.

It can be observed that the experiments carried out before TREC experiments showed that the roots are better in retrieval than all words or stems (Abu El-Khair et al., 2013). This may be due to the small size of the newly created Corpora group, as the search by using words in this small group indicates the difficulty of finding Matching due to the complexity of Arabic morphology.

The importance of stemming arose for the Arabic information retrieval. There was no standard stemming algorithm at the beginning. Ultimately, a new algorithm can deal with the broken plurals in Arabic and can be of great benefit for retrieval as mentioned in the conclusion of Abu Al-Khair et al. (2013) in English language:

Arabic broke plurals are similar to the irregular noun plural forms in English e.g., man (sg.), men (pl.): (sg.), (pl.)": They do not follow a specific pattern. Arabic has many such forms and an efficient algorithm to reduce them to their singular form would improve retrieval results significantly.

Some challenges in the Arab information retrieval system

As seen by Ashkar (2002):

- It has a lot of Vocalization
- The use of diacritic marks can give different meanings to words.

Almعلم	Ambiguous	غامض Ghamid
Alamعَلَم	Flag	علم Alam
Elmعِلم	Science	عِلمElm
Alamaعَلَمَ	Taught	علّم Alama

Regular Nouns	
(Moa'lim معلّم)(teacher, masculine)	Moa'lmin معلّمين (teacher, masculine)
مدرّبة (Modariba)(trainer, feminine)	Modribatمدرّبات (trainer, feminine)

Irregular Nouns	
Tifl طفل (child)	Atfal أطفال (child)
امرأة Imra'a (woman)	Nisaa'نساء (woman)

Fourth: Some Recent Developments in evaluating and measuring Arabic information retrieval Systems

Three of the researchers in the field, Abdel Ali, Comy and Soliman (2004) examined in depth the perspectives of the Arab information retrieval. The three authors dealt with some of the available resources for testing the Arab information retrieval systems where there are evaluation criteria since 2000 by CLEF and TREC.

Arabic Information Retrieval Tools

These tools can be divided into two categories

a) Full Form based-IR

Most of the commercial search engines are of this kind such as Sakhr web search engine www.alidrisi.com ، www.ayna.com and other multilingual UNI code search engines such as: www.google.comor www.alltheweb.com

b. Morphology based-IR

Efforts in the academic environment to evaluate more complex systems shed light on the next generation of Arab search engines, with several approaches including morphology, stem and roots the depend on light stemming (larkey et al., 2002) and also there are non-statistical stemmers or N-gram models in general. It can be said that the use of stemmers improves both Recall and Precision. Larkey's experiments (Larkey et al., 2002) showed the light stemmers and normal stemmers.

Large groups, Corpora

These large groups were found in the information retrieval tests with the introduction of TREC. There is LCD

group (869 megabytes) of Arabic news articles, which include more than 383,872 documents from the Agency of French Press (AFP) which are used in the evaluation of TREC and lexicons.

Evaluation Measures and Tools:

The Arabic language has a high degree of Inflection, because the Arabic language morphology is a very challenging science concerning information retrieval. This complexity may be due to the lack of spaces between words, as well as the overlap in some letters and words. For example, the letter "Alif" can be "Aa" or "Ee".also. The Prefixes and Suffixes the diacritics marks can be two, three, or even four such as Lio'alimanikaha (ليعلمانكها) to teach you, Ta'lamonihin (تعلمونهن) you teach them (female),or Saio'limaniha (سيعلمانها) they will teach it to you.. you teach them. To solve this problem, there are two main trends used to build the morphological analyzers, the table that depends on the morphological analyzers and grammar (Khoja, 2001), or the morphological analyzer of Beesley (1998). By analyzing the contemporary information retrieval with TREC analysis to evaluate the Arabic language, it can be said that performance of the documents retrieval systems from the Arabic group is similar to the performance of the contemporary systems in the other languages. The TREC evaluation uses LDC for a group of AFP articles from 1994 to 2000and the language used was the Modern Standard Arabic (MSA). Ziph's law proved that this group is adequate, complete and representative. If this group is small for syntax and morphology besides spelling of the translated sentences which were written in other letters, it will be irregular in Arabic language.

In conclusion, it should be admitted that there are many problems to be solved, including what criteria the beneficiary adopts when judging the validity of the documents for the Judge Reference question because these criteria may affect the ranking strategy and interfaces between the user and the system Interfaces. The problem with the web is clear and unique because the web user often does not know what he really wants and has difficulty in formulating his questions and how this affects the ranking with the small size of indexes maintained by the search engines. In Baeza-Yates & Ribeiro (1999) see that these indexes reach only 2% of the total number of pages and proposes a solution which is search engines about search engines (Meta search engine)

Fifth: Some theses in the field of Arab information retrieval from the beginning of the twenty-first century until 2013

This period is characterized by the use morphological analysis with semantic methods.

In his PhD thesis from De Montfort University (2000), Al-Tayar introduced

Stemming logarithm and used Morpho-Semantic Method which depends on the semantic links of the morphological Forms AIRSMA. In his abstract, Al-Tayar proposed some additional pieces of information as follows: Arabic Text Retrieval Systems are considered a new phenomenon. These systems introduced topics such as Truncation and Stemming. Arab retrieval systems use three search methods: word, stem, and root. The word in matching is used as a term. The other two terms are included in the morphological analysis. It is noted that each method has drawbacks. The word and Stem may miss the relevant records for morphological expressions and the root can retrieve irrelevant recordings. Thus, Al-Tayar (2000) used a new method which is the Morpho Semantic Method which depends on the Semantic Links of the Morphological forms where a representative sample of the Arabic morphological forms (nouns and adjectives) is selected which constitutes most of word forms so as to improve the effectiveness of words and stems in retrieval besides improving the Root method by not excluding the recordings that can be retrieved 590 recordings were used as a database and Recall and Precision were used as a scale and evaluation of the impact of both the word, stem, and root on the semantic morphological method. The result was excellence in the semantic method and improvement in the retrieval performance of the word and stem for recall which means retrieving more recordings and improving Precision (less recordings that are not valid for retrieval).

To verify the retrieval performance, the Al-Tayar (2000) collected a sample of 590 Arabic registrations and then entered them into a pilot database created using the Prolog language. As indicated above, The results of the experiment showed that the morphological method achieved the highest recall rate (91%) , the root ranked second (88%) , stem technique (79%) ,and word technique (54%).

The study of Al-Malki (2001) for his PhD degree from the University of New Mexico dealt with the experiment of improving the information retrieval in a monolingual environment by applying morphological analysis of the Arabic language as well as another experiment to retrieve information depending on dictionaries for translation between Arabic and English languages. The system successfully tested the news articles in the newspapers of AL Raya in Qatar and AL Watan in the UAE through twenty questions in which full words were used (without the morphological analyzer) and another group where the morphological analyzer was used. The results about the system evaluation were briefly stated.

Darwish and Ord (2002) developed a morphological analyzer called Sebawai. The system used ALPNET dictionary to estimate the frequency of occurrence of the patterns and the initials of the word, prefixes and the final letters of the word, suffixes. The aim of this system is to increase the coverage through the automatically

constructing Lexicons. The system uses a list of paired words and roots that are automatically extracted using the morphological analyzer, ALPNET. Two lists of pair words have been passed on the ALPNET. 280074 pair words were successfully obtained which were used to estimate the frequency of occurrence of prefixes, suffixes, and patterns. This is because the morphological analyzer detects the roots by examining the words and determining the availability of the structures containing prefixes, suffixes, and roots. These stems are then compared to the stems list consisting of 10,000 stems to insure the proper stem. Sebawai was successful in analyzing about 93% of the words analyzed by ALPNET.

It can be concluded that due to the nature of the Arabic language, most of the works published on Arabic Information Retrieval (AIR) deal with the morphological analysis of the Arabic language until the entry of the Arab track in the TREC2001 system. The main aims of the previous systems focused on extracting the root of the Arabic word. The experiments proved that when small groups are used, they indicate that the root indexing is more effective than both stems indexing and words indexing.

Building Arabic Stemmer for Arabic Information Retrieval

Chen and Gey (2002), two researchers from Berkeley group, California University, Berkeley, participated in Cross-Language Information Retrieval (CLIR) through English-Arabic retrieval track by experimenting on one Arabic track for Arabic language and three tracks through Arabic and English languages. The aim of retrieval through Arabic was to translate English topics into Arabic by using mechanical systems of translation on the direct line from English into Arabic. The names of the runs were (BKY Mon, BKY CL1, BKY CL2 and BKY CL30).The first run (BKY Mon) is the Mono-lingual Arabic track and the other three tracks are through Arabic and English.

The research study dealt with building the Arabic stop list and two stemming systems for Arabic. The experiments were conducted on the mono-language retrieval of Arabic and the linguistic retrieval from English into Arabic. It's notable that the two researchers adopted the track that combined the queries with the documentation language through using the translation systems via two machines whose their tracks of retrieval through language achieved 87.94% of the mono-lingual retrieval performance. Then, they set two systems: the stemmer which depends on M.T- based Arabic Stemmer system and Light Stemmer system of Berkeley group. The performance of the latter was better than the former; this means that the findings of the experiment indicate that the extension of the question leads clearly to a better performance of retrieval. Finally, it should be noted that this research was sponsored by the Defense Advanced Research Projects Agency (DARPA) in America and was also done inside the Office of Information Technology by the same agency. Larkey et al. (2002) used the statistical approach on the Arabic stemming without using any n-grams; this means that this approach depends on analyzing the co-occurrence of the terms in the Arabic text. As a start, he used the Arabic text in the reduction process through light stemmer system. They concluded that the statistical approach added notable improvement on light stemmer including Light8, Light2. However, that improvement did not happen with Khoja Stemmer. They recommended not using n-grams in the Arabic text retrieval.

In 2006, Ibrahim Abul- Khair published an article on the impact of the exclusion of the stop words on the Arabic information retrieval system: a comparative study published in the International Journal of Computer Science and Information. The stop words represent 80% of the documents in the group which are of no importance for retrieval (Baeza-Yates & Ribeiro, 1999) ,whereas some call them noise words which include prepositions and definite and indefinite articles ,whatsoever, which means that they are of no significance (Ghoneim, 2003). Ibrahim Abul- Khair started his study stressing that most researches in the field of information retrieval focused on English language. Recently, there have been efforts to develop the information retrieval systems in other languages other than English. The Arabic research studies and experiments in this field are still new and limited in their fields of study.

The present study aims to improve retrieval effectiveness to check the weight of the term 'stop words' on Arabic retrieval system and compare it by using Lemur Toolkit.The best matching logarithm was (BM25) with the general stop words list. Setting a new weight logarithm using the BM25 properties can lead to more improvements. Lemur proposed many applications out of which this study used only a small percentage to identify their performance in the Arabic language.

There are only some of the trial samples of performance improvement of Arabic information retrieval systems through the morphological analysis. However, the final conclusion of Abdul Salam & Nwesri (2008) was as follows: "the morphological analysis in the distinction between the prefixes and suffixes in the Arabic words with the deep analysis of the Arabic words revealed that these efforts were helpful for the Arabic words retrieval. Moreover, the morphological analysis needs lists (almost incomplete) from the roots, stems and rules which are laid before the experiment of most analyzers of prefixes and suffixes, and as a result of lack of diagnosis, there is failure in finding the roots. Based on this, the morphological rules were used to support stemming and exclude prefixes and suffixes.

In short, the approaches included in this paper represent an important step towards a highly effective Arabic

text concerning retrieval. The researcher compared the performance of the systems of the current Arabic information retrieval. The Light 10 Stemmer showed the best performance among the other systems except the Buck walter stemmer when using relevant feedback. The empirical results show that the technical methods that are followed exclude prefixes, and therefore they enhance retrieval effectiveness. The researcher also used three forms of Light 10 Stemmer, which are Light11, Light12 and Light13, so as to insure the exclusion of suffixes.

Six: Some research studies of contemporary universities and especially those whose researchers are with no prior knowledge of Arabic language.

The Use of TREC 10 of the Arabic Information Retrieval System at Massachusetts University:
The researchers at Massachusetts University depended on TREC-10 without any prior experience in Arabic language and the aim of their study was to apply some standard approaches. Moreover, they did not have any resources about training data, cross-language electronic dictionaries or stemmers (no space: one paragraph). The researchers introduced three tracks including mono-language and cross-language tracks. They began with models description, the technical methods and the resources used. Provided that the formal tracks played a good role at the average, the Normalization and Stemming improved after these results beside the dictionary structure, queries and the language models (Larkey et al., 2006)

The used information retrieval engines
The researchers used the inquiry search engine in two of the three Monolingual tracks and in the Cross-Language track in addition to the Language Modeling (LM) of one of the Simple Monolingual Models. Thus, 383.872 Arabic documents were changed into CP 1256.

Stemming
The researchers got a Stemmer from Khoja (2001), the researcher at Lancaster University, Computer Department. Stemming tried to find roots for the Arabic words and the Stemmer excluded the stop words as well. After the TREC, the researchers prepared a Light Stemmer which helped in stripping the definite articles (ال AL /بال BEL /كال KAL /قال QAL) and stripping "WAW (و)" from the prefixes. As for the excluded words, the word Light means Light Stemmer.

The Dictionaries used
The researchers used dictionaries to find all the available translation works from English into Arabic words and phrases such as:
(1) Ectaco Dictionary.
(2) The Sakhr Multilingual Dictionary.
(3) Sakhr Set Machine Translation
(4) Place Name Lexicons.
(5) Small and Large Lexicons
In these studies, the researchers indicated that without stemming, the results were very poor, but with the introduction of Stemming, the precision improved by 49% (Larkey, 2006).

TREC-10 Experiments at the University of Maryland: CLIR and Video
The researchers from the University of Maryland: CLIR and Video participated in each of the following:

A. Arabic – English as cross-languages.
B. The video tracks of TREC 10 and the main aim of CLIT track were to identify the technical methods of the effect of Arabic mono-language on information retrieval in addition to the effectiveness of the translation questions from English into Arabic (the cross- languages in information retrieval).In the section of the mono-language, different concepts are used including words, stems and roots with finding out n-grams too.

It is noted that the researchers argue for CLIR track in selecting the Arabic index terms, the use of morphology, and the translation in the form of Transliteration (Darwish et al., 2007) which is the material written in words that do not exist in the dictionary in addition to recognizing TREC tracks which include small Arabic groups called ZAD. They also support TREC experiments and effects of different index terms on the Arabic monolingual and English to Arabic cross Language retrieval results

Methodology and Discussion
The users from the University of Maryland (Darwish et al., 2007) prepare the methodology and discussion in two parts. The first part is about CLIR track (six full pages) and the second part is about the Video track99 seven full pages). The researcher will only introduce the sub-titles of each track and illustrate the discussion of each part as

follows:

CLIR Track Methodology

- 1 / 2 Introduction
- 2 / 2 Methodology
- 1 / 2 / 2 Arabic Index terms
- 2 / 2 / 2 Arabic Morphology
- 3 / 2 / 2 Translation & Matching
- 3/2 Experiment design
- 1 /3 / 2 Mono-language Arabic track
- 4 / 2 Results
- 5 / 2 Discussion
- 6 / 2 CLIR references (9 references)

CLIR Track Discussion

The few important results were as follows:

- The art of translation was used effectively in the official results, where the precision averages were in the CLIR track, where the ratio reached 89% in comparison with the relation of the precision average of the Arabic track.
- The results for individual questions were better in official tracks compared with the Median within 10 questions, and within 18 cross language tracks.

The results of CLIR were negatively affected by n-grams, and the use of bigrams for roots seems to be an improper idea, especially with CLIR tracks.

TREC Video Track

- 1 / 3 Introduction
- 2 / 3 Show Boundary Detection
- 1 / 2 / 3 Interview
- 2 / 2 / 3 Approach
- 3 / 3 / 3 Experiments
- 4 / 3 / 3 Discussion
- 4/3 Video References (22 references in English)

Video Track Discussion

There are no reference rules for the researcher, but there are some important observations, including: big Semantic Gap in the analysis of the video and the output of the image and the hints, and that there is some improvement when combining other keys such as light and text that are focused on the research being done and what makes the problem more worse (Problem wave) the lack of automated systems that make the problem more precise (Darwish et al., 2007).

The Focal Analysis of the retrieval of Arabic information

A team of five researchers (Diekema et al., 2005) working in the Natural Language Processing Center of the College of Information Studies at Syracuse University in America, conducted this study. The study indicated that both the English-Arabic and the Cross-Language Information Retrieval Environment were created and the analyst was able to prepare questions for an Arabic base in English as well as retrieve a set of relevant Arabic documents. Translation of Arabic documents was retrieved into English for easy reading in English. The correct names of people, places and organizations were extracted from retrieved documents and mapped from Arabic to English; this helped the analyst to prepare a brief summarization of the retrieved documents. Here are some aspects of the study:

Arabic Cross-Language Retrieval

Under this heading, the five researchers noted that the AIR system was the same as the English-Arabic Information Retrieval System, where the analyst can search for Arabic documents by trying to search for a query in English. The Cross-Language Information Retrieval (CLIR) is itself a special case for retrieving information where the retrieval is not restricted to the language of the query language but by questions in one language for retrieval of documents in other languages (Oard & Diekema, 1998). It is noted here that the use of the Arabic language in the system is called the Modern Standard Arabic (MSA), which is the official Arabic language used in the Arab world for news, radio and newspapers.

.

Future Results and Research

The five researchers concluded that the Arabic Information Retrieval System (AIR) was the same as the English-Arabic Information Retrieval System, which allows analysts to search and retrieve information from relevant Arabic language documents, even if they do not know Arabic. Moreover, the researchers will try to find additional techniques for CLIR to reach the Monolingual Arabic Retrieval level.

The Arab Educational, Cultural and Scientific Organization Project Bouânaga Souad, Paddy Soham, Badi Safieh (March 2010), the Arab Educational, Cultural and Scientific Organization

What follows is part of the organization's project related to the Interactive Dictionary of Arabic language and analysis of its free software projects.

The Interactive Dictionary of Arabic Language

The Arab and international interest in the system of derivation and Arabic morphology urged the organization to adopt the project of The Interactive Computer Dictionary of Arabic Language which is a mono-language dictionary. It is expected to include the automatic processing at different levels such as vocabulary, morphological, grammatical, semantic, phonetic, and statistical levels. When accomplished, the dictionary will fill the gap in the language nature of the contemporary Arabs as it will automatically provide them with the mentioned language systems on dealing with computer.

It will also be useful for the many applications of language and computer, and in the language industry, and it will help to provide computer language tools that contribute to enriching and facilitating digital Arabic content. The King Abdul Aziz City for Science and Technology expressed its willingness to cooperate with the organization in the implementation of the project, and provide the necessary financial support for its implementation. A cooperation agreement was signed, in particular, with the King Abdul-Aziz City for Science and Technology .The project was included in King Abdullah's initiative to enrich Arabic content on the Internet. The online interactive dictionary is designed to do searches, and to add vocabulary, meanings, relevant linguistic information, examples and multimedia files as follows:

• "http://www.almuajam.org" contains information on the stages of work in the dictionary.
• "http://almuajam.hiast.edu.sy" includes direct access to the dictionary.
• "The website http://arabicdictionary.kacst.edu.sa" includes direct access to the dictionary.

The Arab Organization for Education, Culture and Science (ALECSO) has made continuous efforts (according to its potential and available resources) to raise awareness of the free and open source software, survey and coordinate efforts in Arab countries in this area**.**

.

Analytical Viewpoint of the Free Software Project of the Arab Educational, Cultural and Scientific Organization

We have to overview a lot of the relevant issues of free and open source software and Arabic. The first question is about the reasons why there is delay in the adoption of modern technology in the Arab region. The Arab region does not occupy a proper status in the world. It does not keep up with the developed world especially in the Information and Communication Technology (ICT) job opportunities, the establishment of technology companies, the innovation of technologies relevant to Information Technology (IT), and the sale of IT products and services at the international level. These things scarcely exist in the Arab world in spite of available potentials. We do not deny the fact that there is a tangible development in the region and that the competent people began to hold responsibilities (Mekawi, 1997). Thus, what are the causes that prevented the Arab countries from giving due interest to this technology?

- Lack of the culture of the information technology awareness among people, which makes the citizen unwilling to use this technology when necessary.
- Lack of the qualified human resources who deal with modern technologies since the success of any IT employment project depends on the trained manpower rather than on the provision of modern tools. Besides, it is often rare to find such qualified manpower.
- Poor use of ICT compared with international standards of developed countries, and they mainly depend on the import policy of the ready-made technology.
- The national human resources are not given sufficient attention in relation to the process of building their scientific and mental capacities and directing them to participate and contribute to the localization of technology in line with the conditions and needs of the society.
- The absence of the role of civil community and active associations in this field, which should play an important role in the success of the projects of modern technologies and support the concerned associations and spread the culture of technology provided that they do not restrict their activity to some seasonal events, but they should build on the government's efforts in some of the places which are still deprived of Internet connectivity within the country.

- Misjudgment of the budget required and allocated for technological activity.
- Therefore, Arab countries should realize that the use of information technology has become one of the criteria that measure the standard of living and development, especially that "the secrets of information technology are in the hands of a small number of countries that control information industry, operation, storage and retrieval, and they own the channels through which this information passes "(Nabil, 2003).

The Free and Open Source Software and its Support for the Arabic Language

It has been pointed out that the limited support for the free and open source software in the Arab region hinders the application of information technology. Software industry is considered one of the most promising industries of the future, and it is the main industry on which the Arab digital content industry is based. Thus, what is the reality and future of the software industry in the Arab world?

The open source software is freely accessible and freely available for developers to review and test them and consequently participate in its enrichment in comparison to exclusively owned software developed by a particular company and monopolize its code, thereby undermining the chances of innovation testing. The free and open source software is suitable to obtain various operating systems and applications at a low cost. Therefore, this software must be adapted to the Arabic language to be a means of disseminating Arabic digital content. The Arab applications that rely on open source software are restricted to limited areas such as Arabic language and offering services on websites.

Results and Recommendations

Ghoneim (2003) shows that there is no deficiency in the Arabic language itself, it is valid as an input or output language, but according to the researcher, there is deficiency in the other languages when trying to translate the Arabic language as we noted at the beginning of this study.

For the following can be noted:

- The study showed some challenges in the research of the Arab information retrieval system when presenting new sources of information.
- The extent of adapting the search engines to suit the characteristics of the Arabic language
- The Arabic language with its classical and colloquial form and writing from right to left has distinctive characteristics of the sematic phonetics, grammatical rules, and semantic rules - the difficulty of processing it in the search engines -carries about five million words derived from 11,350 roots. This represents the retrieval in comparison to the English language which contains about 1.3 million words out of which there are 400,000 key words. While the English language occupies about 67% of the Internet, the Arabic content is about 1%, Arab search engines are still in a weak position. The future must witness advanced Arab search engines such as Google.
- Some of the Developments Introduced into Retrieval Research in general
- There are many attempts to reach the Matching stage between the beneficiary's questions and the documents and how to adapt them to reach the highest recall with the reduction of the invalid documents. Mathematics, Stemming, filtering and documents processing, and extensions of questions are introduced. In the Arabic language, the adaptation in the use of the word, root, and stem for better retrieval is introduced.
- Before the entry of TREC in 2000, the roots achieved the best retrieval of the word and stem. This may be due to the small size of the research group in the various Arab researches. However, after the entry of TREC, the stems could achieve the best results on the roots and words. Future research studies the term 'logarithms', especially with language and probability models in Arabic, and the need to reach the criteria of the Stemming logarithms, stop words and their lists as well as the transformation from small groups to large groups to reach satisfactory results. Before TREC in 2000, the roots achieved the best retrieval of the word and stem. This may be due to the small size of the search group in the various Arabic researches. However, after TREC, the stems could achieve the best results on the roots and words, in addition to the necessity that future research studies should include 'logarithms' to modify concepts, especially with language and probability models in Arabic, and the need to reach the standards of the Stemming logarithm, stop words and their lists as well as the transformation from small groups to large groups to reach satisfactory results.
- The twenty-First Century Researches are introduced into important new areas such as the use of semantic morphology method and the experiment of information retrieval based on the English and Arabic languages dictionaries, developing the morphological analyst till the Arabic track is introduced in TREC 2001 system and comparison with the English tracks in addition to the use of the statistical methods in the reduction of the stems of the Arabic text and its overlap with n-grams and finally (the retrieval techniques affecting the Arabic text and designing a Lexicon that depends on Light Stemming

improvements.

- It was found by extrapolating the references of this study that there were thirteen PhD dissertations obtained by Arabs holders from universities in Cairo (2), Mustansariya (1), and then from American universities: Pittsburgh (4) New York (3) and New Mexico in addition to Loughborough University in England (1) and Australia (1). The current research also noted that some universities (Maryland, Massachusetts and others) focused on using TREC 10, the Arab information retrieval system. We hope that the future will see more doctoral and master's studies in Arab universities in the field of Arabic information retrieval systems.
- The study showed that there were many studies in the field of Arabic information retrieval which were conducted independently without any prior experience in Arabic language.

References

Abdel Ali, A. Comy, A. & Soliman, H. (2004). Arab information retrieval. *Arab Language Processing, 19.*

Abu El-Khair, I. (2003). Effectiveness of document processing techniques for Arabic information retrieval. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh.

Abu El-Khair, I. (2006). Effects of stop words elimination for AIR: Comparative study international J. of computing and information science v.4 (3) On-line 119-132.

Abu El-Khair, I., El-Sobhy, M., Fahim, A., Younes, S., Al-Afeef, Y. & Abdulrahman A. (2013) Self Study Report According to AALE Guidelines. Technical Report. Makkah: University of Umm Al-Qura, Faculty of Social Sciences - Department of Information Science. http://acl.ldc.upenn.edu/w/w02-0506.pdf.

Abdul Salam, F. & Nwesri, A. (2008) Effective retrieval techniques for Arabic text. Thesis. Doctor of Philosophy (PhD). Computer Science and Information Technology, RMIT University.

Ali, N. (1988). The Arabic Language and Computer: A research Study by Al Kholy, A. (in Arabic). Kahira Ta'reeb for Publishing, p.591

Ashkar, J. (2002). Approaches to Arabic information retrieval building on Arabic stemmer for information retrieval TREC vol 2002, Gap.

Al-Atram, M.A. (1990). *The efficiency of Arabic language in indexing and retrieving the Arabic texts*: final report (in Arabic). Riyadh: King Abdul-Aziz City for Science and Technology.

Al-Dayel, A. and Mourad, Y. (2013). Arabic user's attitudes toward web searching using paraphrasing mechanisms. Journal of Computer Science and information systems, 2 (2), p 34-39.

Al-Kharashi, I. A. (1994). The techniques of web search systems under free open source software. Paper submitted to a conference on highways of information: Technology in the service of the community. 16-18 March. P. 77-84.

Al-Malki, A. M. M. (2001). Recent trends in library and information sciences. Al-Warraq for Publishing and Distribution. Jordan: Amman.

Al-Tayar, M.S. (1998). *The efficiency of morphological analysis in Arabic texts retrieva* (in Arabic). King Fahd National Library magazine, 4 (1), p.23-7.

Al-Tayar, M.S. (July 2000). Arabic information retrieval system based on morphological analysis: A comparative study of word, stem, root and morph semantic methods. Ph.D. in computer science, Department of Computer and Information Science, De Mont Fort University, U.S.A.

Bamaflah, F. S. (2000). *The bases of electronic information retrieval* (in Arabic). Riyadh: King Fahd National Library, p.17329.

Baeza-Yates and Ribeiro, N. (1999) Modern information retrieval pearson: Addison Wesley. URL:www.theatlantic.com/doc/194507/bush

Beesley, K. (1998). Arabic morphological analysis on the internet. In proceedings of the 6th international conference and exhibition on Multi-lingual Computing, Cambridge.

Chen, A. and Gey, F. (2002). Building an Arabic stemmer for information Retrieval in: Proceeding of TREC 2000.

Croft, W.B. & Lafferty (2003). Language modeling for information retrieval. Springer.

Darwish, K. & Oard, D. (2002). CLIR experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. The Eleventh Text Retrieval Conference (TREC 2002), 703-710. Retrieved January 25, 2006, from http://trec.nist.gov/pubs/trec11/papers/umd.darwish.pdf.

Darwish, K, doermann, D; Jones, Ryan, Oard, D and Rautiainen, Mika (2007). TREC-10 Experiments at the university of Maryland: CLIR and Video.

Darwish, k (2007). Adapting Morphology for Arabic information retrieval http://www.google.com.eg/20/5/08/22

Diekema, A. R.; Hannouche, J; Ingensoll, G; Oddys R., Liday, E. (2005). Analyst focused Arabic information retrieval. The School of Information Studies, Syracuse University.

Gey, F. & Oard, D. (2002). The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries. The Tenth Text Retrieval Conference (TREC 2001), 16-25. Retrieved

from http://trec.nist.gov/pubs/trec10/papers/clirtrack.pdf.

Ghoneim, M. S. (2003). *Arabic information retrieval systems*: Aspects of ambiguity and horizons of solutions (in Arabic). King Fahd National Library, p. 474 (PhD dissertation 2003) Cairo University.

Hersh, W. (2006). TREC: Genomics Track Overview In: Woorkees & Buchlard (2006).

Khoja, S. (2001). Khoja's Arabic stemmer (version 1.0). London: Khoja.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. SIGIR 2002: proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 275-282.

Larkey, L. S., and Connell, M. E. (2006). Arabic information retrieval at UMass in TREC-10. Center for intelligent information retrieval UMass. Amberst MA, USA, 01002.

Mekawi, H. I. (1997). Modern communication technology in the information age. Cairo: Egyptian Lebanese House. Pp. 109-123

Moukdad, 2004. Image retrieval: theory and research. The scarecrow press. Lanham. MA and Oxford. pp 35-44

Nabil, A. Arab mind and knowledge society. *World of Knowledge Series*, Kuwait.

Qassem, H. (1978). Arabic in specialist information systems. Unpublished doctoral dissertation, University of London.

Oard. D. and Diekema, A. (1998). Cross-language information retrieval. ARIST v.33, pp 223-256.

Suwina'a, A. S. (1994). Information retrieval in Arabic Language (in Arabic). Riyadh: King Fahd National Library, p.176.

Voorhees, E. (2006). Overview of TREC 2006 In: Voohees and Bucklard (2006).