On the Design and Implementation of an Efficient Information Retrieval System for Arabic Language

Mohammad O.Wedyan
Al-Balqa' Applied University, Salt, Jordan
Email: Wedyan56@ gmail.com

Aarti Singh
Maharishi Markandeshwar University, Mullana, Haryana, India
Email:singh2208@gmail.com

***Abstract*—an efficient stop-word removal technique is needed in many natural languages processing application such as: spelling normalization, stemming and stem weighting, Question, Answering systems and an Information Retrieval system (IRS). Most of the existing stop-word removal techniques are based on a dictionary that contains a list of stop-word which is very expensive and takes much time for searching process and required much space to store these stop-words. So, the main objective of this paper is to design and implement an efficient Information Retrieval System for Arabic language. The proposed technique is based on analyzing document in a statistical way through two approaches: manual methods for IRS and automatic methods for IRS. The new proposed Arabic removal stop-word technique has been tested using a set of 500 Arabic abstracts chosen from a set of data chosen from different resources and it gives impressive results. Our work presented in this paper is based on analyzing document in a statistical way through two approaches: (1) Manual method for (IRS) in which we should counting the number of repetition of the word in the document then comparing the degree of similarity between the counted repetition words through using some static database contain specific subject. (2) Automatic methods for (IRS) in which we should counting the number of repetition of the word in the document after that calculating the weight of the counted repetition words through using two different algorithms using mathematical formulas.**

***Index Terms*—automatic indexing, evaluation, information retrieval system, stop word.**

I. Introduction

The world witnesses a huge informational revolution that brings out lots of information and researches. Researching for this information without using search engines is a hard task or it is impossible to gain accurate information without them. Alongside, the importance of information retrieval sciences has risen. This science depended only on libraries in the past, but with the widespread of internet, there was an increasing need for programs and software that facilitate accessing the information accurately and quickly. The importance of this paper lies in that it covers a very important subject of matter, the process of organizing and utilizing from information especially in this age that witnesses a huge information revolution, so it is necessary to use informational Technology to work perfectly with Arabic especially in information retrieval that has some link with Arabic. [1]

There are various models have been developed to retrieve information like: the Boolean model, the Statistical model, which includes the vector space and the probabilistic retrieval model, and the Linguistic and Knowledge-based models. The first model is often referred to as the "exact match" model; the latter ones as the "best match" models.

Queries generally are less than perfect in two respects:

First, they retrieve some irrelevant documents. Second, they do not retrieve all the relevant documents. The following two measures are usually used to evaluate the effectiveness of a retrieval method. The first one, called the precision rate, is equal to the proportion of the retrieved documents that are actually relevant. The second one, called the recall rate, is equal to the proportion of all relevant documents that are actually retrieved. If searchers want to raise precision, then

They have to narrow their queries. If searchers want to raise recall, then they broaden their query. In general, there is an inverse relationship between precision and recall. Users need help to become knowledgeable in how to manage the precision and recall trade-off for their particular information need.

II. Statistical Model

The vector space and probabilistic models are the two major examples of the statistical retrieval approach. Both models use statistical information in the form of term frequencies to determine the relevance of documents with respect to a query. Although they differ in the way they use the term frequencies, both produce as their output a

Manuscript received January 5, 2013; revised March 3, 2013.

doi: 10.12720/ijeee.1.1.5-9

list of documents ranked by their estimated relevance. The statistical retrieval models address some of the problems of Boolean retrieval methods, but they have disadvantages of their own.

A. Boolean Model

In this model each document is connected with a group of index terms and each query to be the form of a Boolean Expression. This expression consists of a number of index terms that are connected together with a Boolean Operator (and, or, not). This system retrieves the documents tat the query determines [2]. This model has certain properties like: Easy to understand. [3] Easy to apply. [2] Can be processed quickly concerning the time needed to process the Query [2]. On the other hand, this model has the following disadvantages. [3] Like: It has a binary decision. Either relevant or irrelevant, many users cannot formulate a correct query.

B. Term Weighting

Term weighting has been explained by controlling the exhaustively and specificity of the search, where the exhaustively is related to recall and specificity to precision. The term weighting for the vector space model has entirely been based on single term statistics. There are three main factors term weighting: Term frequency factor, Collection frequency factor and Length normalization factor. These three factor are multiplied together to make the resulting term weight. A common weighting scheme for terms within a document is to use the frequency of occurrence as stated by Luhn, The term frequency is somewhat content descriptive for the documents and is generally used as the basis of a weighted document vector. It is also possible to use binary document vector, but the results have not been as good compared to term frequency when using the vector space model. There are used various weighting schemes to discriminate one document from the other. In general this factor is called collection frequency document. Most of them, e.g. the inverse document frequency, assume that the importance of a term is proportional with the number of document. Experimentally it has been shown that these document discrimination factors lead to a more effective retrieval, i.e., an improvement in precision and recall. The third possible weighting factor is a document length normalization factor. Long documents have usually a much larger term set than short documents which make long documents more likely to be retrieved than short documents.

C. Evaluating IRS.

Any retrieval system is usually evaluated according to its efficiency and effectiveness. There are two aspects of efficiency, they are time and space. Time is the speed of matching the in-use queries with the document descriptions. Space is the space needed in a disk that the system needs.[4] Efficiency is determined according to the ability of the system to return documents relevant to the user query. The perfect status of the system is referring all the files that are relevant to the process of query and never referring any irrelevant files. The difficulty lies in the determination of relevance because the process of determining relevance of documents is a subjective one. The decision of the person depends much on many factors; experience [5], for example. Any professional in a certain field may see the general information retrieved from a system as irrelevant while any amateur (beginner) sees it as fully relevant. This may lead to increasing in the determination of relevance. In research, researchers usually consider the process of determination of relevance as an objective process. We suggest here that evaluation process is objective and previously agreed on.

D. Related Works

Despite the very little Arabic efforts in developing thesauruses, the theoretical efforts supported and opened new paths for building Arabic thesaurus, even though very limited, the first trials in this field were translation of foreign thesauruses, example of this is the list of Arabic Idioms prepared by Industrial Development Center for the Arab World in 1970, and the Islamic thesaurus which was built manually [6]. There are also some studies in IRS and in building thesauruses. Abu Salem (1992) for example, studies the IR in Arabic Language. His study was based on 120 documents he received from the Saudi Arabian National Computer Conference and on 32 queries. in his research, he studied indexing by using full words and by using the roots only. He found that using the roots is superior to other ways. He also built a manual thesaurus using the relation between expressions to test the possibility of supporting an IRS through this thesaurus. He found that the thesaurus makes IR much better.

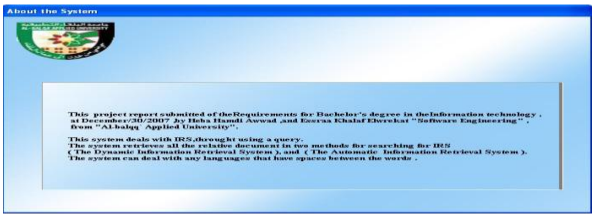
Kanaan, (1997) [7] compared statistical automated indexing and linguistic automated indexing. This study was based on 242 documents taken from Saudi Arabian National Computer Conference and he used 60 queries. His study found that Statistical automated indexing is a bit better than Linguistic automated indexing. The General Thesaurus presented by UN Aid Program – the Program of Authorization in the Arabic World (2003). This one uses initially synonyms that help the researcher to choose his expressions that he has to look for. This thesaurus includes also the relations of origin and branches and those of contextualization between expressions. This helps in boarding the search, if the search has no matches when using a certain expression, the researcher can use either broad terms or narrower ones. Synonyms are the first step in this thesaurus.

Al-Shalabi et al, (2004) suggested a method for determining and deleting stop words in Arabic texts. This method is an Algorithm that depends on Finite State Machine. As deleting stop words is needed in many Natural Languages processing applications and in IRS the researchers DISPLAY THE Algorithm they depended on. They tested the system using the 242 documents that were presented in the Saudi Arabian National Computer Conference in addition to some verses taken from the Holy Quran. They reached an accuracy reaching up to 98%. Al-Zoman et al, (2001) tried to show some difficulties that prohibit the widespread of Arabic on the internet especially in Domain Names and in Arabic URLs. There are some difficulties mentioned by the researchers such as TASHKEEL and KASHEEDEH that is used to make letters longer in shape and the researchers depended on questionnaire through which they tried to take the suggestions of those interviewed in the subject matters. Darweesh (2002) suggested a quicker way to find the roots of Arabic words. He limited that in one day. It depended on statistical and automated methods. The researcher named the program (Sebawei). This program is based on inputting a pair of words and roots. The program determines the additions that occurred to the root. Then another word is input and the program brings pout all the possible roots for the new word. The researcher succeeded in inputting 9000 words taken from texts called ZAD, the other list contained 560000 words taken from texts (Linguistic Data Consortium – Arabic Collection) (LCD) he was able to input 270,000 words of them. The researcher used ALPNET program to find the roots. Any word the program couldn't find its root was neglected for evaluation purposes of the program. When he used the list containing 270,000 words to be input in the program and the group of 9606 words to evaluate the system, Sebawei succeeded in bringing out the roots for 292,000 words, the system succeeded in finding the roots for 128,000 wore and failed in the rest. When using the 9606-word list as a feeding list and inputting 270,468 words to evaluate the system, it failed in finding the roots for 84,421 words and succeeded in finding the roots of the others. When he used 292,212 words Sebawei was able to deal with 92,929 words and failed in dealing with the rest.

Carlberger et al (2001)[1] studied the effect of using infinitives (abstraction) in Swedish on the criterion of precision. After the results they gained, which showed better precision in English, Dutch and Slovenian languages when abstraction was used, they applied the same on Swedish language and calculating the effect of that on information retrieval. They used a sample taken from (KTH News) consisted of 54487 essays. The results were when they used abstraction they gained better results on precision 15.2 % and 18% on recalling. Rasha (2006) Stop Word Removal Algorithm for Arabic Language, preparing documents by deleting punctuation marks and Stop Words. of this mechanism aims at eliminating the unnecessary components such as punctuation marks and stop words which have a great effect on increasing the precision of the system and also increasing the speed of building the thesaurus and making the space size less.

III. Proposed IRS Software and its Interfaces

The proposed system deals with IRS through using a query, supposed to be a group of words " text document" , the system search and retrieve all the relative documents in two methods , one through using Manual method IRS and Automatic methods IRS. This system can deal with any language as there are spaces between the words, and any language can be used in the queries as the system deals with the words statistically and depends on the repetition of the word in one document. The designed program is flexible, easy to be modified as it consists of interfaces and the code is divided into easy procedures. Visual Basic Dot net (.NET) language used in programming. It is preferred because its deal .net framework with simple, easy components and with high integrity with office applications VB.net was also linked with MS Access 2003 as it deals with small number of tables. SQL was used in inputting data, deletion and in modification. The system begins by the first interface which consists of information about the system; this information is displayed in the window shown in Fig. 1.



This project report submitted of theRequirements for Bachelor's degree in theInformation technology , at December/30/2007 ,by Heba Hamdi Awwad ,and Esraa Khalaf Elwrekat "Software Engineering" , from "Al-balqq` Applied University".

This system deals with IRS,throught using a query.
The system retrieves all the relative document in two methods for searching for IRS
(The Dynamic Information Retrieval System), and (The Automatic Information Retrieval System).
The system can deal with any languages that have spaces between the words .

Figure 1. Interface information of system

Then, after this window has shown, the second window in Fig. 2 has shown which consists of:-

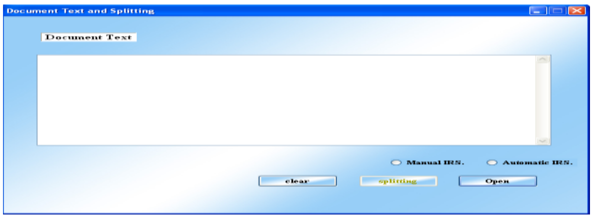
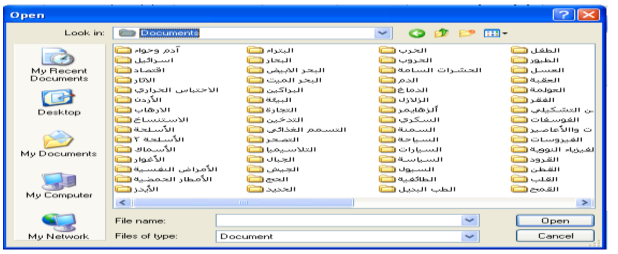


Figure 2. Windows of consist of the interface system

1. Radio Button of “Automatic IRS”. Through this radio button we can choose the Automatic methods of IRS.
2. *Radio Button of “Manual IRS”. Through this radio button we can choose the Manual method of IRS.*
3. Button of “Open “.
4. Button of “Splitting “.
5. Button of “Clear”. Through this button we can delete the textbox content.

Through the “Open” button we can open and load the document text by click the OpenFileDialog shown in Fig. 2 then Fig. 3 appears.

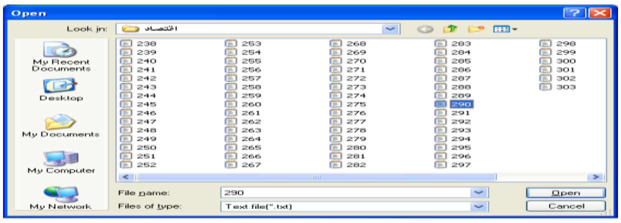


Look in: Documents

File name:

Files of type: Document

Figure 3. The open dialog Window



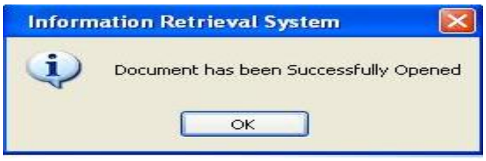
Look in: اقتصاد

File name: 290

Files of type: Text file(*.txt)

Figure 4. The open dialog Window

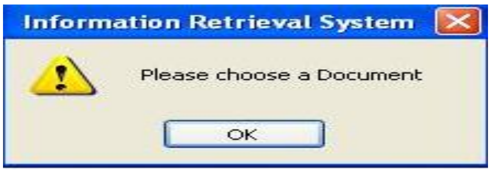
When the chosen text from Fig. 4 to be opened on textbox Fig. 6 this message box shown in Fig. 4 will be displayed. When the chosen text from Fig. 3 to be opened on textbox Fig. 2 this message box shown in Fig. 4 will be displayed.


Information Retrieval System

Document has been Successfully Opened

Figure 5. The message box of the open dialog box

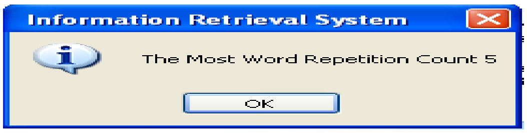
But if there is no text chosen from Fig. 3 to be opened this message box shown in Fig. 5 will be displayed.


Information Retrieval System

Please choose a Document

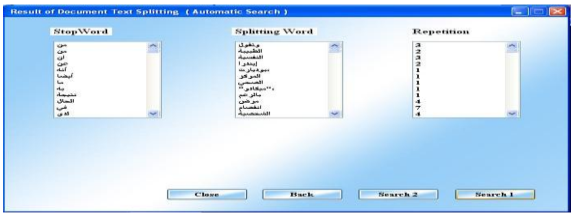
Figure 6. The message box of the open dialog box

Firstly, we choose the type of IRS, through click the "splitting" button we can split the word and calculate the repetition of the splitting word and find the maximum repetition display in message box. Fig. 6, and display.window shown in Fig. 7, Fig. 7 consists the results of clicking this button.


Information Retrieval System

The Most Word Repetition Count 5

Figure 7. The message Box of the maximum repetition


Result of Document Text Splitting (Automatic Search)

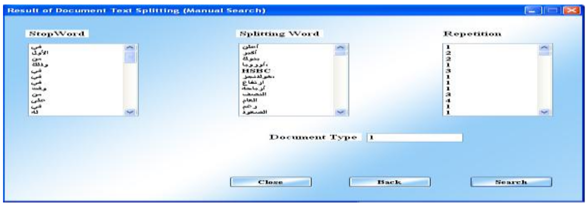
StopWord

Splitting Word

Repetition

3
2
3
2
1
1
1
1
1
4
7
4

Figure 8. The message box of consists the results of clicking this button.


Result of Document Text Splitting (Manual Search)

StopWord

Splitting Word

Repetition

1
2
2
1
3
1
1
1
3
4
1
1

Document Type 1

Figure 9. The message box of consists the results of clicking this button.

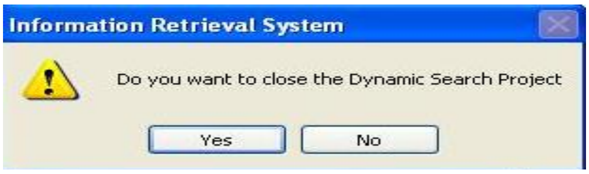
The Automatic methods IRS in Fig. 7. consist of:

Button "Search 1".

Button "Search 2".

Button "Back". Through this button we can go one step back.

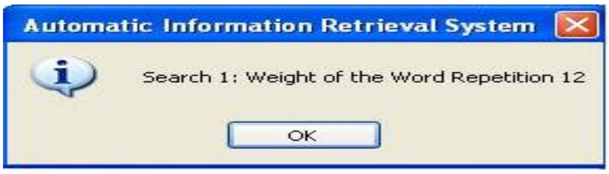
Button "Close". Through this button we can close the system, but when click on this button there is message box Fig. 8 displayed to ensuring if close the system or not.


Information Retrieval System

Do you want to close the Dynamic Search Project

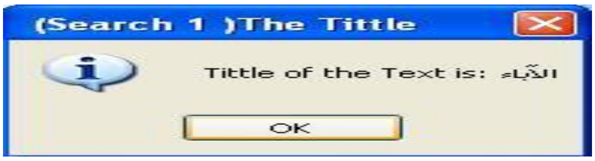
Figure 10. The message box to ensure if close the system.

The "Search 1" button contains method 1 for the Automatic IRS to find the suitable retrieved word that describe the document content from the first fifteen splitting words . And when click on this button the method 1 run and give the splitting words the same weight at the beginning, then update the first fifteen slitting words weight the high weight expecting the rest of splitting words weight. Finally if the maximum repetition found in the updated fifteen splitting weight, then give the maximum repetition weight be the highest weight. But if the maximum repetition not found in the updated fifteen splitting weight, and then give the maximum repetition weight be the weight that the splitting words have at the beginning. The results displayed in message boxes as shown Fig. 9 and Fig. 10.


Automatic Information Retrieval System

Search 1: Weight of the Word Repetition 12

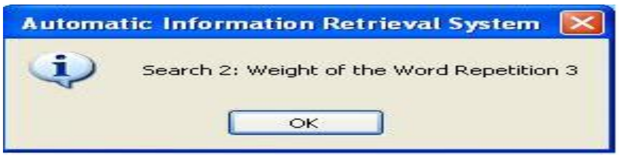
Figure 11. The message box of weight of the word repetition.


(Search 1)The Tittle

Tittle of the Text is: الآباء

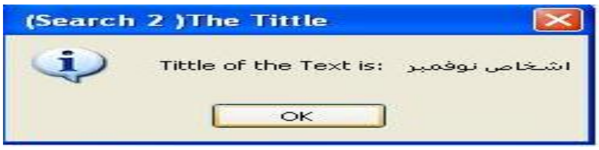
Figure 12. The message box of title of the text .

The "Search 2" button contains method 2 for the Automatic IRS to find the suitable retrieved words that describe the document content from the splitting words. And when click on this button the method 2 run and give the splitting words the same weight, then give the maximum repetition weight be the highest weight. The results displayed in message boxes as shown Fig. 11 and Fig. 12.


Automatic Information Retrieval System

Search 2: Weight of the Word Repetition 3

Figure 13. The message box of weight of the word repetition.


(Search 2)The Tittle

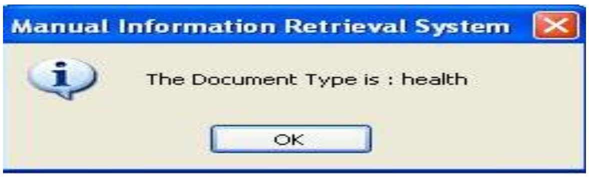
Tittle of the Text is: اشخاص نوفمبر

-

Figure 14. The message box of title of the text.

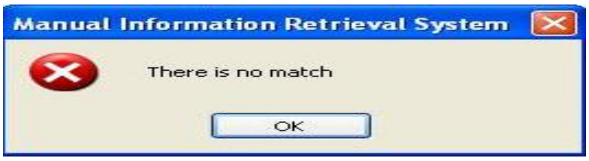
The Manual IRS in Fig. 7 consists of:

1. Button “Search”.
2. Button “Back”. Through this button we can go one step back.
3. Button “Close”. Through this button we can close the system, but when click on this button there is message box Fig. 8 displayed to ensuring if close the system or not. The “Search” button contains method for the Manual IRS to find the suitable retrieved words that describe the document content from the splitting words.
4. Button “Close”. Through this button we can close the system, but when click on this button there is message box Fig. 8 displayed to ensuring if close the system or not. The “Search” button contains method for the Manual IRS to find the suitable retrieved words that describe the document content from the splitting words. And when click on this button the method run and Comparing the degree of similarity between the maximum repetition words through using some static database contain specific subject. The results are giving the document type that has relation with the maximum repetition words displayed in message box as shown in Fig. 13. are give the document type that have not relation with the maximum repetition words displayed in message box as shown in Fig. 14.



The Document Type is : health

Figure 15. The message of the document type that have maximum repetition words.



There is no match

Figure 16. The message box of document type that have not relation with the maximum repetition words

REFERENCES

[1] G. Kanaan and M. Wedyan, ”Constructing an automatic thesaurus to enhance arabic information retrieval system,” in *Proc. 2nd Jordanian International Conference on Computer Science and engineering*, 2006, pp. 89-97.

[2] W. Frakes and R. Yates, *Information Retrieval Data Structures & Algorithms*, PTR Prentice Hall, New Jersey, 1992, ch. 12, pp. 264-292.

[3] R. Yates and B. Neto, *Modern Information Retrieval* , Addison-Wesley, New-York, 1999, ch 2, pp. 19-69

[4] M. Lassi, “Automatic thesaurus construction,” *GSLT Course Linguistic Resources University Collage of Boras*, pp. 2-10, autumn 2002.

[5] G. Salton and M. McGill, *Introduction to Modern Information Retrieva*l, McGraw-Hill, New York, 1983, ch 5, pp 157-191.

[6] A. Rahman, “The use of a system consultant in building thesauruses,” *Scientific Record of the Symposium on the Use of Arabic in Information Technology Organized by the King Abdul Aziz Library \Public*, Riyadh, Saudi Arabic, pp. 22-35, june 1993.

[7] G. Kanaan, "Comparing automatic statistical and syntactic phrase indexing for arabic information retrieval", Ph.D. Dissertation, University of Illinois, Chicago, USA. 1997.

Mohammad Wedyan was born in Jordan in February 1980; He earned his master in 2005 in computer science from Al-Al-Bayt University, and he had a B.Sc. in Computer Sciences from the Mutah University (2002). He is presently LECTUER at the computer science department at Al- Balqa’ Applied University Jordan. Wedyan published many research papers in many topics such as: Information Retrieval, image processing, and artificial intelligence. He is a reviewer for several journals and conferences.

Dr.Aarti Singh has a credible academic record, with various degrees like Ph.D.(Computer Science) from Maharishi Markendeshwar University, Mullana, India in 2011, M.Phil. (Computer Science) from Madurai Kamaraj University, India in 2007, MCA from Kurukshetra University Kurukshetra in 2004, M.Sc.(Computer Science) from Kurukshetra University Kurukshetra in 2002 and B.Sc. (Computer Science & App.) from Kurukshetra University Kurukshetra in year 2000.

She is presently ASSOCIATE PROFESSOR at MMICT &BM (MCA Deptt.), M.M. University, Mullana, Haryana, India.

Dr Singh also owe the credit of 24 published research papers in various national & International Journals of repute, with one paper awarded as the Best Paper in an IEEE Conference. Dr Singh have also participated in many International conferences within India and abroad.

Dr Singh’s research interests include Semantic web, Agent Technology, Web Mining and Intrusion Detection Systems.