Arabic News Classification using Field Association words

Abstract:

Text classification is a popular problem that has been studied extensively in the last four decades. Since many classification schemes can be used, the question of how to choose the best one among many for a designated task remains. In this work, we design a method for classification the Arabic news, the classification system that best fits data given a certain representation. We present a new method for Arabic news classification using field association words (FA words). The document preprocessing system will generate the meaningful terms based on Arabic corpus and Arabic language dictionary. Then, the field association terms will be classified according to FA word classification algorithm.

Keywords: Arabic information retrieval, field association words, Document classification.

1. Introduction:

The retrieval, information science is the science that is careful with searching for documents, information, and data from documents; and also for metadata relating to those documents also will search the databases and the Internet. The process of retrieving information depends on a lot of sciences like: computer science, mathematics, libraries and information science, linguistics and information architecture, statistics, physics, cognitive psychology and other sciences.

Automatic retrieval information systems have been used to reduce the dumping informational process. At the present time there are a lot of universities and public

libraries that use such systems for saving time that was spent to access to books and scientific journals as well as other documents. The most important examples of information retrieval systems, search engines, where such systems are used criteria for measuring the quality of the results of the research process in terms of accuracy and review.

These systems are started to be used at the computer systems in the fifties and sixties. And in the seventies, it was proven that these techniques work well with ammunition for small texts such as Cranfield group. With this progress, it was begun to use techniques that deal with large text systems in the early seventies. Of the most important examples of retrieval systems in that period was to Lukahad Dialog system.

At the beginning of the nineties the US defense department in collaboration with the National Institute of Standards and Technology (NIST) support Text Retrieval Conference (TREC), which was considered as part of the script Tibstr program. The purpose of this conference is to help the new technologies and supplying them with the infra structures of methods and techniques to assess new ways and new approaches to large-scale collections of text documents. As well as the existence of search engines and their appearance was one of the biggest motivating factors for the spread and development of information retrieval systems.

Therefore the use of digital methods for storage and retrieval of information has led to the emergence of digital suggestion, which means that the digital source is no longer available because the hard-disc, the reader, the hardware, or electronic packages are no longer available or unavailable.

Arabic is one of the languages are widespread with an estimated number of 400 million native speakers. And is also, as in the other languages have the inflections and vocabulary and the order of the syntax (subject-verb-object and verb-subject-object), and the use of vowels and also words that are originally derived from the

roots, whether these roots is composed of two characters, three or four, and triple roots are the most common and also the diacritical marks that are often omitted when writing, and names that may be single, collect, or double and masculine and feminine.

Note the increase of Arabic digital documents, whether on the Internet or electronic media, This is making us desperately need to find a retrieval system means in Arabic and their distinctive properties and is able to deal with it.With the existence of programs and databases in Arabic, but it is noticeable that there are problems hindering the process of search and retrieval, and these problems, is the main reason is due to the nature of the Arabic language, which has different characteristics from the rest of the languages, in terms of semantic and complex structural characteristics that affect the accuracy and the efficiency of the recovered data. However, observe that efforts to restore Arab documents are still incomplete and lacked the precision and efficiency and not such efforts in other languages.

Occupies the Arabic language seventh largest language on the Internet, and is one of the languages of the fastest growing in the last decade in terms of users, and because of the rate of use of Arabic Internet speaking, it must be the fourth-largest number of users on the Internet by the year 2020, and this is something which emphasizes the importance of language Arab and the need to retrieve information accurately and effectively. The main goal of text categorization is used to classify the documents into a number of pre-define classes. Text categorization is a research area in information retrieval and machine learning. A lot of supervised learning algorithms have been applied to the text categorization using a training data set of categorized documents. This consists of a training phase and a text classification phase. The previous includes the feature extraction process and the

indexing process. The vector space model has been used as the conventional method for text representation.

Field Association (FA) is a limited set of the conditions terms that can identify Document fields. The Notion of FA words can recognize the subject of many documents fields by finding only some specific words without reading a document field and can be ranked as a super-field and a sub-field. FA terms have five different Stages to associate with the field. FA words are used especially for classifying Arabic documents. These words are extracted from the documents used at the classification process to get the FA word candidates. The reset of the paper are formulated as follows. Section 2 give an outline for the previous work. Moreover , the term of FA words are described in section 3 in detail. Section 4 discuss the Arabic document classification. In addition, section 5 explain our new idea for Arabic news classification using FA words. Section 6 is the expriemental evaluation for new algorithm. Conclusion and future work are presented in section 7.

2. Previous work

There are techniques and algorithms used to classify Arabic documents. On paper [1] discuss the Effect of Stemming on Arabic Text Classification, Stemming the use of several algorithms, including (SVM) The results showed when not in use. Support vector machine achieved (SVM) ranked the highest classification accuracy, using two test methods with 87.79% and 88.54%. On the other hand, when the use of stem impacted negatively on the accuracy where SVM using two test modes accuracy dropped to 84.49% and 86.35%.

At [18] An Automatic Filtering Method for Field Association Words by Deleting Unnecessary Words Delete unnecessary words using the information on the categories and experimental results, it turns out that unnecessary words are deleted automatically at 25% from 38,372 FA word candidates using the presented

method. Furthermore, Precision and F-Measure are improved by 26% and 15%, respectively, over the traditional method. In the [23] Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm There validation test evaluation set which consists of 10 documentsoverall classification accuracy achieved over all categories is 62%, and that the best result by category reaches 90%. There are also [17] Arabic Text Classification Using Maximum Entropy Preprocessed data using Natural language processing, such as tokenizing, stemming, part of- techniques Speech. After that, we used the method of maximum entropy to classify the Arabic Documents. We experimented with our approach using real data, then we Compare the results with other existing systems. This paper deals with classification Arab News using field association.On The paper [8]a strategy for building a morphological matching dictionary of English that infers Meaning of derivations by considering morphological afixes and their semantic classification. An eficient method for selecting compound Field Association (FA) terms from a large pool of single FA On The[14]. Document classification to assign a document to one or more on on The categories based on its contents. This paper suggests the use of Field Association (FA) words Algorithm with Naïve Bayes Classifier to the problem of document categorization of Arabic language.

3. Field Association Words

It is natural people to identify the field of document when they notice peculiar words. These peculiar words are referred as Field-Associating words (FA words); specifically, they are words that allow us to recognize intuitively a field of text or field-coherent passage. Therefore, FA terms can be used to identify the field of a passage, and can be also used to classify various fields among passages. For these causes FA words can be used as a clue to identify a passage field [13]. FA

words can be either words or phrases. For example, the word “President” can indicate the document filed <Politics News>.

Since the basic concept behind FA words involves the choice of a limited set of words that match a given document best, they describe a set of discriminating words. Moreover, FA words are not the same as subject words.

FA word is a minimum word which cannot be divided without wastage Semantic meaning [9]. Based on specific FA word information, topics of documents. All the previous studies are based on FA words in English and Japanese. FA terms are categorized as single FA terms or compound FA terms. Their definitions provided below. Single FA Term: A single FA term is an FA term which is formed by “independent, meaningful, inseparable and smallest unit” usually consisting of a single word. In this paper, two or more words separated by hyphens, but not by white spaces are treated as a single FA term for the purpose of automatic extraction. For example,“ medicine” and “patient” are treated as single FA terms in < Medical News>.

Compound FA Term: An FA term that consists of more than one word. In this paper, only terms consisting of words separated by white spaces are treated as compound FA terms for the purpose of automatic extraction [8,13]. For example, “Presidential vote” is a compound FA term of <POLITICS>.

Field Tree: A field tree is a scheme that represents relationships among document fields. A document field is defined as basic and common knowledge useful for human communication, Leaf nodes in the field tree correspond to terminal fields, nodes connected to the root are super-fields and other nodes correspond to median fields. For example, the path <SPORTS/Water Sports/Swimming> describes super-field <SPORTS> having subfield <Water Sports>, and terminal field <Swimming> [8,11]. In figure (1) the super field "Arabic News" , and medium fields "Medicine ,

Policy, Sport, Education, Economy, Community, Weather, Water sports", and terminal fields "long distance, short distance".

Figure (1- a): Arabic field tree

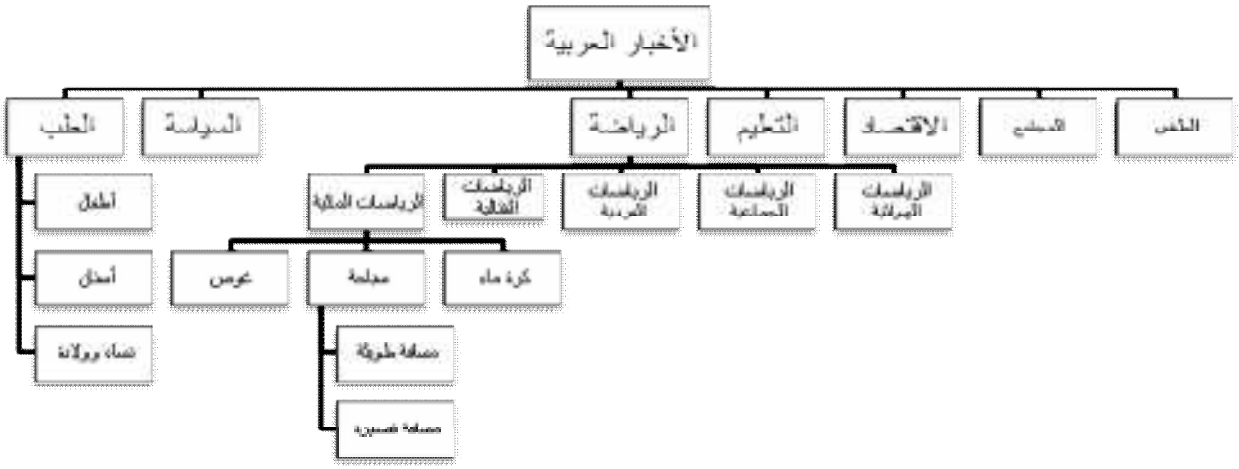
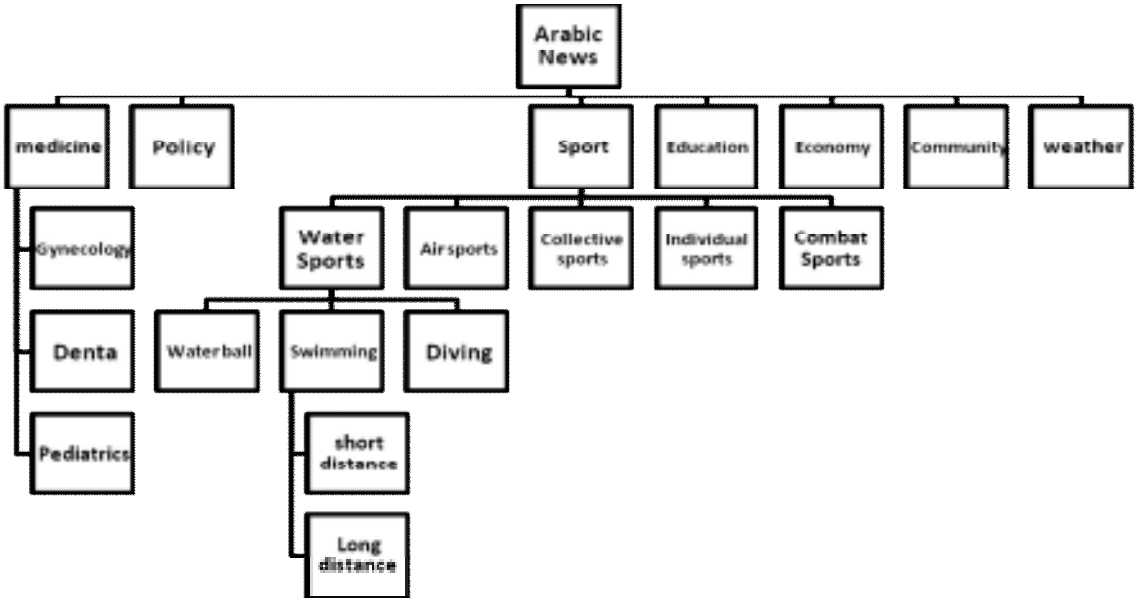


Figure (1- b): the translated field tree in English



Definition 2.1

FA words have various scopes to associate with a field; five precision levels are used to classify FA words to document fields.

(Level 1): Perfect-FA words (PFA) associate with one terminal field.

(Level 2): Semi-perfect FA words (SPFA) associate with more than one terminal field in one medium field.

(Level 3): Medium-FA words (MeFA) associate with one medium field only.

(Level 4): Multiple-FA words (MuFA) associate with more than one terminal field and more than one medium field.

(Level 5): Nonspecific FA words (NSFA) do not specify terminal fields or medium. Fields. Nonspecific FA words include stop words (e.g. Articles, prepositions, pronouns).

Table1: examples of field association words

Levels	FA word	Field association path
1- perfect FA words.	"(korhكره" Means Football in English).	mean > -al akhbar \ Alriyadih < الرياضيه \ الاخبار>(< In English.> news\sports<
2-medium FA words.	Alnwaw"النووي" (Means Nuclear in English).	>)news\ Arabic< Means>-al arabih\ al akhbar < العربيه \ الاخبار>(< In English. Mean>-aldolih\ al akhbar <الدوليه \ الاخبار>(< In English.> news\ International <
3-super FA words.	Jehaz "جهاز"(means Device in English).	\altoknolojia Means >-al akhbar <التكنولوجيا\الاخبار>(< In English.> Technology\news<
4-multiple FA words.	nesbh"نسبه"(means ratio in English).	>) Economy \ stocks< Means>-ashum\ al egtesad < اسهم \الاقتصاد>(< In English. Mean>- aldolih \ alsyash <الدوليه \ السياسه>(< In English. > Politics \ International <
5-non-FA words	hum"هم"(means They in English).	

Table 1 explained every word associated with the field by example in level (1) key word " كوره"(korh Means Football in English)associated with one subfield Which is<الرياضه>(Alriyadih means sports In English). In level (2) key word "النووي "(Alnwawi Means Nuclear in English) associated with a few subfield Which is <العربيه> (al arabih Means Arabic In English)and < الدوليه> (aldolih Mean International In English) of super-field <الاخبار> (al akhbarMean news In English). In level (3) key word "جهاز" (Jehaz means Device in English)associated with one super-field< التكنولوجيا>(altoknolojia Mean Technology In English). In level (4) key word "نسبه" (nesbh means ratio in English) associated with a few subfield Which is <اسهم> (ashum means stocks In English) and < الدوليه> (aldolih Mean International In English)of super-field < الاقتصاد > (al egtesad means Economy In

English) and <السياسه> (alsyash means Politics In English)**.** In level (5) key word "هم"((hum means Them in English)unable to specify the fields.

The new idea for use FA it can be applied on different earlier techniques such as, the vector space model, probabilistic model and language model to modify it and became efficient and suitable for Arabic language.

4. Arabic document classification

Arabic has 28 letters, written from right to left. Contradiction English and Arabic, and Arabic have singular and dual and plural forms. Arabic language in the pre processing stage more complex than it was in the case of the English language [4,6,15]. Arabic three genders: feminine, masculine, neutral.[20] Arabic words are generally classified into three main groups: the names, verbs and names of characters in the Arabic language can be derived from other names and deeds and letters. Verbs in the Arabic language are divided into a perfect, perfect duty. Character category includes pronouns, adjectives, weather, kindness, prepositions and input Interrogative[21]. Based on the patterns of "Awzan". Most of the Arabic words can be obtained from the stem or root word [2, 3, 5, 7, 20, 22, 24].

classification is a division of documents to collect all of them shared a recipe similar groups, as a prelude to order them and save them under a single label indicates it [26].

5. Arabic news classification using FA words

Classification of text techniques is used in many applications, including e-mail filtering, mail routing, filtering spam and watch the news and sorting through digital archive, the indexing mechanism, scientific articles, and the classification of the news and search for interesting information on www [13]. These systems are designed to deal with documents written in English. Does not apply to documents written in Arabic . In this paper, we design an algorithm for

classified Arabic documents using field association words. First we need to extract field association words using algorithm 1, after that classify Arabic document using algorithm 2.

Algorithm 1: extract field association words

Let N is the field root, F is the super field , T is the frequency and R is represent the word , let

Normalization (R,< T >)=[$\frac{(R,T)}{<T>}$](1)

Concentration (R,F) = $\frac{\text{normalization}(R,<F>)}{\text{normalization}(R,<N>)}$ (2)

input

(a) R , for FA , word

(b) normalization(R,<F>) for R and for <F>

(c) threshold a ,to judge FA word ranks

(d) find tree

Output

FA word and their ranks for R.

Step 1 : Select of perfect FA words.

For the root= <N> ,the child field = <N/F> of the field tree

If (R ,<N>) $\geq \alpha$ (3)

So, AF word its perfect , if formula (3) is full field , <N/F> is perfect by <N>

And the same referred is carried out on the field <N/F> .

by repeating the Same selection operation , if <N/F> becomes terminal field, R is selected as perfect FA word in the field <N/F>. if the field <N/F> cannot full field the conation in formula(3) , operation enter to step 2 .

Step 2: selection of semi- perfect FA word if R is not selected as a perfect FA word in the field <N/F>, terminal field has not been reached. therefore ,the field <N> should be as medium field and has at least 2 or more (m ≥2) F. From all F <N/Fi> (1< i <m)of the medium field <N > , Calculate the average value of i times F including word R as in the following :-

$$\left[\frac{\sum_{i=1}^{m} normalization(R,<Fi>)}{m}\right] (4)$$

Accumulated concentration (R , <N/Fi>) ratio for F has higher normalized frequencies then the average value formula (4) .

If the accumulated concentration ratio of i times(1< i<m) exceeds α and the F<N/Fi> are all terminal fields ,R is judged as a semi-perfect FA word in field <N/Fi>, if accumulated value does not exceed the threshold α , R is selected as a medium FA word of field <N>.

Algorithm2: FA word classification algorithm.

Input:

a)T={t1,t2,.....,tn} collection of FA word.

b)B={b1,b2,.....,bm} set of not sorted document.

Output:

Y the classification of B

Method:

1 Run Algorithem1 to get the set of FA words T.

2 T=T union of collection of derivation .

3 Determine Y= {}.

4 Determine Yi= {}.

5 For each Fi belong to F

6 For each Bk belongs to B, m

7 If Ti belong to Bk ,copy Bk to Ti .

8 Y=Y union Yi

9 Else goto step4

10 Return Y.

Example : Consider FA word candidates" دكتوراه" (Doctora- which means PhD in English). as in Figure 1.The number of children fields in <root> is 15 field. We choosed ,< تعليم > (talim - which means Education in English) are subfields A Threshold value α was chosen to be 0.90. In (Step 1), suppose that r is"دكتوراه" and < N> is <root>. The word "دكتوراه" appears the most frequently in the selecting field ,< تعليم > then calculate the concentration ratio of the field <F> = <تعليم> on the field <N/F> = <root/ ,< تعليم > >

Concentration(< تعليم >, "دكتوراه") =0,90

Repeating the same process, select terminal field <الدكتوراه > (AL Doctora- which means PhD in English). in the medium field <التعليم> where " دكتوراه"

appears the most frequently. As the determination is made only in the terminal field <F> =<الدكتوراه> and the concentration ratio is (0,40).

If Concentration(R,F) = $\frac{\text{normalization}(R,<F>)}{\text{normalization}(R,<N>)} \geq \alpha$

then, r is a perfect FA word, means R is associate with only one subfield.

Else

if (conc (R,<N>) $\geq \alpha$ ^ conc (R,<N/F>)) < α

then, R is a semi perfect FA word, means R associate with more than one subfield.

Else

R is a medium FA words if it is associated with one super-field.

So the word "دكتوراه" is determined as a semi perfect FA word in the terminal field .

So, when applying Algorithm 2 after this algorithm all documents that will check and have the same word as a semi perfect FA word return to the same field. Otherwise, if a document has the word as a perfect , a semi perfect or medium FA word then one or more children field will appear .

6. Experimental Evaluation Using FA Words Classification Algorithm

Our experiments trained the system using Arabic news documents collected from the Internet. It mainly collected from Al-jazeera Arabic news channel which is the largest Arabic site, Al-Ahram newspaper, Al-watan newspaper, Al Akhbar, Al Arabiya and Wikipedia the free encyclopedia. The documents categorized into 8 super-field and 52 subfields. The number of files in our corpus is 786 file and it is about 5.6 MB.

For experimental evaluations, we used software written by JAVA with three versions from paper [14]. A classification on Arabic text using FA words was made. The application window is shown in Figure (3).

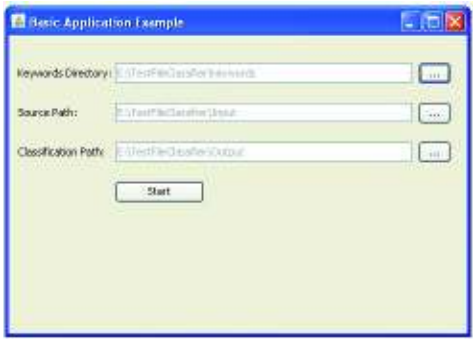


Figure (2) Application window

Simulation results for classification

Input data: (keywords, text)

Output: classified data according to keywords.

We have used about 150 keywords selected by human from corpus.

Precision, Recall and F-measure are used to estimate relevancies of the presented methods and defined as follows :

$$\mathrm{Re}call(R) = \frac{Correct \cdots Classified \cdots Documnts}{Totall \cdots Corrected \cdots Classified}$$

$$\mathrm{Pr}ecison(P) = \frac{Correct \cdots Classified \cdots Documnts}{Totall \cdots \mathrm{Re}trieved \cdots Classified}$$

$$F - measure = \frac{2 \times P \times R}{P + R}$$

Precision, Recall and F-measure for six super-fields are measured using FA words. From the evaluation results it turns out that the best performance is recorded in classification with FA-words as shown in Table 2.

Table 2: Classification using FA words

Name of field	Precision	Recall	F- measure
الطب(al Teb- which means the Medicinein English)	0.72	1	0.8
الرياضة(al Ryadah- which means sport in English)	0.74	0.69	0.71
السياسة(al Siasa- which means the Policy in English)	0.67	1	0.8
التكنولوجيا (al tecnologia- which means technology in English)	0.44	0.1	0.6
التعليم(al Taleem- which means the Education in English)	0.5	0.9	0.64
الاقتصاد(al Iqtesad which means Economy in English)	0.98	0.6	0.74

According to F-measure in table 2 the medium for F-measure equal to 75%

7. Conclusion and future work

FA words are used to classify Arabic documents. Words are extracted from these document corpora to get FA word candidates. Furthermore, we used the FA classifier with our modification to refine Arabic document classification. From the experiential results, and the presented software can be automatically classifying Arabic news. F-measure is 81% of classification using FA words.

Future work could focus on automatic building of Arabic field association words using morphological analysis.

References

[1]Abdullah Wahbeh, Dakota State University, USA, Mohammed Al-Kabi, Yarmouk University, Jordan , Qasem Al-Radaideh, Yarmouk University, Jordan , Qasem Al-Radaideh, Yarmouk University, Jordan, Izzat Alsmadi, Yarmouk University, Jordan, "The Effect of Stemming on Arabic Text Classification", 54 International Journal of Information Retrieval Research, 1 (3), 54-70, July-September (2011).

[2]Arthur W. & Saad M., "Arabic Text Classication Using Decision Trees",Proceedings of the 12th international workshop on computer science and information technologies CSIT, pp. 75-79,(2010).

[3]Arthur W. & Saad M., "Arabic Morphological Tools for Text Mining",(2010).

[4]Al-Harbi S., Almuhareb A., Al-Thubaity A., Al-Rajeh A. & Khorsheed M.,"Automatic Arabic text classication",(2008).

[5]Aljlayl M. & Frieder O.,"On Arabic search improving the retrieval elctiveness via a light stemming approach", Proceedings of the eleventh international conference on Information and knowledge management, pp. 340-347, ACM,(2002).

[6] Al-Refai M., Duwairi R.& Khasawneh N.,"Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization",Proceedings of 4th International Conference on Innovations in Information Technology, pp. 446-450, IEEE, (2007).

[7]Al-Salamah A. I. & Tayli M.,"Building bilingual microcomputer systems",Communications of the ACM, vol. 33, no. 5, pp. 495-504,(1990).

[8] Atlam E., Morita K., Fuketa M. & Aoe J. "A new method for selecting English field association terms of compound words and its knowledge representation", Information Processing and Management,(38),807-821,(2002).

[9] Atlam E., Fuketa, M., Morita, K. & Aoe J., "Document Similarity measurement using Field association terms', Information Processing & Management Journal, 39(6), 809-824, (2003).

[10]Atlam, E.,Elmarhomy G., Morita K.,Fuketa M., & Aoe J, "A new algorithm for construction specific field terms using co-occurrence words information", 8th International conference on knowledge-based intelligent information & engineering systems, Wellington, New Zealand, Part 1 (pp. 530–540), (2004).

[11]Atlam E., Elmarhomy G., Fuketa M., Morita K., Sumitomo T. & Aoe J."An automatic filtering method for field association words by deleting unnecessary words", International Journal of Computer and Mathematics,Vol. 83, No. 3, pp 247-262, March(2006).

[12]Atlam E., G. Elmarhomy, Fuketa, M., Morita, K. & Aoe, J.,"Automatic Building of New Field Association Word Candidates Using Search Engine", Information Processing & Management Journal, 42 (4),951-962, (2006).
[13]Atlam E,M. Abd El-Monsef, M. Amin, O. El-Barbary," Arabic Document Classification:A Comparative Study",JOURNAL OF COMPUTING, VOLUME 3, ISSUE 4, APRIL (2011).
[14]Atlam E.,Abd El-Monsef M.,Amin M.,El-Barbary O.,"Field Association words with Naive Bayes Classifier based Arabic document classification",International Journal of Computer Science Issues,Vol. 8, Issue 3, No. 2, May (2011).
[15]Darwish N., Hegazy N., Said D.& Wanas N.,"A study of text preprocessing tools for arabic text categorization" , Proceedings of The Second International Conference on Arabic Language, pp. 230-236,(2009).
[16]Darwish K., & MagdyW,"Arabic Information Retrieval", Foundations and Trends R in Information Retrieval, vol. 7, no. 4, pp. 239-342, (2013).
[17] El-Halees A. M. (2007). "Arabic Text Classification Using Maximum Entropy ". In The Islamic University Journal (Series of Natural Studies and Engineering) .Vol 15, No. 1 Jan. pp. 157-167.
[18]Elmarhomy Ghada, Elsayed Atlam, Masao Fuketa, Kazuhiro Morita and Jun-ichi Aoe," An Automatic Filtering Method for Field Association Words by Deleting Unnecessary Words" Department of Information Science and Intelligent Systems University of Tokushima, Tokushima 770-8506.
[19]Feldman R.,Sanger J.,"the text mining handbook: advanced approaches in analyzing unstructured data", (2007).
[20]Khoja S., Garside R., & Knowles G.,"A tagset for the morphosyntactic taggingof Arabic" Proceedings of the Corpus Linguistics. Lancaster University (UK), vol. 13, (2001).
[21]Khoja S., "APT: Arabic part-of-speech tagger", Proceedings of the Student Workshop at NAACL, pp. 20-25, (2001).
[22]Mesleh A., "Support Vector Machine Text Classier for Arabic Articles:Ant Colony Optimization-Based feature selection,(2008).
[23]Mohamed EL KOURDI, Amine BENSAID, Tajje-eddine RACHID, Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm, School of Science & Engineering Alakhawayn University
[24] Nwesri A., Scholer F.& Tahaghoghi S.,"Capturing out-of-vocabulary words in Arabic text", Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 258-266, Association for Computational Linguistics,(2006).
[25] Vannevar B.,"As we may think", Atlantic Monthly, 176:101-108, July (1945).
[26] www.abahe.co.uk "Arab British Academy for Higher Education .