Toward an ARABIC Stop-Words List Generation

A. Alajmi
Communication & Electronics Dept., Faculty of Engineering, Helwan University, Egypt.

E. M. Saad
Communication & Electronics Dept., Faculty of Engineering, Helwan University, Egypt.

R. R. Darwish
Mechatronic Dept.,Faculty of Engineering, Helwan University, Egypt

ABSTRACT

Over the past decades systems for automatic management of electronic documents have been one of the main fields of research. Text processing is a wide area that includes many important disciplines. In the processes of organizing unstructured text in order to implement a mining technique, preprocessing has to be applied. One of the most important preprocessing techniques is the removal of functional words which affects the performance of text mining tasks. In this paper, a statistical approach is presented to extract Arabic stop-words list. The extracted list was compared to a general list. The comparison yield an improvement in an ANN based classifier using the generated stop-words list over the general list.

General Terms

Natural Language Processing, Text Processing.

Keywords

Arabic Text Processing, Stop-word List Generation.

1. INTRODUCTION

Preprocessing are the processes of preparing data for the core text mining task. These processes convert the documents from original data source into a format which is suitable for applying various types of feature extraction methods against these documents to create a new collection of documents fully represented by concepts [1]. The preprocessing phase includes all those routines, processes and methods required to prepare data for a text mining system which is the core of knowledge discovery operations.

Text mining preprocessing operations are centered on the identification and extraction of representative features for natural language documents. There are two goals of preprocessing phase. First, is to identify features in a way that is most computationally efficient and practical for pattern discovery. Second, is to capture the meaning of a document accurately; on the semantic level [1].

In order to transform from irregular form in to structured representation; features must be identified. There are a vast number of words, phrases, sentences, typographical elements and layout artifacts that a short document may have. Furthermore, it is necessary to filter out noise from important text; noise is an extraneous text that is not relevant to the task at hand [2]. Stop-words are example of noise in data (functional words and general common words of the language that usually do not contribute to the semantics of the documents and have no read added value.

This paper is organized such that section 2 illustrates characteristics and description of the stop-word lists and discusses stop-word elimination techniques. Section 3 reviews the research work devoted to stop-words list generation. Section 4 presents the technique for generating an Arabic stop-words list. The simulation results are described in section 5. Finally, this paper is concluded in section 6.

2. STOP-WORD ELIMINATION

Stop words are words having no significant semantic relation to the context in which they exist [3]. Stop words are the terms that occur frequently in most of the documents in a given collection. They are extremely common words that would appear to be of little value in helping select documents that matches a user need. Thus, they must not be included as indexing terms. Most of those words are irrelevant to the categorization task and can be dropped with no harm to the classifier performance, and may even result in improvement due to noise reduction [1]. However, stop-words cannot be included in the feature space, because words are isolated and taken out of their context when text is represented by the bag-of-words method.

An efficient stop-word removal technique is needed in many natural languages processing application such as: spelling normalization, stemming and stem weighting, and in Information Retrieval systems (IR) [4]. Most TC systems remove the stop- words, and many systems perform a much more aggressive filtering, removing 90 to 99 percent of all features [1].

The elimination of stop words also reduces the corpus size typically by 20 to 30% which leads to higher efficiency [3]. The general trend in IR systems has been the use of quite large stop lists (200–300 terms) - due to morphological richness of the language; the list contains all possible morphological variants of each stop-word- to very small stop lists (7–12 terms) to no stop list whatsoever [4,5].

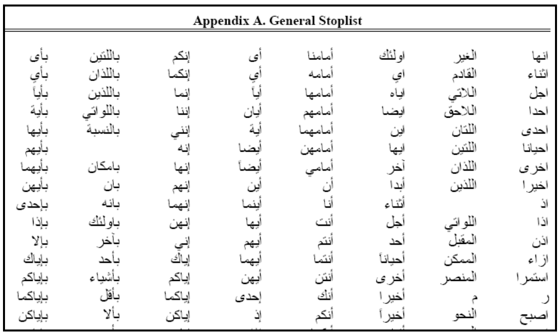
For example, in English articles the propositions such as “the,” “on,” and “with” are usually stop words. Stop-words may also be document-collection specific [3], for example, the word “blood” would probably be a stop word in a collection of articles addressing blood infections, but certainly not in a collection describing the events of World Cup. Subsequently, many words that occur frequently are eliminated.

Eliminating such words from consideration early in automatic indexing speeds processing, saves huge amounts of space in indexes, and does not damage retrieval effectiveness [4]. Two related facts were noticed in the early days of IR [6]. First, a relatively small number of words account for a very significant fraction of all text’s size. Words like "IT", "AND","THE" and "To" can be found in virtually every sentence in English-based documents. Secondly, these words make very poor index terms, with which users are indeed unlikely to ask for documents.

The general strategy for determining a stop list is to calculate the total number of times in which each term appears in the document collection then sort the terms by *collection* *frequency*, and then to take the most frequent terms. The selected terms are often hand-filtered for their semantic content relative to the domain of the documents being indexed, and marked as a *stop list*. Stop list are then discarded during indexing. Figure 1 depict an example of English stop-word list, whereas, figure 2 shows an Arabic stop-word list example.

a	an	and	are	as	at	be	by
for	from	has	he	in	is	it	its
of	on	that	the	to	was	were	will

Figure 1 A stop list of twenty-five semantically nonselective English words that are common in Reuters-RCV1.



Appendix A. General Stoplist

بأى	باللتين	إنكم	أى	أمامنا	اولئك	الغير	انها
بأي	باللذان	إنكما	أي	أمامه	اي	القادم	اثناء
بأياً	باللذين	إنما	أياً	أمامها	اياه	اللاتي	اجل
بأية	باللواتي	إننا	أيان	أمامهم	ايضا	اللاحق	احدا
بأيها	بالنسبة	إنني	أية	أمامهما	اين	اللتان	احدى
بأيهم		إنه	أيضا	أمامهن	ايها	اللتين	احيانا
بأيهما	بامكان	إنها	أيضاً	أمامي	آخر	اللذان	اخرى
بأيهن	بان	إنهم	أين	أن	أبدا	اللذين	اخيرا
بإحدى	بانه	إنهما	أينما	أنا	أثناء		اذ
بإذا	باولئك	إنهن	أيها	أنت	أجل	اللواتي	اذا
بإلا	بآخر	إني	أيهم	أنتم	أحد	المقبل	اذن
بإياك	بأحد	إياك	أيهما	أنتما	أحياناً	الممكن	ازاء
بإياكم	بأشياء	إياكم	أيهن	أنتن	أخرى	المنصر	استمرا
بإياكما	بأقل	إياكما	إحدى	أنك	أخيرا	م	ر
بإياكن	بألا	إياكن	إذ	أنكم	أخيراً	النحو	اصبح

Fig 2. Example of Arabic stop-word list.

Generally, stop-word means high frequency and low discrimination and should be filtered out in the IR system. In the same way, the concept of stop-word in the text mining is similar, but the ability to characterize a document is attached much more importance to judge whether a word is a stop-word or not. However, it is not the best practice to extend the stop-word list as large as possible [7], but on the opposite, to increase the recall rate in IR. As to the text mining process, all the word identified as stop-word should be filtered out to improve the efficiency and accuracy, therefore the main concept is to improve the efficiency and accuracy of the text mining task after the stop-words were filtered out. Two aspects [7] are involved:

- The accuracy of text mining should not be decreased if the stop-words were deleted.
- The dimensionality of the text feature space should be reduced if the stop-words were deleted.

Other ways to construct a stop-word list [7] includes artificial pattern and the entropy calculation. In this section a review of selected publications related to stop-words list generation techniques.

3. RELATED WORK

Stop-words are functional, general, and common words of the language that usually do not contribute to the semantics of the documents and have no read added value. Many categorization systems attempt to exclude stop-words from the list of features to reduce feature space, and to increase classifier performance. Here, a review on research related to sop-words list generation.

Yao and Zen-wen [7] constructed a Chinese stop-word list. The stop list is obtained by merging the classical stop-word list with the stop-words depending on the different domain of the text document corpus. A customizing Chinese-English stop-word list containing 1289 words was constructed.

Savoy [8] defined a general stop-word list for those words which serve no purpose for retrieval, but are used very frequently in composing the documents. They establish a general stop word list for French following the guidelines described in (Fox, 1990). First, all the word forms appearing in their French corpora is sorted according to their frequency of occurrence and extract the 200 most frequently occurring words. Second, all numbers (e.g., "1992", "1"), plus all nouns and adjectives more or less directly related to with the main subjects of the underlying collections is removed. Third, some non-information-bearing words, even if they did not appear in the first 200 most frequent words are included (i.e. they added various personal or possessive pronouns (such as "moi" (me), "tien" (yours)), prepositions ("dessus" (upon)) and conjunctions ("cependant" (however)). The suggested French general stop-word list contains 215 words, and by using such a stop-word list, the size of the inverted file was reduced by about 21% for one test collection, and about 35% for the second corpus.

Hao and Lizhu [9] gave a refined definition for stop words in Chinese text classification from a perspective of statistical correlation theory. The identification of the stop-words list was based on the weighted Chi-squared statistic.

Myerson [6], stated that two conditions a word must satisfy in order to be a stop-word. First it should have a high document frequency (DF). Second, the statistical correlations with all the classification categories should be small. The $\chi 2$ (weighted Chi-squared statistic) was used to measure statistical correlation between a word and classification categories. X^2 for the words is calculated then ordered increasingly. Consecutively, the first word in the ordered list has the minimum value of weighted Chi-squared statistic, i.e. it has a higher document frequency and lesser correlations with all the categories. Chinese corpus of the Mayor's Public Access Line Project texts were used to evaluate and, compare results of classifiers constructed by deleting and retaining the stop words. Thus concluded that the stop word list involving 500 words constructed can reduce the words by 43% of all the words in corpus, and the micro-average F1 improves nearly 7% from 81.39% to 88.76%.

Zheng and Gaowa [10] proposed a method for constructing stop-words list based on entropy calculation for Mongolian language. First, is to determine initial stop word lists then the entropy of every word is calculated and then ordered ascending to entropy. The second step is to combine results with the Mongolian part of speech to produce the final stop-word list.

Zou et al. [11] used an aggregated model to measure both the word frequency characteristic by statistical model and its information characteristic by information model. An approach has been developed based on the idea that stop words are ranked at the top with much larger frequency than the other words. And at the same time they have a stable distribution in different documents. A combination of these two observations redefines the stop words as those words with stable and high frequency in documents. The generated list was compared with other existing lists and showed an improvement over others.

Elkhair [12] conducted a comparative study on the effect of stop words elimination on Arabic IR. Three stop lists were used in the comparison. General stop-list, corpus based stop-list and, a combined stop-list. First list, the general stop-list was created based on Arabic language structure characteristics without any additions and, it consists of 1,377 words. Second lists, corpus statistics has 359 words. It depends on words frequency, such that words occurring more than 25,000 times are added to the list. Third list, combines general and corpus-based stop-list together and, it results in 1529 words. It was concluded that general stop-list performed better than the other two lists.

Sinka and Corne [14] developed new word Entropy based stop lists. Three stop-words list were extracted one from random web pages, second from the BankSearch dataset and, third obtained from unsupervised clustering experiments. It was concluded that existing stop-lists perform well, but are sometimes outperformed by the new stop-lists, especially on hard classification tasks. The generated stop list was compared with Van Rijsbergen's stop list and Brown's stop list. The top 10 words were; *'the', 'and', 'to', 'of', 'in', 'for', 'on', 'is', 'with'* and *'by'*. The Experiment was run on different stop-list size n. *n* = 319, n = 425, no stop list, *n* = 500 and *n* = 1000. Results when using k-means clustering where *k* = 2 on a collection of 2,000 documents containing 1,000 documents. The results show poor clustering accuracy when no stop-list used. Furthermore, the Van Rijsbergen and Brown stop lists leads to good clustering performance. Finally, the highest achieved accuracy from the total of 1,400 trials was attained when using the top 50 words of the suggested stop-list.

Alhadidi and Alwedyan [4] implemented a hybrid stop-word removal technique for Arabic language based on a dictionary and an algorithm. The proposed technique has been tested using a set of 242 Arabic abstracts chosen from the Proceedings of The Saudi Arabian National Computer conferences, and another set of data chosen from the Jordanian Alrai Newspaper.

4. PROPOSED ARABIC STOP-WORD LIST

Arabic is very rich in lexical tokens, that means stop-words are available in big quantities. Stop-words in Arabic have certain properties [4].

- They have no meaning if they are used separately.
- Appear many times in a text.
- Necessary for the construction of the language.
- Mostly adjectives.
- General words and not particularly used in a certain field.
- Not used as a search keyword.
- Never form a full sentence when used alone.
- Stop-words in Arabic include some of grammatical links such as the definite article (AL)(the), attached and separate prepositions, conjunctions, interrogative words, negative words, exclamations and calling letters, adverbs of time and place, also they include all the pronouns, demonstratives, subject and object pronouns, the Five Distinctive Nouns, some numbers, additions and verbs. Stop-words may be separate or attached ones in a form of prefixes or suffixes.

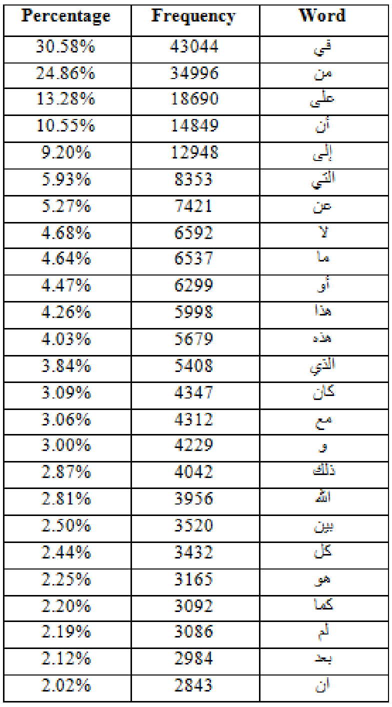
There exists a general Arabic stop-words list; however, due to the highly inflectional nature of Arabic language those words may come in different forms according to prefixes and suffixes attached to them.

This work exploit the generation of an Arabic stop-word list that is based on [11].

Step 1: Word Frequency Calculation

Word frequency is the number of times a word occurs in a document. A list of words and their frequencies are shown in Table 1. The list is sorted in descending order according to the frequency. Notice how functional word ("في", in) appears on top of the list.

Table 1. Top 25 Arabic words with highest Frequencies



Percentage	Frequency	Word
30.58%	43044	في
24.86%	34996	من
13.28%	18690	على
10.55%	14849	أن
9.20%	12948	إلى
5.93%	8353	التي
5.27%	7421	عن
4.68%	6592	لا
4.64%	6537	ما
4.47%	6299	أو
4.26%	5998	هذا
4.03%	5679	هذه
3.84%	5408	الذي
3.09%	4347	كان
3.06%	4312	مع
3.00%	4229	و
2.87%	4042	ذلك
2.81%	3956	الله
2.50%	3520	بين
2.44%	3432	كل
2.25%	3165	هو
2.20%	3092	كما
2.19%	3086	لم
2.12%	2984	بعد
2.02%	2843	ان

Step 2: Mean and variance Calculation

First, we measure the mean of probability (MP) of each word in individual document. Suppose there are M (140781) distinct words and N (1002) documents all together. Denote each word as w_j (j=1… M), and each document as D_i (i=1… N). For each word w_j , calculate its frequency in document D_i denoted as $f_{i,j}$. Then normalize the document length, by calculating the probability $P_{i,j}$ of the word w_j in document D_i.

$$P_{i,j} = \frac{\text{Frequency in the document } D_i}{\text{The total number of words in document } D_i}$$

For each word w_j, the MP among different documents is summarized in formula (1):

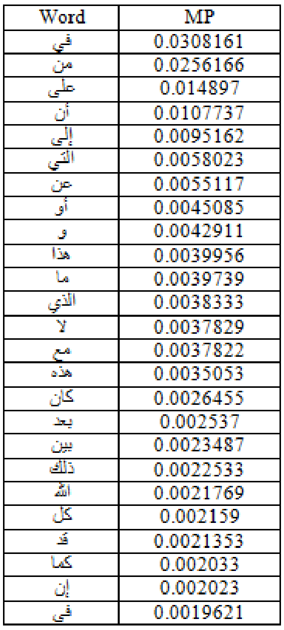
$$MP(w_j) = \frac{\sum_{i=1}^{N} P_{ij}}{N} \qquad (1)$$

Second, the variance of probability (VP) of each word is calculated. The calculation comes from the fact that stop-words should have high MP as well as stable distribution. Based on the calculation of probability, the VP can be defined by formula (2)

$$VP(w_j) = \frac{\sum_{i=1}^{N}(P_{ij} - \overline{P_{ij}}\ )^2}{N} \qquad (2)$$

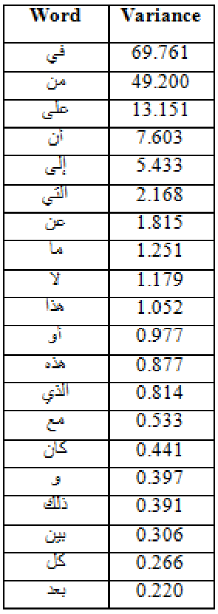
Where $\overline{P}_{i,j}$ = frequency of w/ total number of words over all documents

Table 2. Top 20 Arabic words Mean Probability



Word	MP
في	0.0308161
من	0.0256166
على	0.014897
أن	0.0107737
إلى	0.0095162
التي	0.0058023
عن	0.0055117
أو	0.0045085
و	0.0042911
هذا	0.0039956
ما	0.0039739
الذي	0.0038333
لا	0.0037829
مع	0.0037822
هذه	0.0035053
كان	0.0026455
بعد	0.002537
بين	0.0023487
ذلك	0.0022533
الله	0.0021769
كل	0.002159
قد	0.0021353
كما	0.002033
إن	0.002023
فى	0.0019621

Table 3. Top 20 Arabic words Variance Probability.



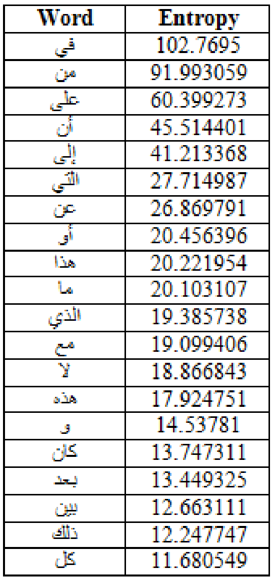
Word	Variance
في	69.761
من	49.200
على	13.151
أن	7.603
إلى	5.433
التي	2.168
عن	1.815
ما	1.251
لا	1.179
هذا	1.052
أو	0.977
هذه	0.877
الذي	0.814
مع	0.533
كان	0.441
و	0.397
ذلك	0.391
بين	0.306
كل	0.266
بعد	0.220

Step 3: Entropy Calculation

Entropy is a measure of information, and is invaluable throughout speech and language processing [13]. Thus Entropy measures the information value of the word w_j. Probability $P_{i,j}$ is the word frequency in document D_i divided by the total number of words in document D_i. The entropy value (H) for word w_j is calculated as in formula (3):

$$H(w_j) = \sum_{i=1}^{N} P_{ij} \times \log\left(\frac{1}{P_{ij}}\right) \qquad (3)$$

Table 4. Top 20 Arabic words based on Entropy calculation



Word	Entropy
في	102.7695
من	91.993059
على	60.399273
أن	45.514401
إلى	41.213368
التي	27.714987
عن	26.869791
أو	20.456396
هذا	20.221954
ما	20.103107
الذي	19.385738
مع	19.099406
لا	18.866843
هذه	17.924751
و	14.53781
كان	13.747311
بعد	13.449325
بين	12.663111
ذلك	12.247747
كل	11.680549

Once the entropy of each word in the dataset has been calculated, the resulting list can be ordered by ascending entropy to reveal the words that have a greater probability of being noise words . Similarly to statistical model, one ordered list is prepared for further aggregation. The higher entropy the word has, the lower information value of the word is. Therefore, the words with lower entropy are extracted as candidates for stop words [11].

Step 4: Aggregation

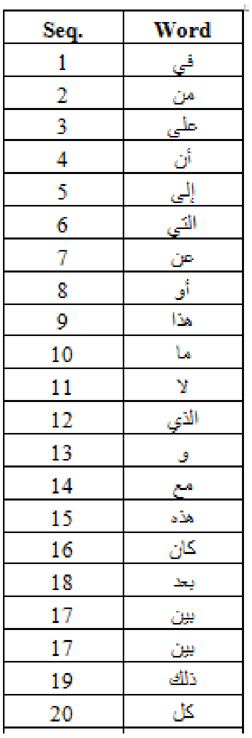
The features of stop words are revealed in different aspects by the three generated ordered lists. One of the popular solutions to it should be Borda's Rule [6], which covers all the binary relations even when many members of a population have a cyclic reference given a set of voters.

The Borda count is a single-winner election method in which voters rank candidates in order of preference. The Borda count determines the winner of an election by giving each candidate a certain number of points corresponding to the position in which he or she is ranked by each voter. Once all votes have been counted the candidate with the most points is the winner. The number of points given to candidates for each ranking is determined by the number of candidates standing in the election. Thus, under the simplest form of the Borda count, where there are n candidates a candidate will receive n points for a first preference, n – 1 points for a second preference, n – 2 for a third, and so on, as shown in the following example:

Ranking	Candidate	Formula	Points
1st	Andrew	(n)	5
2nd	Brian	(n – 1)	4
3rd	Catherine	(n – 2)	3
4th	David	(n – 3)	2
5th	Elizabeth	(n – 4)	1

When all votes have been counted, and the points added up, the candidate with most points wins.

Table 5 Top 100 Arabic words after applying "Borda" ranking

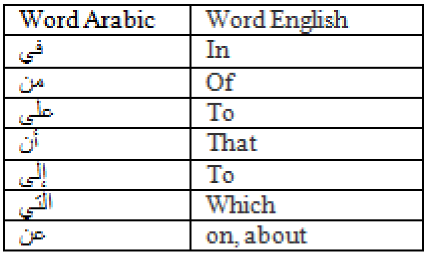


Seq.	Word
1	في
2	من
3	على
4	أن
5	إلى
6	التي
7	عن
8	أو
9	هذا
10	ما
11	لا
12	الذي
13	و
14	مع
15	هذه
16	كان
18	بعد
17	بين
17	بين
19	ذلك
20	كل

5. EXPERIMENTAL RESULT

The experiment used a corpus which contains N=1002 documents. The document collection contains over 700,000 words and includes M=140781 different words. Documents text was tokenized, and non-Arabic alphabets were removed. It was noticed that words like ("في", In), ("من", who) ,("على", on) ,("أن", that),("إلى", to), and ("التي",which) are in the top of the three lists (Mean probability, Variance, Entropy) table 6.

Table 6 Top stop words in the three lists



Word Arabic	Word English
في	In
من	Of
على	To
أن	That
إلى	To
التي	Which
عن	on, about

The top 100 words ranked by Borda count are shown in table 5. Some words are written in a different form such as (who - "اللذان" and "اللّذان") with and without "Shada". Comparing the top 25, 50, and 100 words in the Borda list with the general list in Arabic and English are found in table 7.

Table 7 a comparison between the generated stop list and generalized Arabic and English stop list

Number of words at the top of the list	Overlapping of Arabic General list and proposed Stop List	Overlapping of English stop list and proposed Stop List
25	100%	96%
50	94%	95%
100	94%	92%

One of the advantages of the proposed stop-word list is that it captures the inflection occurring on a word. For example a word "هو" also exists with the prefix "و". Another Example is the word ("كان", was) is also captured by different inflections (تكون -she is, يكون -he is, كانت -she was). However, some words are found which may not be considered as a potential stop word such as (العربيةArabic -Saudi -السعودية -united المتحدة). The previous words might be stop words in some documents but not all; therefore it was removed from the list.

A comparative study was conducted on the use of the general stop list and the generated stop lists to a classification process. The document text was stemmed using the stemmer in [15]. Then classified using Artificial Neural Network classifier. The result shows that the top 200 words generated list (manually edited) outperformed the general list with a 96% efficiency of the classifier, versus 90% when using the generalized list.

6. CONCLUSION

Preprocessing is an important task for text processing systems. The nature of the processed documents forces the need for different techniques to clean and structure the textual data. A corpus of 1000 document was used in this experiment. Those documents comprise different categories of Arabic text. Stop-words are functional and general words of the language that usually do not contribute to the semantics of the documents and have no read added value. The removal of such words should contribute to the improvement of classifier efficiency. In this paper a method based on information statistics was examined to create an Arabic stop-word list. Comparison with a general Arabic stop-word list shows that the presented list outperforms the generalized list in term of text categorization task.

7. REFERENCES

[1] R. Feldman, and J. Sanger, "The text mining handbook", Cambridge university press, 2007.

[2] R. Nisbet, J. elder, G. Miner, "Handbook of statistical analysis and data mining applications", academic Press, Elsevier, 2009.

[3] M. Khosrow, "Encyclopedia of Information Science and Technology", Information Sci, Second Edition, 2009.

[4] B. Alhadidi, and M. Alwedyan,"Hybrid Stop-Word Removal Technique for Arabic Language", Egyptian Computer Science Journal Vol. 30 No. 1 January 2008.

[5] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval", Cambridge university press, 2008.

[6] R.B. Myerson, "Fundamentals of social choice theory", Discussion Paper No.1162, 1996.

[7] Z. Yao, and C. Ze-wen, "Research on the construction and filter method of stop-word list in text Preprocessing", Fourth International Conference on Intelligent Computation Technology and Automation, 2011.

[8] J. Savoy, "A Stemming Procedure And Stopword List For General French Corpora", Journal of the American Society for Information Science, 50(10), 1999, 944-952.

[9] L. Hao, and L. Hao, "Automatic Identification of StopWords in Chinese Text Classification", International Conference on Computer Science and Software Engineering,2008.

[10] G. Zheng, and G. gaowa, "The Selection of Mongolian Stop Words", IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), 2010.

[11] F. Zou, F. L. Wang, X. Deng, S. Han, and L. S. Wang, "Automatic Construction of Chinese Stop Word List", Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015).

[12] I. A. El-Khair, "Effect of Stop Words Elimination for Arabic Information Retrieval: A comparative Study" , International journal of Computing & Information Sciences, Vol. 4, No. 3, December 2006.

[13] Y. Kadri; J.Y. Nie, "Effective Stemming for Arabic Information Retrieval," International conference at the British Computer Society, London, 23 October 2006; pp.68-74.

[14] M.P. Sinka, and D.W. Corne, "Towards Modernised and Web-Specific Stoplists for Web Document Analysis", Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI’03) ,2003.

A.Alajmi, E. Saad, and M. Awadallah, "Arabic Verb Pattern Extraction", 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)