Field Association words with Naive Bayes Classifier based Arabic document classification

M. E. Abd El-Monsef [1], El-Sayed Atlam[2], M. Amin[3], O. El-Barbary[4]

[1, 2, 4] Mathematics Department, Faculty of Science, Tanta University
Tanta, Egypt

[2] Faculty of Engineering, Tokishima University
Tokishima, Japan

[3] Mathematics Department, Faculty of Science , Menofia University
Shebin El-Kom, Egypt

Abstract

Document classification aims to assign a document to one or more 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 Our experimental study shows that using FA algorithm with Naïve Bayes (NB) Classifier gives the ~ 79% average accuracy and, using compound FA words with NB classifier gives ~ 89% average accuracy of the training documents.

Keywords: *Field association words; compound Field association words; Arabic Information retrieval; precision; recall; F-measure; Naïve Bayes theorem.*

1. Introduction

For to the rapid growth of the number of online, Several document retrieval systems have been developed along in many areas such as keyword retrieval, similar file retrieval, automatic document classification, cross language retrieval and document summarization [8, 10, 11, 12, 15, 16, 18, 19, 20].

Arabic Information Retrieval (AIR) has recently become a focus of research and commercial development. Arabic documents, automatic Arabic document categorization become an important task. Very few standards for evaluation of such tools are known and available. A concrete evaluation for AIR systems is necessary for the advancement of this field.

Document Categorization (DC) is a significant due to several information systems have to deal with various numbers of documents; for this, there must be techniques and tools that can be used to classify those documents to their categories efficiently. Naïve Bayes algorithm classifier is DC techniques, which are used to classify documents in different languages. Most of the work in this area was carried out for the English text, on the other hand very few researches have been carried out for the Arabic text. [4] use K-NN algorithm and Naïve Bayes algorithm in order to make a practical comparison between them and previous studies, also [17] suggests the use of rooting algorithm with Naïve Bayes Classifier and reports the algorithm performance in terms of error rate, accuracy, and micro-average recall measures. And the experimental study shows that using rooting algorithm with (NB) Classifier gives ~62.23% average accuracy. [15] Used Naive Bayes (NB) to classify non-vocalized Arabic web documents and the overall classification achieves about 62%.

In this paper we use the naïve bayes classifier with FA words to Arabic document classification. Field Association (FA) word is a word that can be related to the field. Based on specific FA word information, topics of documents (e.g. computer science, algorithm) can be recognized. Some keywords appear frequently in texts and relate strongly to text topics. The research documented in this paper focuses on finding Arabic FA words automatically. Moreover, using it in naïve bayes classifier to classifying Arabic document. To do so, a large quantity of document data for building FA word candidates is necessary. This paper presents a system for automatically finding FA words by adding new FA word candidates which change over time to an FA word dictionary. There are several research developed in FA words area, some of these researches introduced this concept and gave algorithms to find FA terms in a specific field [10], other papers used FA terms in document similarity [6] and other papers used FA terms in passage retrieval.

The reset of paper organized as follows. Section 2 of this paper presents Arabic word formation. Also, defines Compound FA words in Arabic language. Section 3 defines

Arabic document field tree. Section 4 identifies Field association words, its levels and how we determine it. Section 5 is a comparison with traditional classification methods. Section 6 illustrates how we classify Arabic document using FA words and compound FA words. Section 7 discusses Arabic document classification using Naïve Bayes (NB) classifier. Section 8 represents the experimental evaluation and the comparison with other traditional approaches. Section 9 focuses on conclusion and future work.

2. Document Field Association Terms

A Single Field Association (FA) word is a minimum unit (word) which cannot be divided without losing semantic meaning. A Compound FA term consists of two or more single FA term is regarded as being single if it loses field information when divided [1, 2, 3].

Single FA Term Levels

A document field can be ranked as: super-fields, sub-fields or terminal fields. FA terms are grouped according to how well they indicate specific fields. FA terms have different rank to associate with document fields, so five Field levels can be used to classify FA terms according to document fields.

Definition 1

FA words have different level to associate with a field; three useful precision levels are used to classify FA words to document fields, they are: perfect if FA words are associated with one subfield uniquely, semi perfect if FA words are associated with a few subfields of one super- field, medium if FA words are associated with one super-field.

Definition 2

Let (<T>) be the total frequency of all words in the terminal field <T>; let (w, <T>) be the frequency of the word w in the terminal field <T>, the (Normalization (w, <T>)) can be defined as follows:

$$\text{Normalization (w, <T>)} = \left[\frac{Frequency(w,T)}{Total\_Frequency(<T>)} \right] \quad (1)$$

The normalized frequency defines how much a specific word is concentrated in a specific field.

Definition 3

For the parent <S>, the child field <C>, the concentration ratio (Concentration (w, <C>)) of the FA word w in the field <C> is defined as in the following:

$$concentration(w,c) = \left[\frac{Normalization(w,<c>)}{Normalization(w,<s>)} \right] \quad (2)$$

The following algorithm determines FA words by considering their ranks.

Algorithm 1

FA words determination algorithm

Input: W is a set of words, $W = \{w_1, w_2, w_3, \ldots, w_n\}$
Threshold α, to judge FA words ranks,
Document field tree, set of documents ordered as a field

Output: classified FA words according to its field.

Method

If $concentration(w,c) = \left[\frac{Normalization(w,<c>)}{Normalization(w,<s>)} \right] \geq \alpha$

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

Else

if $(\text{conc}(w,<S>) \geq \propto \wedge \text{conc}(w,<S/c>)) < \propto$

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

Else

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

End

3. Naïve Bayes Classifier

A Naïve Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naïve) independence assumptions [13]. In simple terms, a naïve bayes classifier assumes that the presence (or absence) of particular feature of class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round and about 4 in diameter. Even though these features depend on the existence of other features, a naïve bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple [19].

Depending on the precise nature of the probability mode, naïve bayes classifier [9] can be trained very efficiently in a supervised learning setting. This type of learning is called supervised learning because a supervisor (the human who defines the classes and labels training documents) serves as a teacher directing the learning process.

We use the multinomial naïve bayes or multinomial NB model introduced in [7], of probabilistic learning method to normal keyword, FA word and compound FA word.

The probability of a document d being in class c is computed as

$$p(c/d) \alpha \ p(c) \prod_{1 \leq k \leq n_d} p(t_k / c) \quad (5).$$

where, $p(t_k / c)$ is the conditional probability of the term t_k occurring in a document of class c. We interpret $p(t_k / c)$ as a measure of how much evidence t_k contributes that c is the correct class. $p(c)$ is the prior probability of a document occurring in a class c.

In text classification, the goal is to find the best class for the document. The best class in NB classification is the most likely or maximum a posteriori (MAP) class C_{map} :

$$C_{map} = \arg\max_{c \in C} \hat{p}(c/d) = \arg\max_{c \in C} \hat{p}(c) \prod_{1 \leq k \leq n_d} \hat{p}(t_k / c) \quad (6).$$

We write $\hat{p}$ for p because we do not know the true values of the parameters $p(c)$ and $p(t_k / c)$.

In Equation (6), many additional probabilities are multiplied one for each position $1 \leq k \leq n_d$. This can result in a floating point underflow. It is therefore better to perform the computation by adding logarithms of probabilities instated of multiplying probabilities.

$$C_{map} = \arg\max_{c \in C}[\log \hat{p}(c) + \sum_{1 \le k \le n_d} \log \hat{p}(t_k / c)] \quad (7).$$

where, $\hat{p}(c) = \frac{N_c}{N}$, N_c is the number of documents in class c and N is the total number of documents. We estimate $\hat{p}(t/c)$ as the following:

	Doc ID	Words in document	Inc= انفلونزا الخنازير
Training set	1	(12) فيروس (vairoos which means Virus in English) (9) انفلونزا (anfluanza - which means flu in English) (5) الخنازير (al khanazeer - which means the swine in English) (3) مرض (marad - which means disease in English)	Yes
	2	(6) فيروس (vairoos which means Virus in English) (9) انفلونزا (anfluanza - which means flu in English) (6) الخنازير (al khanazeer - which means the swine in English) (8) تطعيم (tataeem - which means vaccination in English	Yes
	3	(10) فيروس (vairoos which means Virus in English) (11) انفلونزا (anfluanza - which means flu in English) (9) مرض (marad - which means disease in English) (9) الصحة (al seha - which means health in English) (4) لقاح (leqah - which means inoculation in English)	Yes
	4	(6) تطعيم (tataeem - which means vaccination in English (1) وباء (wabaa - which means epidemic in English (1) جرعة (gora - which means dosage in English (3) لقاح (leqah - which means inoculation in English)	No
Test set	5	(2) انفلونزا (anfluanza - which means flu in English) (2) الخنازير (al khanazeer - which means the swine in English) (2) وباء (wabaa - which means epidemic in English (3) جرعة (gora - which means dosage in English (5) لقاح (laqah - which means inoculation in English)	?

$$\hat{p}(t/c) = \frac{T_{ct}}{\sum_{t' \in v} T_{ct'}} \quad (8).$$

Where T_{ct} the number of occurrence of t in the training document of the class c , and $T_{ct'}$ is the number of occurrence of t with remain documents of the class c including multiple occurring of a term in a document. If a term t_k in the training document only occurred in one document; then the estimates for the other classes will be zero. To eliminate zeros, we simply add one or using Laplace smoothing, which simply adds one to each count.

$$\hat{p}(t/c) = \frac{T_{ct}+1}{\sum_{t' \in v}(T_{ct'}+1)} = \frac{T_{ct}+1}{\sum_{t' \in v} T_{ct'} + B'} \quad (9).$$

Where $B' = |v|$ is the number of terms in the vocabulary. Add one smoothing can be interpreted as a uniform prior (each term occurs one for each class) that is updated as evidence from the training data aims in.

Naïve Bayes algorithm (multinomial model): Training and testing

TRAIN MULTINOMINAL NB(C,D)

V ← Extract vocabulary(D)
N ← Count Docs(D)
for each $c \in C$
do Nc ← count Docs in class(D,C)
prior[c] ← Nc/N
textc ← concatenate text of all Doc in class(D,C)
for each $t \in V$
do Tct ← count token of each term(textc,t)
for each $t \in V$
do condprob[t][c] ← $\frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
return V, prior, condprob.

Apply multinomial NB(C,V, prior, condprob,d)

w ← extract token from Doc(V,d)
for each $c \in C$
do score[c] ← logprior[c]
for each $t \in w$
do score[c]+ =log condprob[t][c]
return $\arg\max_{c \in C} score[c]$

4. Arabic FA words with NB classifier

In this section, we apply our new method on modified Naïve Bayes (NB) classifier algorithm. That is extracted Arabic FA words and used it innaive bayes classifier.in the following we will illustrste our method by the following example.

Example 4: Naïve Bayes for normal word of Table 1, the multinomial parameters we need to classify the test document as the priors $\hat{p}(c) = 0.75$ and $\hat{p}(\bar{c}) = 0.25$ and the conditional probabilities are given in Table 1.

Table 1: Data for parameter estimation examples

When we apply **NB(C, D)** algorithm on Example (3) we get;

$\hat{p}(c/$ حاقل $) = 0.045$, $\hat{p}(c/$ ازنولفنأ $) = 0.273$,
$\hat{p}(c/$ ريخذانخلا $) = 0.109$, $\hat{p}(c/$ عرجة $) = 0.009$,
$\hat{p}(c/$ ءابو $) = 0.009$

$\hat{p}$ (c/d5)= $0.75 \times (0.045)^5 \times (0.273)^2 \times (0.109)^2 \times (0.009)^2 \times (0.009)^3 = 8 \times 10^{-21}$

Also

$\hat{p}(\bar{c}/$ حاقل $) = 0.2$, $\hat{p}(\bar{c}/$ ءابو $) = 0.1$,
$\hat{p}(\bar{c}/$ عرجة $) = 0.1$, $\hat{p}(\bar{c}/$ ازنولفنأ $) = 0.05$,
$\hat{p}(\bar{c}/$ ريخذانخلا $) = 0.05$,

$\hat{p}$ ($\bar{c}$ /d5)= $0.25 \times (0.2)^5 \times (0.05)^2 \times (0.05)^2 \times (0.1)^2 \times (0.1)^3 = 5 \times 10^{-15}$.

From the above results, the classifier transformed the document d5 from class c to $\bar{c}$. But the true location of d5 is in c. Therefore, we suggest the modification of the train multinomial algorithm as below:

Modifying TRAIN MULTINOMINAL NB(C,D)

V ← Extract vocabulary(D)
N ← Count Docs(D)
for each $c \in C$
do Nc ← count Docs in class(D,C)
prior[c] ← Nc/N
textc ← concatenate text of all Doc in class(D,C)
1 for each $t \in V$
do Tct ← count token of each term(textc,t)

	Doc ID	Words in document	Inc= انفلونزا الخنازير
Training set	1	(2) فيروس انتش 1 ان 1 (vairoos H1N1- which means virus H1N1 in English) (5) أنفلونزا الخنازير (anfluanza al khanazeer - which means swine flu in English) (1) منظمة الصحة العالمية (monazamet al seha al alamya - which means the International Health Organization in English)	Yes
	2	(1) فيروس انتش 1 ان 1 (vairoos H1N1- which means virus H1N1 in English) (6) أنفلونزا الخنازير (anfluanza al khanazeer - which means swine flu in English) (1) حملة تطعيم (hamlat tateem- which means vaccination expeditions in English)	Yes
	3	(1) فيروس انتش 1 ان 1 (vairoos H1N1- which means virus H1N1 in English) (6) أنفلونزا الخنازير (anfluanza al khanazeer - which means swine flu in English) (1) حملات التطعيم باللقاحات (hamalat tateem bel leqahat- which means vaccination expeditions with inoculations in English)	Yes
	4	(2) منظمة الصحة العالمية (monazamet al seha al alamya - which means the International Health Organization in English) (2) حملات التطعيم باللقاحات (hamalat tateem bel laqahat- which means vaccination expeditions with inoculations in English) (3) حملة تطعيم (hamlat tateem- which means vaccination expedition in English) (1) مرض فيروسى (marad vayrossy- which means virus disease in English)	No
Test set	5	(1) فيروس انتش 1 ان 1 (vairoos H1N1- which means virus H1N1 in English) (6) أنفلونزا الخنازير (anfluanza al khanazeer - which means swine flu in English) (3) حملة تطعيم (hamlat tateem- which means vaccination expedition in English) (1) منظمة الصحة العالمية (monazamet al seha al alamya - which mean the International Health Organization in English)	?

for each $t \in V$

if $\hat{p}(t/c) = 0$

do condprob[t][c] $\leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$

else, do $\frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$

return V, prior, condprob.

By applying the modifying algorithm to the same example we get:

$\hat{p}(c/\text{حاقل}) = 0.039$, $\hat{p}(c/\text{ازيولفنأ}) = 0.29$,

$\hat{p}(c/\text{ريجذانخلا}) = 0.10$, $\hat{p}(c/\text{عابسو}) = 9\times10^{-3}$,

$\hat{p}(c/\text{ةعرج}) = 9\times10^{-3}$, $p(c/d_5) \approx 40905\times10^{-13}$.

$\hat{p}(\bar{c}/\text{حاقل}) = 0.272$, $\hat{p}(\bar{c}/\text{ازيولفنأ}) = 0.05$,

$\hat{p}(\bar{c}/\text{ريجذانخلا}) = 0.05$, $\hat{p}(\bar{c}/\text{عابسو}) = 0.09$,

$\hat{p}(\bar{c}/\text{ةعرج}) = 0.09$, $p(\bar{c}/d_5) \approx 1.3736\times10^{-14}$.

Thus, the classifier assigns the test document to c= أنفلونزا الخنازير.The reason for this classification decision is that the occurrence to the positive indicator أنفلونزا الخنازير .

Example 5: Applying the classifier to compound FA words. Naïve Bayes for compound FA words of Table 5, the priors $\hat{p}(c) = 0.75$, and $\hat{p}(\bar{c}) = 0.25$ and the conditional probabilities are given in Table 2,

When we apply NB (C, D) algorithm on Example (4) we get;

$\hat{p}(c/$ فيروس أنتش1 أن 1/ $)=0.25$, $\hat{p}(c/$ أنفلونزا الخنازير $)=0.625$,

Table 2: parameter estimation examples with compound FA words

$\hat{p}(c/$حملة تطعيم$)=0.04$, , $\hat{p}(c/$منظمة الصحة$)=0.032$, $p(c/d_5) \approx 57\times10^{-9}$.

$\hat{p}(\bar{c}/$فيروس أنتش1 أن 1/ $)=0.05$, $\hat{p}(\bar{c}/$ أنفلونزا الخنازير $) = 0.05$,

$\hat{p}(\bar{c}/$حملة تطعيم$)=0.09$, $\hat{p}(\bar{c}/$منظمة الصحة$) =0.18$, $p(\bar{c}/d_5) \approx 7.9\times10^{-15}$.

By using compound FA words instated of using single FA words, tends to more accurate result and better classification.

Table 2: parameter estimation examples with compound FA words

5. Experimental

The new method for using FA words can be applied to classify any Arabic document such as web documents, scientific paper, articles, news and others. In this paper the experimental is formed on a collection of web documents because, its unstructured data and hardly to classify it.

Our experiments trained the system using Arabic 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 16 super-field and 137 subfields. The number of files in our corpus is 1,819 file and it is about 26.4 MB. For experimental evaluation, we download a source code written in JAVA from http://nlp.cs.byu.edu/mediawiki/index.php/CS601R:Project_1_Guidelines. In addition, we modified it to be suitable for the new NB algorithm. Also, we prepare the system according to http://nlp.cs.byu.edu/mediawiki/index.php/How_to_prepare_your_system.

5.1 Experimental Evaluation for FA Naïve Bayes Text Classifier

System measurements (Recall, Precession, F-Measure) have been used as defined in the following, for 6 super-fields with NB classification using keyword, FA words and compound FA as shown in Table 3, Table 4 and Table 5.

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

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

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

From the estimation results, the performance for each method is calculated and the best performance is recorded for NB classification using compound FA-words as shown in Table 6. Moreover, the calculation of F-measure for NB classification

for each class separately using compound FA-words are more accurate than normal keywords and FA words.

Name of field	Precision	Recall	F-measure
الاستنساخ (al estensakh- which means the cloning in English)	0.3	0.66	0.4125
الأمراض (al amrad- which means diseases in English)	0.59	0.79	0.6755
البيئة (al beaa- which means the environment in English)	0.24	0.9	0.378
التكنولوجيا (al tecnologia- which means tecnology in English)	0.33	0.87	0.4785
جسم الإنسان (gesem al ensaan- which means the human body in English)	0.18	0.75	0.29
مقالات علمية (makalat elmiah- which means scientific articles in English)	0.22	0.89	0.35

Table 3: NB classification using Normal Keyword

Name of field	Precision	Recall	F-measure
الاستنساخ (al estensakh- which means the cloning in English)	0.73	0.95	0.825
الأمراض(al amrad- which means diseases in English)	0.82	0.77	0.794
البيئة (al beaa- which means the environment in English)	0.69	1	0.81
التكنولوجيا (al tecnologia- which means technology in English)	0.86	0.84	0.849
جسم الإنسان (gesem al ensaan- which means the human body in English)	0.67	0.98	0.796
مقالات علمية (makalat elmiah- which means scientific articles in English)	0.54	1	0.70

Table 4: NB Classification using FA Words

Name of field	Precision	Recall	F-measure
الاستنساخ (al estensakh- which means the cloning in English)	0.78	0.98	0.86
الأمراض(al amrad- which means diseases in English)	0.89	1	0.94
البيئة (al beaa- which means the environment in English)	0.84	1	0.91
التكنولوجيا (al tecnologia- which means the technology in English)	0.80	1	0.88
جسم الإنسان (gesem al ensaan- which means the human body in English)	0.73	1	0.84
مقالات علمية (makalat elmiah- which means scientific articles in English)	0.86	1	0.92

Table 5: NB Classification using Compound FA Words

	Classification with keywords	Classification with FA words	Classification with compound FA words
Precision	0.31	0.71	0.81
Recall	0.81	0.92	0.99
F-measure	0.43	0.79	0.89

Table 6: The Comparison of Medium for Results of NB

In conclusion, the results of F-measure of both classifiers using in this study for each class separately using compound FA words are more accurate than normal keywords and FA words. Figure 1 show that NB classification using compound FA words more accurate than others.

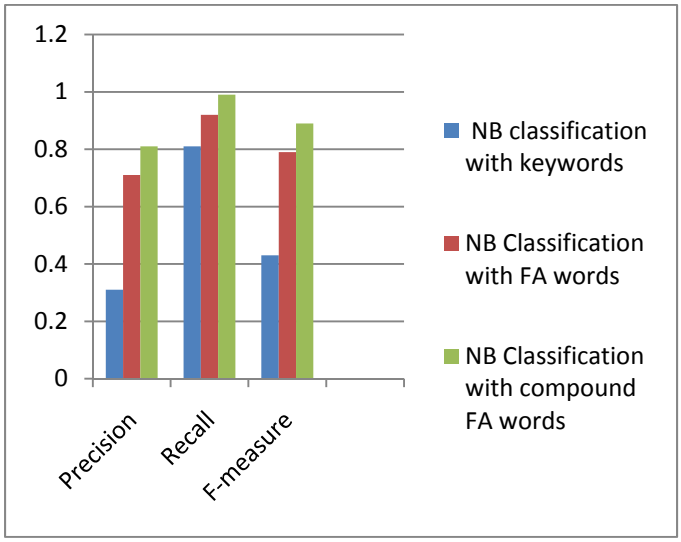

Figure 1: NB classification comparison

6. Conclusion

With increasing popularity of the internet and tremendous amount of online text, automatic document classification is important for organizing huge amounts of data especially Arabic documents. Moreover, document fields can be decided efficiently if there are many FA words and if the frequency rate is high. FA words are used to classify Arabic documents. Words are extracted from these document corpora to get FA word candidates. Furthermore, we used the NB classifier with our modification to refine Arabic document classification. From the experiential results, the NB classifier using FA words recorded of F-measure is 79% and, using compound FA words of F-measure is 89%.

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

References

[1] **Atlam, E.-S., Morita K., Fuketa, M. and Aoe, J. A** New Method For Selecting English Compound Terms and its Knowledge Representation. Information Processing &Management Journal, 38, (2000), 807-821.

[2] Atlam, E.-S., Fuketa M., Morita, K. and Aoe, J. Document Similarity measurement using Field association terms". Information Processing & Management Journal, Vol. 39, pp. 809-824, 2003.

[3] E.-S. Atlam, G. Elmarhomy, M. Fuketa, K., Morita and Jun-ich Aoe., "Automatic building of new Field Association word candidates using search. Information Processing & Management Journal, 42(4), (2006), 951-962.

[4] **Mohammed J. Bawaneh, Mahmud S. Alkoffash and Adnan I. Al Rabea . "**Arabic Text Classification using K-NN and Naive Bayes", Journal of Computer Science 4 (7): 600-605, 2008 ISSN 1549-3636

[5] **Cavnar W. B. and Trenkle J. M.** "N-Gram Based Text Categorization," Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, (**1994).**

[6] **Chinchor N.** "Named Entity task definition", In Proceedings of the Seventh Message Understanding Conference, **(1998).**

[7] **Christopher D., Raghavan P. and Schütze H. "**Introduction to Information Retrieval", ISBN 978-0-521-86571-5, Cambridge University Press ,**(2008).**

[8] **Ciravegna F., Gilardoni L., Lavelli A., Ferraro M., Mana N., Mazza, S., Matiasek J., Black W. and Rinaldi F.** "Flexible Text Classification for Financial Applications: the FACILE System", In Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000, **(2000).**

[9] **El-Kourdi M., Bensaid A. and Rachidi T.** "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm". 20th International Conference on Computational Linguistics. August, Geneva. **(2004).**

[10] **Forman G.** "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research 3 pp. 1289–1305, **(2003).**

[11] **Fukumoto F., and Suzuki Y.** "Automatic clustering of articles using dictionary definitions". In Proceedings of the 16th international conference on computational linguistic (COLINGO'96) (pp. 406–411), **(1996).**

[12] **Fuketa M., Lee S., Tsuji T., Okada M. and Aoe J.** "A document classification method by using field association words", Information Sciences 126 57-70, **(2000).**

[13] **Fuhr N. :**Models for retrieval with probabilistic indexing. Information Processing and Retrieval, 25(1), 55–72, **(1989)**.

[14] **Gu Q. and Zhou J.**, "Learning the Shared Subspace for Multi-task Clustering and Transductive Transfer Classification," icdm, pp.159-168, 2009 Ninth IEEE International Conference on Data Mining, **(2009).**

[15]**Kimoto H. "** Automatic indexing and evaluation of keywords for Japanese newspapers". Transaction on Information and Systems, IEICE of Japan, J74-D-I(8), 556–566 (in Japanese), **(1991**).

[16] **Mladenic D. and Grobelnic M.** "Feature Selection for Unbalanced Class Distribution and Naïve Bayes". In: Proceedings of the 16th International Conference on Machine Learning 258–267, **(1999)**.

[17] **Noaman, H.M.; Elmougy, S.; Ghoneim, A.; Hamza, T.** "Naive Bayes Classifier based Arabic document categorization". The 7th International Conference on Informatics and Systems (INFOS), p 1 – 5, ISBN: 978-1-4244-5828-8, 2010

[18] **Peng F., Huang X., Schuurmans D. and Wang S.** "Text Classification in Asian Languages without Word Segmentation", In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003), Association for Computational Linguistics, Sapporo, Japan **(2003).**

[19] **Sawaf H., Zaplo J. and Ney H. "** Statistical classification methods for Arabic news articles". Natural Language Processing in ACL2001, Toulouse, France **(2001).**

[20] **Tang X. and Han M.** "Ternary reversible extreme learning machines: the incremental tri-training method

for semi-supervised classification", Knowledge and Information Systems Volume 23, Number 3, 345-372, DOI: 10.1007/s10115-009-0220-4, (**2010**).

M . E. Abd El-Monsef M . E. Abd El-Monsef received the B.S&ED degree in Mathematics from Assuit University, Egypt, in 1968 and the B.S degree from Faculty of Science, Assuit University, Egypt, in 1973. He received the MS degree in Mathematics from Al Azhar University, Cairo, Egypt, in 1977. He received his PhD degree in Mathematics from Tanta University, Tanta, Egypt, in 1980. He was assistant professor in the Department of Mathematics, Faculty of Science, Tanta University, from 1984. He was a professor of Mathematics in the Department of Mathematics, Faculty of Science, Tanta University, from 1988. He worked as Vice Dean of Faculty of Science, Tanta University, for postgraduate and researches affair from 1991 to 1996. He worked as Vice Dean of Faculty of Science, Tanta University for students affairs from 1996 to 1999. He worked as Dean of Faculty of Science, Tanta University from 1999 to 2005. He was Chairman of the Scientific Committee of Promotion to Assistant Professors Post in Mathematics of the Scientific Council of the Egyptian Universities from 2001 to 2004. Member of the Scientific Committee of Promotion to Professors Post in Mathematics from 2004 to 2008. He is a member of National Committee for Mathematics, a member of National Committee for History and Philosophy of Science, member of the Board of Directors of Egyptian mathematics Society, member of the editorial board of the Journal of Egyptian Mathematical Society, member of the Egyptian Society for the Arabization of Sciences. He Was the Editor of the Delta Science Journal .He is Representative of Tanta University in the League of Islamic Universities, member of Editorial Board of the international scientific journal Science Echoes and Applied Mathematics & Information Sciences. He is a member of the Supreme Advisory Committee of the Centre for Development of the Delta Region of the Academy of Scientific Research and Technology. He participated in more than 90 Scientific Conference and seminar specialist. He Also supervised over 51 PhD and about 47 Master. Sovereignty to a lot of researches in the fields of general topology and fuzzy topology about 100 research papers published in scientific journals, interior and exterior prestigious. Was awarded the University of Tanta estimated in the basic sciences for the year 2001/2002. His research interests include General Topology, Rough Sets, Digital Topology and Fuzzy Sets.

Dr. El-Sayed Atlam: Received B.Sc. and M. Sc. Degrees in Mathematics from, Faculty of Science, Tanta University, Egypt, in 1990 and 1994, respectively, and the Ph.D. degree in information science and Intelligent systems from University of Tokushima, Japan, in 2002. He has been awarded by a Japan Society of the Promotion of Science (JSPS) postdoctoral Fellow from 2003 to 2005 in Department of Information Science & Intelligent Systems, Tokushima University; He is currently Associate professor at the Department of information science and Intelligent systems from University of Tokushima, Japan. He is also Associate professor at the Department of Statistical and Computer science, Tanta University, Egypt. Dr. Atlam is a member in the Computer Algorithm Series of the IEEE computer society Press (CAS) and the Egyptian Mathematical Association (EMA). His research interests include information retrieval, natural language processing and document processing.

Mohammed Amin was graduated in mathematics in 1983 at Menoufiya University. He studied computer science from 1986 to 1989 at Ain Shams University in Cairo and received the M.Sc. degree in 1990 and the Ph.D degree in computer science in 1997 at the University of Gdansk, Poland. He is associate professor of computer science at the faculty of science, Menoufiya University, and research visitor to the faculty of Philosophy and sciences of the Silesian University, Opava, Czech Republic. Hisresearch area in formal languages and their application in compiler design. Cooperating/distributed systems, web information retrieval, Petri nets and its applications, and finite automata and cryptograph.

O. El-Barbary received B.Sc. Degree in Statistics and computer science from, Faculty of Science, Tanta University, Egypt, in 2004 and M.Sc. in computer science from, Faculty of Science, Menofya University, Egypt, in 2007, and now she is studying Ph.D. in Faculty of Science, Menofya University, Egypt in the field of information retrieval and natural language processing.