International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 5, May 2015
ISSN (Online): 2409-4285 www.IJCSSE.org Page: 114-120

Arabic Text Classification on Full Word

Rasha Elhassan[1] and Mahmoud Ahmed[2]

[1] Computer science Department Mathematical Science collage, University of Khartoum, Khartoum, 11111, Sudan
[2] Dean of Mathematical Science collage, University of Khartoum, Khartoum, 11111, Sudan

[1]rashamamoun@yahoo.com , [2]mali@uofk.edu.sd

ABSTRACT

Text classification aims to extract the useful information from a large data. The documents may contain unnecessary data which may affect the accuracy of the classifier. Data preprocessing phase aims to clean the texts by removing unnecessary information. The main object of this paper is to explain and determine the effectiveness of the data preprocessing on full word in the accuracy of both training model and classifier. This will be done by two approaches: first the observation of data set contain and second the stop word estimation technique. In the experiment, the Sequential Minimal Optimization (SMO), Naïve Bayesian (NB) J48 and K-nearest neighbors (KNN) were used to build the training models. By implement the two approaches and measured the accuracy by precision, recall and f- measure, the results showed that the SMO classifier outperforms the three other classifiers as a training model and a classifier.

Keywords: *Text classification, Arabic Text Classification, Stops Words, Data Preprocess Classifier, ESWF, SWF.*

1. INTRODUCTION

Text classification is utilized to give useful information from the large amount of data. It is one of the important research issues in the field of data mining. Based on the content of the text, categorization is define as the process of grouping this text into one or more predefined categories based on linguistic features [1]. In general, the terms 'Text categorization' and 'Text Classifi-cation' refers to the same meaning. Text categorization is sometimes taken to mean sorting documents by content, while text classification is used to include any kind of assignment of documents to the specific classes like sorting by author, by publisher, or even by language [2]. Text classification is the many important research problems in information retrieval (IR), data mining and natural language processing [3]. It is the primary requirement of text retrieval systems, which retrieve texts in response to a user query.

There are many applications of text classifications, such as e-mail filtering, news monitoring, spam filtering, sorting archives, automated indexing of scientific articles in e-libraries, classification of news stories and searching for interesting information on web.

The rest of this paper is organized as follows: section 2 give a brief review of researches in the area of Arabic text classification which used the stop words estimation as technique of preparing the data without stemming. Section 3 explains the more important point in the Arabic language structure. Section 4 in details, explains the proposed techniques to enhance the accuracy of the classifier. Highlight the structure of the text classification algorithm that used to test the data given in section 5. The experiment is given in section 6. Finally section 7 explains the conclusion and future work.

2. RELATED WORK

By implemented the Sequential Minimal Optimization (SMO), Naïve Bayesian (NB) and J48 (C4.5) algorithms using Weka program, the authors [4] comparing between the algorithms in accuracy and time to get the result. The data preprocessing includes elimination stop word and normalization. The results show that the (SMO) classifier achieves the highest accuracy and the lowest error rate, followed by J48 (C4.5), then the (NB) classifier. But the second part of the results (time) shows that the time needed to get the results, the SMO model is the faster one, followed by NB model, and then J48 classifier which takes a longest time.

K-nearest neighbor (KNN) algorithm and support vector machines (SVM) algorithm evaluated on a collection of news articles. The authors used the full word features. They had considered the tf.idf as the weighting method for feature selection and CHI statistics as a ranking metric. Experiments showed that both methods were of superior performance on the test corpus while SVM has better Micro recall than KNN. The Micro precision values of both classifiers are similar, with KNN a bit better. When the number of features is low, classifiers have similar performance, with advantage given to SVM. But SVM

outperforms KNN clearly as the number of features increases. The prediction time of SVM is better also than KNN [5].

By used support vector machine (SVM), Naïve Bayesian (NB) classifier and k-nearest neighbor (KNN) the author [3] had implemented text classifier for Arabic articles .He investigated the effectiveness of six state-of-the art feature subset selection (FSS) method. Implementation of an Ant Colony Optimization Based-Feature Subset Selection (ACO Based-FSS) method was present. The results were shown in term of Precision, recall, F1 accuracy measure. The author used an in-house Arabic that consists of 1445 documents which classified in 9 categories and referred as the result of unavailable Arabic corpus. The SVM classifier outperformed, 90%, Naïve Bayesian (NB) classifier and k-nearest neighbor (KNN) classifier in the experiment result.

By investigated Naïve Bayesian method (NB) and Support Vector Machine algorithm (SVM) on different Arabic data sets.. The authors preprocessed the data by removed digits, punctuation marks, non Arabic letters, Arabic function words and normalization some Arabic letters. The experimental results revealed that SVM algorithm outperforms the NB with regards to all measures [6].

3. ARABIC LANGUAGE STRUCTURE

Arabic language is one of the widely spoken languages in the world. It is the sixth most spoken language about 200.000.000 people [7, 8]. Arabic language belongs to the Semitic family of languages and it is the official language of the region of Arabian people. It is the language of the Holly Quran. There are two varieties of language: standard al-fushá (الفصحى), and Modern Standard Arabic (MSA) which is based on the classical language [9].

In contrast with Latin-based alphabets, Arabic language written from right to left. There are 28 characters in the language and there is no capitalization in Arabic script. Arabic letters connected by preceding or following letter except six letters, are called non-connectors, connected only when it is in the medial of the word or the end and just by it right side [10]. There are three letters, semivowels letters ا ,و and ي, sometimes are act as consonants depending on context and other times are act as vowels. The rest of the letters are constants.

There are two genders: masculine (مذكر) and feminine (مؤنث) which present by attached the suffix (ة) to the end of the word. Numerical there are singular (مفرد) , dual(مثنى) and plural (جمع) numbers. Plurals numbers are regular (جمع الصحيح) or broken (جمع تكسير).

Arabic grammarians divided the word into three types: noun, verb, particles. Grammatical there are three cases: nominative (الرفع), accusative (النصب) , and genitive(الجر) . Noun has the genitive case when it is the object of a preposition; nominative case when it is subject; and accusative when it is the object of a verb [11, 12, 13, 14].

In general, the Arabic documents prepared by removing digits, punctuation and non-Arabic words as first step. After that stop words removed. Stop words are the words that have not any meaning in the content of the document like (so لذلك, and for بالنسبة). The rest of the words were referred as features or keywords of the documents. For the richness of the Arabic language there is a need to normalizing some writing forms that include "ة" to "ه" and "ي" to "ى" and "إ, أ, آ" to "ا" [3] [15].

The number of the feature is usually very big and it increases by the length of the documents. Some filtering techniques applied to reduce the features number. One of these techniques is the root extraction. Root extraction or stemming step aims to decrease the number of the features and remove the redundant features. The root can be extracted by two approaches: letter weight and stemming algorithms. Letter weight grouped the letters in some ranks and weights, so each letter had a product weight and the three letters with the smallest product value give the root of the word. The second one Stemming algorithms can be divided in three classes: the root-based stemmer; the light stemmer; and the statistical stemmer. Theses algorithms include morphological analysis, removing the prefixes, suffix and infixes of the words and string similarities measures [14] [16].

4. PROPOSED MODEL OF ARABIC DATA PREPROCESSING

This section describes the main phases of the proposed Arabic data preprocess model. The data preprocessing includes two approaches: the first approach observation of data set contain and stop words estimation technique.

The observation approach done by determined was this document belongs to the category or not, even it classified in the newspaper to it, using the human judgment. This step repeated when building the training model until there is no observation.

The following steps were done to build the initial corpus:

1. Each article was converting from HTML to Txt format.
2. Each article was processed to remove the formatting tags, digits, punctuation marks, special marks and Latin-base words.
3. Normalizing needed letters as mention above.

For the above steps the small Visual Basic too (Text Extraction), as shown in figure 1, designed to extract the text from the web page with the advance feature to remove punctuation marks, formatting tags, digits and non-Arabic words. This program created the initial corpora from the first corpora.

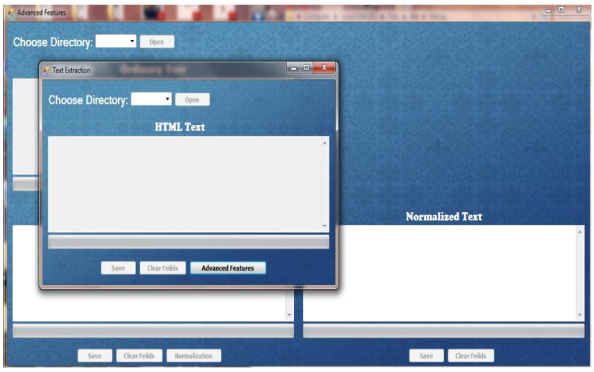

Fig. 1. Extraction program with its features.

First approach: After created the initial corpora, the observation step will start. The documents filtered instead to remove any duplicated and unrelated documents. Observation step will done manual to create the initial corpora which used to build the second training model. This step repeated until no observation found. Figure 2 explain the steps of optimizes the corpus.

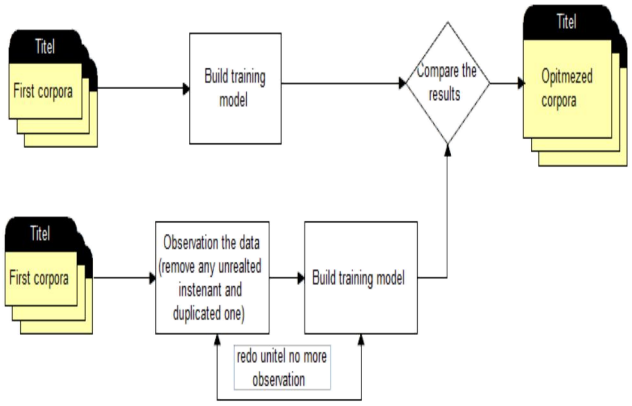

Fig. 2. Data Preprocess first approach: Observation

Table 1 explains the comparison between the training models of initial corpora and the optimized one.

Table 1: Training models Correctly Classified Instances average.

Classifier	*First corpora*	*Optimized corpora*
SMO	92.95	94.09
NB	88.95	89.52
J48	83.23	84.57
IK1	47.8	48.19

Second approach: this approach aims to improve the stop words file (SWF) by . There were two types of SWF:

1. The Basic stop words file (BSWF) which contains about 216 words. These words were the most particles words (preposition, Adverbs, Conjunction, and Interjection like البته, افه, ايها, فى, على, تحت, امس, حول, حيثما) which extract from [11].
2. The second one which called it extended stop words file (ESWF) contains, in addition to the basic stop word, the most appearances word in the training file. This word can be like the stop words because it appears in all the instance like قال which appeared 1386 times. The file contains about 240 words.

To compare the accuracy of the removal files, this step will be done with the training model building. Training models will be building by three cases: without using stop words remove, using basic stop words and finally by using extended stop words remove.
Table 2 explains the time taken to build the training models. It determines the improvement in time taken by classifiers after applied the two approaches. SMO classifier's time is the faster one to build the model.

Table 2: Time Taken to Build Training models in second.

Classifier	*WITHOUT STOP WORDS.*	*BSWF.*	*ESWF.*
IK1	1.1	1.48	1.03
SMO	1.83	1.31	1.31
NB	1.93	1.64	1.60
J48	11.46	10.28	10.24

Figure 3 explains the accuracy of the training models after applied the two approaches. It determines the improvement in accuracy of classifiers after applied the two approaches.

Weighted Avg.	SOM			BN			j48			IB1		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Without stop words	0.939	0.937	0.938	0.904	0.895	0.896	0.85	0.844	0.845	0.637	0.516	**0.519**
with basic stop words	0.943	0.941	0.941	0.904	0.895	0.896	0.854	0.846	0.847	0.607	0.482	0.483
with extanded stop words	0.946	0.945	**0.945**	0.906	0.897	**0.898**	0.865	0.848	**0.849**	0.595	0.484	0.483

Fig. 3. Training model weighted average

The results showed that the two approaches improved the accuracy of the training models in almost time. The exception is the IB1 classifier which decreased the weighted average even there is improve in the correct classified instances.
IB1 is a lazy function and fully dependent on training set. The feature in the training set were very large, so this effective the result of the training mode. The accuracy of classification affect by the process of the feature selection space, the training data set used, and the value of k [17].

5. CLASSIFIERS

In this paper four classifiers were used to classified Arabic text. The tool of data mining that used to expert the model is Weka. Weka open source software was used to build the training model that used to classify the test data set. Weka was developed at the University of Waikato in New Zealand. It is a collection of machine learning algorithms and data preprocessing tools which provides implementations of learning algorithms that you can easily apply to the dataset [18].

5.1. Support Vector Machine classifier

Support vector machine (SVM) is a statistical algorithm based on the Structural Risk Minimization principle. The general idea of SVM is to find a hypothesis that can guarantee the lowest true error by find the optimal separating hyper-plane that has the maximal margin to both sides. The need of both positive and negative training set is the main characteristic of SVM comparing to all other text classification algorithms. SVM been recognized as one of the most effective text classification methods [3] [19, 20]. It needs a large number of features and large set of training to work properly. This is due to the clearer optimal hyper plane obtained by the increasing number of features [5] [21, 22].
The classifier implemented by using Sequential Minimal Optimization (SMO). SMO is the Java implementation of SVM in Weka and it is a fast training SVM.

5.2. K-Nearest Neighbor classifier

Key nearest neighbor (KNN) is a simple statistical algorithm and a well known pattern recognition algorithm. The algorithm stores all classified feature of the classes and classify the new features according to some measures to specific category. The probability of categorized some document to some category, measured by the distribution of the document from that category of the nearest neighbors [19, 20].

The success of classification is dependent on the value of k, so that selecting the value of k is a major drawback of the classifier [23]. For this experiment, to determine the best value of k, a series of experiments with different k values should be conducted. Started with k=1 increased it to 15 as first fifteenth experiments following by k=20 and k=40. Figure 4 explained the relation between the value of k and the correctly classified instance.

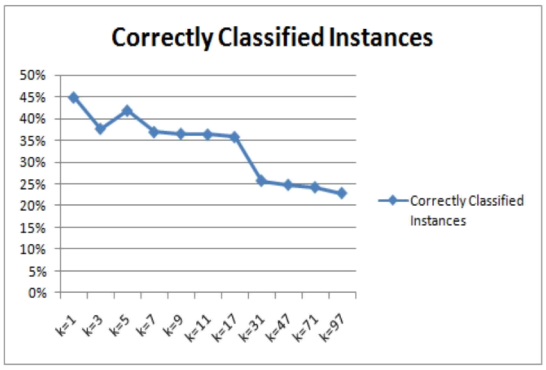

Fig. 4. Correctly classified instances by used different k values.

As determined from the results the best result was determine when k=1.
IB1 is the java implementation of KNN when k=1 algorithm in the Weka program. KNN will be mentions as "IB1" in remind of this paper.

5.3. Naïve Bayes (NB)

Naïve Bayes used probabilistic classification approach based on Bay's Theorem. Naïve Bayesian classifier assumes that features of the input feature vector are statistically independent. NB classifier is a very simple classifier which works very well on numerical and textual data. It is very effective and it is non-parametric in nature [24, 25].

5.4 J48 classifier

J48 is the java implementation of C5.0 algorithm in the Weka program. The j48 classifier based on decision trees. It used divide-and-conquer algorithm to grow an initial tree. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. C4.5 uses information gain to minimize the total entropy of the subsets and the default gain ratio which divides information gain by the information provided by the test outcomes [25, 26, 27].

6. EXPERIMENT

The experiment was conducted by collected 750 documents from local newspaper (Akhir Lahza and Alyoum Altali), and international newspaper (Al-Raya, Asharq Al-Awsat, and Al-Hayat) web sites during the period from January 2001 to January 2015. These documents divided in 5 categories: economy, political, religion, sport and technology. Every category contains 150 documents and the document assigned to one category only. For every category, 105 used for trained the classifier and the rest used for tested it.
The accuracy of the text classifiers have been compared in term of F1 measure, recall and precision where it's computed based on the following rules as in equation [28].

$$\text{Precision (P)} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall (R)} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F –measure} = \frac{2*P*R}{P+R} \quad (3)$$

Where:

- TP: is the number of correct classifications of the positive examples (relevant examples).
- FN: is the number of incorrect classifications of positive examples (irrelevant examples).
- FP: is the number of incorrect classifications of negative examples (irrelevant examples rejected).
- TN: is the number of correct classifications of negative examples (relevant examples rejected).

Cross validation is a model evaluation method. The basic idea of it is to remove some data before training set and used it to test the performance of the model. K-fold is one of cross validation which divided the data to k sub set and one of it used as test set. This will repeat k times for to test all sub set (Bing liu, 2007).
In this paper 10-fold cross-validation was used to evaluate the training model and test the data. Table 3 gives the F1, Recall, and Precision results generated by the four classifiers SVM, NB, J48 and kNN in the same data sets where in each categories using ten-fold cross-validation.

Category	SMO			NB			J48			IB1		
	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
Economy	0.833	0.888	0.859	1	0.8	0.888	0.829	0.755	0.79	0.666	0.266	0.38
Political	0.84	0.822	0.83	0.781	0.955	0.859	0.627	0.822	0.711	0.269	0.955	0.419
Religion	1	1	1	0.808	0.977	0.884	0.973	0.822	0.891	1	0.755	0.86
Sport	1	0.955	0.976	0.95	0.844	0.893	0.902	0.822	0.86	0.625	0.555	0.587
Technology	1	1	1	1	1	1	0.978	1	0.988	1	0.155	0.268

Fig. 5. Precision, Recall and F1 results for categories

As showed in figure 5, the category technology reached the highest Precision and Recall and this returned to specific and specialized feature of it. By reexamined the classifier output, the misclassified documents from the category economy was classified in political category. This is due the fact that economy today effect by the political issues in the real life.
Figure 6 the weighted average of the categories in the four used classifiers. It shows that the precision, recall and f1 reaches its lowest value (0.733, 0.409 and 0.525) when KNN classifier was used and reaches its highest value (0.938, 0.933 and 0.963) when SVM classifier was used. It is clear that SMO has the best accuracy following by NB and J48. The IB1 was the worst classifier in all the measures

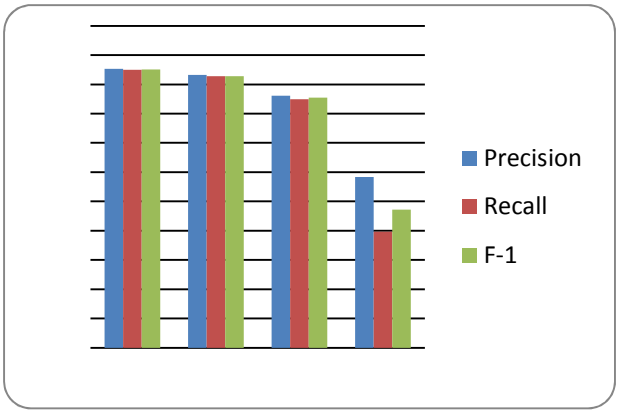

Fig. 6. weighted results for classifiers.

7. CONCLUSION

In this paper the problem of automatically classifying Arabic text documents was discussed. By used in house corpus; the corpus consists of 750 documents that belong to 5 categories from the local and international newspaper web sites. The Sequential Minimal Optimization (SMO), Naïve Bayesian (NB) J48 and K-nearest neighbors (KNN) techniques were used to handle the classification problem. Documents in the data set were preprocessed by used two approaches to improve the accuracy of the four classifiers the training and test stages: observation the data set and extended stop words remove.

Firstly to build the corpus small Visual basic tool used to removing the formatting tags, digits, punctuation marks, special marks Latin-base words. Normalization was used to some letters to decrease the number of features. The experiment shows that the two approaches enhanced the accuracy of the training models. The average of accuracy measures indicated that the SVM algorithm outperformed all the others algorithms in the training stage.

Secondly, the classifiers were tested by used the rest of the corpora. The accuracy of the classifiers was measures by using precision recall and f1 measures. The average of accuracy measures indicated that the SVM algorithm outperformed all the others algorithms regard to F1, Recall and Precision measures.

The two stages of classification indicated that use full word techniques with enhance the training stage give a high accuracy.

In the future work the investigation of stemming approaches will be using to compare it with this recent experiment.

ACKNOWLEDGMENTS

The authors would like to thanks their colleagues, from Garden City Collage - Information Technology department, Mr. Khobib from for his great helpful in programming techniques and Miss. Salma for her valuable comments.

REFERENCES

[1] X. Yabo, W. Bin, T. L. Jin and J. Hongfang, “Recent Trends in Image Processing and Pattern Recognition”, Information Retrieval Tec, IJCA Special Issue on RTIPPR, 2009.

[2] J. Peter and M. Isabelle, Natural language processing for online applications: text retrieval, extraction, and categorization, John Benjamins Publishing Company, Amsterdam, 2002.

[3] M. A. Abdelwadood, "Support Vector Machine text Classifier for Arabic Articles: Ant Colony Optimization-based Feature Subset Selection. PhD Thesis, Faculty of information systems and technology, The Arab Academy for banking and financial Science. Oman, Jordon, 2008.

[4] I. H. Majed, O. Fekry, A. Minwer and Sh. Ahlam, “Arabic Text Classification using SMO, NAÏVE BAYESIAN, J48 Algorithms”, IJRRAS 9 (2), 2011, available at www.arpapress.com/Volumes/Vol9Issue2/IJRRAS_9_2_15.pdf.

[5] H. Ismail, H. Bilal, and E. Eyas, “Performance of KNN and SVM classifiers on full word Arabic articles”, Elsevier Ltd., 2007, available online at www.sciencedirect.com.

[6] A. Saleh. "Automated Arabic Text Categorization Using SVM and NB." Inter-national Arab Journal of e-Technology, 2011, Vol. 2, pp. 124-128.

[7] www2.ignatius.edu/faculty/turner/languages.htm

[8] www.un.org/depts/OHRM/sds/lcp/Arabic/index

[9] Ryding, K. C. (2008). A reference grammar of Modern Standard Arabic (review). Language (Vol. 84, pp. 204–206). Cambridge University Press. doi:10.1353/lan.2008.0050

[10] H. Moukdad, "Stemming and root-based approaches to the retrieval of Arabic documents on the Web." Webology, 3(1), Article 22, 2006, Available at: http://www.webology.org/2006/v3n1/a22.html

[11] A. O. Abu Bashr (Sibawayhi), Al-Ketab, Alamiria Great Publisher, Egypt, 1899.

[12] M. S. Abu Bakr, Alosol in AlNaho, Alresala Library, Beirut, 1988.

[13] Wright, William, R. S. William, and M J de Goeje. A Grammar of the Arabic Language: Translated from the German of Caspari and Edited with Numerous Additions and Corrections. Vol.1. Cambridge: Cambridge University Press, 1896. Available at https://archive.org/details/AGrammarOfTheArabicLanguageV1. at April 2014.

[14] D.Rahab, R. "Arabic Text Categorization," The International Arab Journal of Information Technology, 2007, Vol. 4, pp.125-131.

[15] M.S. Khorsheed and, A.O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Springer Science + Business Media Dordrecht 2013 Published online.

[16] M. A. Abdelwadood., "Support Vector Machines Based Arabic Language Text Classification System: Feature Selection Comparative Study," Advances in Computer and Information Sciences and Engineering, Springer Science + Business Media B.V., 2008.

[17] A. Riyad “Arabic text categorization using KNN algorithm,” available at the University of California Irvin data collections repository: http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

[18] W. H. Ian, F. Eibe and A.H. Mark, Data Mining Practical Machine Learning Tools and Techniques, USA, The Morgan Kaufmann series in data management systems, 2011.

[19] C D. Shweta, I. Maya, K. Parag, "Empirical Studies on Machine Learning Based Text Classification Algorithms," Advanced Computing: An Inter-national Journal (ACIJ), Vol.2, 2011.

[20] K. Aurangzeb, B. Baharum, H. L. Lam, K. Khairullah, "A Review of Machine Learning Algorithms for Text-Documents Classification," Journal of Advances in Information Technology, Vol. 1, 2010.

[21] F. G. Tarek, B. H. Mena and T. F. Zak, "Arabic Text Classification Using Support Vector Machines" http://www.home.cs.utwente.nl/~badiehm/PDF/ ISCA 2009.pdf
[22] P. Sachintha, L. Cen, S. Su,. "A Comparative Study of Text Classification Approaches for Personalized Retrieval in PubMed," IEEE International Conference on Bioinformatics and Biomedicine Workshops, 2011.
[23] M. M. Syiam, Z. T. Fayed & M. B. Habib, an intelligent system for Arabic text categorization, IJICIS, 2006, Vol.6, No. 1.
[24] N. Suguna, and K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010, p.p 18 – 21.
[25] W. Xindong, K.Vipin, Q. J.Roos, G. Joydeep, y. Qiang, M. Hirroshi, J.M. Geoffrey, N. Angus, L. Bing, S. Y. Philip, Z. Zhi-Hua, S. Micheal, J. H. David, S. Dan, "Top 10 algorithms in Data Mining," Knowl Inf Syst , 2008, 14:1–37
[26] R.P. Tina, S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, 2013-open access
[27] Y. P. Pratiksha and S. H. Gawand "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, 2012, p.p 423-426.
[28] L. Bing, Web Data Mining Exploring Hyperlinks Contents and Usage Data, Data Centric Systems and Applications, Springer-Verlag Berlin Heidelberg, USA, 2007.