A Hybrid Feature Selection Approach for Arabic Documents Classification

Mena B. Habib, Zaki T. Fayed and Tarek F. Gharib

Faculty of Computers and Information Sciences
Ain Shams university, Cairo, Egypt
email: tgharib@asunet.shams.edu.eg

Abstract

Text Categorization (classification) is the process of classifying documents into a predefined set of categories based on their content. Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge number of features. Feature selection tries to find a set of relevant terms to improve both efficiency and generalization. There are two main approaches for feature selection, local and global. In Arabic text categorization it was found that using global feature selection gives higher results but may affect some documents in a way so that they do not show any terms in the set of selected features. On the other hand local feature selection is used to overcome this problem but gives lower classification rate. In this paper a hybrid approach of global and local feature selection technique is proposed and compared with both local and global feature selection techniques. Results are reported on a set of 1132 document of six different topics showing that the proposed hybrid feature selection overcome the disadvantages of both of feature selection approaches.

Keywords: Document Classification, Feature Selection, Text Mining

1. Introduction

Text categorization (classification) is the process of classifying documents into a predefined set of categories based on their content. This assignment can be used for classification, filtering, and retrieval purposes. Machine learning approaches are applied to build an automatic text classifier by learning from a set of previously classified documents [1]. Documents are represented used ‘bag-of-words’ [2] scheme (vector space model), in which the structure of a document and the order of words in the document are ignored. The feature vectors represent the words observed in the documents. The document in text categorization system must pass through a set of steps: document conversion which converts different types of documents into plain text, stop word removal to remove insignificant words, stemming to group words sharing the same root, feature selection/extraction, construction of super vector which is the vector containing all terms that appears in all the documents in the corpus, feature weighting, classifier construction, classification, and evaluation of the classifier [3]. In this paper we focused on the document indexing phase which involves construction of the super vector and term selection which can be seen as a form of *dimensionality reduction* (DR) by selecting a subset of terms from the full original set of terms in the super vector according to some criteria, this subset is expected to yield the best effectiveness, or the best compromise between effectiveness and efficiency.

There are two quite distinct ways of viewing DR, depending on whether the task is approached locally (i.e. for each individual category, in isolation of the others) or globally:

- *Local feature selection*: for each category c_i, features are chosen in terms of which the classifier for category c_i will operate [4, 5, 6, 7]. Conceptually, this would mean that each document d_j has a different representation for each category c_i; in practice, though, this means that different subsets of d_j’s original representation are used when categorizing under the different categories;
- *Global feature selection*: features are chosen in terms of which the classifier for all categories $C = \{c_1, . . . , c_m\}$ will operate [7, 8, 9].

In our previous work in Arabic text categorization we used global dimensionality reduction [3], there was a problem that some documents when being represented as vectors, all their terms weight is zero (i.e. it contains no term from the super vector). This problem leads to think about using local term selection which solves this problem but with low classification rate that of global feature selection.

This paper proposes a hybrid feature selection approach and compares it versus both local and global feature selection. Results reported on a set of 1132 documents of six different topics show that the proposed hybrid feature selection surpass the local and global feature selection in selecting effective terms and improves the Arabic documents categorization efficiency.

This paper is organized as follows. Section 2 describes our previous work in Arabic text categorization. Feature selection approaches and the proposed hybrid feature selection approach are discussed in section 3. In section 4 the achieved experimental results are presented .Finally, the conclusion is presented in section 5.

2. Previous work

Many text categorization systems have been developed for English and other European languages, however few researchers work on text categorization for Arabic language. In a previous work, Arabic text categorization systems have been proposed [3].

Figure 1 shows the different phases of the system. The document in text categorization system must pass through a set of steps: document conversion which converts different types of documents into plain text, stop word removal to remove insignificant words, stemming to group words sharing the same root, feature selection/extraction, super vector construction, feature weighting, classifier construction, classification and evaluation of the classifier.

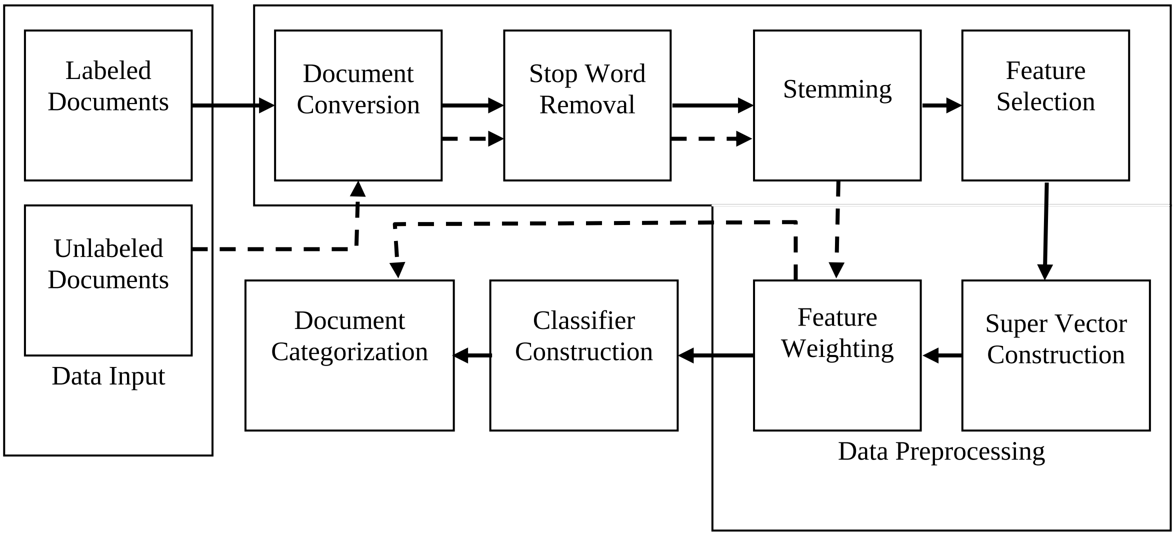

Figure 1. Arabic text categorization system.

2.1. Stop words removal and stemming

2.1.1. Stop words removal

Stop words like prepositions and particles are considered insignificant words and must be removed. A list of 165 word was prepared to be eliminated from all the documents.

2.1.2. Stemming

Stemming is the process of removing all affixes from a word to extract its root. It is essential to improve performance in information retrieval tasks especially with highly inflected language like Arabic language. There are three different approaches for stemming: the root-based stemmer; the light stemmer; and the statistical stemmer.

Root-based stemmer [10] uses morphological analysis to extract the root of a given Arabic word, while the aim of the light stemming approach [11] is not to produce the root of a given Arabic word, rather is to remove the most frequent suffixes and prefixes. In statistical stemmer [12], related words are grouped based on various string similarities measures; such approach often involves n-gram which is a set of n consecutive characters extracted from a word. The main idea behind this approach is that, similar words will have a high proportion of n-grams in common.

An improvement has been performed to statistical stemmer by applying light stemmer before performing similarity measure in order to maximize the performance of the statistical stemmer. Results show that the hybrid approach of light and trigram stemming with is the most suitable stemming approach for Arabic text categorization.

2.2. Document indexing

After stop words removal and words stemming, documents are indexed and represented as a vector of weighted terms. In true information retrieval style, each document is usually represented by a vector of *n* weighted terms; this is often referred to as the bag of words approach to document representation [2]. In this approach the structure of a document and the order of words in the document are ignored. A global super vector is constructed. It consists of all the distinct words (also called terms) that appear in all training samples of all classes after removing the stop words and words stemming.

2.2.1 Term selection

Typically, there can be thousands of features in document classification. Hence, a major characteristic, or difficulty of text categorization problems is the high dimensionality of the feature space. For this, term selection techniques are used to select from the super vector terms a subset of terms that are deemed most useful for compactly representing the meaning of the documents. Term selection is also beneficial in that it tends to reduce over fitting, (i.e. the phenomenon by which a classifier tends to be better at classifying the data it has been trained on than at classifying other data). Usually, term selection techniques consist of scoring each term in the super vector by means of a term evaluation function *f* (TEF) and then selecting a set of terms that maximize *f*. Many term evaluation functions have been introduced for term selection for English text categorization [13, 14, 15]. These functions are Document Frequency Thresholding, Information Gain, CHI Square, Odds Ratio, NGL Coefficient and GSS Score.

A hybrid approach between document frequency thresholding and information gain is used. Document frequency is used to remove rare terms and information gain to select most informative terms from the remaining list.

2.2.2 Term weighting

After selecting the most significant terms in the super vector, each document is represented as a weighted vector of the terms found in the super vector. Every word is given a weight in each document. There are many suggested weighting schemes [15] such as Boolean weighting, Term Frequency (TF) weighting, Term Frequency Inverse Document Frequency (TFIDF) weighting, and Normalized-TFIDF weighting. Normalized-TFIDF schema is chosen as the best schema for term weighting.

2.3 Classification

Two different non-parametric classifiers have been used; k-NN and Rocchio classifiers. Results show that Rocchio classifier is superior over k-NN classifier in both time and accuracy.

3. Feature selection approaches

There are two approaches for feature selection: local and global selection [1]. Let a set of training examples D={d_1, d_2, …, d_n}, where each document belongs to one of a set of categories C={c_1, c_2, …, c_m}.

3.1 Local feature selection

In local feature selection, feature set *f* is extracted from each category of interest (positive class) of which the specific classifier will operate. Feature set extracted from c_1 thus will be differed from feature set derived from category other than c_1. This would mean that each document d_j has a different representation for each category c_i; in practice, though, this means that different subsets of d_j 's original representation are used when categorizing under the different categories.

The feature set *f* can be harvested in two ways: Selecting terms that belongs only to the interested class using relevant documents only (local dictionary); or combining features of both the positive and negative classes using relevant and irrelevant documents (universal dictionary) [16]. In this study, universal dictionary is used as suggested in [16].

There are many feature selection measures used in literatures of text categorization. These measures are Document Frequency Thresholding, Information Gain, CHI Square, Odds Ratio, NGL Coefficient and GSS Score.

3.2 Global feature selection

In global feature selection a feature set *f*, is extracted from all the classes C={ c_1, c_2, …, c_m }. Selected set must preserve and obtain if possible every category-specific significant feature that may be important to classification task and can only safely removes features that will not be relevant to classification task. In this approach all documents have the same representation for all classes.

3.3 Hybrid feature selection

3.3.1 Problem description

In a past proposed [3] Arabic text categorization system we faced a problem in the feature selection phase. When global feature selection approach was used, most of the documents did not contain any term in the list of the selected terms (empty documents). In other words, term evaluation functions select terms with rare appearance in the data set (i.e. terms with very low document frequency). This problem motivated the use of a hybrid approach between document frequency thresholding information gain. Document frequency is used to remove rare terms and information gain to select most informative terms from the remaining list. It was suggested also to use local feature selection.

In this study local feature selection is used. It was noticed that the number of empty documents is reduced but still it gives classification rate less than that of global feature selection.

3.3.2 Combining local and global feature selection

As discussed before, each of local and global feature selection has an advantage and disadvantage. Global feature selection gives high performance but generates a large number of empty documents, and local feature selection gives less performance but reduces the number of empty documents dramatically. The two approaches are combined to gain their advantages and discard their disadvantages. Figure 2 shows the proposed hybrid feature selection algorithm.

The proposed algorithm combines the two feature selection approaches. It selects set of global features and sets of local features for each class and represents the documents with the two representations. In classification phase the global vector of the document is used first if it is not one of the empty documents, if the documents in an empty one then the local feature vectors are used. By this way we gain the advantages of the two approaches: the high classification rate of global approach the low number of empty documents of the local approach. A result in the coming section illustrates this conclusion.4 Experimental results and analysis

```
Step0: Perform preprocessing routines for the documents. This routines includes stop word removal and stemming.

Step1: Select a global set of features for all classes.

Step2: For each class $c_i$ Select a local set of features.

Step3: Represent each document $d_i$ of the training documents as a weighted vector of the global feature set and as a weighted vector for each of the $m$ classes local feature set.

Step 4: In classification phase:
For each document $d_i$
        If($d_i$ is not empty according to the global vector)
        Then Classify $d_i$ using its global vector
        Else Classify $d_i$ using its local vectors
```

Figure 2. Hybrid feature selection approach

4. Experimental results and analysis

4.1 Text Collection

We have collected our own text collection which consists of 1,132 documents and contains 39,468 word collected from the three main Egyptian newspapers El Ahram, El Akhbar, and El Gomhoria during the period August 1998 to September 2004. These documents cover 6 topics. Table 1 shows the number of documents for each topic. Documents have average size of about 117 words before stemming and stop words removal. Documents chosen represent the first paragraph of an article. This choice was preferred because it usually contains an abstract to the whole article.

Table 1. Number of documents for each topic in the text collection.

Topic	No. of documents
Arts	233
Economics	233
Politics	280
Sports	231
Woman	121
Information Technology	102

4.2 Experimental Results

The system discussed in section 2 is used for classifying the documents. Stemming is performed using combination of light and n-gram statistical stemmers. Global and local features are selected using document frequency thresholding and information gain measures. Different thresholds for document frequency are used. Rocchio classifier is used for classification and evaluation performed using commonly used MacroAverageF1 evaluation criterion. Rocchio classifier is used for its simplicity and because it takes into consideration positive and negative exemplars which is equally to using the universal dictionary in feature selection. In the classification phase we consider the empty document as a misclassified document.

Experiments performed on our data set show that the proposed hybrid feature selection gives a high classification rate equivalent to that of global feature selection besides reducing the number of empty documents.

Figures 3 and 4 show the macroAverageF1 measure of the three feature selection techniques with different document frequency threshold. Terms with document frequency below the threshold are removed. It is clear that hybrid approach gives high classification rate.

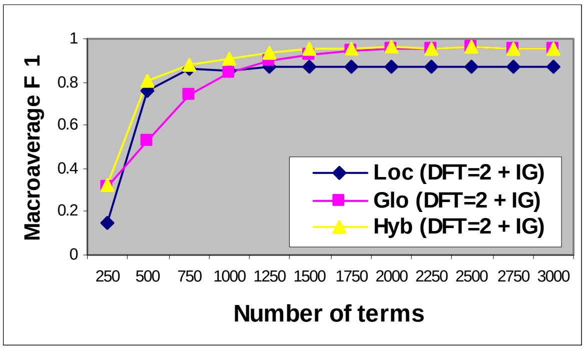

Figure 3. Macroaverage F1 values for the three feature selection approaches using Document frequency threshold=2 combined with information gain

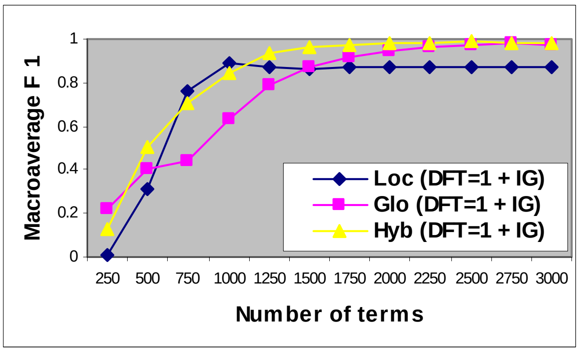

Figure 4. Macroaverage F1 values for the three feature selection approaches using Document frequency threshold=1 combined with information gain

Figures 5 and 6 show the number of empty documents of the three feature selection techniques with different document frequency threshold. It can be observed that hybrid approach reduces the number of empty documents and thus approaching those of local approach.

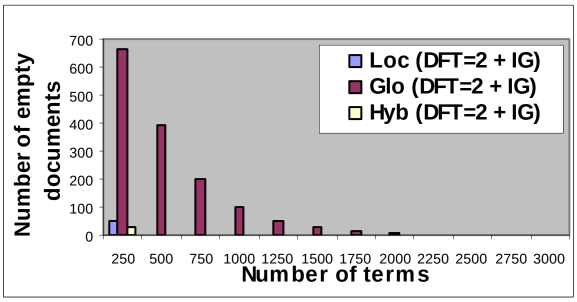

Figure 5. Number of empty documents for the three feature selection approaches using Document frequency threshold=2 combined with information gain

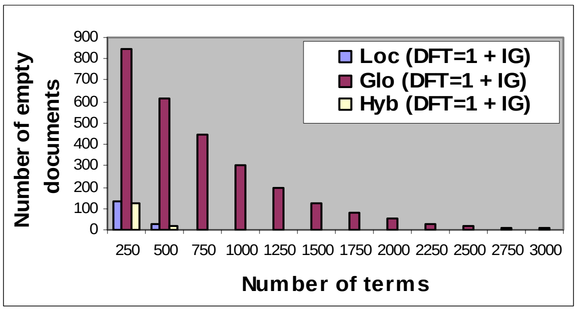

Figure 6. Number of empty documents for the three feature selection approaches using Document frequency threshold=2 combined with information gain

5. Conclusion

This paper proposed a hybrid approach for feature selection. This approach combines local and global feature selection to select effective features to improve efficiency. This approach gives a high classification rate comparativelly with global feature selection and, in the same time, reduces the number of empty documents.

References

[1] F. Sebastiani, “Machine learning in automated text categorization**”,** *ACM Computing Surveys, 34(1),* 2002, pp. 1-47.

[2] T. Mitchell, *Machine Learning*, McGraw-Hill, New York. 1997.

[3] Mostafa M. Syiam, Zaki T. Fayed, and Mena B. Habib, "An Intelligent System for automated Arabic Text Categorization" *International journal of intelligent computing and information systems IJICIS, 6(1*), 2006, pp. 1-19.

[4] H. T. Ng, W. B. Goh, and K. L. Low, "Feature selection, perceptron learning, and a usability case study for text categorization", *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, Philadelphia, US, 1997, pp. 67–73.

[5] Z. Zheng, R. Srihari, and S. Srihari, "A Feature Selection Framework for Text Filtering", *Proceeding of ICDM'03, Third IEEE International Conference on Data Mining,* 2003, pp. 705-708.

[6] B. C. How, and K. Narayanan, "An Empirical Study of Feature Selection for Text Categorization based on Term Weightage", *Proceeding of WI'04, IEEE/WIC/ACM International Conference on Web Intelligence,* 2004, pp. 599-602.

[7] B. C. How, and W. T. Kiong, "An Examination of Feature Selection Frameworks in Text Categorization", *Lecture Notes in Computer Science, Springer Berlin / Heidelberg, volume 3689 / 2005,* 2005, pp. 558 – 564.

[8] Y. Yang, "An evaluation of statistical approaches to text categorization", *Information retrieval, 1-2(1),* 1999, pp.69–90.

[9] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization", *Proceedings of ICML-97, 14th International Conference on Machine Learning,* Nashville, US, 1997, pp. 412–420.

[10] S. Khoja, “Stemming Arabic Text”. Lancaster, U.K., Computing Department, Lancaster University. 1999.

[11] M. Aljlayl and O. Frieder, “On Arabic search: improving the retrieval effectiveness via a light stemming approach”. *Proceedings of ACM CIKM 2002 International Conference on Information and Knowledge Management*, McLean, VA, USA, 2002, pp. 340-347.

[12] S. H. Mustafa and Q. A. Al-Radaideh, “Using N-grams for Arabic text searching”, *Journal of the American Society for Information Science and Technology, 55(11),* 2004, pp. 1002–1007.

[13] M. Rogati and Y. Yang, “High-Performing Feature Selection for Text classification”, *Proceedings of the eleventh international conference on Information and knowledge management CIKM'02,* 2002, pp 659 - 661.

[14] T. Liu, S. Liu, Z. Chen and Wei-Ying Ma, “An Evaluation on Feature Selection for Text Clustering”, *Proceedings of the 12th International Conference ICML 2003,* Washington, DC, USA, 2003, pp. 488-495.

[15] K. Aas and L. Eikvil, “Text categorisation: A survey”, *Technical report, Norwegian Computing Center,* 1999.

[16] Z. Zheng, , and R. Srihari, "Optimally Combining Positive and Negative Features for Text Categorization", *ICML 2003 Workshop*, 2003.