Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies
Mourad Abbas, Kamel Smaïli, D Berkani

To cite this version:
Mourad Abbas, Kamel Smaïli, D Berkani. Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies. 3rd International Conference on Arabic Language Processing, May 2009, Rabat, Morocco. hal-01586533

HAL Id: hal-01586533
https://hal.archives-ouvertes.fr/hal-01586533
Submitted on 13 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—The aim of this study is topic identification by using two methods, in this case, a new one that we have proposed: TR-classifier which is based on computing triggers, and the well-known k Nearest Neighbors. Performances are acceptable, particularly for TR-classifier, though we have used reduced sizes of vocabularies. For the TR-Classifier, each topic is represented by a vocabulary which has been built using the corresponding training corpus. Whereas, the kNN method uses a general vocabulary, obtained by the concatenation of those used by the TR-Classifier. For the evaluation task, six topics have been selected to be identified: Culture, religion, economy, local news, international news and sports. An Arabic corpus has been used to achieve experiments.

Keywords — TR-classifier; k Nearest Neighbors; Arabic corpus, topic vocabulary.

I. INTRODUCTION

Topic identification has been sufficiently studied for Indo-European languages. Generally, the methods used are those of text categorization: Bayesian classifiers [1, 2, 3], decision tree [2, 3, 4], neural networks [5, 6], kNN “k Nearest Neighbors” [7, 8], etc. Nevertheless, for Modern Standard Arabic, few works have been carried out [9, 10, 11, 12].

The aim of this study is to evaluate two text categorization methods on Arabic corpora. The first one is TR-classifier [13], a new method based on computing triggers, and the second one is the k Nearest Neighbors.

The motivation behind the conception of the TR-classifier is that the information present in the longer-distance history is significant [14]. Indeed, for a topic identification task, the presence of the term “guitar” could trigger another list of terms: “music”, “dance”, etc. So, the main idea is to represent each topic by a number of triggers which allow characterizing each topic, and then facilitating the identification.

The second method that we have used, is the well-known kNN. As this method is considered in [15] as one of the top-performing classifiers, we selected it to evaluate TR-classifier by comparing performances of the two methods.

We should note that small sizes of topic vocabularies are chosen for TR-classifier. Whereas, a general vocabulary is needed for kNN, consequently it has been constructed by the concatenation of the topic ones.

The Arabic corpus used in our experiments is downloaded from the website of the Omani newspaper Alwatan; it is composed of more than 9000 articles.

In section II, we give some details both about TR-classifier and kNN method. We talk, in Section III, about the corpus, the representation of documents and the vocabulary construction. Experiments and results are exposed in section IV.

II. METHODS DESCRIPTION

A. TR-Classifier

We start by giving a definition of Triggers as TR-classifier is based on computing them. So, triggers of a word \( w_k \) are the ensemble of words that have a high degree of correlation with it [13, 14, 16]. The main idea of the TR-classifier is based on computing the average mutual information of each couple of words which belong to the vocabulary \( V_i \). Couples of words or "triggers" that are considered important for a topic identification task, are those which have the highest average mutual information (AMI) values [17, 18]. Each topic is then endowed with a number of selected triggers \( M \), calculated using the training corpus of the topic \( T_i \).

Identifying topics by using TR-method consists in:

• Giving corresponding triggers for each word \( w_k \in V_i \), where \( V_i \) is the vocabulary of the topic \( T_i \).

• Selecting the best \( M \) triggers which characterize the topic \( T_i \).

• In the test step, we extract for each word \( w_l \) of the test document, its corresponding triggers.

• Computing \( Q_i \) values by using the TR-distance given by (1):

\[
Q_i = \frac{\sum_{k=1}^{M} AMI(w_k, w_k')}{\sum_{l=0}^{n-1} (n-l)}
\]

Where \( i \) stands for the \( i^{th} \) topic. The denominator presents a normalization of AMI computation. \( w_k' \) are triggers included in the test document \( d \), and characterizing the topic \( T_i \).

• A decision for labeling the test document with topic \( T_i \) is obtained by choosing \( \operatorname{arg\ max} Q_i \).

1 http://www.alwatan.com
B. k Nearest Neighbors

kNN has been applied to text categorization before two decades [19, 8, 20, 15]. Indeed, Yang compared it to a set of text categorization methods using the benchmark Reuters corpus (the 21450 version, Apte set) [15]. It has been found that KNN is one of the top-performing methods after SVM [15]. Many other researches have found that the kNN method achieves very good performances by using different data sets [15, 21, 22].

The strategy of the kNN algorithm is quite simple, so that, to identify a topic-unknown document \( d \), kNN ranks the neighbors of \( d \) among the training document vectors, and uses the topics of the k Nearest Neighbors to predict the topic of the test document \( d \). The topics of neighbors are weighted using the similarity of each neighbor to \( d \). In order to measure this similarity, the cosine distance is used, although other measures are possible, as the Euclidean distance. The cosine similarity is defined by (2).

\[
sim(D_j, D_i) = \frac{\sum_{k=1}^{V} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{V} d_{jk}^2 \sum_{k=1}^{V} d_{ik}^2}}
\]

where \( D_j \) is the \( j \)th training document and \( D_i \) is the \( i \)th test document. \( V \) is the size of the general vocabulary. \( d_j \) and \( d_i \) represent the weights of the words belonging respectively to \( D_j \) and \( D_i \).

To assign the test document \( d \) to the correct topic, a cutoff threshold is needed [15].

III. CORPUS REPRESENTATION

We started by downloading Arabic texts from the archives of the Omani newspaper Alwatan of the year 2004. The size of the extracted corpus is about 10 millions terms which correspond to 9000 articles, distributed over six topics, in this case: Culture, religion, economy, local news, international news and sports.

90 % of these articles are reserved to training and the rest to the evaluation.

We should note that we have realized some elementary operations for topic identification, as eliminating insignificant words that do not bring any information, as function words, and also words whose frequencies are less than a definite threshold.

We dress in table 1 the size of the entire corpus, before and after removing insignificant words.

<table>
<thead>
<tr>
<th>Topics</th>
<th>N. words before</th>
<th>N. words after</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>1,339,210</td>
<td>1,013,703</td>
</tr>
<tr>
<td>Religion</td>
<td>3,122,565</td>
<td>2,133,577</td>
</tr>
<tr>
<td>Int. news</td>
<td>855,945</td>
<td>630,700</td>
</tr>
<tr>
<td>Economy</td>
<td>1,460,462</td>
<td>1,111,246</td>
</tr>
<tr>
<td>Loc. news</td>
<td>1,355,635</td>
<td>1,182,299</td>
</tr>
<tr>
<td>Sports</td>
<td>1,423,549</td>
<td>1,067,281</td>
</tr>
<tr>
<td>Total</td>
<td>9,813,366</td>
<td>7,139,486</td>
</tr>
</tbody>
</table>

The construction of the vocabulary has been made by using the term frequency method which gives good results though its simplicity [23]. Other terms selection methods as Mutual Information [24] and Document Frequency lead also to satisfactory performances.

The KNN method uses a general vocabulary, whereas the TR-classifier, uses a vocabulary per topic, i.e., six topic vocabularies are built, in our case.

We should note that these vocabularies are very small; indeed the size of each topic vocabulary is 300 terms. Nevertheless, they are composed of terms ranked from the maximum, to the minimum according to their frequencies. The reason behind the vocabularies size reduction is to make the topic identification process faster.

Documents need to be transformed to a compact vector form, and the dimension of the vector corresponds to the size of the vocabulary. Each word of the document is weighted by a definite value. The weights or vector components are those commonly used in text categorization, particularly for the TFIDF classifier [25]. Hence, after removing insignificant words, we calculated both the frequency of each word, which is called Term Frequency, and the Document Frequency of a word \( w \), that means the number of documents in which the word \( w \) occurs at least once. The weight of each term results then from the product of Term Frequency and Inverse Document Frequency [25, 26, 27].

IV. EXPERIMENTS AND RESULTS

A. TR-Classifier Performances

As we mentioned in section III, TR-classifier uses a vocabulary per topic, and the words of each vocabulary are ranked according to their frequencies. In these experiments we used much reduced sizes of the six topic vocabularies, in this case 300 terms.

The evaluation of the TR-classifier has been made by varying both topic vocabularies sizes and triggers number \( N \). So, we present performances on average in figure 1.
The number of triggers is important to achieve good performances. According to figure 1, we notice that average Recall (R) and Precision (P) rates improve when increasing the triggers number. Indeed, for \( N=20 \), R and P are respectively equal to 54.88 % and 57.30 %. These values continue to be enhanced when \( N \) is augmented to 250, in this case \( R=89.67 \% \) and \( P=86.44 \% \).

**B. kNN Evaluation**

The main computation made by kNN is the sorting of training documents in order to find the k Nearest Neighbors for the test document. The value of k is usually optimized by several trials on the training and validation sets.

A general vocabulary is used by the kNN method. Thus we constructed a vocabulary by concatenating the six topic vocabularies used in the previous experiment. The resulted size is 800 words.

Performances are lower than those of TR-Classifier by 14 % which is considered as an important difference between the two methods. We present in figure 2, performances in terms of Recall of the two classifiers, for the six studied topics.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>76</td>
<td>49.78</td>
</tr>
<tr>
<td>Religion</td>
<td>75.33</td>
<td>94.95</td>
</tr>
<tr>
<td>Economy</td>
<td>68.66</td>
<td>81.74</td>
</tr>
<tr>
<td>Local</td>
<td>69.33</td>
<td>70.27</td>
</tr>
<tr>
<td>International</td>
<td>80</td>
<td>85.11</td>
</tr>
<tr>
<td>Sports</td>
<td>84.66</td>
<td>92.70</td>
</tr>
<tr>
<td>Average</td>
<td>75.66</td>
<td>70.09</td>
</tr>
</tbody>
</table>

**TABLE II. PERFORMANCES OF THE KNN METHOD**

In perspectives, we aim to enhance TR-Classifier performances by using superior sizes of vocabularies, though it outperforms kNN by 14 %, which is considered as a satisfactory result.

**REFERENCES**


