# TPT: An Empirical Term Selection for Arabic Text Categorization

**Mourad Abbas**
High Council of Arabic Language
Algiers, Algeria
abb.mourad@gmail.com

**Mohamed Lichouri**
Algiers, Algeria
medlichouri@gmail.com

## Abstract

In this paper, we will investigate an empirical term selection method for text categorization, namely Transition Point (TP) technique, and we will compare it to two other widely used methods: Term Frequency (TF) and Document Frequency (DF). For evaluation, we have used the well-known TFIDF technique. Experiments have been conducted by using the Arabic corpus Khaleej-2004 which is composed of 4 categories. The results obtained from this study show that performance is almost the same for the three techniques. However, we should note that TP is advantageous since it uses a vocabulary much smaller than the ones used in TF and DF.

## 1 Introduction

Up to now, all studies about text categorization that have been carried out by using different approaches and algorithms led to relatively satisfactory results, but did not lead to a hopeful and significant improvement, and so did many systems which are based on machine learning and statistical approaches. This is due in part to term selection methods which are not powerful enough. More studies and experiences must be done intensively to achieve this goal. In this regard, we focus our interest on comparing a relatively recent technique namely Transition Point to the two well-known Term Frequency and Document Frequency methods. These last two are used as a reference in this work since they are ranked, in many studies, among the most efficient feature selection methods. The TP technique has been used in different works like: text categorization (Moyotl Hernández and Jiménez Salazar, 2004), (Moyotl-Hernández and Jiménez-Salazar, 2005), summarization (Bueno et al., 2005), clustering of short texts (Jimenez et al., 2005), keyphrases extraction (Tovar et al., 2005), and weighting models for information retrieval systems (Cabrera et al., 2005). After using this feature selection technique in the aforementioned works, we believe that using it as a term selection process for text categorization could yield good performance. The focus in this paper is mainly evaluating the TP technique by using an Arabic corpus and comparing it to TF and DF. Section 2 is dedicated to related work. Section 3 describes succinctly the TP technique. Section 4 presents the experiments and the results. Finally, we conclude in section 5.

## 2 Related Work

Many works on feature selection for text categorization, particularly Term Frequency and Document Frequency had been achieved. In fact, Term frequency had been used in (Yang and Pedersen, 1997), (Abbas et al., 2010), (Abbas et al., 2011) and yielded good results. The selected words are those of high frequency of occurrence. Indeed, the stop words are not included in this selection. For Arabic, Term Frequency had been tested using Khaleej-2004 corpus (Abbas and Smaili, 2005) and the corresponding results are shown in Figure 3. For Document Frequency, the basic idea is that rare terms are non-informative for category prediction. Hence, terms whose document frequency are less than a predetermined threshold are removed. Yang and Pedersen show in (Yang and Pedersen, 1997) that reducing the training corpus by a factor of 10 did not affect performance, and caused only a slight degradation when reducing it by a factor of 100. They stated that the performance achieved by using DF is good and approximates that of Information Gain. In (Brun, 2003) it had been shown that the performance of the TFIDF technique using two vocabularies [1] obtained from TF and DF were about 74.3% and 83.1%, respectively[2]. However, only few works are dedicated to Transition Point. Indeed, Pinto et al. proposed a procedure to cluster abstracts of scientific texts by applying the

[1] The size of the 2 vocabularies are 30000 distinct words.

[2] Here, performance is in terms of Recall.

transition point technique during the term selection process (Pinto et al., 2006). They found that TP outperforms TF and TS (Term Strength). From the research conducted by Moyotl and Jimenez (Moyotl Hernández and Jiménez Salazar, 2004) in which they tested DF, Information Gain and $\chi^2$ in combination with TP, it shows that the DF-TP pair gives the best result. They concluded also that selecting terms lesser than the transition point discarded noise terms with maintaining the performance of categorization. We consider these preliminary encouraging results as the main motivation for testing TP. In the following, we will describe the TP method.

## 3 Transition Point Technique

TP technique is based on the Zipf Law (Zipf, 2016) and also on the studies of Booth (Booth, 1967). In these works, it has been shown that terms of medium frequency are narrowly related to the content of a document (Pinto et al., 2006). This is the motivation for using the terms whose frequency is closer to TP as indexes of a document. It should be noted that TP is a frequency that splits the vocabulary of a document into two groups of terms, with respectively high and low frequency. It can be calculated by using the formula (1):

$$TP_V = \frac{\sqrt{8I_1 + 1.} - 1}{2} \tag{1}$$

$I_1$ stands with the number of words occurring once in the text $T$. According to Booth's law (Booth, 1967), $TP_V$ can be determined by identifying the lowest frequency, among the highest frequencies, that is not repeated. Hence, the first task to realize is to extract a list of terms with their corresponding frequencies, from the text $T$. The result is a frequency-sorted vocabulary given by: $V = [(t_1, f_1), ..., (t_n, f_n)]$, with $f_k \geq f_{k-1}$, then $TP_V = f_{k-1}$ if $f_k = f_{k+1}$ . After identifying TP, the most important words would be those that frequencies are the closest to TP value (Pinto et al., 2006). These words are presented by the expression:

$$V_{TP} = \{t_k | (t_k, f_k) \in V, U_1 \leq f_k \leq U_2\}, \tag{2}$$

$U_1$ and $U_2$ are respectively lower and upper threshold, they can be calculated by using the formulas:

$$U_1 = (1 - NTP).TP_V \quad (NTP \in [0,1]) \tag{3}$$

$$U_2 = (1 + NTP).TP_V \quad (NTP \in [0,1]) \tag{4}$$

## 4 Experiments and Results

Our experiments are carried out by using the well-known TFIDF method. This type of technique is based on the relevance feedback algorithm proposed by Rocchio (Rocchio, 1971). The idea of the TFIDF algorithm is to represent each document $d$ by a vector $D$ = ($d_1$, $d_2$,..., $d_v$) in a vector space. The vector elements are calculated as the combination of the term frequency $TF(w, d)$, which is the occurrence number of the word $w$ in the document $d$, and the inverse document frequency $IDF(w)$ (Salton, 1991; Rosenfeld and Huang, 1992). $DF(w)$ is the number of documents in which the word $w$ occurs at least once. The value $d_i$ is called the weight of word $w_i$ in the document $d$, and is given by : $d_i = TF(w_i, d)*IDF(w_i)$ with $IDF(w_i) = log(DF(w_i)/N)$, where $N$ is the total number of documents. In order to calculate the similarity between a document $D_i$ and the category $D_j$ we used the equation 5. A document is assigned to the category which gives the highest similarity.

$$Sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{|V|} (d_{jk})^2 \sum_{k=1}^{|V|} (d_{ik})^2}} \tag{5}$$

### 4.1 Khaleej-2004 corpus

We built Khaleej-2004 corpus[3] by downloading thousands of articles from an online arabic newspaper. The corpus is divided into four categories, namely:*Sports*, *International news*, *Local news* and *Economy*. Table 1 shows the number of documents for each category. We carried out the usual operations for data preprocessing such as removing all signs of punctuation and stop words. An overview on the size of the corpus before and after removing stop words is presented in Table 2. The size of the resulted corpus becomes 2.172.000 words, i.e reduced by 23.90%.

### 4.2 Terms Representativity

High frequencies of words usually indicate that they are more informative (except stop words).

[3] Khaleej-2004 corpus had been released in 2010, it can be downloaded from:
(http://sites.google.com/site/mouradabbas9/corpora).
(http://sourceforge.net/projects/arabiccorpus/files).

Table 1: Khaleej-2004 corpus

| Topic | Documents | Words |
|---|---|---|
| Economy | 909 | 578.000 |
| Int.news | 953 | 754.000 |
| Loc.news | 2398 | 893.000 |
| Sports | 1430 | 628.000 |
| Total number | 5690 | 2.853.000 |

However, some words of high frequency of occurrence are not informative at all since they belong to more than one category and their frequencies are not very different from each other. For example, the word *year* \' *A m*\, which is considered among the most frequent words in *Economy* category, is not representative because its frequency is also high in the other categories. The 11 most frequent words in each category are extracted from their related corpora and presented in Table 3, written in International Phonetic Alphabet (IPA) and translated to English. Of course, choosing this limited number of words presents simply an illustration to give an overview on the advantages of term frequencies and their limits of being informative in the representation of categories. Other words, in contrast, represent faithfully and rigorously their categories. For example, as presented in Figure 1, the word *Match* \m b A r A t\ which is the most frequent word in the *Sports* category is very rare to find in other categories. It is the same for the word *American* \' m r I k y h\ that we found only in the *International news* category [4], because America is frequently present on the international scene.

In Figure 1, subfigures (a), (b), (c) and (d) present the distribution of the top selected terms extracted from the categories *Local news*, *Sports*, *International News* and *Economy*, respectively. We define this distribution as relative frequencies ($RF$) of terms, given by the values $RF = F_w/N_c$, where $F_w$ stands for the frequency of the word w in the category $C$ and $N_c$ represents the total number of the category $C$.

In each of these subfigures, four curves are presented. One curve concerns the distribution of the words of the category in question, and the three other ones deal with the distribution of the same words over the remaining categories.

## 4.3 Experiments on Transition Point Technique

Booth presented interesting ideas about occurrences of words (Booth, 1967) and tried to extract a law which purpose is to explain and illustrate the case of words of very low frequency of occurrence. For instance, he studied the ratios $I_1/D$, i.e. the ratio of the number of words occurring once to the number of different words for each of the texts. He equally investigated *the remarkable constancy* of this ratio. All the experiments realized by Booth have used English texts [5]. However, he stated that *there is no reason to suppose that the rather arbitrary assumption used to deduce $I_1$ would be equally valid in languages other than English.*
This is another motivation for us to test and evaluate TP technique on Arabic corpora. Terms of high frequency of occurrence are known to be more representative, and then allow to have good results for text categorization tasks. However, The statistics of terms' representativity presented in subsection 4.2 show that some terms are of high frequency of occurrence in all the 4 categories, which means that they are not representative even they are highly occurring. -see Figure 1-.
Relying on Booth assumptions and on the results presented in subsection 4.2, the TP could be viewed as an idea which allow to extract efficient features. Indeed, the idea to find a value $TP_V$ and then extract the terms localized around it seems to be efficient and outperforms the methods based only on high frequencies of terms.
We obtained different vocabularies by using different values of $NTP$. Figure 2 plots recall and precision values given many $NTP$ values. As can be seen in this figure, it shows that performance curves related to the four categories increase with $NTP$. Global values of Recall and Precision for $NTP = 0.9$ are equal to $90.75\%$ and $90.5\%$ respectively. The best result was for DF, indeed Recall was about $91.75\%$ and Precision $91.50\%$. While TF outperforms TP very slightly (Recall=$90.80\%$ and Precision=$91.20\%$). The most important point to be mentioned is that the result achieved by TP is obtained by using a vocabulary size of about 2500 distinct words, which is a very small size in comparison with the vocabularies obtained from TF and DF which sizes attain

[4] Within the eleven extracted words.

[5] Three texts (Western Reserve University), and the sampling of newspaper English published by Eldridge (Eldridge, 1911).

Table 2: The corpus before and after stop words removing.

| Topic | Economy | Int. news | Loc. news | Sports |
|---|---|---|---|---|
| Corpus before | 578.000 | 754.000 | 893.000 | 628.000 |
| Corpus after | 440.000 | 567.000 | 680.000 | 485.000 |

Table 3: Most frequent words for each category

| Sports | | Int. News | |
|---|---|---|---|
| English | IPA | English | IPA |
| Match | \m b A r A t\ | President | \r ’ I s\ |
| Team | \f r I q\ | Iraq | \’ i r A q\ |
| Center | \m r k z\ | Year | \’ A m\ |
| Championship | \b ṭ U l h\ | United | \m t ḥ d h\ |
| Team | \m n t kh b\ | Forces | \q w A t\ |
| Year | \’ A m\ | Government | \ḥ k U m h\ |
| First | \’ w l\ | American | \’ m r I k y h\ |
| Olympic | \’ U l m b y h\ | Past | \m A ḍ I\ |
| second | \th A n I\ | States | \w l A y A t\ |
| Tournament | \d w r h\ | Council | \m zh l s\ |
| Ball | \k r h\ | Elections | \’ n t kh A b A t\ |
| Loc. News | | Economy | |
| English | IPA | English | IPA |
| Year | \’ A m\ | Year | \’ A m\ |
| Work | \’ m l\ | Countries | \d w l\ |
| Ministry | \w z A r h\ | Work | \’ m l\ |
| Festival | \m h r zh A n\ | Turf | \q ṭ A ’\ |
| Activities | \f ’ A l y A t\ | Oil | \n f ṭ\ |
| General | \’ A m h \ | Million | \m l y U n\ |
| Health | \s ḥ y h\ | Market | \s U q\ |
| Instruction | \t ’ l Y m\ | Company | \sh r k h\ |
| Center | \m r k z\ | Companies | \sh r k A t\ |
| Number | \’ d d\ | Council | \m zh l s\ |
| Zone | \m n ṭ q h\ | Trade | \t zh A r h\ |

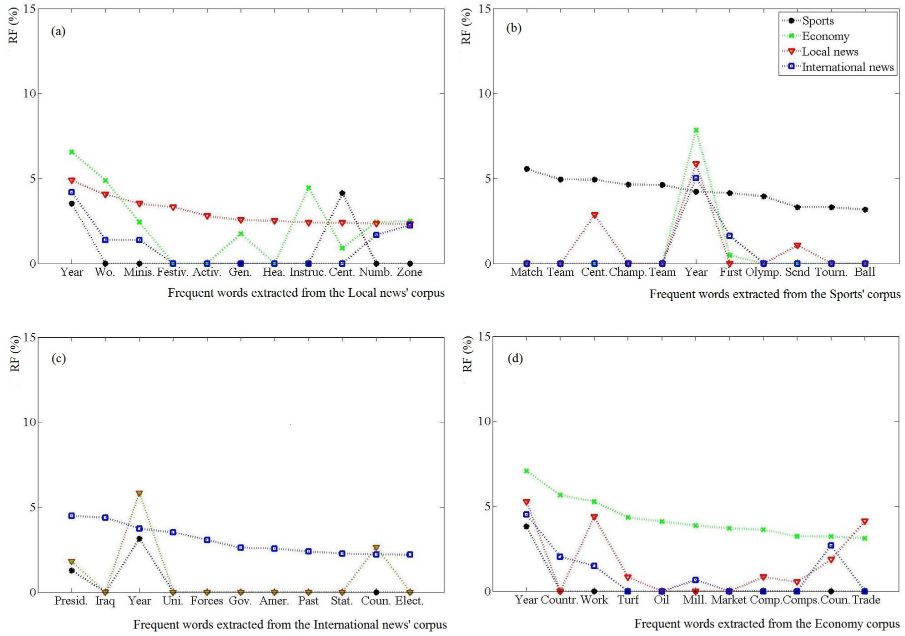

Figure 1: Behavior of RF values of each set of words - presented in Table 3-

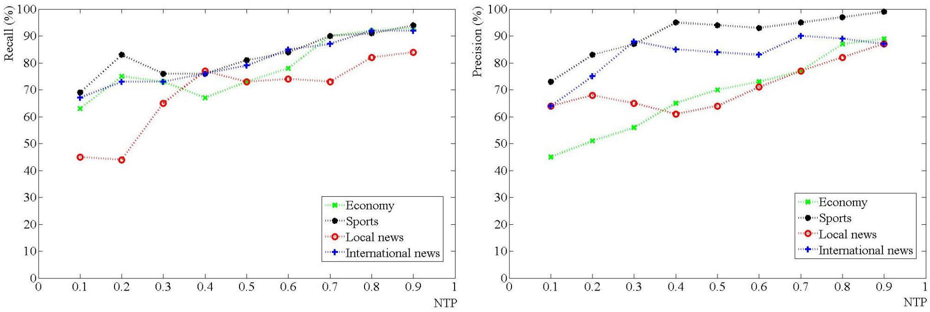

Figure 2: Recall and Precision versus NTP values

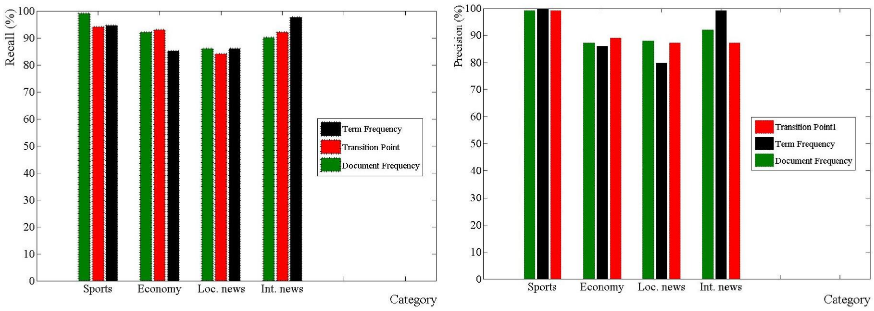

Figure 3: Performance of TP, TF and DF for each category

40000 words. Figure 3 presents the performance of TP, TF and DF related to each category.

## 5 Conclusion

The work presented in this paper is a contribution for the evaluation of the TP technique by using an Arabic corpus. Based on the findings, the obtained results seem to be consistent with other research (Moyotl Hernández and Jiménez Salazar, 2004), (Pinto et al., 2006). The strong point of TP is that we achieved good performance by using a very small corpus. TF and DF are used widely for extracting features. Indeed, in the case of TF, the selected ones are those of high frequency of occurrence. However we presented in section 2 some examples of non-informative words though they are highly frequent. This, we believe that TP could be a good feature selection method for the reasons that we mentioned above. Nevertheless, many other experiments on TP should be realized by using different text collections in order to prove its efficiency. In addition, more efforts must be carried out to find the best method which allows to extract TP because, up to now, it is computed empirically.

## References

M Abbas, K Smaili, and D Berkani. 2011. Evaluation of topic identification methods for arabic texts and their combination by using a corpus extracted from the omani newspaper alwatan. *Arab Gulf Journal of Scientific Research*, 29(3-4):183–191.

Mourad Abbas and Kamel Smaili. 2005. Comparison of topic identification methods for arabic language. In *Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP*, pages 14–17.

Mourad Abbas, Kamel Smaïli, and Daoud Berkani. 2010. Efficiency of tr-classifier versus tfidf. In *2010 First International Conference on Integrated Intelligent Computing*, pages 233–237. IEEE.

Andrew D Booth. 1967. A "law of occurrences for words of low frequency. *Information and control*, 10(4):386–393.

Armelle Brun. 2003. *Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole*. Ph.D. thesis, Université Henri Poincaré-Nancy 1.

Claudia Bueno, David Pinto, and Héctor Jiménez. 2005. El párrafo virtual en la generación de extractos. *Research on Computing Science*, 13:83–90.

Rubı Cabrera, David Pinto, H Jimenez, and D Vilarino. 2005. Una nueva ponderación para el modelo de espacio vectorial de recuperación de información. *Research on Computing Science*, 13:75–81.

RC Eldridge. 1911. *Six Thousand Common English Words: Their Comparative Frequency and What Can Be Done with Them*. Clement Press.

Héctor Jimenez, David Pinto, and Paolo Rosso. 2005. Selección de términos no supervisada para agrupamiento de resúmenes. In *proceedings of Workshop on Human Language, ENC05*, pages 86–91.

Edgar Moyotl Hernández and Héctor Jiménez Salazar. 2004. An analysis on frequency of terms for text categorization. *Procesamiento del lenguaje natural, nº 33 (septiembre 2004); pp. 141-146*.

Edgar Moyotl-Hernández and Héctor Jiménez-Salazar. 2005. Enhancement of dtp feature selection method for text categorization. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 719–722. Springer.

David Pinto, Héctor Jiménez-Salazar, and Paolo Rosso. 2006. Clustering abstracts of scientific texts using the transition point technique. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 536–546. Springer.

Joseph Rocchio. 1971. Relevance feedback in information retrieval. *The Smart retrieval system-experiments in automatic document processing*, pages 313–323.

Ronald Rosenfeld and Xuedong Huang. 1992. Improvements in stochastic language modeling. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Gerard Salton. 1991. Developments in automatic text retrieval. *science*, 253(5023):974–980.

Mireya Tovar, Maya Carrillo, David Pinto, and H Jimenez. 2005. Combining keyword identification techniques. *Research on Computing Science*, 14:157–162.

Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35. Nashville, TN, USA.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.