ARABIC TEXT CATEGORIZATION ALGORITHM USING VECTOR EVALUATION METHOD

Ashraf Odeh¹, Aymen Abu-Errub², Qusai Shambour³ and Nidal Turab⁴

¹Assistanat Professor, Computer Information Systems Department, Isra University, Amman, Jordan
²Assistanat Professor, Part-Time Lecturer, Computer Science Department, University of Jordan, Amman, Jordan
³Assistanat Professor, Software Engineering Department, Al-Ahliyya Amman University, Amman, Jordan
⁴Assistanat Professor, Business Information Systems Department, Isra University, Amman, Jordan

ABSTRACT

Text categorization is the process of grouping documents into categories based on their contents. This process is important to make information retrieval easier, and it became more important due to the huge textual information available online. The main problem in text categorization is how to improve the classification accuracy. Although Arabic text categorization is a new promising field, there are a few researches in this field. This paper proposes a new method for Arabic text categorization using vector evaluation. The proposed method uses a categorized Arabic documents corpus, and then the weights of the tested document's words are calculated to determine the document keywords which will be compared with the keywords of the corpus categorizes to determine the tested document's best category.

KEYWORDS

Text Categorization, Arabic Text Classification, Information Retrieval, Data Mining, Machine Learning.

1. INTRODUCTION

Generally Text Categorization (TC) is a very important and fast growing applied research field, because more text document is available online. Text categorization [1],[7] is a necessity due to the very large amount of text documents that humans have to deal with daily. A text categorization system can be used in indexing documents to assist information retrieval tasks as well as in classifying e-mails, memos or web pages.

Text categorization field is an area where information retrieval and machine learning convene, offering solution for real life problems [10]. Text categorization overlaps with many other important research problems in information retrieval (IR), data mining, machine learning, and natural language processing [11]. The success of text categorization is based on finding the most relevant decision words that represent the scope or topic of a category [23].

TC in general depended on [35] two types of term weighting schemes: unsupervised term weighting schemes (UTWS) and supervised term weighting schemes (STWS). The UTWS are widely used for TC task [24],[25],[26],[13]. The UTWS and their variations are borrowed from
information retrieval (IR) field, and the most famous one is tf.idf (term frequency and inverse document frequency) proposed by Jones [27]. Robertson [28] tried to present the theoretical justifications of both idf and tf.idf in IR.

TC can play [36] an important role in a wide variety of areas such as information retrieval, word sense disambiguation, topic detection and tracking, web pages classification, as well as any application requiring document organization [10]. The text categorization applications: Automatic Indexing [22], Document Organization [22], Document Filtering [22],[21], Word Sense Disambiguation [29].

The goal of TC is to assign each document to one or more predefined categories. This paper is organized as follows; Section 2 tells about the related work in this area, Section 3 discussed our proposed algorithm Section 4, experimental result and Section 5 conclusion.

1.1. Text Categorization Steps

Generally, text categorization process includes five main steps: [3], [35].

1.1.1. Document Preprocessing

In this step, html tags, rare words and stop words are removed, and some stemming is needed; this can be done easily in English, but it is more difficult in Arabic, Chinese, Japanese and some other languages. Word's root extraction methods may help in this step in order to normalize the document's words. There are several root extraction methods, including morphological analysis of the words and using N-gram technique [40].

1.1.2. Document Representation

Before classification, documents must be transformed into a format that is recognized by a computer, vector space model (VSM) is the most commonly used method. This model takes the document as a multi-dimension vector, and the feature selected from the dataset as a dimension of this vector.

1.1.3. Dimension Reduction

There are tens of thousands of words in a document, so as features it is infeasible to do the classification for all of them; also, the computer cannot process such amount of data. That is why it is important to select the most meaningful and representative features for classification, the most commonly selection methods used includes Chi square statistics [4][38], information gain, mutual information, document frequency, latent semantic analysis.

1.1.4. Model Training

This is the most important part of text categorization. It includes choosing some documents from corpus to comprise the training set, performs the learning on the training set, and then generates the model.

1.1.5. Testing and Evaluation

This step uses the model generated from the model training step, and performs the classification on the testing set, then chooses appropriate index to do evaluations.
1.2. Characteristics of Arabic Language

Arabic accent is spoken 300 times more than English language; Arabic characteristics are assorted in abounding aspects. In [30],[37] Nizar summarized most of the Arabic language characteristics.

Arabic script or alphabets: Arabic is accounting from appropriate to left, consists of 28 belletrists and can be continued to 90 by added shapes, marks, and vowels [31]. Arabic script and alphabets differ greatly when compared to other language scripts and alphabets in the following areas: shape, marks, diacritics, Tatweel (Kashida), Style (font), and numerals, and Distinctive letters and none distinctive letters.

Arabic Phonology and spelling: 28 Consonants, 3 short vowels, 3 long vowels and 2 diphthongs, Encoding is in CP1256 and Unicode.

Morphology:

- Consists from bare root verb form that is trilateral, quadrilateral, or pent literal.
- Derivational Morphology (lexeme = Root + Pattern)
- Inflectional morphology (word = Lexeme + Features); features are
- Noun specific: (conjunction, preposition, article, possession, plural, noun)
- Verb specific: (conjunction, tense, verb, subject, object)
- Others: Single letter conjunctions and single letter prepositions

2. RELATED WORKS

Franca Debole et al. [3] presents a supervised term weighting (STW) method. Its work as replacing idf by the category-based term evaluation function tfidf .They applied it on standard Reuters-21578. They experimented STW by support vector machines (SVM) and three term selection functions ( chi-square, information gain and gain ratio). The results showed a STW technique made a very good results based on gain ratio.

Researchers in Man LAN [5] conduct a comparison between different term weighting schemes using two famous data sets (Reuters news corpus by selected up to 10000 documents for training and testing, 20 newsgroup corpus which has 20000 documents for training and testing). And also create new tf:rf for text categorization depended of weighting scheme. They used McNemar's Tests based on micro-average and break-even points performance analysis. Their experimental results showed their tf:rf weighting scheme is good performance rather than other term weighting schemes.

Pascal Soucy [6] created a new Text Categorization method named ConfWeight based on statistical estimation of importance word. They used a three well known data sets for testing their method, these are: Reuters-21578, Reuters Corpus Vol 1 and Ohsumed .They compare the outcome results from ConfWeight with KNN and SVN results obtained with tfidf. They recommend using ConfWeight as an alternative of tfidf with significant accuracy improvements. and also it can work very well even if no feature selection exists.
Xiao-Bing Xue and Zhi-Hua Zhou [8] expose the distribution of a word in the document and effect it in TC as frequently of word appears in it. They design distributional features, which compact the appearances of the word and the position of the first appearance of the word. The method depending on combining distributional features with term frequency and related to document writing style and document length. They modeled it by divided documents to several parts then create an array for each word in parts and its number of appearance. They concluded that the distributional features when combined with term frequency are improved the performance of text categorization, especially with long and informal document.

Youngjoong KO, Jinwoo Park [9] try to progress the TC accuracy by create a new weighting method named TF-IDF-CF which make same adaptation on TF-IDF weighting on vector space depended on term frequency and inverse document frequency. Their method considers adding a new parameter to represent the in-class characteristic in TF-IDF weighting. They experiment the new method TF-IDF-CF by using two datasets Reuters-21578 (6088 training documents and 2800 testing documents) and 20newsgroup(8000 training documents, 2000 testing documents) and compared the result of their method with Bayes Network, Naïve Bayes, SVM and KNN, then they prove TF-IDF-CF achieved very good results rather than others methods.

Suvidha, R. [12] present a method that based on combined between two text summarization (one used TITLE and other used TERMS) and text categorization within used Term Frequency and Inverse Document Frequency (TFIDF) in machine learning techniques. They find out that the text summarization techniques are very useful in text categorization mainly when they are combined together or with term frequency.

Sebastiani, F. [13] explains the main machine learning approaches in text categorization, and focused on using the machine learning in TC research. The survey introduce of TC definition, history, constraints, needs and application. The survey explains the main approaches of TC that covered by machine learning models within three aspects; document representation, classifier construction and classifier evaluation. The survey described several automatic TC techniques and most important TC applications and introduced text indexing, by a classifier-building algorithm.

Al-Harbi S. et al. [14] introduces an Arabic text classification techniques methodology by present extracting techniques, design corpora, weighting techniques and stop lists. They try to create model for training datasets for Arabic text classification used the different seven datasets; Saudi Press Agency (SPA) 1,526 documents, (SNP) 4,842 documents, WEB Sites 2,170 documents, Ten writers 821 documents, Discussion Forums 4,107 documents, Islamic Topics 2,243 documents and Arabic Poems 1,949 documents and implemented a tool for Arabic Text Classification (ATC Tool).

El-Halees A. [15] focus on Arabic text documents classification. He built his classifier called ArabCat to work in corpus that he built real datasets from eight classes (politics, sports, culture, arts, sciences, technology, economy and health). He said his Arabic text documents classification is more performance than all Arabic classification systems.

Kourdi et al. [16] used statistical machine learning algorithm Naive Bayes (NB) to classified non-vocalized Arabic web documents. They built own datasets from five categories consist of 300 web documents for each category. Their method results showed that the average accuracy was 62%.
Sawaf H. et al. [17] achieved classification on large Arabic corpus, named as Arabic NEWSWIRE corpus using statistical methods. Their corpus covers several categories as (politics, culture, economy and sports) with documents size 33 KB. They showed even with no morphological analysis statistical methods for document clustering obtain satisfied results and it very robust and reliable and it could be capable for Arabic Text.

Syiam et al. [18] present an intelligent Arabic text categorization system which used Machine learning algorithms for stemming and selection. They used normalized-tfidf schema for Arabic text categorization. They applied the Arabic text categorization on corpus consists of 1,132 documents from three Egyptian newspapers: El Ahram, El Gomhoria and El Akhbar. They cover 6 categories as Arts 233 documents, Economics 233 documents, Politics 280 documents, Information Technology 102 documents, Woman 121 documents and Sports 231. They suggested hybrid method consist of combining between Document Frequency and Information Gain is the suitable stemming for Arabic text gives average accuracy of 98%.

Hmeidi I., et al. [19] describe study of Arabic text categorization using two machine learning methods these K nearest neighbor (KNN) and support vector machines (SVM). They create their own corpus based on a collection of news articles for training and testing. They showed that these methods results are most efficient but SVM better in prediction.

Hua Li C., Cheol Park S. [20] present a used of Back propagation neural network (BPNN) for text categorization. They improved BPNN to be efficient in Text Categorization approaches by solved some defects such as slow convergence. They showed that the improved model of BPNN was achieved high efficient for categorization within experiment it on standard Reuter-21578.

Abu-Errub A. [38] introduces a new Arabic text classification algorithm using TFIDF and Chi square measurements. The tested document is compared with main categories documents using TFIDF method to determine the tested document's category, and then the Chi square measurement is used to determine the tested document's subcategory. The researcher examines the proposed algorithm using 1090 training documents categorized into 10 main categories and 50 subcategories, the examination of 1100 different documents shows that the proposed algorithm is capable of classifying the tested document into its appropriated subcategory.

Research by Abu-Errub A. et al.[39] introduces a method of extracting the roots of Arabic words. The proposed algorithm consists of stemming the words of the tested document by deleting the prefixes and suffixes of each word. Then the resulted stemmed word is compared with its morphological weights, and finally the extra letters are deleted from the word to produce the root of the original word.

3. PROPOSED ALGORITHM

The corpus that is used for checking process consists of 38081 words. After deleting the repeated words 8112 different words are left. Figure 1 represents the suggested Text Categorization steps:

Step1: Delete stop word from the documents: the proposed algorithm eliminates stop words from the document using the Arabic stop words list which contains 162 words shown in table1:
Step 2: Normalize the rest of the words: such as removing punctuation, converting words to lowercase, stripping numbers out, as following [38]:

- Remove punctuation.
- Delete numbers, spaces and single letters.
- Convert letters (ء ) , (ؤ ) , (ئـ ) , (أ ) , (إ ) into (ا ) , and (ة ) into (ه ) .

Step 3: Apply stemming on words: by applying techniques to find basic form the root of words by removing affixes (suffixes and prefixes) attached to its root.

Step 4: Calculate the weight of each word in the tested document using weighting scheme function that is the product of the term occurrence frequency (tf) and the inverse document frequency (idf). The inverse document frequency of the ith term is commonly defined as log (n/df) where n is the number of documents in the document set, and df is the number of documents in which the term appears, the weight function as shown below

\[ W_{ij} = tf_{ij} \times \log \left( \frac{n}{df_j} \right) \]

Step 5: Choose two keywords that have the largest weight in the tested document after calculating the weights of all words.

Step 6: Find the most suitable category: by comparing the two keywords chosen in step 5 with key words of each category.

Step 7: Return the name of category the has high match percentage.
4. PROPOSED ALGORITHM

To examine the proposed algorithm, a set of Arabic articles covering different topics were chosen. Researchers created eleven categories, a corpus containing 982 documents with varying size and content about (Agriculture, Astronomy, Business, Computer, Economics, Environment, History, Politics, Sport, Tourism, and Religion) used these documents for learning the proposed algorithm system. As shown in Table 2.

<table>
<thead>
<tr>
<th>Category</th>
<th>No. of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>107</td>
</tr>
<tr>
<td>Astronomy</td>
<td>98</td>
</tr>
<tr>
<td>Business</td>
<td>70</td>
</tr>
<tr>
<td>Computer</td>
<td>95</td>
</tr>
<tr>
<td>Economics</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2. Category and number of documents in each category
After that we tested our algorithm using 1000 testing documents. A sample of the results is shown in Table 3 below. As shown in the table, document #1, for example, has 93.7% matching with the environment category, so it can be categorized as an environmental document. Also, document #396 was categorized as a tourism document by 86.4%, and as a business document by 85.3%. Finally, document #741 has 89% matching with the agriculture category, but has 0.0% matching with the business category.

Table 3. Sample of proposed algorithm results

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Doc. 1</th>
<th>Doc. 57</th>
<th>Doc. 265</th>
<th>Doc. 396</th>
<th>Doc. 463</th>
<th>Doc. 596</th>
<th>Doc. 601</th>
<th>Doc. 741</th>
<th>Doc. 823</th>
<th>Doc. 983</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>6.1%</td>
<td>24.7%</td>
<td>1.7%</td>
<td>77.7%</td>
<td>29.8%</td>
<td>19.9%</td>
<td>15.2%</td>
<td>89.0%</td>
<td>1.9%</td>
<td>18.5%</td>
</tr>
<tr>
<td>Astronomy</td>
<td>0.0%</td>
<td>29.8%</td>
<td>72.3%</td>
<td>73.3%</td>
<td>11.6%</td>
<td>32.5%</td>
<td>46.9%</td>
<td>58.9%</td>
<td>2.0%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Business</td>
<td>34.7%</td>
<td>65.2%</td>
<td>44.8%</td>
<td>85.3%</td>
<td>27.3%</td>
<td>96.0%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.9%</td>
<td>54.4%</td>
</tr>
<tr>
<td>Computer</td>
<td>7.0%</td>
<td>10.3%</td>
<td>0.0%</td>
<td>21.0%</td>
<td>40.2%</td>
<td>5.7%</td>
<td>80.3%</td>
<td>17.0%</td>
<td>9.6%</td>
<td>49.5%</td>
</tr>
<tr>
<td>Economics</td>
<td>7.5%</td>
<td>14.6%</td>
<td>0.4%</td>
<td>45.2%</td>
<td>26.7%</td>
<td>42.3%</td>
<td>6.6%</td>
<td>1.3%</td>
<td>7.3%</td>
<td>53.0%</td>
</tr>
<tr>
<td>Environment</td>
<td>92.7%</td>
<td>0.9%</td>
<td>1.6%</td>
<td>2.7%</td>
<td>50.1%</td>
<td>1.0%</td>
<td>29.4%</td>
<td>88.4%</td>
<td>87.4%</td>
<td>93.3%</td>
</tr>
<tr>
<td>History</td>
<td>0.0%</td>
<td>37.5%</td>
<td>98.8%</td>
<td>17.0%</td>
<td>93.9%</td>
<td>26.3%</td>
<td>8.7%</td>
<td>15.3%</td>
<td>1.1%</td>
<td>12.3%</td>
</tr>
<tr>
<td>Politics</td>
<td>26.5%</td>
<td>3.7%</td>
<td>33.7%</td>
<td>61.1%</td>
<td>68.9%</td>
<td>75.0%</td>
<td>18.7%</td>
<td>26.2%</td>
<td>35.6%</td>
<td>22.8%</td>
</tr>
<tr>
<td>Religion</td>
<td>70.7%</td>
<td>75.1%</td>
<td>33.5%</td>
<td>13.1%</td>
<td>68.9%</td>
<td>8.0%</td>
<td>54.3%</td>
<td>22.0%</td>
<td>66.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Sport</td>
<td>14.8%</td>
<td>5.5%</td>
<td>85.8%</td>
<td>6.1%</td>
<td>6.1%</td>
<td>38.1%</td>
<td>38.2%</td>
<td>22.8%</td>
<td>88.2%</td>
<td>47.5%</td>
</tr>
<tr>
<td>Tourism</td>
<td>42.4%</td>
<td>18.3%</td>
<td>66.6%</td>
<td>86.4%</td>
<td>37.6%</td>
<td>12.2%</td>
<td>4.1%</td>
<td>42.2%</td>
<td>0.2%</td>
<td>5.7%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS AND COMPARISON

Text categorization is the process of grouping similar documents into categories based on their contents, which makes information retrieval process easier. This paper introduces a new Arabic text categorization method using vector evaluation method. The proposed method determines the key words of the tested document by weighting each of its words, and then comparing these key words with the key words of the testing corpus. As shown in the experimental results, the proposed algorithm prepared the documents to ensure a better selection of the maximum amount of documents after testing the contents of the proposed was 98% in category and in another category was 93%. The proposed method proves the ability to categorize the Arabic text documents into the appropriate categories with a high precision rate.
REFERENCES


