Final Report for Project P11418EO01

Automatic Multi-Level Categorization of Arabic Text Documents

October 5, 2017

Fawaz Al-Anzi
Department of Computer Engineering
College of Engineering and Petroleum
Kuwait University
PO Box 5969, Safat 13060
Kuwait

and

Dia AbuZeina
Department of Computer Engineering
College of Engineering and Petroleum
Kuwait University
PO Box 5969, Safat 13060
Kuwait

Introduction

This is the final report for the project P11418EO01 entitled "Automatic Multi-Level Categorization of Arabic Text Documents". The starting and ending dates of the project are 1/1/2015 and 30/09/2017.

The exponential growth of online textual data calls for efficient information retrieval (IR) methods for text categorization. In particular, multilevel Arabic text categorization is an active research area that recently received significant attention. It has been long observed that the flat (i.e. one level) text categorization dominates the Arabic text categorization contributions; however, less attention has been devoted to multilevel type of categorization. This research is a comprehensive practical study that focuses on Arabic text categorization with a particular attention to the multilevel type. In this work, we consider different aspects such as the categorization methods, the up-to-date machine learning toolkits, the textual feature extraction methods, and the difficulty in obtaining Arabic textual corpora. The main objective of this research project is to compile a suitable large enough hierarchical Arabic textual corpus that can be used to investigate the current multilevel text categorization methods as well as introducing innovative categorization methods. Of course, the prepared corpus will be made available for other researchers in natural language processing (NLP) and text mining fields. As a result, the objectives of this research have been successfully achieved. The rest of this report explores the works as well as the publications.

Abstract

This project considers a major type of text categorization process that is multilevel text categorization for the Arabic language. In fact, it has been long observed that the difficulty in obtaining hierarchical (i.e. multilevel) Arabic text corpora degrades the research in this active area. Recently, there has been growing interest in NLP and text mining communities to address the corpora availability problem in order to exercise the innovative information retrieval (IR) algorithms. Therefore, we decided to promote the research in this area by compiling a hierarchical corpus, as it is the cornerstone to facilitate the tasks of the Arabic language researchers. That is, the first objective of this project was to prepare a large hierarchical multilevel Arabic text corpus. The first objective was fulfilled by preparing a relatively large corpus that contains about 15,789 documents for training and testing distributed into three levels. The second objective was to use the prepared corpus to investigate the current and new categorization methods for multilevel text categorization. To fulfill the second objective, we performed the categorization task using two methods; the top-down approach using the cosine similarity measure and the top-down approach using the Markov chain method. The top-down approach using the cosine similarity measure is a well-known text categorization method that is widely used in other language such as English, but not yet for Arabic. However, the second approach that is based on the Markov chain is a novel approach in text categorization field. The experimental results shows that the Markov chain based method is a promising approach since its performance outperforms the well-known top-down approach that is based on the cosine similarity measure.

1. Background

Text categorization (TC) is an important research domain that has been quite success in text mining and search engines. However, the exponential growth of digital data requires efficiently handling documents with thousands of categories defined over a large hierarchical structure (also called taxonomy). In this report, categorization and classification will be used interchangeably. At first, this research project serves the Arabic language that is spoken by more than 380 million speakers who extends over large geographical areas in northern Africa and Middle East, Mubarak and Darwish (2014). The literature shows that the research in hierarchical Arabic texts categorization is little compared to other languages such as English. In fact, most of the research in this field focuses on flat (i.e. not hierarchical) text. No doubt, obtaining a sizable textual corpus is a major challenge especially for the hierarchical categories type. That is, the shortage of hierarchical Arabic corpora is a major reason for such research delay.

Today, the big-data environment that naturally organized as a class hierarchy highlights the need for efficient information retrieval (IR) algorithms. For example, electronic or digital archiving is an important information treasure for media and news agencies. It helps to utilize such valuable information to produce good and worthy reports. Journalists and editors tend to use the old information and articles in order to produce new valuable materials that can be used in new articles. Therefore, enhancing methods for instantly retrieving data is crucial since the data size is overwhelming increasing.

The data hierarchical structure is widely used to organize data in order to speed up the retrieval time as well as for understanding the data for better knowledge discovery. Examples of hierarchies include the evolutionary tree of species, the federal budget, and business organizational charts, Guerra-Gómez et al. (2015). Sorower (2010) indicated that the idea behind hierarchical

model comes from the goal to reduce the computational complexity at each level and thus making the learning algorithm efficient. In fact, it is no longer justifiable to preserve text documents without a particular hierarchical arrangement even with small data collections. The most successful paradigm for organizing this mass of information is by organizing such huge data into a tree structure in which a document belongs to a topic in a certain level also belongs to all of its parent topics, ancestors. Figure 1 shows a simple hierarchy of two categories with two level each. Figure 1 shows that utilizing hierarchical structure means gradually minimizing the classification task into smaller subtasks (i.e. divided and conquer) each time moving to a lower level. Therefore, significantly reducing the features set.

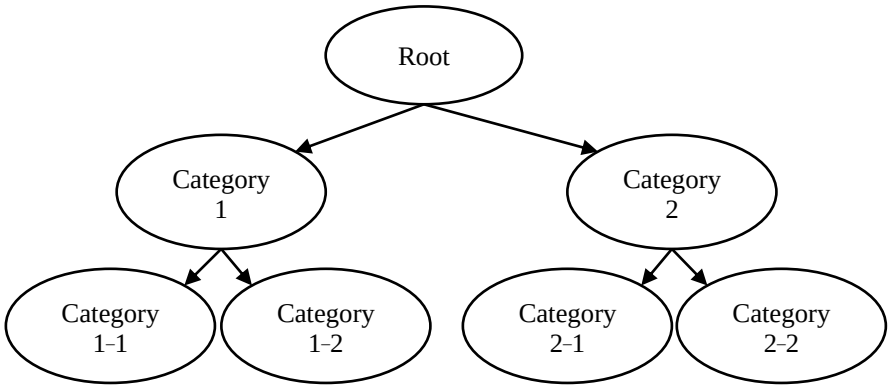

Figure 1. An example of a 3-level tree of two categories

There are many advantages of using hierarchical structure. Such advantages include less computational cost, less retrieval time, better accuracy, etc. In fact, the widespread of huge datasets is probably the main reason of thinking towards hierarchical structurer. In addition, hierarchical structure helps to override hardware limitation as well as satisfying scalability through discarding large portions of data at each level. Moreover, many real-world classification problems already have hierarchical relationships between labels that naturally require hierarchical classification algorithms as indicated by Bi et al. (2012). The same were presented by Silla et al. (2011) as they

indicated that the classes be predicted in many real-world problems are organized into a class hierarchy such as tree or a Directed Acyclic Graph (DAG). Regarding performance, D'Alessio et al. (1998) demonstrate that precision and recall can be significantly improved when considering hierarchy in categorization problem. Joshi et al. (2011) indicated that hierarchical classification aims at increasing the classification accuracy as well as speed. Godbole (2002) indicated that hierarchies provides valuable navigational aid in browsing large text collections like Internet directories. Ying (2011) demonstrated that hierarchical structure can scale well and cope with changes to the category trees. Ruiz et al. (2002) indicated that the use of the hierarchical structure improves text categorization performance with respect to an equivalent flat model. Chakrabarti et al. (1997) indicated that hierarchal databases used for better searching and browsing digital libraries and patent databases. Dhillon et al. (2003) indicated that by using a hierarchy, the classifier utilize a features set that is more relevant to the classifications sub-task at hand and requires only a small number of features.

In general, there are two types of datasets for hierarchical TC, single-label data and multi-label data. In single-label, only one label or category assigned to a document while this constraint is not exist for multi-labels as one or more labels can be assigned to a document. The scope of this research is to perform hierarchical multiclass single-label classification for the Arabic language. The data collection used in this work is category tree that has documents on both the internal and the external nodes. Top-down is the most used strategy for hierarchical classification, Silla et al. (2011). With top-down approach, a classifier used at each hierarchy level to guide the classification process to find the best matched document to be used next as an indicator to narrow down the search space. Svetlana et al. (2006) indicated that top-down approach produces consistent classification when compared to other hierarchical classification methods.

In this work, we used the top-down approach with the cosine similarity measure as well as the top-down approach with Markov chain representation. The proposed method depends on the (first order) Markov chain theory in which the neighbor characters sequences of the documents are used to create the probabilities transition matrices. Hence, each document is represented using a sequence of co-occurrences document characters. Based on the Markov chain method, each category of the corpus is used to create a single probability transition matrix to be used in the classification process. The framework of the Markov chain based method is described in Figure 2. In this figure, each category represented by a Markov chain (a probability transition matrix) that is used to score a document in question. Once scored, a comparison process is performed to find the maximum score (i.e. the most likely category) against the training categories.

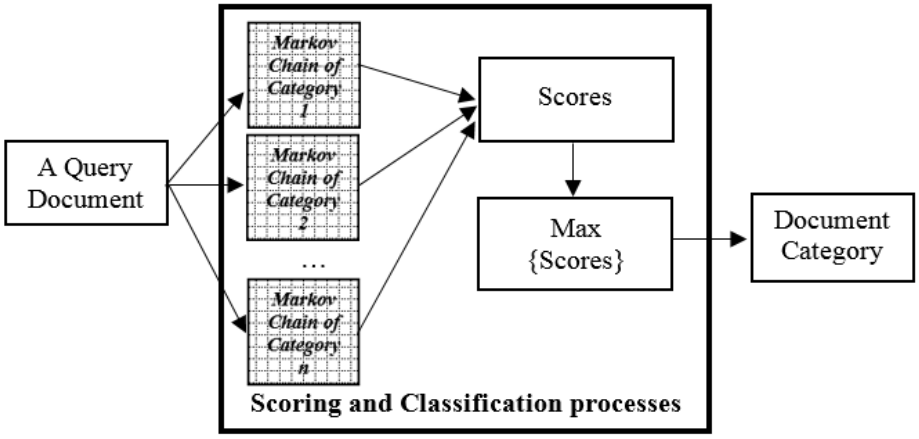

Figure 2. The framework of Markov chain method for one level hierarchy

2. Problem Formulation

Manning and Schütze (1999) defined text classification as the task of classifying texts into one of a pre-specified set of classes based on their contents. According to Sebastiani (2002), text classification is the activity of labeling natural language texts with thematic categories from a

predefined set. However, the distinction in this project is that it handles multilevel text classification using two classifiers; the cosine measure and the Markov chain scores.

Top-down technique is based on using a classifier in each level of the tree as shown in Figure 3. The documents in each level are used for training. The literature has many studies discuss this method. For instances, Chuang et al. (2000) used top-down approach for hierarchical text classification based on a concept hierarchy. Dhillon et al. (2003) presented a feature/word clustering technique for dimensionality reduction purpose. They used top-down strategy for building smaller class models in hierarchical text classification. Chakrabarti et al. (1997) applied a Bayesian hierarchical patent classification system for a taxonomy of 12 subclasses organized in three levels. De Campos et al. (200) developed a Bayesian network-based model for hierarchical text categorization by exploiting the hierarchical and lexical information from the thesaurus. Xue et.al (2008) used top-down approach and support vector machine (SVM) classifier for hierarchical text classification. Secker et al. (2007) used top-down approach for the hierarchical prediction of protein function.

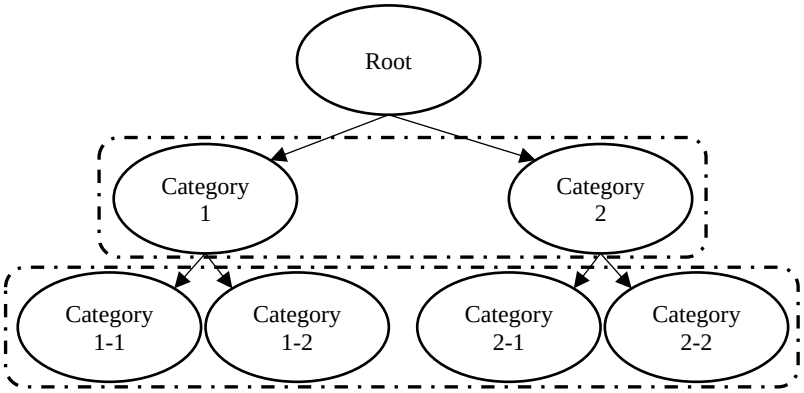

Figure 3. The top-down approach hierarchical text classification

Hence, the problem can be described as shown in Figure 4. Figure 4 shows the process of finding the category of a document in question. The dash line arrow shows the classification path

supposing that the matched document resides in Category 2. The classification process moved to Category 2_1, and then to Category 2_1_1 according the classification output. In Figure 4, the cosine similarity measure is used as a classifier at each level. Feature vectors represent the documents representation in numerical forms based on the selected features.

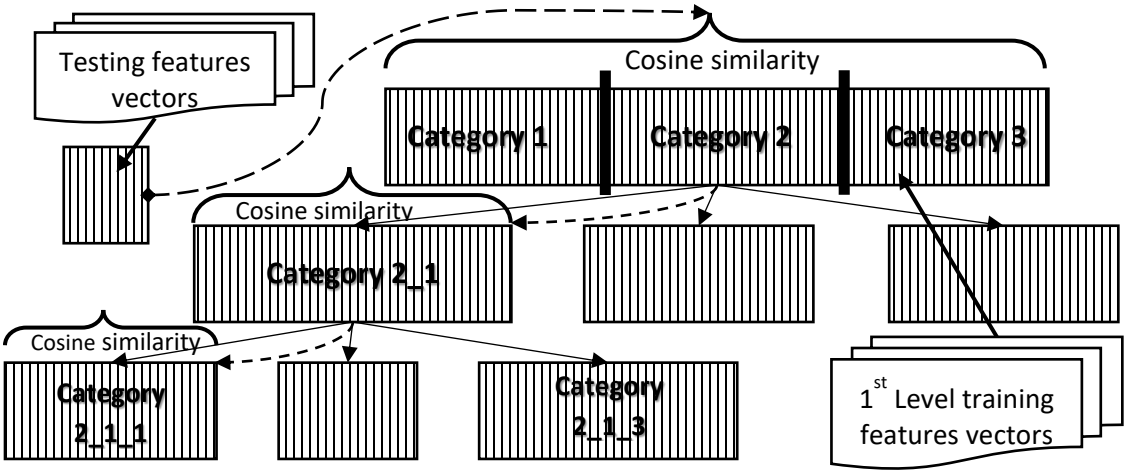

Figure 4. An example of top-down classification process in a 3-level hierarchy

However, the unique contribution of this work is the employment of Markov chains in text classification. Markov chains are increasingly being adopted in real-world computing applications since it provides a convenient way for modeling temporal, time-series data. At each clock tick, the system moves into a new state, which can be the same as the previous one. Markov chains are directed graphs (a graphical model) that are generally used for sequential data-mining tasks. Such tasks are characterized by relatively long sequences for many purposes such as prediction, classification, clustering, information sciences, internet applications, pattern discovery, etc. Rabiner (1989) indicates that there are two reasons for the Markov chains popularity, which it is very rich in mathematical structure and works well in practice for several important applications.

In fact, there are many applications that employ the Markov chain, so it is worthy to take note of the study of Von Hilgers et al (2006) regarding the five greatest applications of the Markov chains. They indicated that the applications are: Scherr's application to computer performance evaluation, Brin and Page's application to PageRank and Web Search, Baum's application to Hidden Markov Models (HMMs), Shannon's application to information theory, and Markov's application to Eugeny Onegin. There are many reference for HMM, Leon-Garcia et al. (2008) demonstrate many concepts of HMM and the related applications. To employ the Markov chain for text classification, a Markov chain model has to be created for each category as shown in Figure 5. Figure 5 also shows the scoring process. A testing document is initially compared with all probability transition matrices at the first level. Once finding the most related category, the process goes down to the second level, and then to the third level. That is, it is a top-down approach.

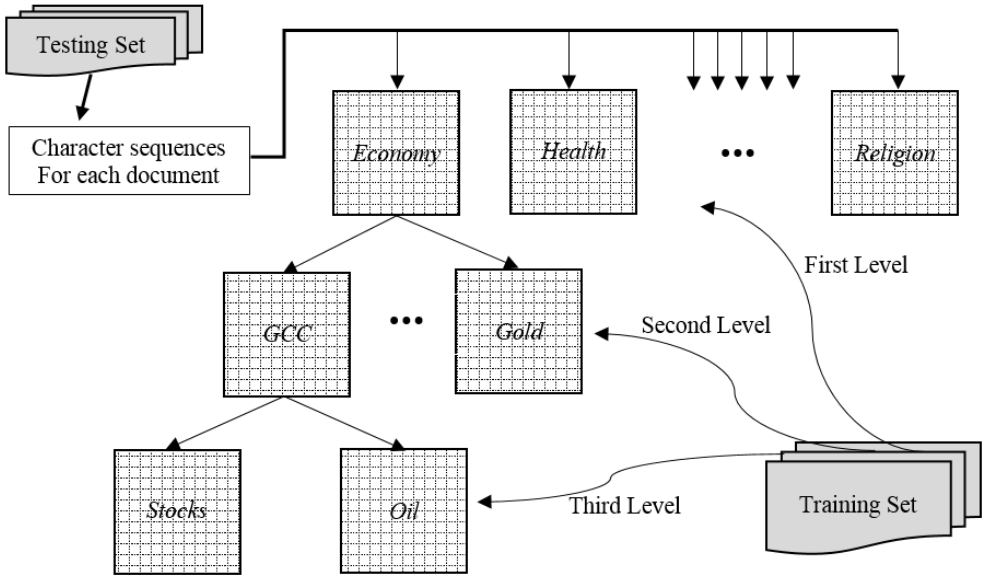

Figure 5. An example of Markov chain classification process in a 3-level hierarchy

3. The Experiments Setup

3.1 Hierarchical Data Collection

For hierarchical data collections, Sun et al. (2003) identified four distinct category structures for hierarchical classification that include virtual category tree, category tree, virtual directed acyclic category graph, and directed acyclic category graph. In our study, we used directed acyclic category graph that allows documents to be assigned into both internal and leaf categories. Towards preparing the proposed corpus, we got 20,179 documents from Alqabas (2016) newspaper - Kuwait. Among the obtained collection, there were many short documents; therefore, we performed filtering that discards the short length documents. That is, all short documents (less than 800 characters) were removed from the corpus. The reason of eliminating such short documents is the desire to have reasonable documents length for training and testing processes. After filtering, the prepared corpus contains 15,789 documents that were split into two sets for training (13,569 documents) and for testing (2,220 documents). The overall document of training set distributed into a 3-levels hierarchy as the following: 3,396 documents in the first level, 5,498 in the second level, and 6,895 in the third level. The training set contains 7,045,457 words that has 236,246 unique words, and the testing set contains 1,126,058 words with 98,648 unique words. Table 1 shows the categories and subcategories labels. In the table, the token "***" means that there is no subcategory for that branch. We emphasize that the categories and the subcategories presented in Table 1 is just an example to be used for this research. However, in real-world applications, different and even complex categories might be prepared and, of course, according to the topics and domains.

Table 1. The corpus information

#	1st Level	2nd Level	3rd Level	#	1st Level	2nd Level	3rd Level
1	Economy (اقتصاد)	GCC economy	Stocks markets	6	Sports (رياضة)	Teams	***
			Oil and gas			Competition	Boxing
		Islamic economy	***				Bicycle
		Industrial countries	Cars				Cars
			Aircrafts				Hybridize
		Gold	***				Football
2	Health (صحة)	Diseases	Aerobic	7	Tourism (سياحة)	Western	USA
			Skin				France
			Cancer			Middle East	Turkey
			Heart				Egypt
		Fat and diabetes	***			GCC	***
		Medical insurance	***	8	Religion (شؤون دينية)	Christian	***
3	Law and justice (امن وقضاء)	Lawyer	***			Islamic	Festivals
		Crime and sanction	Embezzlement				Pilgrimage
			Execution				Fatting
			Jail				Holy Quran
			Drug	9	Parliament (برلمان)	International	***
4	Education (تعليم)	Primary schools	kindergarten			Arabic	Elections
			Elementary				Committee
			Middle	10	Politics (سياسه)	Wars	First
		Higher Education	Scientific Research				Second
			Universities			Kings and presidents	***
		Literacy	***			Refugees	***
5	Technology (تكنولوجيا)	Information security	***			Nuclear weapons	***
		Devices	Apple			Terror	***
			Dell			Peace process - Palestine	***
			Samsung	Total = 15,789 documents			

3.2 The Proposed Method for Top-Down Cosine Measure approach

The purpose of text categorization is to assign a pre-defined class of an input document. Therefore, each document has to be converted into the appropriate form for classification. Bag-of-words is the common textual representation form in text categorization systems. This work is a supervised learning task that has three main steps to perform classification; preprocessing, training, and testing. The proposed method is summarized in the following procedure:

Step 1: preprocessing

1. The stoplist is defined to remove the common words in the documents. For example, the name of the newspaper (i.e. the documents source). In the prepared corpus, the stoplist is defined as the following entries :{القبس، رقم التسلسل، النص، صورة، الخ }.
2. The ignore characters set is defined. This set includes : ® é \ ² … è ² ™ ¯ › × — ô é ü ÷ ¹ × ‘ ° © ½ ¾ ë ‹› ' [ ] ” _ ; ؛ – } { - • = % ’ * , / : ! « ? » ؟ 0 1 2 3 4 5 6 7 8 9 - A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o b q r s t u v w z y z > < | # @ $ + & ^ ،) (¶.
3. The words normalized by replacing some letters to other forms with maintaining the same meanings. The goal of normalization is to reduce the ambiguity that might appear due to characters differences. In addition, all diacritics are eliminated. The replacement include: أ➔ا, إ➔ا, آ➔ا , ة➔ه , ◌َ➔ , ◌ُ➔ , ◌ِ➔ , ◌ٍ ➔ , ◌َ ➔ , ◌ْ➔ , ◌ّ➔ , ◌ٌ➔ . So, the diacritics replaced with nothing.

Step 2: Training

1. The training set (13,569 documents) used to generate the words dictionary.
2. The words dictionary used to generate features vectors using vector space model (VSM).
3. The VSM features vectors weighted using term frequency – inverse document frequency (TF-IDF).

Step 3: Testing

1. The testing set (2,220 documents) used to generate the TF-IDF features vectors for each document based on the information collected in the training step (i.e. the vocabulary)

2. The cosine similarity measure is used as local classifier at each level to find the best-matched document. According to the obtained result, the classification is moved to the second level, and then to the third level, etc.

3. The accuracy is measured.

3.3 The Proposed Method for Markov chain approach

The proposed Markov chain method is a supervised learning method that requires a training set for learning. Mainly, three steps were performed as following:

Step 1: preprocessing

1. The ignore characters set is defined. This set includes : à ® é \ ² … è ² ™ ¯ › × — ô é ü ÷ ¹ × ‘ ° © ½ ¾ ë ‹› ' [ ] ” _ ; ¦ – } { - • = % ’ * , / : ! « ? » ¦ 0 1 2 3 4 5 6 7 8 9 - A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o b q r s t u v w z y z > < | # @ $ + & ^ ،) (¶.

2. The stoplist is defined to ignore the common words during the classification process. In this method, we defined the stoplist as the common words among the training categories (including all documents in the subcategories). However, there are some words that are important to a particular category, and at the same time appear in all categories, such as the word:”اقتصاد” ➔”economy”. So, to avoid discarding the important words, any word that has more than 50% occurrences in a particular category is considered as an important word in that category, and therefore, removed from the stoplist. An example of words removed from the stoplist is “تكنولوجيا” ➔ “Technology”, in the Technology category. Removing words from the stoplist means considering the word as discriminative words and, hence, involved in the classification process.

3. After declaring the ignore characters and the stoplist, the left words are normalized which include replacing some letters in other forms while maintaining the same meaning. In addition, all diacritics were eliminated. The replacement included: أ➔ا, إ➔ا, آ➔ا , ً➔ , ُ➔ , ِ➔ , ٍ ➔ , َ ➔ , ْ➔ , ّ➔ , ٌ➔ . So, the diacritics replaced with nothing.

4. The Document Frequency (DF) textual feature is used that restrict the words to be added to the words dictionary. In this work, we chose DF as greater than 20. Only the words that appear in more than 20 different documents are added to the dictionary. Of course, a different DF threshold can be used.

5. The prepared dictionary is altered to remove two things: "ال" at the beginning of words and "ية" at the end of words. An example of removing "ال": "الابتكار" ➔ "Innovation" is replaced to "ابتكار" , and an example of removing "ية": "الاجرامية"➔" Criminal" is replaced to "اجرام".

Step 2: Training

1. The previous step (preprocessing) ends up with the dictionary used in the training step. The Arabic alphabet characters used in this work include 34 characters that include:

{ا,آ,ب,ة,ت,ث,ج,ح,خ,د,ذ,ر,ز,س,ش,ص,ض,ط,ظ,ع,غ,ف,ق,ك,ل,م,ن,ه,و,ؤ,ى,ي,ء,ئ}

2. To create Markov chain matrices, a mapping process is used to denote a particular number for each Arabic character as follows:

{'ا':'0','آ':'1','ب':'2','ة':'3','ت':'4','ث':'5','ج':'6','ح':'7','خ':'8','د':'9','ذ':'10','ر':'11','ز':'12','س':'13',
'ش':'14','ص':'15','ض':'16','ط':'17','ظ':'18','ع':'19','غ':'20','ف':'21','ق':'22','ك':'23','ل':'24','م':'25',
'ن':'26','ه':'27','و':'28','ؤ':'29','ى':'30','ي':'31','ء':'32','ئ':'33'}

3. Each word in the dictionary is split for a sequence of characters. The sequence ends by "*" as indicator of the end of the word. The examples of words in the dictionary are shown in Figure 6.

```
...
"المعلمون": "م ع ل م و ن *",
"المعلمين": "م ع ل م ي ن *",
"المعلوماتي": "م ع ل و م ا ت ي *",
"المعلوماتية": "م ع ل و م ا ت *",
"المعماري": "م ع م ا ر ي *",
"المعمارية": "م ع م ا ر *",
"المعوقات": "ع و ق ا ت *",
...
```

Figure 6. A part of the words in the dictionary

4. For each category and subcategory, a probability transition matrix is created. The dimensions of the matrix is 34 × 34, the total number of unique characters appeared in the corpus. For each sequence occurrence, 1 is added in the corresponding cell in the transition matrix. An example how to fill the probability transition matrix is shown in Figure 7, which has a short sentence on Economy.

The sentence to be modelled:
"... مؤتمر المشروعات الصغيرة ..."
Using the dictionary, it translates to:
م ؤ ت م ر * م ش ر و ع ا ت * ص غ ي ر ة
Using the mapping table:
⇨ 25 29 4 25 11 * 25 14 11 28 19 0 4 * 15 20 31 11 27 *
The transitions:
25➔29, 29➔4, 4➔25, 25➔11, 25➔14, 14➔11, 11➔28, 28➔19, 19➔0, 0➔4, 15➔20, 20➔31, 31➔11, 11➔27
Each sequence is used as an index to add 1 in the corresponding cell in the transition matrix

Figure 7. Finding transitions sequences of the training data

5. An example of the probability transition matrix of the Economy category (first level) is shown in Figure 8. The figure shows that the highlighted sequence ت➔ص that usually

appears in the word "اقتصاد"➔"Economy" has a large weight compared to other sequences. It is also noted that the sequence آ➔س also has a large weight, this sequence mainly appears in the words "آسيا"➔ "Asia" which has large occurrences within the Economy category in the training set.

	ا	آ	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص
ا	0.00	0.00	0.02	0.00	0.10	0.00	0.03	0.02	0.01	0.09	0.00	0.14	0.01	0.12	0.01	0.02
آ	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.63	0.00	0.00
ب	0.14	0.00	0.00	0.03	0.03	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.00	0.00	0.00
ة	0.05	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00
ت	0.03	0.00	0.01	0.00	0.01	0.10	0.05	0.03	0.00	0.03	0.00	0.07	0.00	0.01	0.02	**0.23**
ث	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
ج	0.18	0.00	0.06	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.02	0.09	0.01	0.00	0.00
ح	0.16	0.00	0.00	0.10	0.15	0.00	0.00	0.00	0.00	0.07	0.00	0.03	0.01	0.04	0.00	0.04
خ	0.25	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.21	0.01	0.03	0.00	0.03

Figure 8. A part of the probability transition matrix of an economy documents

Step 3: Testing

1. The testing set is used to evaluate the proposed method. As shown in the previous step (step 2), each document is converted to a list of sequences to be utilized in the transition matrices. We chose to find a document score by summation of the weights of sequence. As previously indicated, we use summation instead of multiplication to avoid handling small numbers.

2. A testing document is initially compared with all probability transition matrices at the first level. Once finding the most related category, the process goes down to the second level, and then to the third level.

3. The testing process is started by converting the query document into a sequence of characters. Of course, only the word in the dictionary contributes in the generated sequence. Hence, the query document sequence depends on the total words in the document

that are also listed in the dictionary. Therefore, VSM is not employed in this method. An example of a generated sequence for a short sentence is shown in Figure 9.

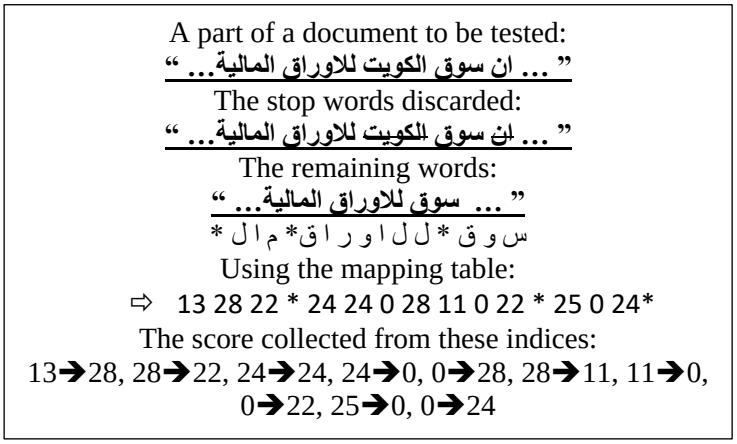

Figure 9. Finding transitions sequences of a testing document

4. The experimental results

Before conducting any experiment, two parameters were set. First, the occurrence of the words. Any word occurs less than three time discarded. That is, a word that appears two times or less eliminated. One reason of that is to eliminate spelling errors that usually appear few times. Second, the length of words to be considered in the training process. Any word less than certain threshold has to be eliminated. This help to discard the small words that have no effectiveness (i.e. noise) in the classification process. Examples of small words in Arabic include "في" and "على" that can be translated to "in" and "on". Al-Anzi and AbuZeina (2016) indicated that eliminating small words enhances the performance. The following subsection demonstrates the obtained results for the tow investigated methods.

4.1 The Results of the Top-down Cosine measure approach

The experimental results shows that the accuracy was good in the first level while it was relatively poor in the third level. This is nature, since the documents at the third level are more specific than the documents in the first level that are more general. In addition, the errors occur at the first level are spread out to the lower levels that might produces more errors in the deep levels. Figure 10 shows the performance achieved for each category, it also shows the overall level performance in terms of recall, precision, F1 value, and accuracy. In the figure, the categories names are listed in Table 1.

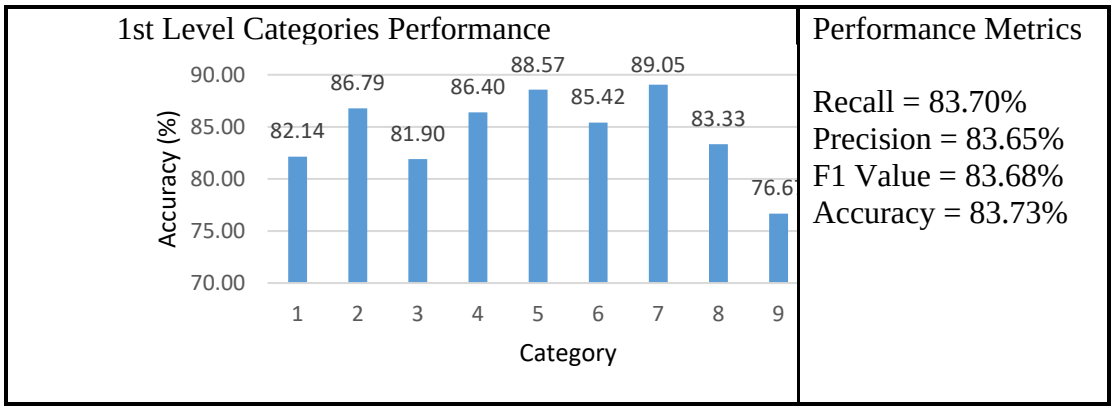

Figure 10. The performance of the first level categories

Figure 10 shows that the category that labelled by (9 ➔ Parliament) has the lowest accuracy. After investigating the confusion matrix, we found that many of this category testing instances were wrongly classified as “Economy” and “Law and Justice” categories. We also investigated the results of the Politics category that is labelled by number 10 in Figure 10. It was found that many of this category testing documents were wrongly classified as “Parliament”,

"Religion", and "Economy". These mixes of classifying reinforces the demand for further analysis to investigate the reasons of such mixing to alleviate its poor effects on the performance.

Since the testing dataset has almost equal partitions of documents for each category. The F1 value and the accuracy metrics are almost same. Hence, we only measured the accuracies for the second and third levels. The second level accuracy scored 71.26% and for the third level was 56.80%. Therefore, the overall all performance of our experiments demonstrated in figure 11.

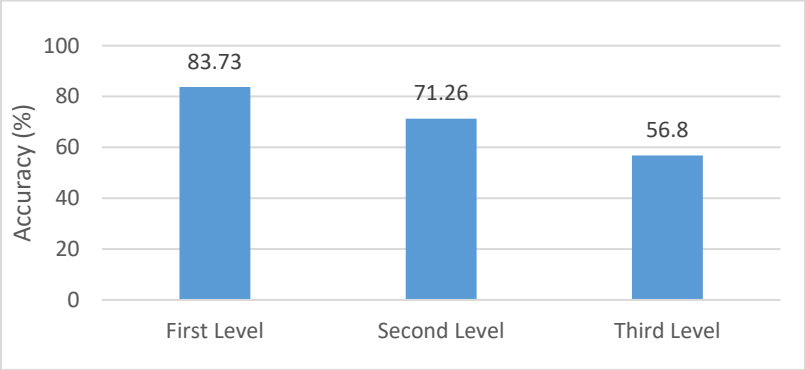

Figure 11. The overall performance of 3-level hierarchy

4.2 The Results of the Markov chain Approach

In this part, we present the result of the novel Markov chain approach for Arabic text classification. The experimental results shows that the performance is good in the first level while it is acceptable in the second level. This is natural as previously indicated. For illustration, the accuracies of the Law-Justice >> Crime and sanction >> Embezzlement, Execution, Jail subcategories are extremely low since the documents contents are very similar. The accuracies are 0.35%, 55%, 37.5% for Embezzlement, Execution, Jail, respectively. In addition, the errors that occurred at the first level are spread out to the lower levels that might produce more errors in the deeper levels. Table 2 shows the performance achieved for the first level in terms of precision,

recall, F1 value, and accuracy. Since the F1 measure and the accuracy are very close, we just measure the accuracies for the second and the third level.

Table 2. The performance of the proposed method (three levels)

1st level performance (%)	2nd level performance (%)	3rd level performance (%)
Precision=90.29, Recall=90.69, F1 =90.49, Accuracy=90.29	Accuracy=77.09	Accuracy=63.33

5. Conclusion

This work is considered as a test bed for further research in Arabic language hierarchical text classification. By preparing a sizable hierarchical corpus, it boosts the research as well as initializing for new research directions. The future research could be more specific by devoting more attention to time and space efficient text classification algorithms. We employed Markov chain for hierarchical Arabic text classification which is the first attempt in this direction. The results are very encouraging as the proposed method found to be better than the top-down cosine similarity measure. The textual features that are generated for Markov chain method are semantic loss; it is worthy considering semantic methods based on the outcomes of this study. It is beneficial to conduct a thorough study to compare the time and the space complexities of the proposed method and other text classification methods. It is also worthy to consider adding log probabilities instead of adding the probabilities when evaluating the query documents.

References

Mubarak, Hamdy, and Kareem Darwish. "Using Twitter to collect a multi-dialectal corpus of Arabic." ANLP 2014 (2014): 1.

Guerra-Gómez, J. A., Pack, M. L., Plaisant, C., & Shneiderman, B. (2015). Discovering temporal changes in hierarchical transportation data: Visual analytics & text reporting tools. Transportation Research Part C: Emerging Technologies, 51, 167-179.

Sorower, Mohammad S. "A literature survey on algorithms for multi-label learning." Oregon State University, Corvallis (2010).

Bi, Wei, and James T. Kwok. "Mandatory leaf node prediction in hierarchical multilabel classification." Advances in Neural Information Processing Systems. 2012.

Silla Jr, Carlos N., and Alex A. Freitas. "A survey of hierarchical classification across different application domains." Data Mining and Knowledge Discovery22.1-2 (2011): 31-72.

D'Alessio, S., Murray, K. A., Schiaffino, R., & Kershenbaum, A. (1998, June). Category Levels in Hierarchical Text Categorization. In EMNLP (pp. 61-70).

Joshi, Shweta, and Bhawna Nigam. "Categorizing the document using multi class classification in data mining." Computational Intelligence and Communication Networks (CICN), 2011 International Conference on. IEEE, 2011.

Godbole, Shantanu. "Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers." Annual Progress Report, Indian Institute of Technology–Bombay, India (2002).

Ying, Cao. "Novel top-down methods for Hierarchical Text Classification."Procedia Engineering 24 (2011): 329-334.

Ruiz, Miguel E., and Padmini Srinivasan. "Hierarchical text categorization using neural networks." Information Retrieval 5.1 (2002): 87-118.

Chakrabarti, S., Dom, B., Agrawal, R., & Raghavan, P. (1997, August). Using taxonomy, discriminants, and signatures for navigating in text databases. InVLDB (Vol. 97, pp. 446-455).

Dhillon, Inderjit S., Subramanyam Mallela, and Rahul Kumar. "A divisive information theoretic feature clustering algorithm for text classification." The Journal of Machine Learning Research 3 (2003): 1265-1287.

Svetlana Kiritchenko, Hierarchical text categorization and its application to bioinformatics, University of Ottawa, Ottawa, Ont., Canada, 2006

Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. Vol. 999. Cambridge: MIT press, 1999.

Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

Chuang, W. T., Tiyyagura, A., Yang, J., & Giuffrida, G. (2000). A fast algorithm for hierarchical text classification. In Data Warehousing and Knowledge Discovery (pp. 409-418). Springer Berlin Heidelberg.

De Campos, Luis M., and Alfonso E. Romero. "Bayesian network models for hierarchical text classification from a thesaurus." International Journal of Approximate Reasoning 50.7 (2009): 932-944.

Xue, G. R., Xing, D., Yang, Q., & Yu, Y. (2008, July). Deep classification in large-scale text hierarchies. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval(pp. 619-626). ACM.

Secker, Andrew, et al. "An experimental comparison of classification algorithms for the hierarchical prediction of protein function." BCS-SGAI Mag (Expert Update) 9.3 (2007): 17-22.

Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77.2 (1989): 257-286.

Von Hilgers, Philipp, and Amy N. Langville. "The five greatest applications of Markov Chains." Proceedings of the Markov Anniversary Meeting, Boston Press, Boston, MA. 2006.

Leon-Garcia, Alberto, and Alberto. Leon-Garcia. Probability, statistics, and random processes for electrical engineering. Upper Saddle River, NJ: Pearson/Prentice Hall, 2008.

Sun, Aixin, Ee-Peng Lim, and Wee-Keong Ng. "Performance measurement framework for hierarchical text classification." Journal of the American Society for Information Science and Technology 54.11 (2003): 1014-1028.

Alqabas. (2016, September). Retrieved from http://www.alqabas.com.kw/Default.aspx

Al-Anzi, Fawaz S., and Dia AbuZeina. "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing." Journal of King Saud University-Computer and Information Sciences (2016).

Appendix A

Journal Papers

- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina. "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing." Journal of King Saud University-Computer and Information Sciences (2016). http://www.sciencedirect.com/science/article/pii/S1319157816300210
- [**Accepted**] Al-Anzi, Fawaz S., Dia AbuZeina, Shatha Hasan. "Utilizing Standard Deviation in Text Classification Weighting Schemes", International Journal of Innovative Computing, Information and Control. Volume 13, Number 4, August 2017.http://www.ijicic.org/contents.htm
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina," Technology in the service of Quranic rhetoric", International Journal on Islamic Applications in Computer Science And Technology- IJASAT. http://www.sign-ific-ance.co.uk/index.php/ijasatarabic/article/viewFile/1557/1418
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, "Employing Fisher Discriminant Analysis for Arabic Text Classification", Computers & Electrical Engineering.
- [**Submitted**] Al-Anzi, Fawaz S., and Dia AbuZeina. "Better Arabic Text Clustering Through Dual Use of Latent Semantic Indexing and Cosine Similarity", the International Arab Journal of Information Technology.
- [**Submitted**] Al-Anzi, Fawaz S., and Dia AbuZeina ,"Beyond Vector Space Model for Hierarchical Arabic Text Classification: A Markov Chain Approach", Information Processing and Management.
- [**Submitted**] Al-Anzi, Fawaz S., and Dia AbuZeina "Hierarchical Arabic Text Classification", The Journal of Scientific Annals of Computer Science.

Conference Papers

- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina ,"Big Data Categorization for Arabic Text Using Latent Semantic Indexing and Clustering", International Conference on Engineering Technologies and Big Data Analytics (ETBDA'2016) Jan. 21-22, 2016 Bangkok (Thailand).
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina. "Stemming impact on Arabic text categorization performance: A survey." 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA). IEEE, 2015.
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina ,"A New Enhanced Variation of TF-IDF scheme for Arabic Text Classification", 3rd International conference on Innovative Engineering Technologies (ICIET'2016) August 5-6, 2016 Bangkok (Thailand).
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, "Reinforcing Arabic Language Text Classification and Clustering: Theory and Application" Conference on Information Technology (ACIT'2016), Sultan Moulay Slimane University, Morocco, December 6-8, 2016.

- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, "Dimensionality Reduction Techniques for Big-Data Text Classification Systems", American Scientific Publishers in Advanced Science Letters (ISSN: 1936-6612 (Print): EISSN: 1936-7317.
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, "Information Technology in the service of the Holy Quran and its Sciences", 4th International Conference on Islamic Applications in Computer Science and Technologies – IMAN 2016 20 – 22 December 2016, Khartoum, Sudan.
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, " An Improved TF-IDF Model using Standard Deviation: An Arabic Text Classification Case", SMC'2017: Data Engineering In Bioinformatics, Image and Data Analysis, 23-25 Mar 2017 Tangier (Morocco)
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, "Arabic Text Classification Using Linear Discriminant Analysis", The IEEE International Conference on Engineering & MIS 2017 Monastir, Tunisia.
- [**Accepted**] Al-Anzi, Fawaz S., and Dia AbuZeina, "A Micro-Word Based Approach for Arabic Sentiment Analysis", 14th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2017 October 30th to November 3rd, 2017