ISSN 2277-3061

The Automated VSMs to Categorize Arabic Text Data Sets

Mamoun Suleiman Al Rababaa, Essam Said Hanandeh

Al albayt University, Mafraq, Jordan
marababaa@aabu.edu.jo
Zarqa University, Zarqa, Jordan
Hanandeh@zu.edu.jo

ABSTRACT

Text Categorization is one of the most important tasks in information retrieval and data mining. This paper aims at investigating different variations of vector space models (VSMs) using KNN algorithm. we used 242 Arabic abstract documents that were used by (Hmeidi & Kanaan, 1997). The bases of our comparison are the most popular text evaluation measures; we use Recall measure, Precision measure, and F1 measure. The Experimental results against the Saudi data sets reveal that Cosine outperformed over of the Dice and Jaccard coefficients.

Keywords

Arabic data sets; Data mining; Text categorisation; Term weighting; VSM.

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 13, No. 1

editor@cirworld.com

www.cirworld.com, www.ijctonline.com

1. INTRODUCTION

Text categorisation (TC) is one of the most important tasks in information retrieval (IR) and data mining (Sebastiani 2005). This is because of the significance of natural language text, the large amount of text is stored on the internet, in addition to the available information libraries and document collection. Further, TC importance rises up since it concerns with natural language text processing and classification using different techniques and procedures , in which it makes the retrieval and other text manipulation processes easy to execute.

Many TC approaches from data mining and machine learning exist such as: decision trees (Quinlan, 1993), Support Vector Machine (SVM) (Joachims, 1998), rule induction (Moulinier et al., 1996), and Neural Network (Wiener et al., 1995). The goal of this paper is to present and compare results obtained against Arabic text collections using K-Nearest Neighbour algorithm. Particularly, three different experimental runs of the KNN algorithm on the Arabic data sets we consider are performed, using three different VSMs (Cosine, Dice, Jaccard).

Generally, TC based on text similarity goes through two steps: Similarity measurement and classification assignment. Term weighting is one of the known concepts in TC. It can be defined as a factor given to a term to reflect the importance of that term. There are a lot of term weighting methods, including, inverse document frequency (IDF), weighted inverse document frequency (WIDF) and inverse term frequency (ITF) (Tokunaga and Iwayama, 1994). In this paper, we compare different variations of VSMs with KNN (Yang, 1999) algorithm using IDF. The bases of our comparison between the different implementations of the KNN are the F1, Recall and Precision measures. In other words, we want to determine the best VSM, which if merged with KNN produces good F1, Precision and recall results. To the best of the author's knowledge, no comparisons have been performed against The Saudi Newspapers (SNP) using VSM.

The paper is organized as follows: , Section two is review of related literature .Section three is the description of TC problem. Experiment results are discussed and explained in section four , and finally conclusions and future works in Section 5.

2. Related Works

As Syiam, et. al., (2006) pointed out that there are over 320 millions Arabic native speakers in 22 countries located in Asia and Africa. Due to the enormous energy resources, the Arab world has been developing rapidly in almost every sector especially in economics. As a result, a massive number of Arabic text documents have been increasingly arising in public and private sectors, where such documents contain useful information that can be utilised in a decision making process. Therefore, there is a need to investigate new intelligent methods in order to discover useful hidden information from these Arabic text collections.

Reviewing the existing related works proved that there are several methods which have been proposed by researchers towards Arabic text classification. For classifying Arabic text sources the N-Gram Frequency Statistics technique is investigated by Khreisat, (2006). This method is based on both Dice similarity and Manhattan distance measures in classifying an Arabic corpus. For this research the Arabic corpus was obtained from various online Arabic newspapers. The data is associated with four categories. After carrying out several pre-processing on the data, and experimentation, the results indicated that the “Sport” category outperformed the other categories with respect to recall evaluation measure. The least category was “Economy” with around 40% recall. In general the N-gram Dice similarity measure figures outperformed that of Manhattan distance similarity.

In (Thabtah et el., 2008), the authors investigated different variations of Vector Space Model using KNN algorithm, They mentioned the following variations: Cosine modulus, Dice modulus and Jacaard modulus, using different term weighting method. The average F1 results obtained against six Arabic data sets indicated that Dice based TF.IDF and Jaccard based TF.IDF outperformed Cosine based TF.IDF, Cosine based WIDF, Cosine based ITF, Cosine based log(1+tf), Dice based WIDF, Dice based ITF, Dice based log(1+tf), Jaccard based WIDF, Jaccard based ITF, and Jaccard based log(1+tf).

(Guo, Y., Shao, Z. & Hua, N. (2010), Active cooperation trees were used to classify Arabic documents automatically. The documents of the corpus were gathered from Arab Web sites. The corpus consists of 6825 articles, varying in length and divided into seven categories as follows: politics, economy, sports, medicine, science and technology, law and religion. We used the book SVM and naive Bayes seeded to make comparison with the active cooperation of trees. It was founded that SVM and naive Bayes seed had the best performance of the algorithms with reference to accuracy. This is because of the method of increasing the strength of a batch that enhances the Performance of C4.5. The disadvantage, however, was the decision of trees algorithm itself. Has investigation naive Bayesian method, SVM classification algorithm based on association rule algorithm for data classification determines Arabic. Data set consists of 5121 Arabic documents of different lengths that belong to the seven categories. The experimental results showed that the classification of algorithm based on association rule outperformed naive Bayesian method and algorithm for SVM regard to f1, precession and recall and measures.

(Saleh Alsaleem, 2011). The author used NB and SVM for automatic classification of Arabic text, he collect a large dataset of 5121 documents of classes. It will be interesting to see the impact of extending the classes for more than 7 classes. The results show that the SVM algorithm outperformed NB.

3. Text Categorisation Problem

TC is the task in which texts are classified into one of predefined categories based on their contents. If the texts (data of the study) are newspaper articles, categories could be, for example, economics, politics, sports, and so on. This task has various applications such as automatic email classification and web-page categorization. Those applications are becoming increasingly important in today's information-oriented society.

TC problem can be defined according to (Sebastiani 2005) as follows: The documents divided in two datasets, for training and testing. Let training data set = {d_1,d_2,…,d_g}, where g documents are used as examples for the classifier, and must contain a good number of positive examples for all the categories involved. The testing data set {d_{g+1},d_{g+2},…,d_n} used to test the classifier effectiveness. The matrix shown in Table 1 represents data splitting into training and testing. A document d_y is considered a positive example to C_k if C_{ky} =1 and a negative example if C_{ky} =0.

Table 1: Representation of text categorization problem

Category	Training data set			Testing data set		
	d_1	…	d_j	d_{j+1}	…	d_n
C_1	C_{11}	…	C_{1j}	$C_{1(j+1)}$		C_{1n}
…	…	…	…	…	…	…
C_m	C_{m1}	…	C_{mg}	$C_{m(j+1)}$	…	C_{mn}

Generally, TC task goes through three mainly steps: Data pre-processing, text classification and evaluation. Data preprocessing phase means making the text documents suitable to train the classifier. Then, the text classifier is constructed and tuned using a text learning approach against from the training data set. Finally, the text classifier gets evaluated by some evaluation measures i.e recall, precisinon, etc . The next two sub-sections are devoted to discuss the main phases of the TC problem related to the data we utilised in this paper.

3.1 Data Pre-Processing on Arabic Data

The data used in our experiments are from The Saudi Newspapers (SNP) (Al-Harbi, 2008), the data set consists of 5121 Arabic documents of different lengths that belong to 7 categories. The categories are in the fields of (Culture "الثقافية " , Economics "الإقتصادية" , General "العامة" , Information Technology " تكنولوجيا المعلومات " , Politics " السياسية", Social " الأجتماعية ", Sport " الرياضة"), Table 2 represents the number of documents for each category.

Table 2: Number of Documents per Category

Category Name	Number of Documents
Culture	738
Economics	739
General	728
Information Technology	728
Politics	726
Social	731
Sport	731
Total	5121

Arabic text is different from English one. In other words, Arabic language is highly inflectional and derivational language which makes the monophonical analysis is a complex task. Furthermore, in Arabic script, some of the vowels are demonstrated by diacritics which usually left out in the text. Moreover, Arabic uses capitalisation for proper nouns that create ambiguity in the text (Thabtah et al., 2008; Hammo et. al. 2002). In the Arabic data set we are using, each document file was saved in a separate file within the corresponding category's directory.

Moreover, we represented the Arabic data set to a form that is suitable for the classification algorithm. In this phase, we have followed (Benkhalifa et al., 2001; Guo et al., 2004; El-Kourdi et al., 2004) data format and processed the Arabic documents according to the following steps:

1. Each article in the Arabic data set is processed to remove the digits and punctuation marks.

2. We have followed (Samir et al., 2005) in the normalization of some Arabic letters such as the normalization of (hamza (إ) or (أ)) in all its forms to (alef (ا)).

3. All the non Arabic texts were filtered.

4. Arabic function words were removed. The Arabic function words (stop words) are the words that are not useful in IR systems .

3.2 Classification Assignment

There are many approaches to assign categories to incoming text such as (SVM) (Joachims, 1998), Neural Network (Wiener et al., 1995) and k-nearest neighbor (KNN) (Yang 1999). In our paper, we implemented text-to-text comparison (TTC), which is also known as the KNN (Yang 1999). KNN is a statistical classification approach, that has been studied in pattern recognition over four decades. KNN has been successfully applied to TC problem, i.e. (Yang and Liu, 1999), (Yang, 1999), and showed promising results comparing with other statistical approaches such as Baysian based Network.

4. KNN Algorithm

Does the algorithm prove its effectiveness in the supervised classification of textual data? The learning phase consists of storing the labelled examples. The classification of new texts is made by calculating the distance between the vector representing the new document and each stored instance of the data set. The Nearest instances are selected and the document is assigned the majority class (the weight of each class may be weighted according to its distance). In order to make a comparative study and because the similarity measure plays a crucial role in the method, we used the three similarities measures.

5. Vector Space Model

The vector space model uses non-binary weights that are assigned for the documents and queries index terms (Salton, 1968). This will suggest a partial matching retrieval instead of the relevant / non-relevant matching. The non-binary weights assigned for both the queries and documents are ultimately used to measure the degree of similarity between each of the documents in store in the system and the user query. Hence, the vector model will also take into consideration documents which match the query terms partially.

The vector model uses the t-dimensional vectors to represent both document and query. For a document **dj** (where j is the document number) and a query **q**, their t-dimensional representations are **dj** and **q** as follows:

The query q representations is :

$$\vec{q} = (w_{1,q}, w_{2.q}, \ldots, w_{t,q})$$

and the document dj representation is :

where $\mathbf{w_{i,q} \geq 0}$ and t is a total number of index terms in the system.

The vector model proposes to evaluate the degree of similarity of the document **dj** with regard to the query **q** as the correlation between the vectors **dj** and **q**. This correlation can be quantified, for example, by the cosine of the angle

$$\vec{dj} = (w_{1,j}, w_{2.j}, \ldots, w_{t,j})$$

between these two vector (Salton, 1968), That is,

$$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

Vector model can use different similarity measures other than cosine similarity as the following Table 3 shows (Salton, 1988) :

Table 3 : Similarity Measures (Salton, 1988):

Similarity Measure	Evaluation for Binary Term Vector	Evaluation for Weighted Term Vector
Cosine	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert^{1/2} \bullet \lvert q \rvert^{1/2}}$	$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$
Dice	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2}$
Jaccard	$sim(d,q) = \frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert - \lvert d \cap q \rvert}$	$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j} \times w_{i,q}}$

Note |d| is the number of term in document d.

Index term weights can be calculated in many different ways, the most popular ways are (Salton & McGill, 1983) :

1- Binary term weights

2- Term Frequency-Inverse Document Frequency (***tf-idf***) term weights which is given by the next formula:

Wij=tf *idf

Let N be the total number of documents in the system and ni be the number of documents in which the index term ki appears. Let ***freq$_{ij}$*** be the raw frequency of term ki in the document dj (i.e., the number of times the term ki is mentioned in the text of the document dj). Then, the normalized frequency ***f$_{ij}$*** of term ***ki*** in document ***dj*** is given by :

$$f_{i,j} = \frac{freq_{i,j}}{\max\ freq_{l,j}}$$

Table 4.: Results F1, Recall, and Precision of Arabic text categorization

Category Name	**Cosine**			**Dice**			**Jaccard**		
Evaluation measures	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Culture	0.756	0.737	0.746	0.733	0.711	0.722	0.721	0.731	0.726
Economics	0.931	0.927	0.928	0.852	0.887	0.869	0.843	0.83	0.836
General	0.497	0.532	0.514	0.451	0.442	0.446	0.339	0.392	0.364
Information Technology	0.917	0.887	0.902	0.842	0.891	0.865	0.952	0.952	0.952
Politics	0.835	0.873	0.854	0.832	0.921	0.874	0.885	0.842	0.863
Social	0.651	0.623	0.636	0.513	0.52	0.516	0.591	0.542	0.565
Sport	0.917	0.979	0.947	0.96	0.934	0.946	0.911	0.94	0.925
Average	0.786	0.794	0.789	0.740	0.758	0.748	0.748	0.747	0.747

where the maximum is computed over all terms which are mentioned in the text

of the document dj. If the term ***ki*** does not appear in the document ***dj*** then $f_{ij} = 0$. Further, let $\boldsymbol{idf_i}$, inverse document frequency for ***ki***, be given by

$$idf_i = \log \frac{N}{n_i}$$

The best known term-weighting schemes use weights which are given by $w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$

Such term weighting strategies are called ***tf-idf*** schemes.

For the query term weights, (Salton and Buckley , 1988) suggest

$$w_{i,q} = \left(0.5 + \frac{0.5\, freq_{i,q}}{\max\, freq_{l,q}} \right) \times \log \frac{N}{n_i}$$

where $\boldsymbol{freq_{iq}}$ is the raw frequency of the term ***ki*** in the text of the information

request ***q***.

6. Experiment Results

Arabic text is different than English, since Arabic language is highly inflectional and derivational language which makes monophonically analysis a complex task. Also, in Arabic script, some of the vowels are represented by diacritics which usually left out in the text and it does use capitalisation for proper nouns that creates ambiguity in the text (Hammo et. al. 2002).

Three TC techniques based on vector model similarity (Cosine, Jaccard, and Dice) have been compared in term of F1 measure, which is shown in equation (1). These methods use the same strategy to classify incoming text i.e. KNN. We have many options to construct a text classification method; we compared techniques using IDF term weighting method. All of the experiments were implemented using Java on 3 Pentium IV machine with 1GB RAM.

The F1 measureed is computed on the following equation:

$$F1 = \frac{2 * \Pr ecision * \operatorname{Re} call}{\operatorname{Re} call + \Pr ecision} \qquad (1)$$

Precision and recall are widely used evaluation measures in IR and ML, where corresponding to Table 5,

$$\Pr ecision = \frac{TP}{(TP + FP)} \qquad (2)$$

$$\operatorname{Re} call = \frac{TP}{(TP + FN)} \qquad (3)$$

Table 5: Confusion matrix

Class	Predicted as Actual Class	Predicted as Other
Actual Correct	True Positive (TP)	False Negative (FN)
Other Classes	False Positive (FP)	True Negative (TN)

Table 4 gives the F1 results generated by the three algorithms (Cosine, Dice and Jaccard) against seven Arabic data sets; where in each data set we consider 70% of documents arbitrary for training, and 30% for testing. *K* parameter in the KNN algorithm was set to 9.

After analysing Table 4, we found that the Cosine categorize outperformed Dice and Jaccard Algorithms on all measures (F1, Precison and recall).

Particularly, Cosine outperformend Dice and Jaccard on 6,5 data sets respectively with regards to F1 results. Also Recall results obtain that the Cosine outperformed Dice and Jaccard on 5,6 data sets respectively. And Precison results obtain that the Cosine also outperformed Dice and Jaccard on 6, 6 data sets respectively.

The average of three measures obtained against seven Arabic data sets indicated that the Cosine dominant Dice and Jaccard.

5. Conclusions and Future Works

This study intended to develop an Arabic text classifier in order to classify Arabic text. We investigated different difference of Vector Space Model, using KNN algorithm. These differences are Cosine coefficients, Dice coefficients and Jacaard coefficients. We also used IDF term weighting method. The obtained average of the three measures against seven Arabic data sets indicated that the Cosine is more dominant than Dice and Jaccard.

References

[1] Al-Harbi, S. (2008) 'Automatic Arabic Text Classification', JADT'08: 9es Journées internationales d'Analyse statistique des Données Textuelles., pp. 77-83.

[2] Benkhalifa, M., A. Mouradi, and H. Bouyakhf. "Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization," Int. J. Intel Syst (16:8), 2001, pp.929-947.

[3] El-Halees A., Mining Arabic Association Rules for Text Classification In the proceedings of the first nternational conference on Mathematical Sciences. Al-Azhar University of Gaza, Palestine, 15 -17 (2006). To be appear.

[4] El-Halees A., Arabic Text Classification Using Maximum Entropy The Islamic University Journal (Series of Natural Studies and Engineering)Vol. 15, No.1, pp 157-167, 2007.

[5] Guo, G., H. Wang, D. Bell, Y. Bi, and K. Greer. "An kNN Model-based Approach and its Application in Text Categorization," In proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistic, CICLing, LNCS 2945, Springer-Verlag, 2004, pp.559-570.

[6] Hammo, B., Abu-Salem, H., Lytinen, S., and Evens, M. 2002. "QARAB: A Question Answering System to Support the Arabic Language". Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. pp. 55-65.

[7] Joachims T. (1998). Text Categorisation with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning (ECML), (pp. 173-142). Berlin, 1998, Springer.

[8] Junker, M., R. Hoch, and A .Dengel. On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy. in Proceedings of the Fifth International Conference on Document Analysis and Recognition. 1999.

[9] El-Kourdi, M., Bensaid, A., Rachidi, T., Automatic Arabic Document Categorisation Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics . August 28th. Geneva (2004).

[10] Laila Khreisat (2006). Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. DMIN 2006: 78-82

[11] Moulinier, I., Raskinis, G., Ganascia, J. (1996) Text categorization: a symbolic approach. Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval,1996.

[12] Quinlan, J. (1993) C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

[13] Samir, A., W. Ata, and N. Darwish. "A New Technique for Automatic Text Categorization for Arabic Documents," 5th IBIMA Conference (The internet & information technology in modern organizations), 2005, Cairo, Egypt.

[14] Sawaf, H. Zaplo,J. and Ney. H. (2001). "Statistical Classification Methods for Arabic News Articles". Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July.

[15] Sebastiani F. (2005) Text categorization. In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109—129

[16] Syiam, M. M., Fayed, Z. T. & Habib, M. B. (2006) 'An Intelligent System For Arabic Text Categorization', IJICIS, Vol.6, No. 1 .

[17] Thabtah F., Hadi W., Al-Shammare G. VSMs with K-Nearest Neighbour to Categorise Arabic Text Data. In The World Congress on Engineering and Computer Science 2008. (pp.778-781), 22-44 October 2008. San Francisco, USA.

[18] Tokunaga, T. , Iwayama, M. (1994). Text Categorisation Based on Weighted Inverse Document Frequency. 1994, Department of Computer Science, Tokyo Institute of Technology: Tokyo, Japan.

[19] Yang. Y. (1999). An evaluation of statistical approaches to text categorization, Journal of Information Retrieval, 1(1/2):67-88, 1999.

[20] Guo, Y., Shao, Z. & Hua, N. (2010) 'Automatic text categorization based on content analysis with cognitive situation models', Information Sciences 180, pp. 613-630

[21] Raheel, S., Dichy, J., andHassoun, M. The Automatic Categorization of Arabic Documents by Boosting Decision Trees. To appear in the proceedings of the 5th International IEEE/ACM Conference on Signal-Image Technology and Internet-Based Systems, IEEE CS Press, Marrakech, Morocco. 2009. Pages: 294-301

[22] Saleh Alsaleem, Automated Arabic Text Categorization Using SVM and NB, International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011

Associate Professor, Mamoun Suleiman Al Rababaa received the degree of B.Sc. in1994, he earned his master degree of M.Sc. in CS, A Ph.D. in Computer Engineering was received in 1999, he joined Irbid National

University in Jordan from 2000-2003 then joined to Al-al Bayt University in 2003 in Jordan. He published more than 25 research paper in international journals and conferences

Assistant Professor, Essam Said Hanandeh received the degree of B.Sc. in1990, he earned his master degree of M.Sc. in IT, A Ph.D. in CIS was received in 2008, he joined Zarqa University in Jordan in 2008. Assistant Professor Essam Said Hanandeh has been worked for 15 years as programmer & System Analyst. He published 6 research paper in international journals and conferences