OPEN ACCESS
Saudi Journal of Engineering and Technology
Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: http://scholarsmepub.com/sjet/

Original Research Article

A New Model for Arabic Text Clustering by Word Embedding and Arabic Word Net

Nehad M. Abdel Rahman Ibrahim*

College of Computer Science and Information Technology Imam Abdulrahman bin Faisal University, Dammam, KSA

DOI: 10.36348/SJEAT.2019.v04i10.001 | **Received:** 07.10.2019 | **Accepted:** 14.10.2019 | **Published:** 17.10.2019

***Corresponding author:** Nehad Mohamed Ibrahim

Abstract

A major challenge in article clustering is high dimensionality, because this will affect directly to the accuracy. However, it is becoming more important due to the huge textual information available online. In this paper, we proposed an Arabic word net dictionary to extract, select and reduce the features. Additionally, we use the embedding Word2Vector model as feature weighting technique. Finally, for the clustering uses the hierarchy clustering. Our methods are using the Arabic word net dictionary with word embedding, additionally by using the discretization. This method are effective and can enhance improve the accuracy of clustering, which shown in our experimental results.

Keywords: Machine Learning, Clustering, CBOW, SKIP-GRAM, Word Embedding, Arabic Word Net Dictionary.

INTRODUCTION

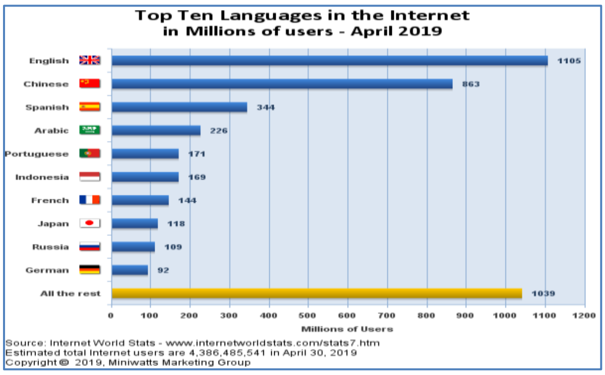

Fig-1: Top ten languages in the internet users in 30 April 2019

Arabic Language is the 5th widely used languages in the world and the 4th language using the internet 30 April 2019 as in Figure 1[1]. There are very large numbers of Arabic texts posted daily on the web. This increase of such amount of data led to various trends towards Arabic text clustering.

Word Embedding

Word2Vec

Word2vec is representation method used to produce word embedding. It represents two-dimensional vectors, with the coordinates x and y each representing one of the vector's values, it depends on the adjacent words in the sentence and how many times it is repeated. Word2vec have become widely used because they provide produce word embedding [7]. Word2vec is not a deep neural network, it convert text into a numerical form that deep learning can understand.

Continuous Bag-of-Words Model

Tomas Mikolov [6] Propose the continuous bag-of-words model is used to represent an unordered collection of words as a vector. It is the method uses for simple document classification; an example of this might be the task of classifying an email as spam. Training complexity is then as seen in formula (1), where V is size of the vocabulary and N are words using 1-of-V coding.

$$Q = N \times D + D \times \text{Log}_2(V). \qquad (1)$$

CBOW different on the standard bag-of-words model, it uses continuity in word distributed representation within the context. The model structure is shown at Figure 3.

Continuous Skip-gram Model

New model to predict the word depend on the context and enhance learning of a word depend on another word within the sentence proposed in [6]. Also, he uses each current word as an input to projection layer and a log-linear classifier to predict words within a frame of words before and after the current word. This will enhance in the word vectors, but the complexity will increase, because the distant between words are usually less than the distant between the current word and its related words.

Related work

This paper is the second part of [2], which proposed the new method for clustering the Arabic articles by using AWNET and diacritics [4, 8]. Propose to solve the problem of dimensionality by learning a distributed representation for words that allows each training sentence to define the semantically neighboring sentences [10]. The researchers try to use deep learning in the text mining and also the content of knowledge [11]. Propose a model measure semantic similarity and achieve clustering, using k-means method to achieve document clustering with semantic feature extraction and perform document vectorization to the Arabic contents web pages and uses the semantic similarities [12]. Use clustering algorithm "Frequent Item set-based Hierarchical Clustering (FICH)" to cluster Arabic. They conducted those experiments on 600 Arabic documents using N-grams based on word level, Trigrams and Quad grams. In [9] Propose extensive experimental, flexible and outdo several widespread clustering methods when tested on three public short text [3], deep neural networks and recurrent neural networks have been applied to fields including speech recognition, audio recognition, computer vision, social network clarifying, bioinformatics, machine translation and natural language processing, where they formed results comparable to and in some cases superior to human experts.

A neural probabilistic language model proposed in [4] to match the shout of dimensionality by learning a distributed representation for words, which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences.

Word2vec is a mainly is representation method used to produce word embedding from raw text. It comes in two approaches, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model (Section 3.1 and 3.2 in [6]).

Regarding to Algorithms, all these models are more similar, except CBOW predicts output words from input context words. Skip-gram does predict source context-words from the target words. This reversal influence seems like randomly, but statistically it has the effect that CBOW smooth over a lot of the distributional information. Skip-gram is handle each context-target pair as a new statement and this tends to do better when we have larger datasets.

Two novel model proposed in [6] architectures that computing continuous vector representations of words from very large data sets, they use the two models proposed in to calculate the words similarity and apply clustering method compare the results with previous results on the same datasets.

Proposed Model

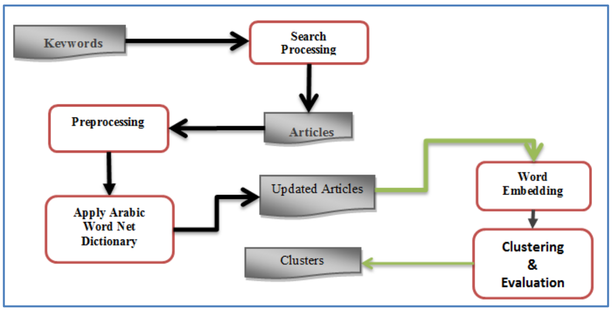

Fig-2: Proposed Model

We proposed a new clustering method by applying CBOW model proposed in [6] to calculate the words similarity. This method is enhanced accuracy by clustering based on semantic similarity. The proposed work divided to five phases: 1st phase for data collection from blogs web sites search processing, 2nd phase for Arabic text preprocessing, 3rd phase to apply Arabic word net dictionary, 4th phase apply embedding word representation CBOW using java language algorithm, and the 5th phase for clustering using hierarchy clustering and the 5th phase for evaluation the clustering by silhouette evaluation method.

DATA COLLECTION

In this phase we develop web tool to use the Google search service regarding to our keywords and define the certain blogs web sites in our search. This tool can read the keywords from our database one by one and use the Google service to search for each keyword and store the related results articles and comments in our database.

Preprocessing

In Arabic language there is a great challenge in the exploration of the texts because of the complexities of language, both in terms of infrastructure and from where conformation. Arabic language is the language of inflectional too many derivatives and her every word as well as the difference in composition alter the meaning of the word. Using special labels are called configuration rather than vowels, and they vary according to the shape of the word, the process is very important and useful word-processing so as to reduce the unusual words and increases the classification accuracy, we perform the following steps on stored articles in our database and that resulted from the data collection phase:

- Removing digits, non-Arabic letters, single letters, punctuations, diacritics and special letters ($, %, &, #, . . .).
- Separate all words by spaces.
- Remove (ها, وا, ان, ون, ين , ات) from the end of the word.
- Remove (لل, ولل, وتال, وال,كال, وكال , تال, ال) from the beginning of word except اللهم, الله, إله لله.
- Normalize some characters by a single one such as (آ, إ , أ) by (ا) , (ئ,ي) by (ى) and (ة) by (ه).
- Normalize characters appear more than one time (e.g.,يااااااااااااااااارب) by a single one (يارب).

Remove all stop words. Eliminating the stop-words from the text helps us in identifying the most important words.

Arabic Word Net

The purpose of this phase is reach the highest similarity ratio between articles' words and this by finding all relations that links those words together and uses these relations with specific similarity ratio which lead to improvement of clustering accuracy. We will enhance the clustering by find all relations between the pairs of words based on Arabic WordNet these relation will enhance the similarity between articles and this will produce clustering more accurate. Each word is represented as a synset. Each item of the synset can be any type of the part of speech (POS): verb, noun, subject, adjective and adverb. For example: the word "شحن" has the synset { "شحن" ,"نقل", …}. Each item of this set can be any POS. For example "شحن" can be noun or verb.

Word Embedding and Clustering

An article is commonly represented as a table of words in a database. The basis of the table of words corresponds to distinct words in an article collection. Each list of words represents one article. The components of the article list are the weights of the corresponding words that represent their relative importance in the article and the whole article collection.

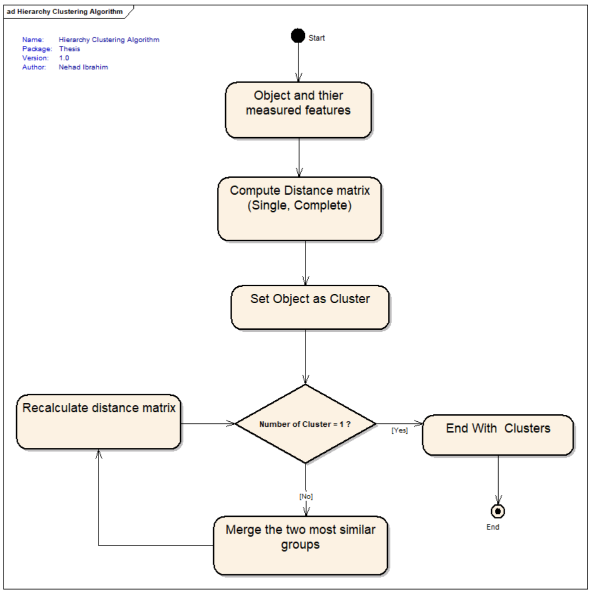
ad Hierarchy Clustering Algorithm
Name: Hierarchy Clustering Algorithm
Package: Thesis
Version: 1.0
Author: Nehad Ibrahim
Start
Object and thier measured features
Compute Distance matrix (Single, Complete)
Set Object as Cluster
Number of Cluster = 1 ?
[Yes]
End With Clusters
End
[No]
Merge the two most similar groups
Recalculate distance matrix

Fig-1: Hierarchal Clustering Algorithm

Implementation of this algorithm by C# and is based on the similarity measurements between pairs of words, creates clusters until the lowest distance of two clusters is greater than distance criterion (maximum distance). Dependent of the metric and the type of data you must have a roughly suggestion of the possible values for maximum distance to get reasonable results as shown in figure 3.

Mathematically, the linkage function – the distance D(X,Y) between clusters X and Y – is described by the expression $\mathrm{D(X,Y)} = \min_{x\in X, y\in Y}(x,y)$, Where X and Y are any two sets of elements considered as clusters, and d(x, y) denotes the distance between the two elements x and y.

Single-linkage clustering is one of hierarchical clustering methods. Which depend on set clusters in hierarchy model, at each step adding two groups, which contains related pair of elements not yet belonging to the same other cluster. In hierarchy agglomerative clustering which is also called as bottom up clustering, each data points are considered to be a separate cluster and the clusters are merged based on a criteria. The merging can be done by using complete link, single link, and centroid or wards method. We implement fusion functions (single-linkage and complete-linkage).

We take two samples of data, the first sample 144 articles containing 13,747 words as in Table (1) and the second sample 718 articles containing 33480 words as in Table (2), use these words as a features for these articles we perform the following algorithm:

- Apply two models (CBOW and Skip-Gram) by enter all attributes (words) as input to the model as in Figure (4).
- Store the vectors results into our database.

- Use models to search about similarity (cosine distance) between pairs of words and store all these distances in the database.
- Apply hierarchy clustering algorithm to cluster our articles in each dataset.
- Evaluate the results from hierarchy clustering by Silhouette evaluation method (the same evaluation method in our previous work to compare between the two methods).

Table-1: Dataset – 1

Method:	**Single Linkage**	**Max Article Length (Chars)**	**2568**
max Distance:	3	Minimum Article Length (Chars)	1000
Articles Count:	144	Words Count	13747

Table-2: Dataset – 2

Method:	**Single Linkage**	**Max Article Length (Chars)**	**231445**
max Distance:	3	Minimum Article Length (Chars)	4
Articles Count:	718	Words Count	33480

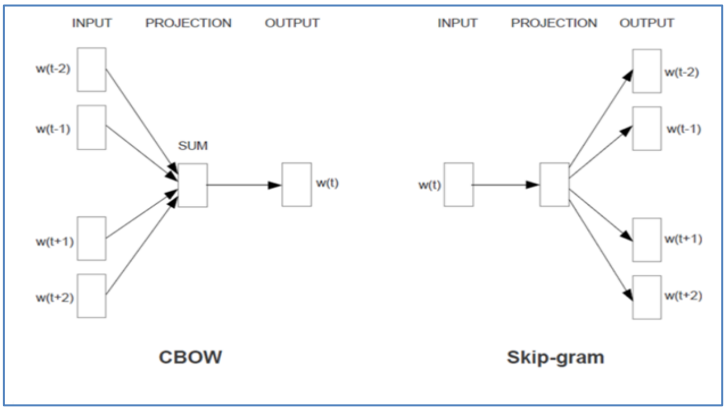

Fig-4: CBOW and Skip-gram structure

Evaluation

We need to evaluate our clustering results to validate if our method is accurate or not. There are several methods for a measure of similarity between two clusters. as the measurement can be used to differentiate between the clusters.

Evaluation Algorithms

The silhouette method is a method of assessment based on the interpretation and examination of consistency within a cluster. The value of a silhouette is calculated by measuring the similarity of the object with its mass (cohesion) and comparing it with other blocks (separation). The silhouette shows a measurement of the points surrounding a single block with the points in neighboring groups.

If the value of the Silhouette pointer is close to 1, the object is matched well with its cluster and does not match well with neighboring groups. The tangent coefficient is calculated using the average mass distance within group A and the average mass distance b (b) per sample. Silhouette Coefficient is defined as in the following formula:

$S(k) = (b(k) – a(k)) / \max \{ (a(k), b(k)) \}$ Where,

- a(k) is refer to the average of k^{th} object dissimilarity to all other objects in the same cluster.
- b(k) is refer to the average of k^{th} object dissimilarity with all objects in the closest cluster.

Silhouette Values ranges

S(k) will with in [-1, 1] and the value of S(k) are based on the following cases:

1. If silhouette value is near to 1, this refer to the sample is well-clustered and already assigned to a very appropriate cluster.
2. If silhouette value is about to 0, this refer to the sample could be assign to another cluster closest to it, and the sample lies equally far away from both the clusters.
3. If silhouette value is close to –1, this refer to sample is misclassified and is merely placed somewhere in between the clusters.

RESULTS & DISCUSSION

By apply dataset - 1 and dataset - 2 on CBOW and SKIP-GRAM models, these experiments given us the following results:

Exp-1: We apply our algorithm on dataset - 1 and use CBOW model the result shown in Table (3):

Table-3: Experiment 1 results

Article Count	**144**
Words Count	13747
Clustering Method	Hierarchy
The Silhouette Coefficient	0.636

Exp-2: We apply our algorithm on dataset - 1 and use SKIP-GRAM model the result shown in Table (4):

Table-4: Experiment 2 results

Article Count	**144**
Words Count	13747
Clustering Method	Hierarchy
The Silhouette Coefficient	0.639

Exp-3: We apply our algorithm on dataset - 2 and use CBOW model the result shown in Table (5):

Table-5: Experiment 3 results

Article Count	**718**
Words Count	33480
Clustering Method	Hierarchy
The Silhouette Coefficient	0.616

Exp-4: We apply our algorithm on dataset - 2 and use SKIP-GRAM model the result shown in Table (6):

Table-6: Experiment 4 results

Article Count	**718**
Words Count	33480
Clustering Method	Hierarchy
The Silhouette Coefficient	0.624

Validation

We can validate from our result as follows sample from our data: Select different keywords to search as (تصنيف النباتات – علم تصنيف النبات – علم النبات الزهرية).

The search result set of articles may be more than ten thousand from articles but in the real we need the most articles related as the following data which grouped in cluster number 23, there are another cluster number 67 that group sample of article given from the search result by keyword (أحمد زويل) as shown in table 7.

Table-7: Sample of clustering result

AID	Article Title	CID
14011	علم تصنيف النبات ـ ويكيبيديا، الموسوعة الحرة	23
14012	تصنيف:تصنيف النباتات ـ ويكيبيديا، الموسوعة الحرة	23
14014	علم تصنيف النبات	23
14015	ملخص لـ علم تصنيف النبات	23
14016	دروس في مقرر تصنيف نبات » جامعة أم القرى	23
14017	أهداف وأسس علم تصنيف النبات » جامعة أم القرى	23
14018	تصنيف النباتات الزهريه – موضوع	23
20946	ما أهمية علم التصنيف – موضوع	23
20949	موسوعة علم التصنيف-منتديات	23
20950	أسس تصنيف الكائنات الحية » جامعة أم القرى	23
20951	علم التصنيف (الكائنات الحية) ـ	23
20744	تاريخ علم التصنيف	23
20943	علم التصنيف ـ ويكيبيديا، الموسوعة الحرة	23
20944	ما هو تصنيف الكائنات الحية – موضوع	23
14255	تصنيف النباتات الزهرية – موضوع	23
1908	أحمد زويل ـ ويكيبيديا، الموسوعة الحرة	67
1909	احمد زويل – ويكيبيديا	67
1910	جولولي | السيرة الذاتية لـ: أحمد زويل	67
1911	أحمد زويل وجائزة نوبل – موضوع	67
1912	أهم أعمال أحمد زويل – موضوع	67
1915	اكتشاف علمي جديد للدكتور أحمد زويل -	67
1916	قصه حياة العالم المصرى د/احمد زويل	67

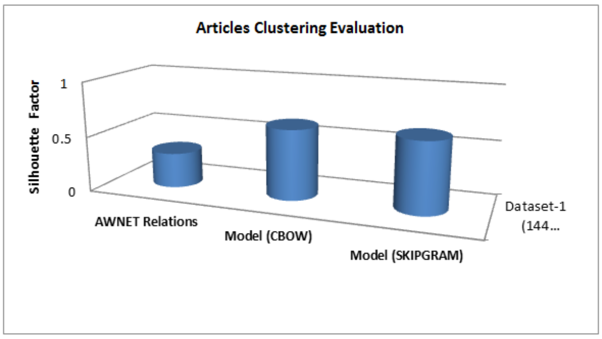

Fig-5: Comparison between the clustering by AWNET and after using Word Embedding (CBOW

Comparison to the previous work & Skip-Gram) in small dataset

In this work we apply set of experiments listed as the following

The result from Exp. - 1 (CBOW) and Exp. - 2 (SKIP-GRAM) on the Dataset-1 presented in Figure 5 which the difference between two models is minor but the difference between two models and the previous work (AWNET) is major.

The result from Exp. - 3 (CBOW) and Exp. - 4 (SKIP-GRAM) on the Dataset-2 as shown in Figure 6 the difference between SKIP-GRAM model and the previous work (AWNET) is major.

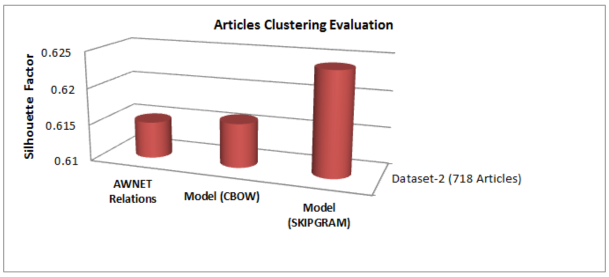

Fig-6: Comparison between the clustering by AWNET and after using Word Embedding (CBOW & Skip-Gram) in big dataset

We apply discretization to the two Arabic datasets by using RDI team in Egypt and apply the previous four experiments, we gotten the result shown in figure 7.

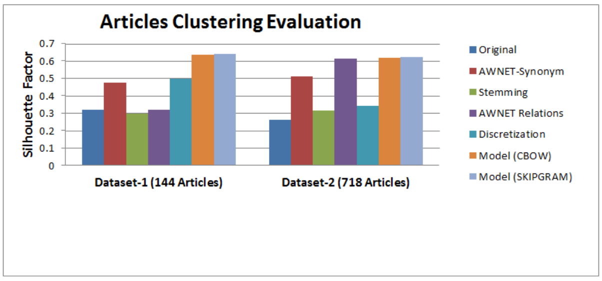

Fig-7: Summary of comparison between original and after applying (ARWNET, Stemming, Discretization, and Embedding words models)

CONCLUSION

In order to cluster the huge number of Arabic content posted daily on the internet and getting better performance and accuracy, we proposed our model that enhanced accuracy by clustering based on word embedding and Arabic word net dictionary, additionally by using the discretization. It also reduced time by using embedding representation and using appropriate threshold (maximum similarity percentage between clusters). Experimental results show that accuracy is enhanced, also show that the enhancement in similarity between articles.

REFERENCES

1. http://www.internetworldstats.com/stats7.htm
2. Ibrahim, E. N. M. A. R., Hammouda, A. A., & Nouh, S. A. H. (2017). New Approach for Text Mining of Arabic on the Web. *IJCSIS*, *15*(3).
3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
4. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.
5. Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards AI. *Large-scale kernel machines*, *34*(5), 1-41.
6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
7. Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011, May). Extensions of recurrent neural network language model. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5528-5531). IEEE.
8. Bengio, Y. (2006). Neural probabilistic language models (sections 6.1, 6.2, 6.3, 6.7, 6.8), Studies in Fuzziness and Soft Computing ,194, Springer, chapter 6.
9. Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., & Zhao, J. (2017). Self-taught convolutional neural networks for short text clustering. *Neural Networks*, *88*, 22-31.
10. Wang, H., Jiang, M., Qi, J., Zhang, X., Wang, Q., Zhou, Y., ... & Pei, Z. (2014, March). Application of Deep Learning in Text Mining. In *2014 International Conference on Mechatronics, Control and Electronic Engineering (MCE-14)*. Atlantis Press.
11. Alghamdi, H. M., Selamat, A., & Karim, N. S. A. (2014). Arabic web pages clustering and annotation using semantic class features. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 388-397.
12. Yahya, A. D. A. A. (2009, June). Clustering Arabic Documents Using Frequent Itemset-based Hierarchical Clustering with an N-Grams. In *The 4th International Conference on Information Technology*. *Al-Zaytoonah University, Jordan*.