*Journal of AI and Data Mining*
***Vol 3, No 2, 2015, 203-208.*** 10.5829/idosi.JAIDM.2015.03.02.09

# Comparing k-means clusters on parallel Persian-English corpus

A. Khazaei and M. Ghasemzadeh*

*Electrical & Computer Engineering Department, Yazd University, Yazd, Iran.*

Received 24 August 2014; Accepted 04 July 2015
**Corresponding author: m.ghasemzadeh@yazd.ac.ir (M. Ghasemzadeh).*

**Abstract**
This paper compares clusters of aligned Persian and English texts obtained from k-means method. Text clustering has many applications in various fields of natural language processing. So far, much English documents clustering research has been accomplished. Now this question arises, are the results of them extendable to other languages? Since the goal of document clustering is grouping of documents based on their content, it is expected that the answer to this question is yes. On the other hand, many differences between various languages can cause the answer to this question to be no. This research has focused on k-means that is one of the basic and popular document clustering methods. We want to know whether the clusters of aligned Persian and English texts obtained by the k-means are similar. To find an answer to this question, Mizan English-Persian Parallel Corpus was considered as benchmark. After features extraction using text mining techniques and applying the PCA dimension reduction method, the k-means clustering was performed. The morphological difference between English and Persian languages caused the larger feature vector length for Persian. So almost in all experiments, the English results were slightly richer than those in Persian. Aside from these differences, the overall behavior of Persian and English clusters was similar. These similar behaviors showed that results of k-means research on English can be expanded to Persian. Finally, there is hope that despite many differences between various languages, clustering methods may be extendable to other languages.

**Keywords:** *Clustering, Mizan English-Persian Parallel Corpus, K-means, Principal Component Analysis (PCA).*

## 1. Introduction

Document clustering is the application of cluster analysis to textual documents and is widely used in the natural language processing (NLP) fields such as information retrieval and automatic text summarization. For example, document clustering has a significant impact on improving the information retrieval precision in search engines [1]. Document clustering automatically assigns each of the documents in a smaller group called clusters. Each cluster should contain documents with similar content. Document clustering input is a document collection while its output is documents grouped based on their similarity. So far, much text clustering research has been done and many clustering methods have been proposed. Is an efficient text clustering method for one language extensible to other languages? In other words, whether the parallel documents clusters obtained by the same clustering method will be similar. Based on document clustering goal, each cluster should contain documents with similar contents. Therefore, it is expected that a document clustering method should earn similar clusters for parallel documents in different languages. On the other hands, different languages usually have many differences in vocabulary, morphology, grammar, syntactic structures, and so on. Thus, clustering quality and its steps can be influenced by documents linguistic characteristics [1].

In this research, we want to know whether the clusters of aligned Persian and English texts obtained by the k-means method are similar. Persian and English languages have many differences that can affect the quality of clusters. In section 3.3, k-means method will be introduced in more details.

English is spoken as a first language by the majority populations in several countries, including the United Kingdom, the United States, Canada, Australia, Ireland, and New Zealand. Modern English is the international language of communication, science, information technology, business, entertainment, diplomacy, etc. Persian is spoken in Iran, and with a different dialect in Afghanistan, Tajikistan, and some other regions which historically came under Persian linguistic influence [2].

The rest of this research paper is organized as follows: related works in this area are dealt with in section 2. Section 3 describes the method. In this section, data selection and feature extraction methods are discussed. Then the PCA dimension reduction and the k-means clustering methods that used in this research are introduced. The experiments and their results are discussed in section 4. Finally, section 5 discusses and concludes the paper.

## 2. Related research works

Clustering is unsupervised learning techniques for grouping samples into clusters. Samples in the same cluster should be as similar as possible and samples in different clusters should be as dissimilar as possible. There are two types of Clustering techniques: hierarchical and partition [3]. Hierarchical techniques can create clusters with better quality but these techniques are relatively slow. The most widely used partition techniques are k-means and its variants [3]. Time complexity hierarchical techniques are higher than partition techniques. For this reason, k-means is still used by researchers. For example, Krishnasamy et al. proposed a hybrid approach for data clustering based on modified cohort intelligence and k-means [4]. In another research, Hang Wu et al. used k-means algorithm in the storm platform [5].

Many studies have focused on English documents clustering. Some researchers have also focused on the Persian documents clustering. For example, Parvin, et al. proposed an innovative approach to improve the performance of Persian text classification and clustering. Their proposed method used a thesaurus as a helpful knowledge to obtain the real frequencies of words in the corpus [6]. In other research, using Brown algorithm, Ghayoomi proposed a word-clustering approach to overcome Persian parsing problems [7].

The number of research on English texts clustering is much more than Persian. Therefore, the proposed English texts clustering methods are more efficient than those are in Persian. Although Persian and English have many differences that may affect the quality of clusters, this paper is to investigate whether an efficient text clustering method for English is extensible to Persian.

## 3. Method description

In the first step of comparing Persian and English clusters, the suitable data should be aggregated. Then, the appropriate features should be extracted. Data selection and feature extraction are discussed in section 3-1. The extracted features are high-dimensional. To increase clustering speed and the quality of clusters, dimension reduction methods were used. In section 3-2, the used dimension reduction methods are explained. The researchers make use of k-means as a clustering method. This method is described in section 3-3.

### 3.1. Data and feature extraction

A parallel English-Persian corpus is required to find out whether the aligned Persian and English texts clusters are similar. A parallel corpus in the simplest case is a collection of texts. They are texts placed alongside their exact translation or translations into one or more other languages. In this study, Mizan English-Persian parallel corpus was used [8].

Mizan parallel corpus has one million aligned Persian and English sentences. Using Mizan parallel corpus, Supreme Council of Information and Communication Technology developed a basic statistical translation system called "Online Translator" in collaboration with Iran University of Science and Technology [8].

In this research 100,000 sentences were selected from Mizan corpus. After selecting suitable data, the appropriate features should be extracted. The feature vectors were created using text mining techniques.

To create feature vectors, in the first step, the researchers extracted the words from Persian and English texts, separately. Then, extracted words were stemmed. Stemming is a process of reducing words to their stems. Stemming reduces different forms of words as well as the length of the feature vectors. Due to Persian and English differences, it is necessary to use different stemming algorithms and tools. The WVT tool was used for stemming English texts [9]. The WVT is a flexible Java library for statistical language modeling. For Persian stemming, Ferdowsi University Natural Language Processing Tool Version 1.1 was used [10].

After word extraction and stemming steps, stop-words are usually removed. Stop-words are words

that almost never have any capability to distinguish documents, such as articles *a* and *the* and pronouns such as *it* and *them*. These common words can be discarded before completing the feature generation process. There are various lists for stop-words. There is no standard stop-words list for Persian or English languages. For example Ranks NL listed different stop-words lists for some languages [11].

Therefore, instead of using predefined stop-words lists, they are built automatically. The most frequent words are often stop-words [1]. The choice of the threshold value for frequent words is very important. There is no precise method to select this threshold. If many words are considered as stop-words, then there is a possibility that relatively informative words have been omitted from the feature vectors. The words that have more than 99,900 frequencies were removed in the present research. It reminds that our data are 100,000 aligned Persian and English sentences. This threshold was chosen empirically and with caution to avoid missing informative words.

On the other hand, the words that have less than 100 frequencies were also removed. The very rare words are often typos and can also be dismissed [1].

After words extraction, stemming, and removing more frequent and very rare words, TF-IDF (Term Frequency – Inverse Document Frequency) values were calculated for remaining words. TF-IDF is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection of documents. TF-IDF formula is

$$f_{ij} \log \frac{\text{number of documents}}{\text{number of documents that include word } i}$$

In this formula *fij*, is frequencies for word *i*in document*j*. In TF-IDF, the term frequency is modulated by a factor that depends on how the word is used in other documents [3]. If the word is in the document, the value of TF-IDF is not equal to zero. Otherwise, its value in the vector is zero.

Figure 1 shows feature extraction steps. The same method was used for the feature vectors construction from Persian and English texts. Length of obtained feature vector for each Persian sentence is 1415 and for each English sentence is 1095 using this feature extraction method. The length of feature vectors is the first difference of the clustering process in Persian and English texts. English is a morphologically poor language, while Persian is morphologically rich [12]. Morphological difference between English and Persian languages caused the larger feature vector length for Persian.

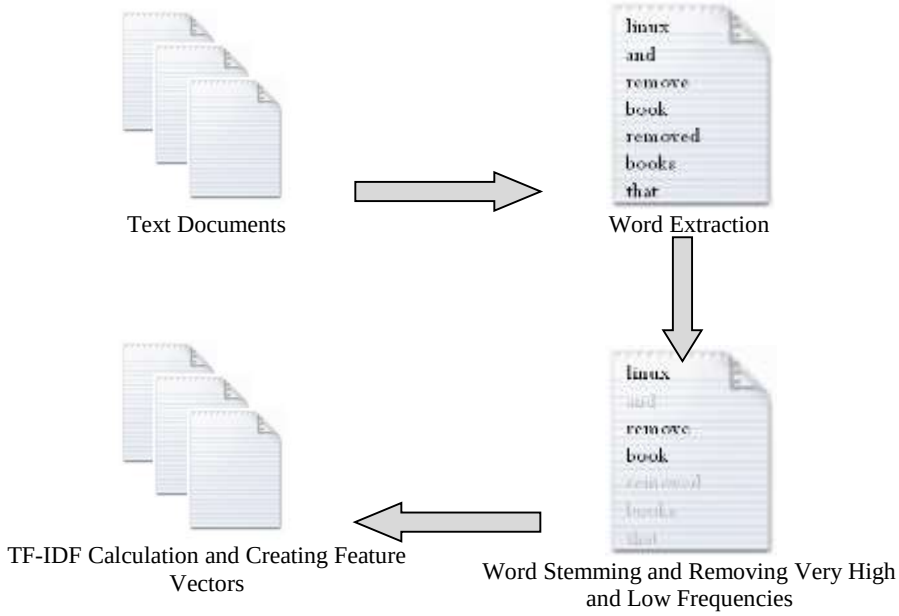


**Figure 1. Feature extraction steps.**

### 3.2. Principal component analysis

To improve feature vectors and reduce their dimensions, Principal Component Analysis (PCA) dimension reduction method was used before clustering. The PCA is a mathematical procedure to convert a set of possibly correlated features into a set of uncorrelated feature values. The number of principal components is less than or equal to the number of original features with minimal loss of information [3]. In many cases, the number of

PCA features may be more than expected number. For example, in this study, the length of feature vectors didn't change after using PCA, and there were no zero coefficients in eigenvector. In these cases, a threshold for more dimension reduction can be considered. This threshold can be the number of features or the maximum information that can be lost. In both cases, the best features are selected with minimal loss of information. Here, both methods have been used to determine threshold values and reduce dimensions of feature vectors (in section 4). Furthermore, MATLAB PCA function was used.

### 3.3. K-means clustering method

K-means method is one of the basic and popular clustering methods in data mining. This clustering method is also used in text clustering. K-means aims at partitioning *n* samples into *k* clusters. Each sample belongs to the cluster with the nearest mean. Final k-clusters should minimize the within-cluster sum of squares. Mean sum of squares is usually a metric for clusters comparison. Mean sum of squares formula is:

$$SS = \sum_{i=1}^{k} \sum_{x \epsilon c_i} \sum_{j=1}^{n} (m_j^i - x_j)^2$$

$$meanSS = \frac{SS}{N}$$

In these formulas, *x* is one sample in $C_i$cluster and $x_j$ is *j*-th feature for *x* sample. The $m_j^i$ is *j*-th feature for $C_i$ cluster center, *k* is the number of clusters, and *n* is the sample numbers.

Here, k-means method has been done several times for each experiment and those with minimum mean sum of squares was selected as the best [13].

The k-means clustering method has two challenges: Computational complexity problem and the appropriate number of clusters (that is *k*). For the computational complexity problem, there are efficient heuristic algorithms that are coverage quickly to local minimum and this problem is almost solved. The user has to provide the *k* value and he does not usually have any clue about it. Until now, many methods have been proposed to find the appropriate number of clusters. Some of them are simple and others are complicated and time consuming [13].

In this research, the optimal value for the number of clusters was not found. The experiments have been done for a few *k* values because in the current research:

1- The dimensions of feature vectors and the number of samples are high and k-means running with large *k* values would be very slow.
2- The number of categories in text categorization is not usually large. Thus, a few *k* values are enough for comparing the Persian and English clusters.

## 4. Evaluation and results

In section 3-1, feature vectors construction was described. The large numbers of samples and dimensions have a negative impact on k-means speed, and the dimension reduction methods can have a significant impact on running speed improvement. Thus, two types of experiments were designed for evaluation and comparison of Persian and English clusters.

In the first type, the same number of features for Persian and English were selected using PCA method. In these experiments, vector dimensions of both languages are equal. Thus, their results are not affected by differences in the length of vectors, but the amount of information loss for these vectors is different.

Table 1 shows these experiments results for several Ks. As mentioned in section 3-3, the mean-SS is our evaluation metric for clusters comparison. As expected, increasing the *k* values decreased the Mean-SS of clusters. Moreover, for each *k* value, increasing the length of the vectors increased the Mean-SS of clusters. Considering table 1, the difference between peer to peer Persian and English Mean-SS values is not significant in most cases. In most of table 1 experiments, English is a bit richer than Persian. Whenever the difference between Persian and English feature vectors information was less than 7%, English clusters were richer than Persian. However, for 800 features (with 7.17% difference in information loss) and 1000 features (with 8.17% difference in information loss), Persian results are a bit richer than English.

In the second type of experiments, the same amount of information loss for Persian and English vectors was considered. These results are not affected by differences in the amount of information loss, but the length of feature vectors for Persian and English are different. Table 2 shows these experiments results.

As table 2 indicates, the difference between peer to peer Persian and English Mean-SS values are more than table 1 results. In all of table 2 experiments, English is richer than Persian. These results were affected by differences of Persian and English vector dimensions.

**Table 1. Comparing Persian and English clusters with equal vectors dimensions for several Ks.**

| Number of features | Persian | | | | | English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sum of percentage of features variance | K=10 | K=20 | K=30 | K=40 | Sum of percentage of features variance | K=10 | K=20 | K=30 | K=40 |
| **10** | 5.2823 | 0.0241 | 0.0168 | 0.0145 | 0.0136 | 6.2098 | 0.0206 | 0.0083 | 0.0060 | 0.0052 |
| **50** | 17.5365 | 0.1455 | 0.1252 | 0.1108 | 0.1011 | 18.4894 | 0.1390 | 0.1181 | 0.0925 | 0.0719 |
| **100** | 27.2206 | 0.2385 | 0.2198 | 0.2038 | 0.1918 | 28.8974 | 0.2378 | 0.2173 | 0.1907 | 0.1711 |
| **150** | 34.7899 | 0.3162 | 0.2965 | 0.2802 | 0.2724 | 37.0705 | 0.3117 | 0.2876 | 0.2739 | 0.2532 |
| **200** | 41.1538 | 0.3799 | 0.3613 | 0.3456 | 0.3291 | 43.8643 | 0.3788 | 0.3530 | 0.3338 | 0.3184 |
| **500** | 65.7843 | 0.6195 | 0.6108 | 0.5996 | 0.5884 | 71.0035 | 0.6112 | 0.6038 | 0.5971 | 0.5865 |
| **800** | 81.0270 | 0.7677 | 0.7578 | 0.7327 | 0.7265 | 88.1936 | 0.7858 | 0.7685 | 0.7653 | 0.7498 |
| **1000** | 88.6143 | 0.8418 | 0.8254 | 0.8126 | 0.7970 | 96.7822 | 0.8689 | 0.8477 | 0.8415 | 0.8107 |

**Table 2. Comparing Persian and English clusters with equal amount of information loss for several Ks.**

| Sum of percentage of features variance | Persian | | | | | English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Number of features | K=10 | K=20 | K=30 | K=40 | Number of features | K=10 | K=20 | K=30 | K=40 |
| **70%** | 572 | 0.6652 | 0.6489 | 0.6373 | 0.6238 | 486 | 0.6275 | 0.6106 | 0.5884 | 0.5783 |
| **80%** | 776 | 0.7571 | 0.7417 | 0.7286 | 0.7167 | 644 | 0.7189 | 0.6989 | 0.6843 | 0.6681 |
| **90%** | 1042 | 0.8510 | 0.8376 | 0.8248 | 0.8101 | 839 | 0.8092 | 0.7894 | 0.7776 | 0.7700 |
| **100%** | 1415 | 0.9511 | 0.9306 | 0.9190 | 0.9006 | 1095 | 0.9066 | 0.8903 | 0.8614 | 0.8503 |

## 5. Discussion and conclusions

Document clustering has many applications and it has been a matter of interest for many years. The goal of document clustering is grouping documents based on their content similarity. If similar documents are grouped in the same cluster, the language of documents should have little impact on the quality of clusters. In other words, an efficient document clustering method, regardless of its documents language, should be extensible to other languages. On the other hand, different languages usually have many differences and they may affect the documents clustering.

This study's purpose was to compare clustering of aligned Persian and English texts using k-means method. Persian and English languages have many differences. The k-means is one of the basic clustering methods and it is of interest documents clustering field researchers. In this paper, the feature extraction method for both languages was the same. The morphological difference between English and Persian languages caused the larger feature vector length for Persian. After feature extraction and using the PCA for dimensions reduction, the clustering was done with k-means method.

The results demonstrated that English clusters are a bit richer than Persian. Despite the slight superiority of English clusters, similar behaviors were observed for two languages in various experiments. These similar behaviors showed that the results of k-means research on English language can be expanded to Persian. Thus, there is a hope that despite the many differences between various languages, clustering methods may be extendable to other languages. Future research could examine whether the other clustering algorithms are extendable.

## References

[1] Sholom, M. W., Nitin, I. & Tong, Z. (2010). Fundamentals of Predictive Text Mining. London: Springer Publishing Company.

[2] Windfuhr, G. (2009). The Iranian Languages. London, UK: Routledge Curzon.

[3] Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers.

[4] Krishnasamy, G., Kulkarni, A. J. & Paramesran, R. (2014). A hybrid approach for data clustering based on modified cohort intelligence and K-means. Expert Systems with Applications, vol. 41, no. 13, pp. 6009–6016.

[5] Wu, Sh., Wang, Zh., He, M. & Dong, H. (2014). Large-Scale Text Clustering Based on Improved K-Means Algorithm in the Storm Platform. Applied Mechanics and Materials, vol. 543-547, pp. 1913-1916.

[6] Parvin, H., Dahbashi, A., Parvin, S. & Minaei-Bidgoli, B. (2012). Improving Persian Text Classification and Clustering Using Persian Thesaurus. Advances in Intelligent and Soft Computing, vol. 151, pp 493-500.

[7] Ghayoomi, M. (2012). Word clustering for Persian statistical parsing. Advances in Natural Language Processing, vol. 7614, pp. 126-137, Springer.

[8] Supreme Council of Information and Communication Technology, Mizan English-Persian Parallel Corpus, (2013). Available: http://dadegan.ir/catalog/mizan [2014-01-01].

[9] The Word Vector Tool, Available: http://wvtool.sf.net [2014-01-01].

[10] Natural Language Processing Tool ver 1.1, Ferdowsi University, Available: https://wtlab.um.ac.ir/index.php?option=com_content&view=article&id=320&Itemid=200&lang=fa [2015-08-08].

[11] Ranks NL website, Available: http://www.ranks.nl/stopwords [2015-15-01].

[12] Mahmoudi, A., Faili, H. & Arabsorkhi, M. (2013). Modeling Persian Verb Morphology to Improve English-Persian Machine Translation. Advances in Artificial Intelligence and Its Applications, vol. 8265, pp. 406-418, Springer.

[13] Duda, R. O., Hart, P. E. & Stork, D. G. (2000). Pattern Classification, Second Edition. Wiley-Interscience Publication.

نشریه هوش مصنوعی و داده کاوی

دانشگاه شاهرود

# مقایسه‌ی خوشه‌بندی کا-میانگین بر پیکره‌ی موازی فارسی-انگلیسی

**عاطفه خزاعی و محمد قاسم‌زاده***

**دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران.**

ارسال ۲۰۱۴/۰۸/۲۴؛ پذیرش۲۰۱۵/۰۷/۰۴

**چکیده:**

این مقاله خوشه‌های متن‌های هم‌طراز فارسی و انگلیسی حاصل از روش کا-میانگین را با هم مقایسه می‌کند. خوشه‌بندی متن کاربردهای بسیاری در حوزه‌های مختلف پردازش زبان طبیعی دارد. تاکنون پژوهش‌های خوشه‌بندی بسیاری برای اسناد انگلیسی انجام شده است. اکنون این سؤال مطرح می‌شود، آیا نتایج حاصل از این پژوهش‌ها قابل بسط به سایر زبان‌ها می‌باشد؟ ازآنجاکه هدف خوشه‌بندی اسناد گروه‌بندی آن‌ها بر مبنای محتوایشان می‌باشد، انتظار می‌رود که پاسخ این سؤال مثبت باشد. از سوی دیگر، تفاوت‌های بسیاری بین زبان‌های مختلف وجود دارد که می‌تواند منجر به پاسخ منفی به این سؤال شود. این پژوهش بر روش کا-میانگین که یکی از روش‌های پایه و محبوب در خوشه‌بندی اسناد می‌باشد، متمرکز است. می‌خواهیم بدانیم آیا خوشه‌های متن‌های هم‌طراز فارسی و انگلیسی حاصل از روش کا-میانگین مشابه یکدیگرند؟ برای یافتن پاسخ این سؤال پیکره‌ی موازی فارسی-انگلیسی میزان به عنوان محک در نظر گرفته شد. پس از استخراج ویژگی‌ها با روش‌های متن‌کاوی و اعمال روش کاهش بُعد PCA، خوشه‌بندی کا-میانگین انجام شد. تفاوت‌های مورفولوژیکی بین زبان‌های فارسی و انگلیسی، منجر به طول بردار ویژگی بزرگ‌تر برای فارسی شد. بنابراین تقریباً در همه‌ی آزمایش‌های انجام شده نتایج زبان انگلیسی کمی بهتر از فارسی بود. گذشته از این تفاوت‌ها رفتار کلی خوشه‌های فارسی و انگلیسی مشابه بود. این رفتار مشابه نشان می‌دهد که نتایج پژوهش‌های کا-میانگین در زبان انگلیسی می‌تواند قابل بسط به زبان فارسی باشد. در پایان این امید وجود دارد که با وجود تفاوت‌های بسیار میان زبان‌های مختلف ممکن است روش‌های خوشه‌بندی قابل بسط به سایر زبان‌ها باشند.

**کلمات کلیدی:** خوشه‌بندی، پیکره‌ی موازی فارسی-انگلیسی میزان، کا-میانگین، تحلیل مؤلفه‌های اصلی (PCA).