Jurnal Teknologi

Full paper

The Hybrid Feature Selection k-means Method for Arabic Webpage Classification

Hanan Alghamdi*, Ali Selamat

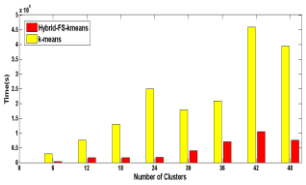
Faculty of Computing, Universiti Teknologi Malaysia (UTM), 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: hanani.alghamdi@gmail.com

Article history

Received :1 January 2014
Received in revised form :
1 June 2014
Accepted :10 September 2014

Graphical abstract



Abstract

The high-dimensional data features found in the enormous amount of Arabic text available on the Internet is an important research problem in Web information retrieval. It reduces the accuracy of the clustering algorithms and maximizes the processing time. Selecting the relevant features is the best solution. Therefore, in this paper, we propose a feature selection model that incorporates three different feature selection methods (CHI-squared, mutual information, and term frequency-inverse document frequency) to build a hybrid feature selection model (Hybrid-FS) for *k*-means clustering. This model represents text data in a high structure (consisting of three types of objects, namely, the terms, documents and categories). We evaluate the model on a set of common Arabic online newspapers. We assess the effect of using the Hybrid-FS with standard *k*-means clustering. The experimental results show that the proposed method increases purity by 28% and lowers the runtime by 80% compared to the standard *k*-means algorithm. We conclude that the proposed hybrid feature selection model enhances the accuracy of the *k*-means algorithm and successfully produces coherent-compact clusters that are well-separated when applied to high-dimensional datasets.

Keywords: Feature selection; Arabic; webpage classification; *k*-means

Abstrak

Ciri data berdimensi tinggi yang terdapat dalam jumlah besar teks Bahasa Arab di Internet merupakan satu masalah yang penting dalam penyelidikan web capaian maklumat. Ciri ini mengurangkan ketepatan pengelompokan algoritma dan memaksimumkan masa pemprosesan. Memilih ciri-ciri yang berkaitan adalah penyelesaian yang terbaik. Oleh itu, dalam artikel ini, kami mencadangkan satu model pemilihan ciri yang menggabungkan tiga kaedah pemilihan ciri yang berbeza (CHI-kuasa dua, maklumat bersama, dan frekuensi songsang jangka kekerapan dokumen) untuk membina model pemilihan ciri berhibrid (Hybrid-FS) untuk kelompok *k-means*. Model ini mewakili data teks dalam struktur yang tinggi (yang terdiri daripada tiga jenis objek, iaitu, syarat-syarat , dokumen dan kategori). Kami menggunakan set surat khabar atas talian berbahasa Arab untuk menilai prestasi model. Kami menilai kesan penggunaan Hibrid-FS dengan piawaian kelompok *k*-means. Keputusan eksperimen menunjukkan bahawa kaedah yang dicadangkan meningkatkan kejituan data sebanyak 28% dan mengurangkan jangkamasa sebanyak 80% berbanding dengan algoritma *k-means* yang piawai. Kami membuat kesimpulan bahawa model hibrid pemilihan ciri yang dicadangkan meningkatkan ketepatan algoritma *k-means* dan berjaya menghasilkan kelompok yang jelas, padat dan dipisahkan apabila digunakan untuk dataset dimensi tinggi.

Kata kunci: Pemilihan ciri; Arab; pengelasan laman web; *k-means*

1.0 INTRODUCTION

The amount of Arabic text available on the World Wide Web (WWW) is large and contains unorganized information. A text clustering technique aims to manage this enormous amount of information by classifying relevant information and therefore improving the structure of the text available on the Internet. The clustering technique applied as unsupervised classification of unlabeled terms. Any classification framework essentially contains a feature extraction procedure that aims to eliminate irrelevant/ redundant features, and keep the features that contain reliable and informative information within a corpus.[1]

A document representation model takes the free running text and produces structured input for a clustering algorithm. The vector space model (VSM) is the main approach to achieve that goal. The VSM employs the "bag of words" (BOW) to express the text; moreover, it represents a document's text as a vector in a feature space, where the frequency value of the vector is considered. The features involved in the VSM are a word (which is commonly used), character or phrase.[2]

A single document is represented by a multidimensional vector in a feature space, where each dimension conforms to the weighted value of a distinct word within the corpus. A collection of documents produces high dimensional data. The large size of feature vectors is a significant problem for text clustering methods. The efficiency of clustering algorithms may be reduced and the processing time might be increased due to the "curse of dimensionality".[3] Consequently, a way to reduce high-dimensional data is needed. Using the feature selection method on the original feature space has many benefits, as it will cut the operation time of the classifier (by removing an irrelevant feature, which will reduce the size of the dataset). In addition, it improves the accuracy of the classifier since removing meaningless features and keeping the significant features helps in classifying text documents. These benefits all result in minimizing the memory size required to process the corpus.

Several attempts have been made to suggest solutions for the high-dimensional data problem. Some studies[4,5] applied dimensionality reduction methods such as principal component analysis (PCA). Other researches[6,7] utilized a hybrid feature selection model to solve this problem. Another study suggested a model to reduce the dimensions through the use of probability distributions.[8]

This paper seeks to represent text data in a high structure model by considering three objects, namely, the term, document and category.[9] The *k*-means document clustering is integrated with a hybrid feature selection scheme in order to select the feature subset and combine all the subsets to get a new set of features.[6] This integration is investigated as a way to improving *k*-means algorithm when utilized in high-dimensional datasets.

This paper is organized as follow: first, the literature reporting studies investigating the problem of high-dimensional data are briefly reviewed; the next section introduces the model proposed in this study; this is followed by the presentation of the experimental work and results; finally, the paper is concluded including some suggestions for future research.

2.0 RELATED WORK

The accuracy of the clustering algorithms used in Arabic webpage classification is significantly affected by the document representation model. Improving a document representation model using a feature selection method may overcome the problems of high-dimensional data. A number of research studies have attempted to analyze these complexities and have recommended some solutions. The suggested algorithms have aimed to apply an improved VSM which is intended to reduce high-dimensional data[4,6,7], or to represent a low-dimensional VSM[8] and exploit the semantic relations between terms.[5]

Napoleon and Pavalakodi (2011)[4] intended to improve efficiency and accuracy in the *k*-means algorithm with high-dimensional datasets by using PCA but they were not able to completely capture the semantic similarity between the terms.[4] In contrast, Farahat and Kamel (2011)[5] made use of the semantic relations between terms by using a generalized VSM hybrid vector representation.[5] In their model, they mapped the statistical correlations between terms into the latent space with latent semantic indexing and PCA. Their model improves the effectiveness of the clustering algorithms, yet it requires a distributed implementation when using a large-scale dataset as a consequence of the difficult calculation required by the semantic kernels compared to VSM. A hybrid feature selection model based on the document frequency (DF), mutual information (MI), information gain (IG) and CHI-squared ($\chi 2$ statistics) methods was offered by Li and Zhang (2012)[6] which sought to merge the advantages of various feature selection models in one model as a way to enhance the Naive Bayes model. However, a limitation in this work arises when it comes to considering the semantic similarity between the terms and demonstrating how these similarities can be combined using feature selection.

In his proposed approach, Gunal (2012)[10] investigated which features or feature combinations were better identifiers for text classification[7] and explored the effect of other factors such as the feature size, the applied classification method, and the success assessment. His study concluded that it is more effective to select features using a variety of methods rather than using a single method. A novel feature selection method, namely, the distinguishing feature selector (DFS), was introduced for text classification.[10] DFS is a probabilistic approach that considers the contributions of the document's terms to the class discrimination and allocates relevance scores to them by bearing in mind some requirements for the term characteristics.

Reducing high-dimensionality by using the probability distribution method was suggested by Li and Zhang (2012).[8] The probability distributions of the categories that the document could belong to are involved as the vectors to represent the document and these distributions are then fed to the classifier. The limitation of this method arises when handling categories with a lot of common keywords, as it suffers from insufficient capability to differentiate between the categories and demonstrate document information.[11] Moreover, it ignores the dissimilarities in the probability distributions of the terms in a document. [12]

As a way to represent low-dimensional VSM, a study by Gharib *et al*. (2012) [3] introduced a WordNet lexical category with the SOM neural network. It uses a semantic text document clustering approach to enhance the performance of document clustering.

However, far too little attention has been paid to improving document representation models for Arabic web page clustering. There is a need for a model that integrates semantic relations, reduces high-dimensionality and lowers the runtime consumption data.

3.0 PROPOSED MODEL

The proposed model is a hybrid model of feature selection with *k*-means (called the Hybrid-FS-kmeans). An example is illustrated in Figure 1 to explain the proposed model. It illustrates the data representation in a three-dimensional structure. In this case there are four categories, namely, Art, Science, Sport and Fashion which differentiated in Figure 1 by four colors, purple, blue, green, and red respectively. The initial dataset contains two documents in each category, and the features contained in the categories are 12 Art, 20 Science, 17 Sport and 16 Fashion (Figure 1(a)). These features can be reduced using the Hybrid-FS model (Figure 1(b)) which is a mixture of multiple feature selection methods instead of using a single feature selection method. The input representation to the clustering process takes into account three kinds of objects–the term, document, and category–as a way to provide better performance.

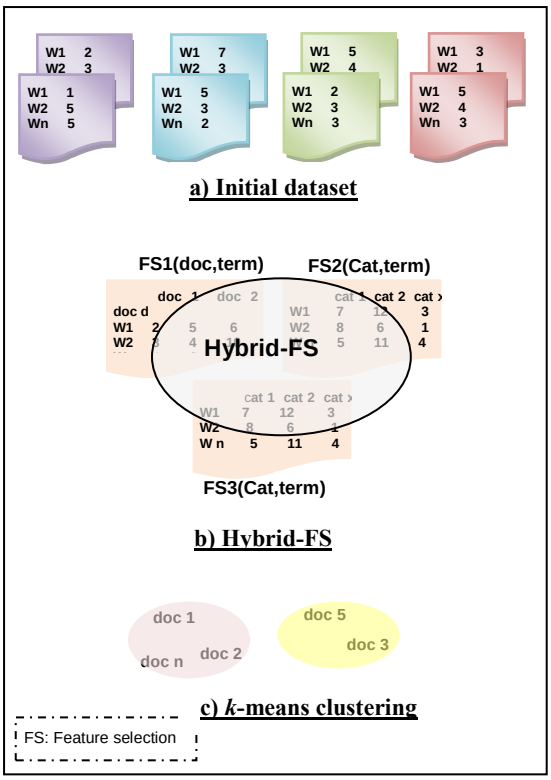


Figure 1 The proposed model

3.1 Hybrid Model for Feature Selection

In the hybrid feature selection model we intend to integrate the advantages of each feature selection in one model. Different models for hybrid feature selection are found in the literature.[7,6] In this paper, we apply a feature selection model that incorporates three different feature selection methods (CHI, MI, TF-IDF) to build a hybrid model.[6] The various selected feature sets from each method are combined to form one set of selected features which is used as the feature space for the text classification process. In the following sections we describe each feature selection method and the hybrid model for these feature selection methods.

1) CHI Squared (χ^2 Statistic)

The CHI feature selection method evaluates the lack of independence involving the text feature and text category. It has been proved to be an effective method for classifying Arabic text when used with the support vector machine.[13]

The χ^2 statistics can be calculated for each term in all categories within the corpus of documents as in the following equation:

$$\chi^2(w_t, C_i) = \frac{N(AD-EB)^2}{(A+E)(B+D)(A+B)(E+D)} \quad (1)$$

where

- w_t is the term extracted from Doc_d ,N is the total number of words in Doc_d, C_i means category i, A means the number of documents containing w_t where $w_t \in Doc_d$, $Doc_d \in C_i$,B means the number of documents containing w_t where $w_t \in Doc_d$, $Doc_d \notin C_i$, E means the number of documents containing w_t where $w_t \notin Doc_d$, $Doc_d \in C_i$, and D means the number of documents containing w_t where $w_t \notin Doc_d$, $Doc_d \notin C_i$.

When the $\chi^2(w_t, C_i)$ results in a value of zero, as term w_t and the category C_i are independent of each other , then we calculate the average score and we keep all terms w_t for which $\chi^2_{avg}(w_t) \geq w_t$ is the threshold and discard all others.

2) Mutual Information

In natural language processing, MI is used to evaluate the compactness of term w_t with the category C_i. A high MI value means that there is a higher correlation between the word and category. A measure of mutual information of the term w_t and the category C_i is defined according to Machova *et al*. (2007)[14] as in following equations:

$$MI(w_t, C_i) = \log \frac{P(w_t \wedge C_i)}{P(w_t) \times P(C_i)} \quad (2)$$

and is estimated using

$$MI(w_t, C_i) \approx \log \frac{A \times N}{(A+E) \times (A+B)} \quad (3)$$

where:

- N is the total numbers of documents, A means the number of documents containing w_t where $w_t \in C_i$,B means the number of documents containing w_t where $w_t \notin C_i$, and E means the number of documents belonging to C_i where $w_t \notin Doc_d$

When the term w_t and the category C_i are independent of each other's, the $MI(w_t, C_i)$ results in value of zero, as. Then we need to calculate the maximum score and we keep all the terms related with these scores.

3) Term Frequency–Inverse Document Frequency

The TF-IDF weighting scheme is used as a feature selection in the literature. [15,16]. This measure combines both the TF (representing the occurrence of every term) and the IDF (representing the general weight of the term over a corpus of documents). IDF gives a lower weight to frequent terms within a document collection, and specifies a higher weight to those terms that occur rarely. In contrast, the TF-IDF score determines the relevant density of a given word in a single document. Thus, terms with a higher TF-IDF value are common in a single document or in a small group of documents and as a result it would be more useful for finding similar documents and enhancing the classification algorithm as follows [17]:

$$TFIDF(w_t, Doc_d) = TF(w_t, Doc_d) \times IDF(w_t) \quad (4)$$

where *TF (t,d)* is the term frequency of w_t in doc_d and *IDF* (w_t) is the inverse document frequency of w_t. *IDF* (w_t) is shown in the following equation where it divides the total number of document, *doc* in group by the *DF*(w_t).

$$IDF(t) = \log \frac{|Doc|}{DF(w_t)} \quad (5)$$

3.2 Hybrid-FS Model

The various feature subsets selected by each method are combined together to form one set of selected features which will be used as

the feature space for the text classification process. Consider the three feature selection methods as FS_1, FS_2, FS_3 and the threshold set as δ_1, δ_2, δ_3, and the resulting feature subset as RF_1, RF_2, RF_3. The steps in hybrid model are as follows (Figure 2):

Step 1: Implement feature selection method FS_x and retain all features which greater than threshold δ_y.

Step 2: Obtain RF_1, RF_2, RF_3 by repeating Step1 for each feature selection method.

Step 3: Combine all features RF_1, RF_2, RF_3 and get a final feature set F as: $F = (RF_1 \cup RF_2) \cap RF_3$. The three feature selection methods are combined by union and intersection. We then represent each document by considering only the retained features.

As shown in Figure 2, the Hybrid-FS is obtained through a union of the sub-features in the feature space (terms and categories) and intersecting the resulting union sub-features in the feature space (terms and documents). The union grouping method output is all the terms that have been selected by each of the two feature selection methods, FS_1 and FS_2. However, the intersection method result is based on the frequent terms found in both feature sets.

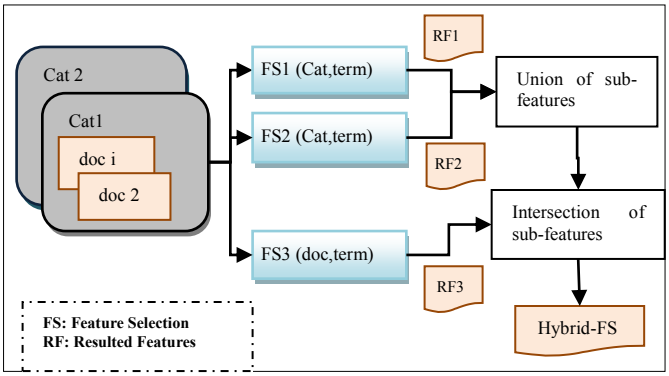


Figure 2 Hybrid-FS model

3.3 Clustering

The *k*-means algorithm consider a simplest and most regularly used for clustering.[18] It aims to grouping a nearest neighbor vectors together in order to compact the document vectors onto a smaller set .The *k*-means splits a documents into the selected number of clusters based on keywords which will nevertheless reflect the similarity of the documents at a semantic level.[19] The steps done for *k*-means algorithm are as in Figure 3 below.

Input: A list of Feature sets $F = \{f_1, f_2, \ldots, f_c\}$, a cluster similarity measure Δ, number of clusters K
Output: Set of clusters $C = \{C_1, C_2, \ldots, C_K\}$
BEGIN
initialize $C_1, C_2 \ldots C_k$ with f_c;// $C_1, C_2 \ldots C_k$ are cluster centres
do
calculate $min\ \Delta(f_{c1}, f_{c2})$; // the similarities using cosine distance
$C_y \leftarrow f_c \cup C_y$;//assign to the nearest cluster centre
re-compute $C_1, C_2, \ldots, C_K$ using Simple Mean function;
until no change in $C_1, C_2, \ldots, C_K$;// convergence is achieved
return $C_1, C_2, \ldots, C_K$;
END

Figure 3 Standard k-means clustering algorithm

The similarity measure used is Cosine Similarity, which is defined as in (6), where $f_{i,1}$ means the total weight of term w_j in feature set f_1.

$$cosine(f_1, f_2) = \frac{\sum_{i=1}^{t} w_{t,1} w_{t,2}}{} \quad (6)$$

4.0 EXPERIMENTAL SETUP

The conducted experiments aim to determine the impact of the proposed hybrid model (Hybrid-FS-kmeans) on improving clustering algorithm performance. The cluster evaluation measures consider the degree to which a cluster encloses documents from a particular category and the well-separateness between the clusters (between cluster-centroid distances). The *k* (number of clusters) for the *k*-means algorithm is set as the equal number of pre-defined categories in the dataset and its multiples. For evaluation purposes, we conducted two experiments for Arabic website classification. The clustering performance of the Hybrid-FS-kmeans was examined during the initial experiment. The runtime consumed was observed in the second experiment. Before we implemented the experiments, we needed to prepare the webpages for the classification algorithms. We applied a pre-processing phase to collect and extract the related webpages. It also reduced the noisy terms (unwanted terms in the text) with the aim to make measuring the weighting of features easier. This phase consisted of collecting the URL seeds of webpages as a dataset using a Web extractor agent, and pre-processing the text. The following sections describe the pre-processing phase, the characteristics of the datasets utilized in the experiments, then discusses the evaluation criteria.

4.1 Pre-Processing Phase

The pre-processing phase aims to transform the collected Arabic text documents into an easily accessible representation of texts that is suitable for the clustering algorithm. In this phase, we have followed work done by Al-shammari (2010)[13] in handling the collected Arabic textual data according to these steps: tokenization, filtering and stemming.

1) Tokenization: It includes removing of punctuation symbols, numbers, non Arabic text and other symbols that can be used throughout the text such as tatweel character "-", used for aesthetic writing in the Arabic texts. In addition, a diacritic which is a feature of Arabic scripts omitted in this step. The Arabic language diacritical marks are: ـً, ـٌ, ـٍ , ـَ, ـُ, ـِ and ـّ. In fact, delete diacritical marks helped on defining the similarity between words.

2) *Filtering*: In this step, stop-words are removed. Stop-words are the words that appear frequently in the text and don't have any semantic meaning such as Arabic conjunction words, pronouns, and prepositions. Although, delete stop-words starting with a prefix 'و' letter, example "بين" ,"وبين" both refer to the same stop-word. In this experiment, the stop-words used counts 467 words and based on (http://arabicstopwords.sourceforge.net/). This step comprises normalization of some Arabic letters as following:

a) Replacing " إ", "ٱ " ," ؤ " and " آ" by alif bar "ا".
b) Replacing " ة" by " ه" at the end of the words.
c) Replacing " ى" by " ي" at the end of the words.
d) Replacing the letter " يء " by "ئ".

3) *Stemming:* In principle, the process of stemming from the Arabic word is to remove suffixes, prefixes and infixes. It plays the significant role in this experiment. The Larkey's stemmer [18] is adopted, but with some changes added. The aim of modifications is mainly to avoid drawbacks with implementing such a stemmer, also to achieve the idea of local stem[19]. The

applied stemmer overcomes many stemming errors caused by unguided removal of a fixed set of affixes associated with Larkey's stemmer, especially where it is hard to distinguish between an extra letter and a root letter[19] .

The implemented stemmer for each .txt document can be found clearly in Alghamdi and Selamat (2012).[18]

4.2 Datasets

In this study, we used an in-house collected corpus from the archives of online Arabic newspapers since there are no common Arabic datasets available to test the proposed classifier. The newspapers were Al-Akhbar[i], Alhayat[ii], Aldostor[iii], Gomhuria online[iv], Akhnar Alarab.Net[v], Alriyadh[vi], and the Saudi Times[vii]. This corpus is commonly used for applications related to Arabic text language. [19,20,13] The collected datasets contained 1554 documents of different lengths. These documents belonged to six categories as presented in Table 1. We used a Web extractor agent to extract the textual data from these websites. The tool used in this study was the Easy Web Extract version 2.7.[viii]

Table 1 Arabic dataset

Category Name	Number of Documents
Political news	395
Economic news	266
Sports News	262
Social News	125
Cultural News	231
Technology & Science	292
Total	1554

4.3 Evaluation Criteria

The quality of a clustering algorithm using the selected datasets was estimated using three evaluation measures, namely, the purity mean intra-cluster distance (MICD) davies-bouldin index (DBI) measures, which are widely used to evaluate the performance of unsupervised classification algorithms.[21,22,27] These evaluation measures are computed as follows:

- The purity measure is used to estimate the coherence of a resulted cluster. In our model, it evaluates the degree to which a cluster encloses documents from a particular category. The purity of a single cluster C_i of size e_i , is formally defined as:

$$Purity(C_i) = \frac{1}{e_i} \max_h e_i^h \quad (7)$$

where the $\max_h e_i^h$ represents the main category in cluster C_i and e_i^h correspond to the number of the documents that are in cluster C_i annotate to category h. In an optimal cluster, which just groups of documents from a single category, its purity rate is 1. The purity value will be between 0-1 and the higher value reflects a better quality of the cluster.

- Mean intra-cluster distance (MICD) is the distance between data vectors and its cluster centre where a low MICD signify a compact cluster and a high MICD is a loose cluster. It is calculated as follows[24]:

$$MICD = \frac{1}{N} \sum_{c_i \in c_k} \|c_i - \mu_k\| \quad (8)$$

where N is the number of pages to be clustered, k is the selected number of the clusters, μ_k represents the center of the cluster C_k, and $\| \ \|$ stands for Euclidean distance.

- Davies-bouldin index (DBI) aims to find a well separated compact clusters .It takes into account within cluster vectors variance and distance between clusters centers. 27 The smaller value of DBI shows a better clustering result. The DBI is calculated as:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j=1\cdots k, j\neq i} \left\{ \frac{\text{diam}(c_i) + \text{diam}(c_j)}{\|\mu_i - \mu_j\|^2} \right\} \quad (9)$$

where $\text{diam}(c_i)$ and $\text{diam}(c_j)$ are average distances of all data vectors in clusters i and j to their respective cluster centroids, μ_i is the center of cluster c_i consisting of N_i points and $\|\mu_i - \mu_j\|^2$ is the Euclidean distance between these centroids .

A good clustering has a small MICD (similar data vectors are grouped together), smaller DBI rate and high purity values.

■5.0 RESULTS AND DISCUSSION

The objective of this experiment was to investigate the effect of using the Hybrid-FS-kmeans clustering. We compared our proposed model with the standard *k*-means, using the Squared Euclidean distance measure for modeling the similarity between documents. As shown in Figure 4 and according to the purity evaluation, our Hybrid-FS-kmeans model produced the highest purity scores while the standard *k*-means performed worst. This indicates that using the model that includes hybrid feature selection with *k*-means is better than using the standard *k*-means.

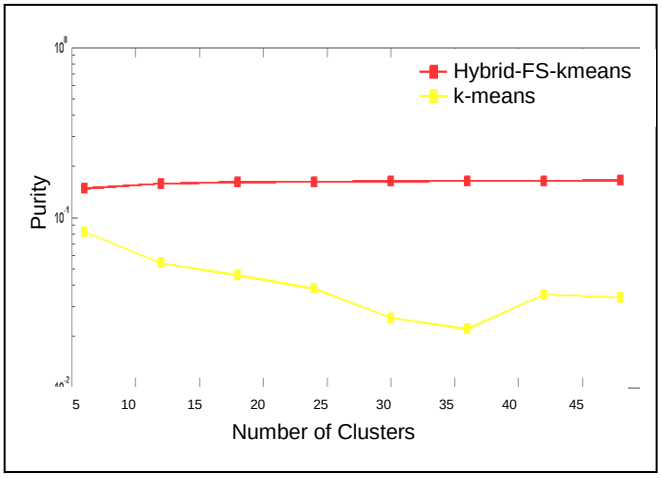


Figure 4 Purity results

[i] http://www.al-akhbar.com/
[ii] http://alhayat.com/
[iii] http://dostor.org/
[iv] http://www.gomhuriaonline.com/
[v] http://akhbaralarab.net/
[vi] http://www.alriyadh.com/section.home.html
[vii] http://www.sauditimes.net/Default.aspx
[viii] Easy Web Extract (http://webextract.net/)

Figure 5 shows the comparison results based on the MICD evaluation. The Hybrid-FS-kmeans produced good clusters with small MICD scores and tended to outperform the standard *k*-means. It helped in producing compact clusters where all the points in the cluster were close to each other.

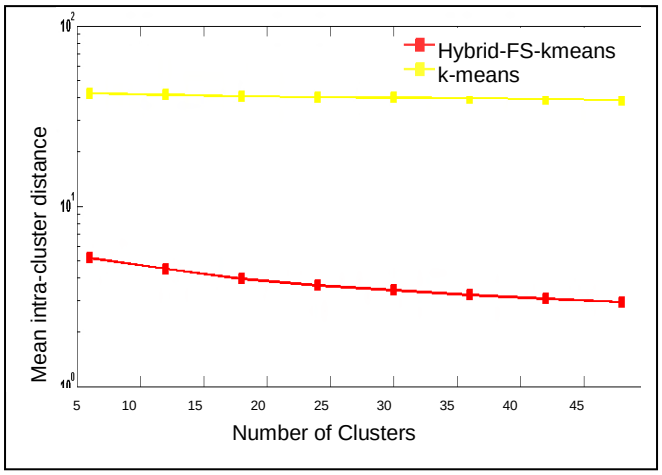


Figure 5 Mean intra-cluster distance (MICD) results

The comparison results based on DBI evaluation shown in Figure 6. Using Hybrid-FS-kmeans model give me smaller of DBI which means that we have a good separation distance between clusters and minimal distances between cluster's vectors and it centre.

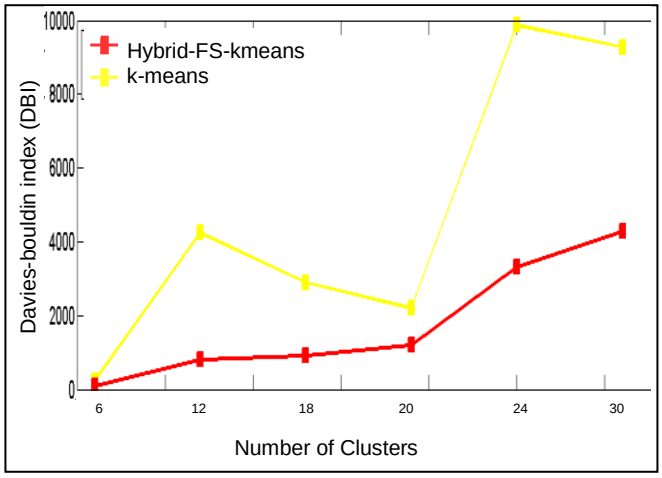


Figure 6 Davies-bouldin index (DBI) results

In conclusion, the proposed model performed an effective classification task that produced compact and well-separated clusters according to the categories as indicated by the high purity and low MICD and DBI scores.

Figure 7 displays the consumed time , calculated in seconds with each of the classification approaches. The computational time of the approach was determined according to the dataset size and the amount of words used in the document representation. Using our method, the dimension of the features reduced from 20390 features to 18000 features. The time elapsed by the Hybrid-FS-kmeans was much shorter than the standard *k*-means.

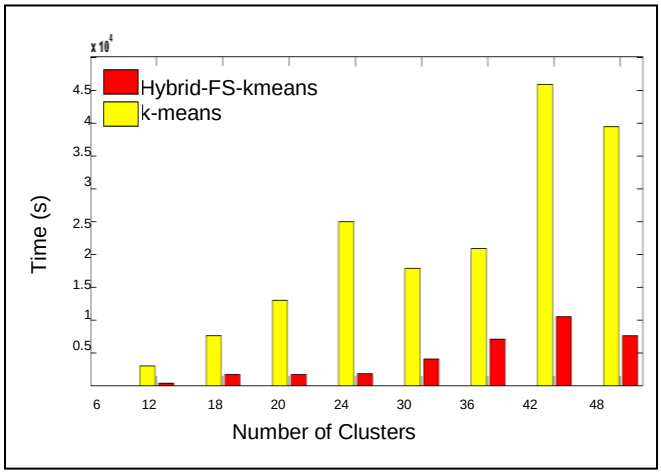


Figure 7 Runtime consumption

The low runtime consumption by our Hybrid-FS-kmeans model was due to the lower dimensions of the dataset. Thus, we can conclude that the Hybrid-FS method participated in a fast classification task, and we can be assured of the effectiveness and efficiency of our proposed Hybrid-FS-kmeans in classifying Arabic webpages.

5.1 Discussion

Using the proposed Hybrid-FS-kmeans model as explained in this paper, we have been able to increase purity, according to average value, by 28% compared to the standard *k*-means algorithm (Figure 4). In addition, we succeeded in decreasing the MICD by 98% compared to the standard *k*-means algorithm (Figure 5). Furthermore, we managed to lower the runtime using the proposed model by 80% compared to the standard *k*-means algorithm (Figure 7). We found that the Hybrid-FS-kmeans performed a perfect clustering task as it able to minimize MICD , increase purity with low runtime. It contributed to a fast classification task. By using the proposed model, we were able to represent text data in a high structure that consists of three types of objects, namely, the term, document and category. Moreover, we improved the effectiveness of the *k*-means algorithm to produce coherent-compact clusters that were well-separated according to the categories.

■6.0 CONCLUSION

An model called the Hybrid-FS-kmeans to classify Arabic webpages was proposed in this paper. The model is intended to reduce the high-dimensionality of datasets as a way to improve text clustering. The model consisted of integrating the *k*-means document clustering method (applied as unsupervised classification) with a hybrid model based on feature selection. A feature selection model that incorporates three different feature selection methods (CHI-squared, mutual information, and term frequency-inverse document frequency) to build a hybrid feature selection model (Hybrid-FS) was proposed. The proposed model was examined on a set of common Arabic online newspapers. As a result, we obtained promising classification results that indicated the method was effective in decreasing the MICD, maximizing purity and minimizing runtime. We revealed that a combination of the features selected by various methods is effective in improving the *k*-means clustering.

For future work, we are interested in carrying out more experiments to compare the outcome of the single feature selection method with the Hybrid-FS. We also plan to propose hybrid feature extraction (latent semantic analysis, independent component analysis and principle component analysis) similar to the Hybrid-FS to be used with *k*-means as another way to lower the dimensionality of the dataset and to decrease the time consumption.

Acknowledgement

The authors would like to extend their thanks to Universiti Teknologi Malaysia (UTM) Research University funding Vot 03H02 and Ministry of Higher Education, Saudi Arabia, for supporting the research.

References

[1] Chang, Y. and K. Lee. 2011. Bayesian Feature Selection for Sparse Topic Model. *IEEE InternationalWorkshop on Machine Learning for Signal Processing*. 1–6.

[2] Zhang, Y. and Q. Zhang. 2006. A Text Classifier Based on Sentence Category VSM. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*. 244–249.

[3] Gharib, T. F., M. M. Fouad, A. Mashat, and I. Bidawi. 2012. Self Organizing Map -based Document Clustering Using WordNet Ontologies. *Int. J. Comput. Sci.* 9(1): 88–95.

[4] Napoleon, D. and S. Pavalakodi. 2011. A New Method for Dimensionality Reduction using K- Means Clustering Algorithm for High Dimensional Data Set. *Int. J. Comput. Appl*. 13(7): 41–46.

[5] Farahat, A. K. and M. S. Kamel.Statistical Semantics for Enhancing Document Clustering. 2011. *Knowl. Inf. Syst.* 28(2): 365–393.

[6] Li ,R. Z. and Y. Sen Zhang. 2012. Study on the Method of Feature Selection Based on Hybrid Model for Text Classification. *Adv. Mater. Res*. 433–440: 2881–2886.

[7] Gunal, S. 2012. Hybrid Feature Selection for Text Classification. *Turkish J. Electr. Eng. Comput. Sci.* 20(2): 1296–1311.

[8] Isa, D., L. H. Lee, V. P. Kallimani, and R. RajKumar. 2008. Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine. *Knowl. Data Eng.* 20(9): 1264–1272.

[9] Jing, L., J. Yun, J. Yu, and J. Huang. 2011. High-Order Co-clustering Text Data on Semantics-Based Representation Model. *Advances in Knowledge Discovery and Data Mining*. 171–182.

[10] Uysal, A. K. and S. Gunal. 2012. A Novel Probabilistic Feature Selection Method for Text Classification. *Knowledge-Based Syst.* 36: 226–235.

[11] Zhou, Y., Y. Yang, W. Peng, and Y. Ping. 2010. A Novel Term Weighting Scheme With Distributional Coefficient For Text Categorization With Support Vector Machine. *IEEE Youth Conference onInformation Computing and Telecommunications (YC-ICT)*. 2–5.

[12] Guru, D. S., B. S. Harish, and S. Manjunath. 2010.Symbolic Representation of Text Documents. *Third Annual ACM Bangalore Conference*. 1–4.

[13] Mesleh, A. 2007. Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. *J. Comput. Sci.* 3(6): 430–435.

[14] Machova, K., A. Szaboova, and P. Bednar. 2007. Generation of a Set of Key Terms Characterising Text Documents. *J. Inf. Organ. Sci.* 31(1).

[15] Yongqing, W., L. Pei-yu, and Z. Zhu. 2008. A Feature Selection Method based on Improved TFIDF. *Third International Conference on Pervasive Computing and Applications*. 94–97.

[16] Qu, S., S. Wang, and Y. Zou. 2008. Improvement of Text Feature Selection Method Based on TFIDF. *International Seminar on Future Information Technology and Management Engineering*. 79–81.

[17] Ramos, J. 1999. Using TF-IDF to Determine Word Relevance in Document Queries. *First International Conference on Machine Learning*.

[18] Andrews, N. O. and E. A. Fox. 2007. *Recent Developments in Document Clustering.*

[19] Jain, A. and M. Murty. 1999. Data Clustering: A Review. *ACM Comput. Surv.* 31(3): 255–323.

[20] Larkey, L., L. Ballesteros, and M. Connell. 2007. Light Stemming for Arabic Information Retrieval. *Arabic Computational Morphology*, no. Ldc, A. Soud, A. van den Bosch, and G. Neumann, Eds. Springer. 221–243.

[21] Al-shammari, E. 2010. Improving Arabic Document Categorization : Introducing Local Stem. *10th International Conference on Intelligent Systems Design and Applications*. 385–390.

[22] Alghamdi, H. M. and A. Selamat. 2012. Topic Detections in Arabic Dark Websites Using Improved Vector Space Model. *4th Conference on Data Mining and Optimization (DMO)*. 6–11.

[23] Al-diabat, M. 2012. Arabic Text Categorization Using Classification Rule Mining. *Appl. Math. Sci.* 6(81): 4033–4046.

[24] Alsaleem, S. 2011. Automated Arabic Text Categorization Using SVM and NB. *Int. Arab J. e-Technology*. 2(2): 124–128.

[25] Huang, A. 2008. Similarity Measures for Text Document Clustering. *The New Zealand Computer Science Research Student Conference*.

[26] Rokach, L. and O. Maimon. Clustering Methods. *Data Mining and Knowledge Discovery Handbook*. O. M. and L. Rokach, Ed. New York.

[27] Rana S., Jasola S., and Kumar R. 2013. A Boundary Restricted Adaptive Particle Swarm Optimization for Data Clustering. *Int. J. Mach. Learn. Cybern.* 4(4): 391–400.