Modified union feature selection method on English translation of hadith text clustering

A F Huda\(^1\), N Priatna\(^1\), Q U Safitri\(^2\) and W Darmalaksana\(^3\)

\(^1\) Mathematic Department, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia
\(^2\) Mathematic Department, Institut Teknologi Bandung, Bandung, Indonesia
\(^3\) Ilmu Hadith Department, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia

*afhuda@uinsgd.ac.id

Abstract. The high feature space (dimension) is one of the main issues to be considered in the text clustering process. Therefore, various dimensional reduction methods have been introduced for selecting informative sub feature. Each method uses a different strategy to select sub feature, and the results are different even if using the same data set. Typically, union methods and intersection methods are used to combine selected sub feature with different reduction methods. The union method selects all feature and intersection only selects the general feature under consideration. Thus, the union approach causes an increase in feature dimensions and the intersection approach causes the loss of some important feature. Therefore, in order to take advantage of a method and reduce its weaknesses, this research proposes new approach, which are called modified union. This approach applies the union methods to select top ranking feature and applies intersection methods to the rest of the feature. In this case, feature selection uses the Term Variance (TV) and Document Frequency (DF) methods to calculate the relevance value of each feature. The effectiveness of the proposed method is tested on the data set of Hadith Shahih Bukhary. The results show that the proposed method improves clustering accuracy over other methods with DB index is 2.7.

1. Introduction

The digital era increases the number of digital documents exponentially, so that the document processing becomes an unavoidable need. Text clustering is one of the digital text processing based on their characteristics. Text clustering is applied to several domain applications, such as organizing search results based on user queries [1], browsing documents in large collections [2], topic detection [3], generating hierarchies of documents on the web [4]. However, a problem arises when the dimensions of the feature space in the text clustering are very large. Space feature in clustering text is the number of terms found in the document. As a result, the larger the document will be clustered, the more features will be generated. This will make the computing costs expensive and reduce the performance of the clustering algorithm because it contains less relevant terms. To overcome this, a dimensional reduction method appears [5].

Bharti et al. introduce the feature reduction method by combining the union with the intersection of different extraction methods in the document sub feature [6]. The union method is an extraction method by combining all features of a document, while intersection only selects the common features considered. As a result, the union method will add data dimensions, while intersection will eliminate some important features. Merging these two methods is to apply the union method to top-level features,
and apply the intersection method to other features. In this study, the feature selection used is term variance (TV) and document frequency (DF).

The results of this dimension reduction will be carried out on the Sahih al-Bukhory hadith translation in English, to be clustered using the k-means algorithm. Next, the clustering results will be validated using the Davies Bouldin Index to find out the quality of the cluster formed. Finally, the performance of the reduction method will be known for the clustering of hadith data.

2. Modified feature selection

2.1. k-Means clustering techniques

Clustering is a technique of grouping objects into similar groups [7]. Clustering has several approaches, including center-based partition clustering. Clustering center-based partitions is the division of objects into groups that are mutually independent, based on a center. If the center is an average object in a cluster, it is called k-means, whereas if the center is a central cluster object, it is called k-medoid.

k-Means is the simplest clustering technique. k-Means have been discovered since 1957 by Lloyd, and continue to be developed by other scientists, such as Forgey on 1965, Friedman and Rubin on 1967, and McQueen on 1967, but were only published in 1982 [5]. Basically, the k-means algorithm will group n objects into k clusters, and each cluster is represented by a "center" obtained from the average data value in each cluster. Each object will be a member of the cluster that has the closest center to it. One way to define this closeness is by euclidean distance. Euclidean distance can be calculated by the equation,

\[ d(x, y) = \sqrt{\sum_{i=1}^{p} |x_i - y_i|^2} \]  \hspace{1cm} (1)

with \( x \) and \( y \) are vectors, and \( p \) is the number of vector attributes.

However, k-means has several disadvantages, that is the number of clusters determined at the beginning, only finding local optimal solutions, and depends on the initial centroid chosen. Even the initial centroid can be chosen randomly. Therefore, the results of clustering with the k-means algorithm require validation to be accepted. One method of validating the results of clustering is Davies Bouldin Index (DB).

Davies Bouldin Index measures cluster quality based on the amount of data proximity to the centroid cluster followed and the distance between the centroid of the cluster. Davies Bouldin index is obtained by equation,

\[ DB = \frac{1}{k} \sum_{i=1}^{k} M_i = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left\{ \frac{\delta_i + \delta_j}{\Delta_{ij}} \right\} \]  \hspace{1cm} (2)

where \( \delta_i = \frac{1}{n_i} \sum_{e=1}^{n_i} \|x_e - c_i\| \), and \( \Delta_{ij} = \|c_i - c_j\| \). \( n_i \) is the number of elements in the \( i \)-th cluster, \( c_j \) and \( c_i \) is center of \( j \)-th and \( i \)-th cluster, respectively.

2.2. Feature selection methods

Feature selection is one of the dimension’s reduction methods by removing irrelevant features based on certain criteria. There are two approach models in the feature selection, namely wrapper and filter. The wrapper approach model uses unsupervised learning algorithms for each sub feature candidate then evaluates them with a validation measure, so that testing is required repeatedly to get suitable features, and requires computationally expensive for large data dimensions [8]. While the filter approach uses statistical analysis to determine the relevance of features without repeatedly testing with machine learning, so that the filter model approach is relatively fast and more efficient [5]. Some ways to calculate the relevance value of a feature with a filter model are as follows:
2.2.1. Document Frequency (DF). Document frequency (DF) obtained by calculating the number of documents where a feature appears. This is the simplest criterion.

2.2.2. Term Contribution (TC). The definition of Term Contribution is the value of feature/term contributions in the process of calculating document persistence. TC is obtained from the dot product weight of the term in documents i-th and j-th, for all j, that is,

\[ TC(t_k) = \sum_{i,j \neq i} w(t_k, d_i) w(t_k, d_j) \]  

with \( w(t_k, d_i) \) is the weight of the k-th term in the i-th document.

2.2.3. Term Variance Quality (TVQ). The value of TVQ \( (q(t_k)) \) requires the frequency of features in the i-th document, as well as the number of documents where the \( t_k \) term appears at least once.

\[ q(t_k) = \sum_{i=1}^{n} f_{ki}^2 - \frac{1}{n} \left( \sum_{i=1}^{n} f_{ki} \right)^2 \]  

(4)

2.2.4. Term Varians (TV). A feature that appears in a few documents or has a general distribution throughout the document will have a low value of \( v(t_k) \). From the equation below, it can be seen that the value of \( v(t_k) \) is strongly influenced by the frequency of occurrence \( (f_k) \) of the feature.

\[ v(t_k) = \frac{1}{n} \sum_{j=1}^{n} \left( f_{kj} - \bar{f}_k \right)^2 \]  

(5)

2.3. Combine feature selection methods

The feature combination is done by the union and intersection method. Suppose that the Term Variance (TV) and Document Frequency (DF) filter methods are used to calculate feature relevance values. Let \( T \) be a collection of original features from the preprocessing results of document set \( D \), with \( T = \{t_1, t_2, ..., t_r\} \). Let \( FS_1 = \{t_{11}, t_{12}, ..., t_{q}\} \) is a sub feature obtained from the TV, with \( q < r \) and \( FS_2 = \{t_{21}, t_{22}, ..., t_{r}\} \) are sub features obtained from the DF, with \( s < r \). Then,

2.3.1. Union. Let \( FS_3 \) is a combination of sub feature \( FS_1 \) and \( FS_2 \), then

\[ FS_3 = FS_1 \cup FS_2 = \{t_{31}, t_{32}, ..., t_{r}\} \]  

(6)

2.3.2. Intersection. Let \( FS_4 \) is a combination of sub feature \( FS_1 \) and \( FS_2 \), then

\[ FS_4 = FS_1 \cap FS_2 = \{t_{41}, t_{42}, ..., t_{r}\} \]  

(7)

2.3.3. Modified union. This method is a combination of union and intersection methods. For example \( FS_5 \) is the result of a combination of both methods, selected constants \( c_1 \) and \( c_2 \) as the percentage composition of the union and intersection methods, respectively, so that,

\[ FS_5 = c_1 \% \{FS_1\} \cup c_2 \% \{FS_2\} = \{t_{51}, t_{52}, ..., t_{r}\} \]  

(8)

3. Experiment

3.1. Data set

In this study, the data set used was 539 hadith from 5 chapters in Sahih al-Bukhary, so that the data to be processed was 539 documents. The data set details are presented in table 1. Comparison of the results of the proposed method, an experimental scenario is designed as in table 2.
Table 1. Dataset.

| Name Chapter                              | Amount Hadith |
|-------------------------------------------|---------------|
| Adzan                                     | 106           |
| Wudhu                                     | 88            |
| Holding Fast to the Quran and Sunnah      | 86            |
| Knowledge                                 | 77            |
| Tawheed                                   | 182           |
| **Total**                                 | **539**       |

Table 2. Experimental scenario.

| Test | Test Scenario                           |
|------|-----------------------------------------|
| 1    | Data set → Pre processing → DF → k-Means |
| 2    | Data set → Pre processing → TV → k-Means |
| 3    | Data set → Pre processing → Union (DF, TV) → k-Means |
| 4    | Data set → Pre processing → Intersection (DF, TV) → k-Means |
| 5    | Data set → Pre processing → Modified Union → k-Means |

3.2. Pre processing

Data pre processing is done by Python version 3.5, by experiencing several stages, i.e reading documents, tokenizing, case folding, stop-word removal and stemming. Tokenizing is the process of breaking a document into a list of words/terms contained in the document. Case folding is done to homogenize the font type in the document. Stopword removal is the process of removing words/terms that are too general and do not contribute to documents. While stemming is the process of changing terms into basic words. From the pre processing results, 2737 terms are generated, meaning that the processed data set can be formed into a matrix of $539 \times 2737$.

3.2.1. Weighting. After the document goes through the pre-processing, then the corpus or dictionary is compiled which contains all the words in the document. Then the weight of each term of each document (term weighting) is calculated.

3.2.2. Feature selection process. The dimension reduction process is done to reduce 50% of all features produced. Feature reduction is based on the value of relevance of DF and TV.

3.2.3. Clustering and index of cluster. k-Means used to cluster data set after feature selection with all variation on experimental scenario. All of experiment use $k = 5$, because of number of chapter in hadith data set. The result of clustering evaluated in Davis Bouldin Index (DB).

4. Results and discussion

Data sets with features that have been selected and grouped using $k$-means with $k = 5$, because the data set is taken from the 5 chapters of the hadith of Sahih al-Bukhary. The DF and TV feature selection results produce a number of fixed features, i.e 1368 features from the initial total feature because they only throw 50% of the term with the smallest DF and TF weight. However, if more than 50% of the smallest features have the same weight value, the system will take 50% from the first list to be discarded, and leave the rest.

For the union and intersection method produces a number of different final features, but tends to decrease significantly from the number of original features. The union method incorporates features from DF and TV, and obtained 1551 features, resulting in increased dimensions compared to the DF and TV methods. While the intersection method produces fewer features than the previous 3 methods. However, this method causes the loss of several important features.
As for the modified union, the final total term varies because there are random terms taken even though using weighting $c_1$ and $c_2$. In addition, the effects of $k$-means clustering that depend on the selection of initial centroids make the results of clustering in this fifth scenario quite vulnerable. As a result, clustering with the feature selection modified union is carried out 5 times for performance compared to the previous four methods. Even so, the method of selection feature modified union produces the most reduction, with the most optimal clustering results. The results for this fifth scenario are presented in table 3, and a comparison of the results of the five methods is presented in table 4.

Table 3. Clustering result with feature selection modified union on hadith shahih Al-Bukhary.

| Experiment | Amount Term | Validation Clustering Result (DB) |
|------------|-------------|----------------------------------|
|            | Initial = 2737 |                                  |
|            | Reduced | Final |                               |
| 1          | 1754   | 983   | 2,786                           |
| 2          | 1765   | 972   | 2,782                           |
| 3          | 1755   | 982   | 2,696                           |
| 4          | 1734   | 1003  | 3,030                           |
| 5          | 1752   | 985   | 2,548                           |
| Average    | 985    | 2,768 |

Table 4. Clustering result with feature selection on hadith shahih Al-Bukhary.

| Experiment Method | Total Term | Validation Clustering Result (DB) |
|-------------------|------------|----------------------------------|
|                   | Initial = 2737 |                                  |
|                   | Reduced | Final |                               |
| DF                | 1369    | 1368  | 3,233                           |
| TV                | 1369    | 1368  | 2,973                           |
| Union (DF, TV)    | 1186    | 1551  | 3,212                           |
| Intersection (DF, TV) | 1550  | 1187  | 3,246                           |
| Modified Union (DF, TV) | 1752 | 985   | 2,768                           |

Table 4 it shows that the method of combining features with the modified union proposed is the best method. This also proves that the number of processed features does not make good clustering results, because a number of features can increase computational complexity and can worsen the performance of clustering methods.

5. Conclusion
Based on the above research, the combining method of DF and TF-based feature selection with a modified union gives the most optimal results on the results of clusters. Even so, this method still has disadvantages because it depends on the value of the parameters $c_1$ and $c_2$ which makes the cluster results unstable. It is hoped that future research will provide a solution to determine the value of these parameters in order to produce good reductions and good clusters.

References
[1] Zamir O, Etzioni O, Madani O and Karp R M 1997 Fast and Intuitive Clustering of Web Documents in Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining 97 287-290
[2] Cutting D R, Karger D R, Pedersen J O and Tukey J W 1992 Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections in Proc. of the 15th annual International
ACM SIGIR Conference on Research and Development in Information Retrieval 318-329

[3] Huang S, Peng X and Niu Z 2011 News Topic Detection Based on Hierarchical Clustering and Named Entity in 7th Int. Conf. on Natural Language Processing and Knowledge Engineering 280-284

[4] Koller D and Sahami M 1997 Hierarchically Classifying Documents Using Very Few Words in Proc. of the 14th Int. Conf. on Machine Learning 170-178

[5] Hani'iah M, Fatichah C and Purwitasari D 2016 Reduksi Dimensi Fitur Menggunakan Algoritma ALOFT untuk Pengelompokkan Dokumen in Scientific Journal of Information Technology 14(2) 207-216

[6] Bharti K K and Singh P K 2015 Hybrid Dimension Reduction by Integrating Feature Selection with Feature Extraction Method for Text Clustering in Expert Systems with Applications 42 (6) 3105-3114

[7] Tan P N, Steinbach M and Kumar V 2006 Introduction to Data Mining (Addison Wesley)

[8] Turban E, Sharda R and Delen D 2011 Decision Support and Business Intelligence Systems 9th Edition (New Jersey: Pearson Education Inc)