Journal of Information Technology and Computer Science
Volume 6, Number 2, August 2021, pp. 117-129
Journal Homepage: www.jitecs.ub.ac.id

Evaluation of TF-IDF Algorithm Weighting Scheme in The Qur'an Translation Clustering with K-Means Algorithm

M. Didik R. Wahyudi*[1]

[1]Informatic Engineering, Faculty of Sains and Technology, State Islamic University Sunan Kalijaga
[1]m.didik@uin-suka.ac.id
*Corresponding Author

Received 12 February 2021; accepted 11 June 2021

Abstract. The Al-Quran translation index issued by the Ministry of Religion can be used in text mining to search for similar patterns of Al-Quran translation. This study performs sentence grouping using the K-Means Clustering algorithm and three weighting scheme models of the TF-IDF algorithm to get the best performance of the Tf-IDF algorithm.
From the three models of the TF-IDF algorithm weighting scheme, the highest percentage results were obtained in the traditional TF-IDF weighting scheme, namely 62.16% with an average percentage of 36.12% and a standard deviation of 12.77%. The smallest results are shown in the TF-IDF 1 normalization weighting scheme, namely 48.65% with an average percentage of 25.65% and a standard deviation of 10.16%. The smallest standard deviation results in a normalized 2 TF-IDF weighting of 8.27% with an average percentage of 28.15% and the largest percentage weighting of 48.65% which is the same as the normalized TF-IDF 1 weighting.

1. Introduction

Al-Qur'an is the Muslim holy book and is used as a guide to the life of Muslims, has a unique structure consisting of two parts, three parts, five parts, seven parts and so on. The Qur'an consists of 114 letters, each letter consisting of several verses. The number of verses in the Qur'an reaches 6236 verses. The Qur'an is divided into 30 sections called juz. Each juz is divided into several curves. The letters in the Qur'an have varying amounts of ruku depending on the number of verses in the letter and the short length of each verse [1]. The grouping aims to make it easier to memorize, learn, and study the Qur'an. In each letter can contain various themes. Certain themes can be in several letters. To easily learn and understand the Qur'an, it can use the translation of the Qur'an by following the language understood.

The Indonesian translation of the Qur'an issued by the Ministry of Religion of the Republic of Indonesia is the main reference in Indonesia, although there are several versions of the Indonesian translation of the Qur'an carried out by various social organizations. The translation of the Qur'an in Indonesian is an interesting object for computer scientists to demonstrate the knowledge, wisdom and law of the verses of the Qur'an in a computer system. Understanding the meaning of the verses of the Qur'an can be done by reading interpretations written by interpreters. However, this has not been enough to give a complete picture of the meaning contained in the

Qur'an. For us to get a complete picture of the various themes in the Qur'an, we must read and understand all parts of the Qur'an.

In the field of computing, the various unique structures of the Qur'an are very interesting to study. One way to research is with text mining. Various text mining methods can be used to group certain data, one of them is clustering. Text clustering is an important part of the text mining method. Text clustering is a classification of documents that divides a collection of text into several subsets called clusters, the text of each cluster having greater similarity than those in different clusters. Clustering is particularly useful for organizing documents to improve information rediscovery and support the browsing process [2].

The quality of clustering is very dependent on the process of removing interference from the pattern used in the clustering process. So we need pre- processing processes such as separating words from documents (tokenization), removing words that often appear but are not relevant (stopword removal) and changing words into basic words (stemming). Each word will be represented by a weighting method based on the frequency of words appearing, namely TF-IDF. TF- IDF is very well used for weighting but has many limitations. Many researchers propose modifications to TF-IDF for the best performance [3][4][5]

2. Related Work

The research entitled Application of the Cosine Similarity Algorithm in Text Mining Translated the Qur'an Based on Topic Linkages [6] was carried out by searching for similarities in the text in the Qur'an translation. Of the similarity groups formed, they are then compared with the Qur'an index compiled by the Ministry of Religion of the Republic of Indonesia. From the comparison of the two groups, the results show that the similarity generated from the Cosine Similarity has a similarity of 46,42% with the Qur'an index made by the Ministry of Religion. These results are felt to be less than optimal, so this research is expected to provide better results.

The Qur'an index compiled by the Ministry of Religion was compiled by expert commentators and has been institutionally recognized in Indonesia as a valid the Qur'an index. Chapters in the Qur'an index are arranged based on the similarities between the verses of the Qur'an, so that they can be used as a reference to test the results of text similarity between text mining. One way to improve the performance of the text mining algorithm is to normalize the algorithm. Normalization can be done by making the right weighting scheme to increase effectiveness [3].

The TF-IDF normalization in a study entitled "Modified TF-IDF Term Weighting Strategies for Text Categorization" conducted by Rajendra Kumar Roul, et al., Succeeded in processing text well, but it was still too simple in processing a text, so many neglected the details of words that were used. actually more meaningful [7]. Optimization of TF-IDF can also be done by using the maximum TF-IDF method. Maximum TF-IDF is a normalization method in which the frequency value is divided by the largest number of words that appear to optimize the best results in the algorithm. This is done so that the K-Nearest Neighbor algorithm gives the best results [8].

Research conducted by S. Albitar et al propose a new measure for assessing semantic similarity between texts based on TF/IDF with a new function that aggregates semantic similarities between concepts representing the compared text documents pair-to-pair using a semantic similarity matrix. Experimental results demonstrate that our measure outperforms other semantic and classical measures with significant improvements in the concept space [4].

Research conducted by Calho, H, propose using Genetic Programming to find a suitable expression composed of TF and IDF terms that maximizes the discrimination of such terms given a reduced bootstrapping set of examples labeled for each region [5]. From the research that has been done, it shows that TF-IDF normalization can

improve the quality of the results for the better. Therefore, this research will normalize the TF-IDF with the hope of obtaining better results.

3. Research Method

The method used in this research is the study of literature. In the literature study, literature search is not only for the initial steps of preparing a research framework but also utilizing library resources to obtain research data [9]. In this type of research, researchers do not have to go to the field and meet with respondents. The data needed in research can be obtained from library sources or documents. The data source used in this study is the translation of digital the Qur'an compiled into a dataset with the required format.

The data that will be used in this study are the text of the translation of the Qur'an in Indonesian and the index of the Qur'an issued by the Ministry of Religion of the Republic of Indonesia contained in the Al-Fatih Manuscripts. The following are the steps for the research:

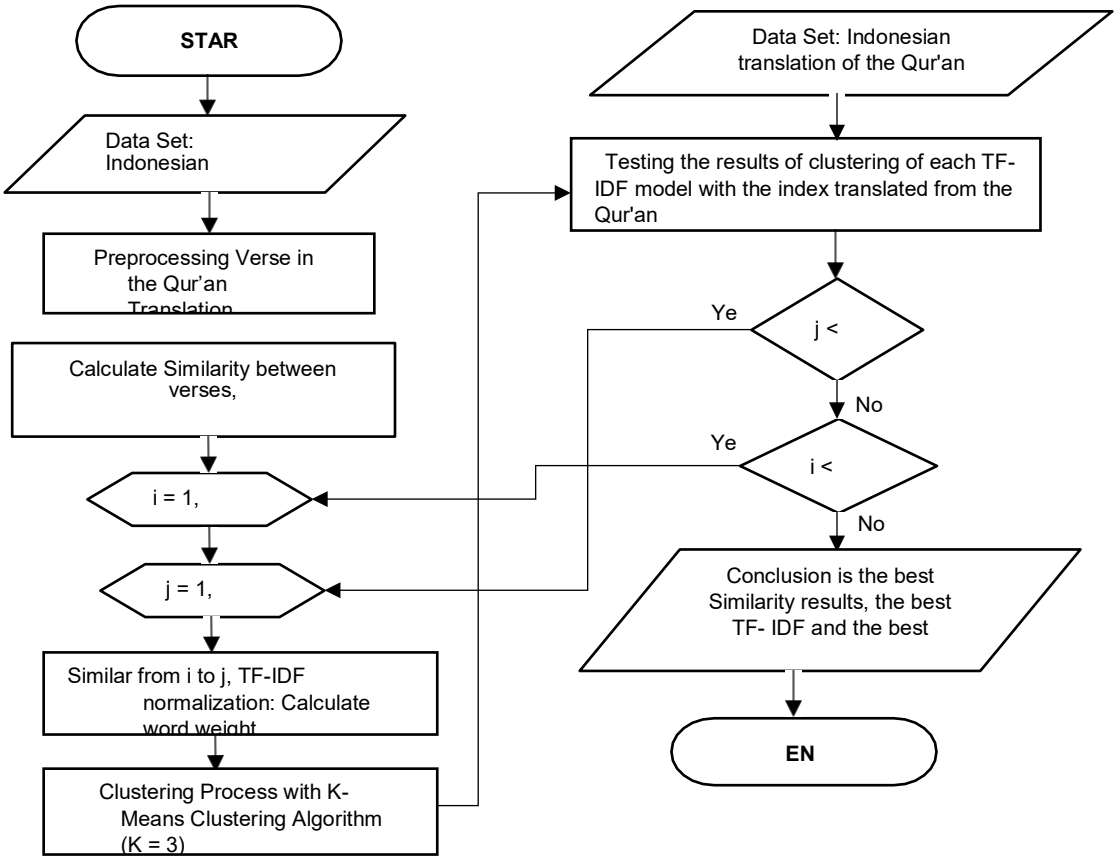

Figure 1: Research Steps

3.1 Preprocessing

Preprocessing in this study includes lemmatizing and stop words are not removed. The process of lemmatizing is a process to return a word to the root word. In this study, the stop word was left alone and not deleted. The removal of the stop word will result in a change in the actual meaning of the translation of the Qur'an. This is actually contrary to the truth value of a verse of the Qur'an.

3.2 Similarity between the translation of verses of the Qur'an

Similarity data between verses used in this study is the result of data processing carried out in previous research [6]. From this data, as many as 6136 groups of similarity were formed. For this research, we take 15 similar groups with the most number of verses which will represent the number of chapters in the Qur'an index and evaluated with the Qur'an index. The similarity group of the verse is shown in Figure 2 below:

KLP-SIM	id1	JML
1	[4:140]	2211
2	[3:152]	2197
3	[2:190]	2135
4	[3:103]	2100
5	[3:160]	2091
6	[2:120]	2064
7	[5:12]	2063
8	[2:80]	2050
9	[4:94]	2037
10	[3:55]	2032
11	[3:20]	2027
12	[2:110]	1991
13	[2:137]	1970
14	[2:187]	1963
15	[3:123]	1962
16	[3:32]	1950
17	[3:99]	1928
18	[2:148]	1893

Record: 1 of 6134 No Filter

Figure 2: List of Similar Groups

The next step is to process each group of similarity with the three TF-IDF algorithm models which will be continued with the clustering process with the K-Means Clustering algorithm. The number of clusters formed is 3 clusters for each group of similar.

3.3 Normalization TF-IDF

In Information Retrieval, TF-IDF (Term Frequency-Inverse Document Frequency) is a product of two statistics, namely Term Frequency (TF) and Inverse Document Frequency (IDF) which shows how important a word is for a document in a corpus collection [10]. There are many ways to determine the two statistical values

This value is often used as a weighting factor in searching in information retrieval, text mining and user modeling. The TF-IDF value will increase proportionally if the frequency of occurrence of words in a document is balanced by the number of documents in the corpus containing the word. TF-IDF is one of the most popular term weighting schemes today. Around 83% of text-based recommendation systems in digital libraries use TF-IDF [11].

Various TF-IDF weighting schemes are often used by search engines as one of the methods in providing value and relevance ranking of documents requested by users. TF-IDF can be used to filter stop words in various subject areas, including summation and text classification. In this study, three TF-IDF models will be used to find the best performance in the clustering that will be carried out. The three TF-IDF models are:

1. Traditional TF-IDF. Traditional TF-IDF is a form of TF ID commonly used with default settings without parameters. This model works quite well but

ignores many details when processing documents such as document length and frequency distribution. The traditional TF-IDF invocation command in the python language used in this study is:

tfidf = TfidfVectorizer ()

2. TF-IDF Normalization 1. In normalization 1, the TF-IDF algorithm used is given additional parameters. The additional parameter is to eliminate words that appear in more than 80% of documents (max_df parameter in TF-IDF) and less than 20% of documents (min_df parameter in TF-IDF). This word removal is done with the assumption that words that appear in more than 80% of documents and less than 20% of documents are words that are not important and have no meaning in the document being processed [12]. The command program for calling TF-IDF normalization 1 in python language used in this study is:

tfidf = TfidfVectorizer (max_df = 0.8, min_df = 0.2, use_idf = True, ngram_range = (1,2), sublinear_tf = True, norm = 'max ')

3. TF-IDF Normalization 2. TF-IDF Normalization 2 still uses the parameters max_df and min_df as normalization 1. It's just that the values of the two variables are different. In normalization 2, the value of max_df = 0.5 and value of min_df = 0.1. The value of the two parameters is adopted from Singhal's research, A., entitled Pivoted document length normalization [13]. The program command for the normalization 2 TF-IDF call in python is:

tfidf = TfidfVectorizer (max_df = 0.5, min_df = 0.1)

3.3 K-MEANS Clustering

Clustering is grouping data items into a number of groups [14]. The two main approaches are clustering with the partitioning approach and clustering with the hierarchical approach. Clustering with the partitioning approach is clustering by sifting through data. Hierarchical clustering groups data by creating a hierarchy in the form of curves that describe clustering clusters where similar data will be placed in adjacent hierarchies.

One clustering algorithm is K-Means. The K-Means algorithm, first introduced by MacQueen JB in 1976, is a method of analyzing data or the method of data mining that performs the process of modeling without supervision (unsupervised) and is one of the methods for grouping data with a partition system. The K-Means method groups data into groups, where the data in one group has the same characteristics and has different characteristics from the data in other groups [15]. The results of the cluster using the K-Means method depend on the center value of the initial group given. Giving different initial values can produce different groups [16].

After weighting with three TF-IDF models, the next process is clustering the results of this weighting into 3 clusters using the K-Means Clustering algorithm. This clustering process is carried out on all **three** TF-IDF models. The clustering command with the K-Means Clustering algorithm in the python language is as follows:

clusterer = KMeans (n_clusters = 3, max_iter = 300, toll = 0.0001)

The use of 3 clusters in this study is because this study uses 3 TF-IDF models. The max iteration value is 300, because this is the maximum number of iterations of the k- means algorithm for one run [17]

3.4 Conversion of The Qur'an Index Data

In this study, the Qur'an index was used by the Ministry of Religion of the Republic of Indonesia. This index data is included in the Al Fatih Manuscripts. Al Fatih Manuscripts is one of the Qur'ans which is quite complete [18]. One of them is the Qur'anic index. In this study, the editor of Al Fatih was willing to provide a soft copy of Al Qur'an index data in the text format as shown in Figure 3.

```
BAB I SEKITAR ARKANUL ISLAM,,
PASAL I : AD-DIEN (AGAMA),,
1,Agama yang diridhai Allah.  ,
,"(2:112,213)(3:19,83,102)(4:125)(5:3)(6:14,70,125,161,162)(27:91)(33:35)(39:11-12,22)",
"(40:66)(41:33)(42:13)(45:18,19)(61:9)(72:14)(98:4,5)(110:1-2).",,
2,Tidak ada paksaan.,
,(2:256)(10:99)(18:29)(22:78)(42:8),
3,Ajakan kepada Islam.,
,"(2:211,285)(5:3)(6:70)(21:92)(23:52)(28:61)(32:18)(39:11-14)(57:16)(87:14)(98:5)",
4,Hakikat Islam.,
,"(1:6-7)(2:112,131-132,135,142,208)(3:19-20,",
"51,67,85,101)(4:125)(5:16)(6:136,153,161)(7:29)(9:33)(10:25)(11:56)(12:40)(16:76)(19:36)(21:92)(22:54,78)
33)(42:13,53)(43:43,61,63)(48:2,20,28)(61:9)(67:22)(72:13)(98:5).",,
5,Ikhlas dalam beragama.,
,"(10:22,105)(29:65)(31:32)(39:2-3,11)(40:14,65)(98:5).",
6,Orang-orang Islam. ,
,"(2:132,136)(3:52,64,84,102)(5:11)(6:163)(10:72)(16:89,102)(21:108)(22:78)(23:52)(27:81,91)(29:46)(30:53)
7,Jahiliyah.,
,"(3:154)(5:50)(6:28,136,140)(33:33)(48:26)",
PASAL II : TAUHID,,
I.,Tauhidullah,
1,Wujudullah.,
"(2:28-29,164)(3:18,190,191)(6:73,80)(7:185)(10:6)(11:7)(13:2-4)(16:48,81)(17:12)(20:54,128)(21:33)(22:18)
1:11,25,31)(36:33,44)(39:38)(40:13)(41:37-40,53)(42:29,32)(43:9,81)(45:3-5)(50:6-11)(64:1-4)(67:3,19,30)(7
2,Pengesaan secara mutlak & meniadakan sekutu.,
,,"(2:255)(3:2,26)(6:18,56,161,163,164,165)(10:32,104-105)(16:51)(20:28)(27:26)(30:30)(37:4)(43:82,84)(64:
3,Wahdaniyatullah (keesaan Allah).,
,,"(2:21,22,28-29,107,115,117,133,163,
165,255)(3:5-6,18,27,62,83,109,129,189)(4:1,87,126,131-132)(5:17,72-77,120)(6:1-2,12-13,17-24,46-47,59-61,
8,36,55-56,66-70,101)(11:7)(13:12-17)(14:19-20,32-34)(15:16-27)"
"(16:2-23,36,48-49,51-52,65-73,78-81)(17:12,40,42-44,111)(19:35,88-91)(21:19-33)(22:31,34,61-66,71)(23:17-
7-9)  (27:25-26,59-65,86,88,93)(28:62-75)(29:19)(30:8-11,40,48-50,54)(31:10-11,25-26,29-31)(32:6-9,",,
"27)(35:3,9,11-13,27-28,41)(36:12,71-73,",,
"77-83)(37:4-6,149-159)(38:65-66)(39:4-6,",,
```

Figure 3. Al Qur'an text index format from Al Fatih Manuscript publisher

Furthermore, this text file is processed to be a dataset with the desired format. Forming this dataset is done manually and generated with a python application. The desired dataset format as shown in table 1.

Tabel 1 : Al Qur'an index dataset format

Chapter	Verse	Chapter	Verse	Chapter	Verse
1	[109:2]	3	[97:3]	8	[98:4]
1	[109:3]	3	[97:4]	8	[7:34]
1	[109:4]	3	[97:5]	8	[10:49]
1	[109:5]	3	[25:30]	8	[15:5]
1	[109:6]	3	[43:88]	8	[16:61]
1	[2:270]	3	[43:89]	8	[17:58]
1	[3:35]	3	[5:44]	8	[35:40]
1	[19:26]	3	[5:45]	8	[35:45]
1	[22:29]	3	[5:47]	8	[36:43]
1	[76:7]	3	[5:50]	8	[71:4]
2	[2:177]	4	[3:7]	9	[2:104]
2	[2:186]	4	[3:18]	9	[4:86]
2	[2:256]	4	[4:83]	9	[17:53]
2	[2:285]	4	[11:24]	9	[19:42]
2	[3:84]	4	[13:16]	9	[19:43]

2	[3:110]	4	[29:43]	9	[19:44]
2	[3:179]	4	[35:19]	9	[19:45]
2	[3:193]	4	[35:28]	9	[19:46]
2	[4:135]	4	[39:9]	9	[19:47]
2	[4:162]	4	[58:11]	9	[19:48]

The Qur'an index dataset in the desired format. Then it is processed to eliminate duplicate verses in the same chapter. This duplication occurs because in each chapter the Qur'an index has several sub-indexes that most likely a verse will be a member of more than one sub-chapter member in the same chapter. In the following table 2 is displayed the number of verses in the Qur'an index both duplicate and non-duplicate:

Table 2. Number of verses in the Qur'anic index

Number	Chapter	Sum of Verses (Duplicate)	Sum of Verses (Not Duplicate)	Sum of Duplicate
1	Arkanul Islam	5861	3413	2448
2	Iman	4101	2346	1755
3	Al Quran	622	502	120
4	Cabang-cabang Ilmu	582	474	108
5	Amal	888	721	167
6	Dakwah Kepada Allah	191	184	7
7	Jihad	698	360	338
8	Manusia dan hubungan kemasyarakatan	1016	661	355
9	Akhlak	973	771	202
10	Peraturan tentang Harta	374	300	74
11	Hukum	254	207	47
12	Negara dan Kemasyarakatan	38	37	1
13	Pertanian dan Perdagangan	27	26	1
14	Sejarah dan Kisah-kisah	782	680	102
15	Agama-agama	458	309	149

3.5 Experiment Description

Classification is carried out on a group of documents so that they are easily understood and studied done in various ways. In this study, the grouping of the translation of the Qur'an is done by several methods to find the best performance. In the past years research, the grouping of the translation of the Qur'an in Indonesian is based on the degree of similarity between the verses [6]. This grouping uses the Cosine Similarity algorithm. The results of the grouping of translations of the verses of the Qur'an are then tested for compatibility with the Qur'an index. Al Qur'an Index is a way of presenting information about the existence of a particular theme or verses, and this is intended to facilitate understanding the Qur'an [19]. This research gives the result that with the similarity between verses by 20%, the similarity level of this similar group is 46.42%. This percentage of match level decreases if the similarity level is raised. This is due to the increasing level of similarity between the translations of the verses of the Qur'an, the number of similar verses decreases.

In this study, the group similarity between verses produced in the preliminary research will be further processed with other methods to see the performance results. The following are the steps of the research:

```
for similar in range(15):
    nmridx ← similar + 1, Ldo ← data['tjclean']
    for i in range(len(Ldo)):
        if (ida[i] == nmridx):
            Ldoc.append(Ldo[i]), Lida.append(id2[i])
TfLdoc ← tfidf.fit_transform(Ldoc)
    clusterer ← KMeans(n_clusters=3, max_iter=300, tol=0.0001)
    cluster_labels = clusterer.fit_predict(TfLdoc)
```

The K-Means clustering parameter in the above program code is taken from the scikit learn library. The contents of the variable max_iter = 300 are maximum number of iterations of the k-means algorithm for a single run. The contents of the parameter toll = 0.0001 are relative tolerance with regards to Frobenius norm of the difference in the cluster centers of two consecutive iterations to declare convergence [12].

3.6 Traditional TF-IDF and Clustering process

The traditional TF-IDF used in the first model gave quite good results, although it ignored many details when processing documents such as the frequency with which a term appeared and the frequency distribution. In the program fragment, it is shown that this process will be repeated 15 times, according to the number of similar groups that are candidates for chapters in the Qur'an index. Each group of similarities will produce a set of weights presented in the form of a sparse matrix (sparse matrix).

The results of the traditional TF-IDF clustering that the number of verses produced in each cluster for each group of similarities tends to be uneven. There are clusters where there are very many verses and there are also very few.

3.7 TF-IDF Normalization 1 and Clustering process

In normalization 1, the TF-IDF algorithm used is given additional parameters to eliminate words that appear in more than 80% of documents (max_df parameter in TF-IDF) and less than 20% documents (min_df parameter in TF-IDF). This word removal is done with the assumption that the word is a word that is not important and has no meaning in the document that is processed [12]. This normalization TF-IDF calling program command 1 will be given by replacing the traditional TF-IDF command.

The results of normalization 1 TF-IDF clustering show the number of verses produced in each cluster for each group of similarities tends to be more balanced and even when compared to traditional TF-IDF calculations. Although there are still clusters where the number of verses is very large and there are also very few, it is relatively more balanced.

3.8 TF-IDF Normalization 2 and Clustering process

TF-IDF Normalization 2 still uses the parameters max_df and min_df as normalization 1. It's just that the values of the two variables are different. In normalization 2, the value of max_df = 0.5 and value of min_df = 0.1. The value of the two parameters is adopted from Singhal's research, A., entitled Pivoted document length normalization [20]. This normalization TF-IDF calling program command 2 will be given by replacing the traditional TF-IDF command contained.

The results of TF-IDF clustering normalization 2 show that the number of verses produced in each cluster for each group of similarities tends to be more balanced and evenly distributed when compared to traditional TF-IDF calculations. When

compared with TF-IDF Normalization 1, then TF-IDF Normalization 1 gives better results.

3.9 Testing of Clustering Results on Chapter Al Qur'an

At this stage, the results of clustering that have been carried out in the previous stage will be tested on each chapter in the Qur'anic index. Each cluster in each group of similarity generated from each TF-IDF model will be tested against each chapter in the Al Qur'an index. The test results show the percentage of the number of verses in each TF-IDF model. The percentage results on the traditional TF-IDF are shown in the following table 3.

Table 3. Traditional TF-IDF Performance

Chap.	Similar	Cluster Number	Sum	Sum of Verse in the index	Percentage of Verses in the index
1	14	0	991	3413	29,04
2	14	0	644	2346	27,45
3	3	2	115	502	22,91
4	5	2	88	474	18,57
5	14	0	258	721	35,78
6	8	0	74	184	40,22
7	3	2	214	360	59,44
8	3	2	200	661	30,26
9	14	0	309	771	40,08
10	3	2	111	300	37,00
11	3	2	84	207	40,58
12	3	2	23	37	62,16
13	5	2	12	26	46,15
14	14	0	139	680	20,44
15	3	2	98	309	31,72
AVERAGE					36,12
STANDARD DEVIATION					12,77

Cluster test results formed from the performance of the traditional TF-IDF algorithm as presented in table 3 show that the largest percentage of the number of clusters contained in the Al Qur'an index chapter is 62.16% contained in chapter 12 on State and Society. The smallest percentage is produced in chapter 4 on branches of science. Of the 15 groups of similarity of verses, only 4 groups of similar gave the best results, with the highest number of verses in the Qur'an index chapter, namely groups of similar 3, 5, 8, 14. The average percentage was 36.12% with standard deviations by 12.77%. The next process is clustering testing of TF-IDF normalization 1 as shown in Table 4.

Table 4: Normalization TF-IDF Performance 1

Chap.	Similar	Cluster Number	Sum	Sum of Verse in the index	Percentage of Verses in the index
1	6	0	647	3413	18,96
2	6	0	452	2346	19,27
3	14	0	97	502	19,32
4	3	2	58	474	12,24
5	6	0	203	721	28,16
6	6	0	54	184	29,35
7	7	2	150	360	41,67
8	6	0	115	661	17,40
9	6	0	225	771	29,18
10	3	1	72	300	24,00
11	6	0	51	207	24,64
12	6	0	18	37	48,65
13	9	0	9	26	34,62
14	2	1	86	680	12,65
15	6	0	76	309	24,60
AVERAGE					25,65
STANDARD DEVIATION					10,16

The results of cluster testing formed from the normalization 1 TF-IDF algorithm performance as presented in table 4 show that the largest percentage of the number of clusters contained in the Al Qur'an index chapter is 48.65% contained in chapter 12 on Country and Society. The smallest percentage is produced in chapter 4 on branches of science. These results are identical to the results of traditional TF-IDF processing. From 15 groups of similarity verses, there are 6 similar groups that give the best results, with the highest number of verses in the Qur'an index chapter, namely similar groups 2, 3, 6, 7, 9, 14. The average percentage is 25.65 % with a standard deviation of 10.16%. The next process is clustering testing of TF-IDF normalization 2 as shown in Table 5.

Table 5: Normalization TF-IDF Performance 2

Chap.	Similar	Cluster Number	Sum	Sum of Verse in the index	Percentage of Verses in the index
1	7	0	841	3413	24,64
2	7	0	517	2346	22,04
3	7	0	146	502	29,08
4	3	1	73	474	15,40
5	5	1	206	721	28,57
6	5	1	60	184	32,61
7	7	0	147	360	40,83
8	1	1	156	661	23,60
9	7	0	229	771	29,70

10	5	1	85	300	28,33
11	5	1	59	207	28,50
12	5	1	18	37	48,65
13	3	1	7	26	26,92
14	7	0	119	680	17,50
15	1	1	80	309	25,89
AVERAGE					28,15
STANDARD DEVIATION					8,27

Cluster test results formed from the performance of the normalized TF-IDF algorithm 2 as presented in table 5 show that the largest percentage of the number of clusters contained in the Al Qur'an index chapter is 48.65% contained in chapter 12 on Country and Society. The smallest percentage is produced in chapter 4 on branches of science. These results are also identical to the results of the traditional TF-IDF processing and normalized TF-IDF 1. Of the 15 groups of verse similarity, there are 4 groups of similar that give the best results, with the highest number of verses in the Qur'an index chapter, namely the similar 1 group, 3, 5, 7. The average percentage is 28.15% with a standard deviation of 8.27%.

4. Summary and Conclusions

Based on research that has been done, it can be concluded that the evaluation of the TF-IDF weighting scheme in clustering the translation of the Qur'an with the K- Means Clustering algorithm was successfully carried out. The results of the grouping of verses based on the level of similarity as done in this previous study were then processed again by using the three TF-IDF weighting scheme models for further clustering using the K-Means Clustering algorithm. The results of clustering over 15 groups of similarity verses show that the number of verses in each cluster and similar groups is relatively uneven. There is a cluster that tends to have a relatively large number of verses and there are clusters that tend to have a small number of verses. However, the results of clustering from TF-IDF normalization 1 are relatively more evenly distributed in the number of verses in the resulting cluster.

The results of this clustering are then matched with the Al Qur'an index to find which group of verses best fits the theme in the Al Qur'an index. From the three TF-IDF weighting scheme models the highest percentage results obtained from the traditional TF-IDF weighting scheme were 62.16% with an average percentage of 36.12% and a standard deviation of 12.77%. The time needed to process this model is 98.09 seconds. The smallest results are shown in normalization 1 TF-IDF weighting scheme which is 48.65% with an average percentage of 25.65% and a standard deviation of 10.16%. The time needed to process this model is 20.69 seconds. The smallest standard deviation resulted in normalization 2 TF-IDF weighting is 8.27% with an average percentage of 28.15% and the largest percentage of weighting is 48.65%, the same as normalization TF-IDF weighting 1. The time needed for the process this model for 14.30 seconds. A more complete comparison of these results is presented in table 6 and figure 3 below:

Table 6: Comparison of TF-IDF weighting

Chapter	TF-IDF Weighting		
	Traditional	Normal 1	Normal 2
1	29,04	18,96	24,64
2	27,45	19,27	22,04
3	22,91	19,32	29,08
4	18,57	12,24	15,40
5	35,78	28,16	28,57
6	40,22	29,35	32,61
7	59,44	41,67	40,83
8	30,26	17,40	23,60
9	40,08	29,18	29,70
10	37,00	24,00	28,33
11	40,58	24,64	28,50
12	62,16	48,65	48,65
13	46,15	34,62	26,92
14	20,44	12,65	17,50
15	31,72	24,60	25,89
AVG	**36,12**	**25,65**	**28,15**
STDEV	**12,77**	**10,16**	**8,27**

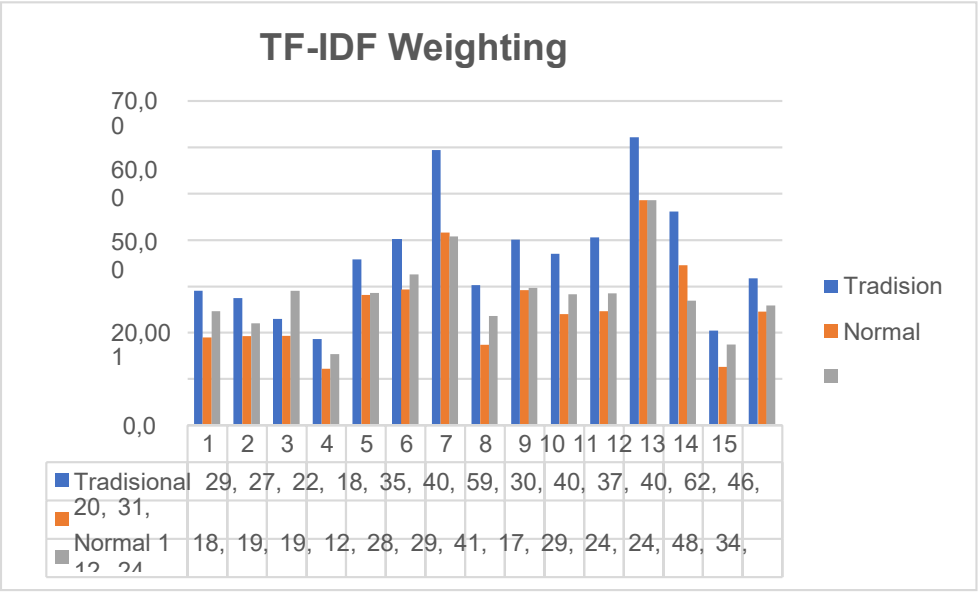

Figure 3 : Comparison of TF-IDF weighting

References

1. T. Khotimah, "Pengelompokan Surat Dalam Al Qur ' an Menggunakan Algoritma K-Means," *J. SIMETRIS*, vol. 5, no. 1, pp. 83–88, 2014.
2. C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432. Springer US, 2013.
3. G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A Comparison of Term Weighting Schemes for Text Classification and Sentiment Analysis with a Supervised Variant of tf.idf," *Commun. Comput. Inf. Sci.*, vol. 584, no. February, p. v, 2016, doi: 10.1007/978-3-319-30162-4.
4. S. Albitar and B. Espinasse, "An Effective TF / IDF-based Text-to-Text Semantic Similarity Measure for Text Classification," no. January 2015, 2014, doi: 10.1007/978-3-319-11749-2.
5. C. Hiram, "Simple TF·IDF Is Not the Best You Can Get for Regionalism Classification," 2014, doi: https://doi.org/10.1007/978-3-642-54906-9_8.
6. M. D. R.Wahyudi, "Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al- Qur ' an Berdasarkan Keterkaitan Topik," *Semesta Tek.*, vol. 22, no. 1, pp. 41–50, 2019, doi: 10.18196/st.221235.
7. R. K. Roul, "Modified TF-IDF Term Weighting Strategies for Text Categorization," no. October, 2018, doi: 10.1109/INDICON.2017.8487593.
8. B. K. Hananto, A. Pinandito, and A. P. Kharisma, "Penerapan Maximum TF-IDF Normalization Terhadap Metode KNN Untuk Klasifikasi Dataset Multiclass Panichella Pada Review Aplikasi Mobile," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 12, pp. 6812–6823, 2018.
9. M. Zed, *Metode Penelitian Kepustakaan*. Yayasan Pustaka Obor Indonesia, 2004.
10. A. Rajaraman and J. D. Ullman, *Mining of massive datasets*, vol. 9781107015. Cambridge: Cambridge University Press, 2011.
11. B. Joeran, C. Breitinger, B. Gipp, and S. Langer, "Research-paper recommender systems: a literature survey," *Int. J. Digit. Libr.*, vol. 17, no. 4, pp. 305–338, 2015.
12. D. Pedregosa, F., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825--2830, 2011, [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html.
13. A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '96*, 1996, pp. 21–29, doi: 10.1145/243199.243206.
14. S. Andayani, "Formation of clusters in Knowledge Discovery in Databases by Algorithm K-Means," *SEMNAS Mat. dan Pendidik. Mat. 2007*, 2007.
15. Y. Agusta, "K-Means – Penerapan, Permasalahan dan Metode Terkait," *J. Sist. dan Inform.*, vol. 3, 2007.
16. B. Santosa, *Data mining teknik pemanfaatan data untuk keperluan bisnis*. Yogyakarta: Graha Ilmu, 2007.
17. F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
18. *Al Fatih:Mushaf Al Qur'an Tafsir Per Kata Kode Arab*. Jakarta: PT Insan Media Pustaka, 2013.
19. Hani M. Atiyyah, *Quranic Text: Toward a Retrieval System*. 1996.
20. A. Singhal, G. Salton, M. Mitra, and C. Buckley, "Pivot Document length normalization," *Inf. Process. Manag.*, vol. 32, no. 5, pp. 619–633, 1996, doi: 10.1016/0306-4573(96)00008-8.