Qur'an Search System for Handling Cross Verse Based on Phonetic Similarity

Intan Khairunnisa Fitriani[1]*, Moch Arif Bijaksana[2], Kemas Muslim Lhaksmana[3]
School of Informatics, Telkom University[1], [2], [3]
Bandung, Indonesia
intankhair@student.telkomuniveristy.ac.id[1], arifbijaksana@telkomuniversity.ac.id[2],
kemaslhaksmana@telkomuniversity.ac.id[3]

***Abstract*— The number of verses in the Qur'an that is not small will be difficult and time consuming if done manually. Building a search system in the Qur'anic verse using the Indonesian Arabic-Latin equivalent will be very helpful for the Muslim community in Indonesia, especially for those who are not familiar with Arabic writing. In this study, a verse search system will be built on the Al-Qur'an based on phonetic similarity, more details about the handling of the verses in the Al-Qur'an. The system was built using the Jaro-Winkler algorithm to calculate the value of similarity and using the N-Grams algorithm for ranking documents. The same study has been done before with the name Lafzi +, with MAP 90% and 93% recall. In previous studies cases such as *nun wiqoyah* at the end of the verse could not be handled, so the system could not handle the search for the entire Qur'an. So to complete the previous research, in this study added rules other than pre-existing rules so that they can handle *nun wiqoyah* at the end of the verse. Rules are added to the phonetic coding process, for example in Surah An-Nisa verses 37 and 38, namely ﴿٣٦﴾ الَّذِينَ يَبْخَلُون فَخُورًا the verse contains nun wiqoyah at the end of the verse, in the Lafzi system + the verse produces a phonetic encoding "fahuranilazinayabhalun", by adding the rules of nun wiqoyah in this system result in the phonetic coding of the verse pieces becoming "pahuranxalajinayabhalun", to match the results of the phonetic coding stored in the database. By applying the Jaro-Winkler method to calculate the value of similarity and N-Grams for ranking documents and adding *nun wiqoyah* rules, this system generates 94% MAP and 92% recall. The results of this study indicate an increase in MAP, this shows that this system can improve the accuracy of systems that have been built before.**

Keywords*— *al-quran, cross-verses, phonetic, jaro-winkler, n-gram.

I. INTRODUCTION

Phonetic search is a method of finding information in a file on where the algorithm is used to find combinations of characters that sound similar to the specified combination [1]. Doing a search based on phonetic similarity will produce better search results because it will tolerate errors in writing, as long as the sound is still the same as the word in query.

The search system for the Qur'anic verses based on phonetic similarity will certainly be very helpful for Muslims, especially for Indonesian Muslims. The contents of the Qur'an are not small, consisting of 30 juz, 114 surahs, and 6666 verse [2], searching manually will be very time-consuming. At present, there are not a few systems for searching Al-Qur'an verses, one of which is Tanzil in 2007, only that Tanzil is still searching through Arabic script queries so that people who are not familiar with Arabic script will find it difficult. Besides the phonetic-based search system using Latin letters also already exist, one of which is Islamicity in 2001 and Lafzi in 2012. For the phonetic search system in Islamicity using international Arab-Latin matching that is different from the matching of Arab-Latin Indonesia. As for the search system, Lafzi has used Indonesian Arab-Latin matching.

A search on the Qur'an is generally carried out when someone remembers a particular verse or is listening to murattal and wants to find a fragment of the verse. The query entered is in the form of a memorized verse or one that comes to mind without knowing whether the query is the correct paragraph or the query crosses two verses. So we need a search system that can handle queries for cross verses. In the search system, Lafzi himself cannot handle cross-verse searches.

In a study conducted by Eki Rifaldi on a cross-verse search system in the Qur'an based on phonetic similarity developed from Lafzi and named Lafzi+, the system produces MAP 0.9 and Recall 0.93 [3]. In this research, the application of the Jaro-Winkler algorithm is still not implemented as a whole, and the system cannot handle cross-verse searches for the entire Qur'an. Cases like *nun wiqoyah* at the end of the verse still cannot be handled by the Lafzi+ system. To complement the previous research, in this journal beside the existing equivalent rules, the rules for handling nun wiqoyah will also be added at the end of the verse. Rules are added to the phonetic coding process, for example in Surah An-Nisa verses 37 and 38, namely ﴿٣٦﴾ فَخُورًا اَلَّذِينَ يَبْخَلُون the verse contains nun wiqoyah at the end of the verse, in the Lafzi system + the verse produces a phonetic encoding "fahuranilazinayabhalun", by adding the rules of nun wiqoyah in this system result in the phonetic coding of the verse pieces becoming "pahuranxalajinayabhalun", to match the results of the phonetic coding stored in the database.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.986, Copyright ©2021
Submitted : September 10, 2020; Revised : December 1, 2020; Accepted : January 8, 2021; Published : February 2, 2021

In the previous system, because there were no rules that dealt with nun wiqoyah, the search for a verse piece containing nun wiqoyah at the end of the verse was not found, whereas by adding the nun wiqoyah rule to the existing rules, the search for cross-verses that contained nun wiqoyah at the end of the verse can be found.

By adding the nun wiqoyah rule to the phonetic coding process and implementing the Jaro-Winkler algorithm in calculating the similarity value and using the number of N-grams in the ranking so it is expected that the MAP and Recall results will be better, and for queries that do not cross verse it is expected that it will not reduce the MAP value and Recall from the previous system.

II. LITERATURE REVIEW

A. *Related Work*

Lafzi is a system of searching verses of the Qur'an based on phonetic similarity using the Indonesian Arabic-Latin equivalent. Lafzi uses the trigram method in searching for words that are similar to user queries and uses 4 ranking methods, namely searching with vocals using number ranking, searching without vowels using number ranking, searching with vowels using position ranking, and searching without vowels using position ranking. With the best system performance, namely searching by vocal using the ranking of the number[4].

Then developed a system of Lafzi which is named Lafzi + with cross verse search feature by producing 90% MAP and 93% Recall[3]. This system uses the Jaro-Winkler method to calculate the value of similarity and ranking of documents using the number of n-grams.

Table 1 is a comparison table for the calculation of MAP and Recall from Lafzi and Lafzi +. In the Lafzi system, it produces MAP and Recall N / A because in the Lafzi system it cannot handle cross-verse searching. Whereas for the Lafzi + system it produces MAP 0.9 and Recall 0.93, with MAP and Recall for queries that do not cross verses produce the same MAP and Recall as Lafzi.

TABLE I. COMPARISON OF MAP AND RECALL FROM LAFZI AND LAFZI+

Sistem	Lintas ayat		Tidak lintas ayat	
	MAP	**Recall**	**MAP**	**Recal**
Lafzi	N/A	N/A	0.9	0.97
Lafzi+	0.9	0.93	0.9	0.97

B. *Rules For Reading Laa Washal At The End Of Verse*

The sign of *laa washal* means that it cannot stop. If the sign *laa washal* is in the middle of the verse, then it is not allowed to stop. But if there is at the end of the verse, then it is allowed to stop.

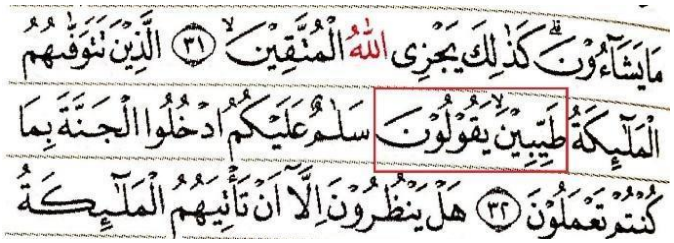

Figure 1. Surah An-Nahl verse 32

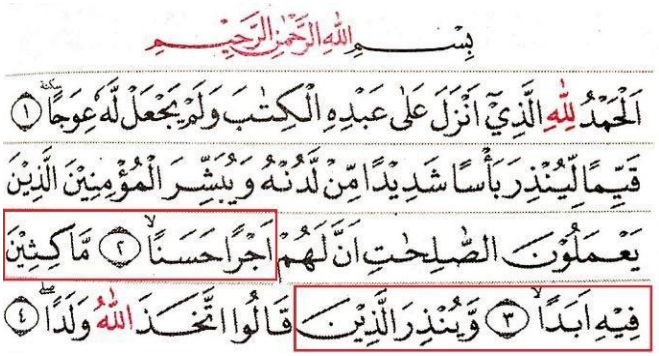

Figure 2. Surah Al-Kahfi verse 1-3

Figure 1, Surah An-Nahl verse 32, is an example of *laa washal* in the middle of the verse. Figure 2 is Surah Al-Kahfi verses 1 to 3 which is an example of *laa washal* at the end of verse.

C. *The Rules for Reading Nun Wiqoyah*

Nun wiqoyah or *Nun Washal* is a small nun placed under the letter *Alif Washal*. *Nun wiqoyah* is the sound of *kasrah nun* ("ni") that occurs when the letters in *tanwin* are read together with *alif washal*. The way to read *nun wiqoyah* is by removing the *tanwin* sound (unread *tanwin* letters) and an additional sound appears "ni" afterward.

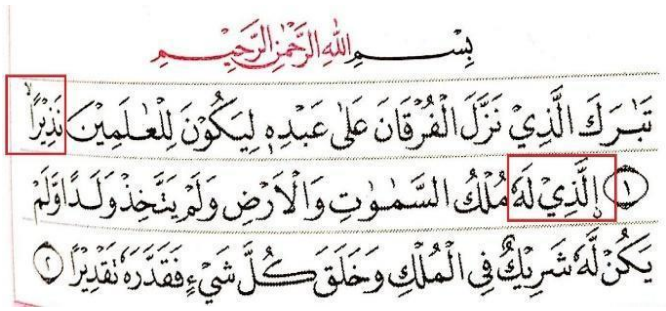

Figure 3. Surah Al-Furqan verse 1 & 2

Figure 3 is Surah Al-Furqan verses 1 and 2. In the section marked read "na-dzii-ra-ni-la-dzii-lahu". Another example in Figure 4 is surah An-Nisa verses 36 and 37. in the section marked read with "fa-khu-ra-ni-la-dzi-na".

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.986, Copyright ©2021
Submitted : September 10, 2020; Revised : December 1, 2020; Accepted : January 8, 2021; Published : February 2, 2021

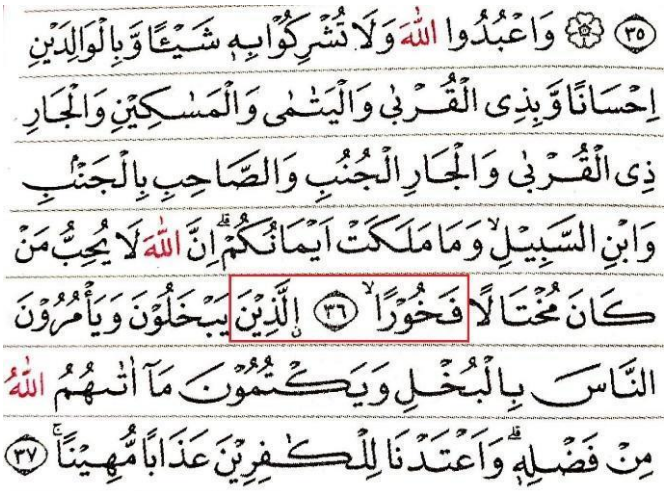

Figure 4. Surah An-Nisa verse 37 & 37

D. N-Grams

N-Gram is a number of N characters in a string. Usually, blank characters are added at the beginning and end of words to help match the beginning of a word and the end of a word[5]. To represent blanks, an underscore ("_") is used. Thus, the word "STRING" will consist of N-grams:

1) Bi-grams : S_ , ST, TR, RI, IN, NG, G_
2) Tri-grams : _ST, STR, TRI, RIN, ING, NG_ , G_ _
3) Quad-grams : _STR, STRI, TRIN, RING, ING_, NG_ _, G_ _ _

In general, strings of length n, filled with blanks, will have n + 1 bi-grams, n + 1 tri-grams, n + 1 quad-grams, and so on. The advantage of matching N-gram-based strings is that writing errors will only affect some n-grams because the string is divided into in-character string strings.

E. Longest Common Subsequence (LCS)

Longest common subsequence is a special case of edit distance. LCS can be solved by dynamic programming and by filling in the table m x n [6]. LCS works by matching each character in the input string with every character in the string in the database. The way it works to solve LCS is to determine the longest common subsequence for all possible prefix combinations of input strings.

String C is the common subsequence of strings A and B if C is the subsequence of A and subsequence of B. String C is the longest common subsequence (LCS) of strings A and B if C is the common subsequence of A and B of the maximum length, in words otherwise there is no common subsequence from A and B which is longer [7].

F. Jaro-Winkler Distance

Jaro-Winkler distance is measuring the similarity between two strings by using the similarity prefix [8]. The higher the value of the calculation for the two strings shows the more similar the string.

The score is normalized so that a value of 0 indicates no similarity and a value of 1 indicates an appropriate match. Jaro Winkler's algorithm has three parts, namely[9]:

1) Calculate word length
2) Determine the same number of characters in two words
3) Find the number of transpositions

To determine the similarity (dj) value between two words s1 and s2, Jaro Winkler uses the equation:

$$dj = \frac{1}{3}\left(\frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{t}\right) \quad (1)$$

In equation (1) where:

m = number of equal characters
| s1 | = length of string 1
| s2 | = length of string 2
t = number of transpositions

III. RESEARCH METHODOLOGY

A. System Overview

The system flow is explained in Figure 5. In the system flow it is explained that there are two processes, namely processing online and offline. Offline processing is only done once to form an index [4]. For online processing that is phonetic coding, N-gram tokenization, N-gram matching, document ranking, then issuing search results.

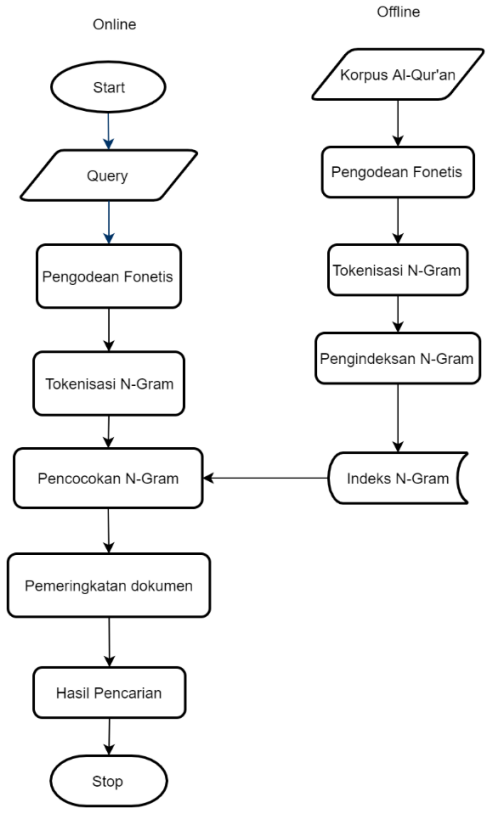

Figure 5. System Overview

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.986, Copyright ©2021
Submitted : September 10, 2020; Revised : December 1, 2020; Accepted : January 8, 2021; Published : February 2, 2021

B. Phonetic Coding

Phonetic coding is a preprocessing stage. Phonetic coding is made taking into account the reading rules in the Qur'an (tajwid), phonetic coding rules follow the rules made before Lafzi. The rules phonetic coding for queries is as follows:

1) Vocal Substitution

By replacing vowel O into vowel A, replace vowel E into vowel I.

2) Consonants

The same double consonant letters are put together.

3) Vocal Integration

The same double vowels are put together.

4) Diphthong Substitution

Dual vowels (diphthongs) are substituted by changing AI to AY and AU to AW.

5) Marking Hamzah Letters

Vowels that meet the following rules are marked as *hamzah*:

- The letters A, I, and U which are at the beginning of a word or after space
- Arrangement of vocal IA or IU, then the letter after I is marked *hamzah*
- Arrangement of vocal UA or UI, letters after U are marked *hamzah*

6) Subtitution of Ikhfa Reading

Literature readings written in NG sounds, the letter G must be removed.

7) Subtitution of Iqlab Reading

Iqlab reading was substituted by changing NB to MB.

8) Substitution of Idgham Reading

Remove the letter N if it meets the letters idgham (Y, N, M, W, L, R).

9) Nun Wiqoyah Rules

Any queries ending in "an", "in", or "un" and followed by *hamzah* will be changed by removing the suffix "an", "in", or "un" and adding "ni" before *hamzah*.

10) Matching to Phonetic Code

Matching to the phonetic code needs to consider the Arabic letters represented by 2 consonant letters. Rules for matching Latin characters to phonetic codes are listed in Table 2.

TABLE II. RULES FOR MATCHING LATIN CHARACTERS TO PHONETIC CODES

Script	Equivalent
2 consonants	
SH, TS, SY	S
KH, CH	H
ZH, DZ	Z
DH	D
TH	T
GH	G
NG(’ain)	X
1 consonant	
F, V, P	F
Q, K	K
J, Z	Z
‘, ‘(apostrophe)	X

C. N-Grams Tokenization

In this study, the N-Grams algorithm applied is tri-gram. Queries entered by users and Al-Qur'an texts in the corpus will be documented to form trigrams. In this study, blanks are not needed at the beginning and end of a verse. The tokenization process uses an overlapping window with 3 characters length[10]. An example can be seen from Figure 6, which is an illustration of n-gram tokenization with n along 3. In the illustration, it can be seen that the clause of surah An-Nisa verses 156 and 157, the trigram of the paragraph fragment "AZIMAWAKAWLIHIM" namely "AZI", "ZIM", "IMA", "MAW", "AWA", "WAK", "AKA", "KAW", "AWL", "WLI", "LIH", "IHI", "HIM".

Figure 6. Illustration of trigram tokenization

D. Trigram Indexing

After the tokenization process, the next process is the formation of inverted indexes. Information stored in the index is the document identifier, the number of certain trigrams in the document, and the position where the first trigram appears in the document.

Table 3 is the Inverted index table of the verse fragments مِن قَبْلُ (٣) ٱلْإِنجِيلَ written in Latin as "INZILMINKABLU", and documented so that it becomes a trigram of "INZ", "NZI", "ZIL", "ILM", "LMI", " MIN ", " INK ", " NKA ", " KAB ", " ABL ", "BLU ". The trigrams will be mapped into a table, as in Table 3.

TABLE III. INVERTED INDEX

Id	Trigram	Document Identifier
816	INZ	{”131”:[86],”135”:[32], ...}
1256	NZI	{”11”:[24,42],”13”:[53], ...}
2157	ZIL	{”11”:[25,43],”68”:[230], ...}
777	ILM	{”7”:[33],”9”:[30], ...}
1040	LMI	{”127”:[125],”134”:[64], ...}
1120	MIN	{”2”:[24],”10”:[12], ...}
810	INK	{”11”:[48],”27”:[143], ...}
1220	NKA	{”11”:[49],”27”:[144], ...}
905	KAB	{”11”:[50],”28”:[50], ...}
8	ABL	{”11”:[51],”28”:[51], ...}
319	BLU	{”32”:[134],”96”:[62], ...}

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.986, Copyright ©2021
Submitted : September 10, 2020; Revised : December 1, 2020; Accepted : January 8, 2021; Published : February 2, 2021

E. Trigram Matching

At this stage, matching each trigram of the query with the trigram in the corpus is performed. Then count the same number of trigrams between the trigram queries and the trigrams of the corpus and then rank the documents. At this stage, matching each trigram of the query with the trigram in the corpus is performed. Then count the same number of trigrams between the trigram queries and the trigrams of the corpus and then rank the documents.

F. Calculate Similarity Value And Document Ranking

To calculate the similarity value, the Jaro Winkler method is used. If the input query is the same as the document, then it will give a value of 1. The resulting value ranges from 0 to 1, with values 1 for both strings that are the same, and values less than 1 for strings that are not the same.

In this system the ranking of documents uses n-grams. Every n-gram will be counted between the input query and the Al-Qur'an corpus, then ranking by ordering from the same n-gram to the fewest. The maximum value is the number of n-grams input query and the minimum value is 1 because the search only looks for documents that have at least 1 n-gram which is the same as the input query.

Verses that have the same similarity value and the same n-gram will be excluded as the best search results.

IV. RESULT AND DISCUSSION

Evaluation is done by calculating the MAP and Recall values. MAP (Mean Average Precision) is the average of Average Precision. Average Precision (AP) is the average maximum precision at different recall values. Precision to measure the correct presentation of guesses or how accurate the predictions are. The following is a formula for calculating MAP:

$$MAP = \frac{1}{|Q|} \cdot \sum_{i=1}^{Q} \blacksquare AP(Q_i) \quad (2)$$

$$AP = \sum_{i=1}^{N} \blacksquare \left(\frac{TP_{-i}}{TP_{rank-i}} \right) \quad (3)$$

Meanwhile, Recall is how many relevant documents we take. The very high recall will cause low precision. Recall calculations are carried out using the following formula:

$$Recall = \frac{|(InformationRetrieval) \cap (OutputSystem)|}{|InformationRetrieval|} \quad (4)$$

A. Result

By adding 20 new data that is a cross-verse and contains nun wiqoyah at the end of a paragraph, the tests carried out are testing on a regular query, a cross-verse query, and a cross-verse query containing nun wiqoyah at the end of the verse.

TABLE IV. EXAMPLE OF QUERIES AND RESULT OF PHONETIC CODING

	Regular query	**Cross-verse query**	**Cross-verse query containing nun wiqoyah**
User input	millahirrahman	nirahim'alhamd	fahuranilazinay abhalun
Result of phonetic coding process	milahirahman	nirahimxalhamd	pahuranilajiyab halun
Result of the phonetic coding after adding the nun wiqoyah rule	milahirahman	nirahimxalhamd	pahuranxalajina yabhalun
presence in system	[1:1],[27:30],[1; 73],...	[1:1,2] [40:7]	[]

The table 4 contains rows of user input, the results of the phonetic coding, result of phonetic coding after adding the nun wiqoyah rule and the presence in the system. In the table, it can be seen that before adding the nun wiqoyah rules the phonetic coding results for the query "fahuranilaziyabhalun", namely "pahuranilajinayabhalun", meanwhile after adding the nun wiqoyah rules the results of phonetic coding become "pahuranxalazinayabhalun" so that they match the results of the phonetic coding stored in the database.

TABLE V. RESULTS OF SYSTEM TESTING FOR CROSS-VERSES SEARCH

		Lafzi	**Lafzi+**	**Lafzi++**
Reguler query	MAP	0.9	0.9	0.9
	Recall	0.97	0.97	0.97
Cross-verse query	MAP	N/A	0.9	0.94
	Recall	N/A	0.93	0.92
Cross-verse query containng nun wiqoyah	MAP	N/A	0.26	0.71
	Recall	N/A	0.25	0.8

B. Analysis of Testing Results

As seen in table 5, testing each query on each system gets different MAP and Recall results. For cross-verse queries, the Lafzi System still cannot handle cross-verse queries so that the MAP and Recall for verse queries are N / A, the Lafzi + system

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.986, Copyright ©2021
Submitted : September 10, 2020; Revised : December 1, 2020; Accepted : January 8, 2021; Published : February 2, 2021

gets 90% MAP and 93% Recall, the Lafzi ++ system gets 94% MAP and Recall 92%. For the Lafzi + system, queries containing cross-verse generate 26% MAP and 26% recall, while for the Lafzi ++ system it produces 71% MAP and 80% Recall. The result of testing of cross-verse queries containing nun wiqoyah shows that the addition of the nun wiqoyah rule to the phonetic coding process can improve the search accuracy for verses containing nun wiqoyah.

V. CONCLUSION

Based on the results obtained in the system testing, this system is proven to be able to handle cross-verse searches that contain nun wiqoyah. The addition of the nun wiqoyah rule can handle cross-verse searches containing nun wiqoyah which could not be handled in the previous system. In the previous system for the query "fahuranillazinayabhalun" the result of the phonetic coding process was "pahuranilajinayabhalun", while in the Lafzi ++ system for the input query "fahuranilazinayabhalun" after going through the phonetic coding process the query was converted into "pahuranxalajinayabhalun stored in the phonetic database" so that it matched the results of the phonetic coding. , so that when the N-Grams matching process is done, the results are greater than the previous system. Greater N-Grams results make for better accuracy of search results.

REFERENCES

[1] D. N. Lapedes, *Dictionary of Scientific and Technical Terms*. New York: McGraw-Hill, 1974.

[2] A. Syarifuddin, *Mendidik anak: membaca, menulis dan mencintai Al-Quran*. Jakarta: Gema Insani, 2004.

[3] E. Rifaldi, M. A. Bijaksana, and K. M. Lhaksamana, "Sistem Pencarian Lintas Ayat Al-Qur'an Berdasarkan Kesamaan Fonetis," *Indones. J. Comput.*, vol. 4, no. 2, pp. 177–188, 2019.

[4] M. A. Istiadi, "Sistem Pencarian Ayat Al-Qur'an Berbasis Kemiripan Fonetis," Skripsi Program Sarjana, Institut Pertanian Bogor, Bogor, 2012.

[5] W. B. Cavnar, J. M. Trenkle, and A. A. Mi, "N-Gram-Based Text Categorization," *Proc. SDAIR-94, 3rd Annu. Symp. Doc. Anal. Inf. Retr.*, 1994, doi: 10.1.1.53.9367.

[6] L. Bergroth, H. Hakonen, and T. Raita, "A survey of longest common subsequence algorithms," 2000, doi: 10.1109/SPIRE.2000.878178.

[7] D. S. Hirschberg, "Algorithms for the longest common subsequence problem," *J. ACM*, vol. 24, no. 4, pp. 664–675, 1977.

[8] M. Rajabzadeh, S. Tabibian, A. Akbari, and B. Nasersharif, "Improved dynamic match phone lattice search using Viterbi scores and Jaro Winkler distance for keyword spotting system," in *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, 2012, pp. 423–427.

[9] F. Friendly, "PERBAIKAN METODE JARO--WINKLER DISTANCE UNTUK APPROXIMATE STRING SEARCH MENGGUNAKAN DATA TERINDEKS APLIKASI MULTI USER," *J. Teknovasi J. Tek. dan Inov.*, vol. 4, no. 2, pp. 69–78, 2018.

[10] A. Nwesri, "Effective retrieval techniques for Arabic text," 2008.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.986, Copyright ©2021
Submitted : September 10, 2020; Revised : December 1, 2020; Accepted : January 8, 2021; Published : February 2, 2021