Verse Search System for Sound Differences in the Qur'an Based on the Text of Phonetic Similarities

Agni Octavia[1]*, Moch. Arif Bijaksana[2], Kemas Muslim Lhaksmana[3]
Bachelor of Informatics Engineering, Faculty of Informatics[1], [2], [3]
Telkom University, Bandung, Indonesia
agnioctv@students.telkomuniversity.ac.id[1], arifbijaksana@telkomuniversiti.ac.id [2],
kemasmuslim@telkomuniversity.ac.id [3]

***Abstract*— Al-Qur'an has a lot of content, so the system of searching for verses of the Al-Qur'an is needed because if it is done manually it will be difficult. One of the search systems for the verses of the Al-Qur'an in accordance with Indonesia's pronunciation is Lafzi. The Lafzi system can search for verse fragments using keywords in Latin characters. Lafzi has been developed into Lafzi +, wherein the Lafzi + system can be used to search verses of the Al-Qur'an with different sounds on stop signs. However, the Lafzi+ can only overcome the difference in the sound of the stop sign and cannot be applied throughout Al-Qur'an. Based on these problems, the system needs to be developed to overcome the differences in sound in the middle of the verse and can be applied throughout the Al-Qur'an. The method used in the process of searching for the verse is the N-gram method. The N-gram used in this research is trigram. The process flow of this system is first normalized in the phonetic coding process after normalized then tokenization of trigrams and then trigrams are matched between the query and the corpus and entered into the ranking process to get an output candidate. In the making process, the LIS (*Longest Increasing Subsequence*) method is used to get an orderly and strict trigram sequence. The highest order score will be the top output. The results of this study obtained a recall value of 100% and MAP of 87%.**

***Keywords*— phonetic search, n-gram, string matching**

I. INTRODUCTION

Al-Qur'an is a holy book that is used as a way of life for Muslims. The Al-Qur'an consists of 30 chapters, 114 surahs, 6,236 verses, and 77,845 words [1]. With the contents of the Al-Qur'an very much, a search system for the verses of the Al-Qur'an is needed, because if you do a manual search it will be difficult.

Research on the system of searching verses in the Al-Qur'an has been done for a long time such as Tanzil, IslamiCity, and Lafzi. Tanzil [2] is one of the application systems that provides facilities for searching verses of the Al-Qur'an. However, to search for verses, users must use keywords that contain Arabic characters. This makes it difficult for users who cannot use keywords using Arabic characters, especially for Muslims in Indonesia. To overcome this problem, a search system needs to be developed based on phonetic similarity. The search system is based on phonetic similarity, namely by making Arabic alphabetical equivalents and pronunciation in Latin script. IslamiCity [3] is a system that can search for verses based on phonetic similarity. However, matching the Arabic-Latin script used by the system is Latin-International, where there is a slight difference from the Latin-Indonesian language. Lafzi [4], one of the applications that have developed a system of searching verses in the Al-Qur'an based on phonetic similarity using matching Indonesian-Indonesian characters.

The Lafzi system can search verses in the Al-Qur'an using keywords in Latin script that are in accordance with Indonesian pronunciation. However, the application has not been able to overcome the search for verses to change the sound at the stop sign optimally. Some words in the Al-Qur'an followed by *waqaf* or at the end of the verse usually occur in differences in reading it. For example in the word "*Wa aâanahu*" if *waqaf*, then the way to read it becomes "*Wa aanah*". Research developed by Naufal Rasyad [5] about searching for verses of the Al-Qur'an with different sounds on stop signs (Lafzi +) and the system produces a recall of 100% and MAP of 84% but in previous studies still not implemented throughout the Al- Qur 'an.

Therefore, in this final project research will complete previous research on the system of searching verses in the Al-Qur'an to distinguish sounds based on phonetic similarity. The difference between current and previous research is that the system in this study will be implemented throughout the Al-Qur'an and not only overcome the search for verses to change the sound at the stop sign but apply to sound changes in the middle of the verse so that in the previous system there are still verses that cannot be found. The search method used is a search with N-gram that is applied to the phonetic code. N-gram used in this research is a trigram. Nurhanifah [6] uses the trigram index method to inexact matching data on English from agricultural documents and gives good results. The

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i3.935, Copyright ©2020
Submitted : 19 Juli 2020, Revised : 3 Agustus 2020, Accepted : 25 Agustus 2020, Published : 30 Agustus 2020

advantage of using N-gram, in general, can be indexed by the tokenization of N-gram in the corpus to build data structures of inverted index data so that searching for certain terms can be done quickly [7].

System testing is done to determine the accuracy of how well the system can do the search. The accuracy used is MAP and recall. The following is the relationship between the search process method and the system testing evaluation used.

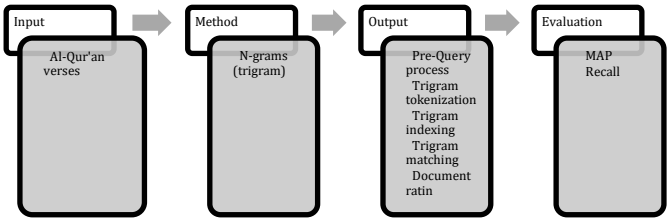

Fig. I. Process Flow

II. LITERATURE REVIEW

A. Research Summary

Lafzi is a system of searching verses in the Al-Qur'an based on phonetic similarities that are in accordance with Indonesian pronunciation, making it easier for Muslims in Indonesia to search for verses in the Al-Qur'an. The search method in Lafzi uses the trigram search applied to the phonetic code.

Lafzi + is a development system of Lafzi. The development carried out on Lafzi + can overcome the search for verses in the Al-Qur'an with different sounds on the stop sign. The flow of the Lafzi process with Lafzi + is no different except that the addition of rules to the phonetic coding process is added to the rules for sound changes.

Table I shows the comparison of results from the Lafzi and Lafzi + systems [5]. The results of Lafzi + get a recall value and MAP value of 100% and 84%, this result is better than the Lafzi system which only gets a recall value and MAP value of 81% and 65%.

Table I. Comparison Table of Lafzi and Lafzi +

System	Sound Difference		Original Reading	
	Recall	MAP	Recall	MAP
Lafzi +	100%	84%	85%	58%
Lafzi	81%	65%	85%	58%

B. How to Waqaf Readings

In reading Al-Qur'an, it is recommended to beautify the voice, and reading it in one breath is also required to produce a perfect reading. It means the reader can to complete reading and stop at a stop sign or stopping it in the middle of the verse. However, the reader has to re-reading it from the previous sentence which still has a connection meaning. There are some rules of how to *waqaf* reading [8] among them:

- If the last letter has a *harakat sukun*, it doesn't change the reading.
- If the last letter is *fathah*, *kasrah*, and *dhammah*. The last letter is read as *sukun*.
- If the last letter *ta marbuthah* whether it is in the middle or at the end of the sentence, it is read as replacement of ha' that is *sukun*
- If the last letter is alive (*harakat*), but it is preceded by sukun, then those two letters are read sukun and the last one is read in a low voice
- If at the end of a sentence is preceded by the original mad or mad layyin (mad which previous letter is fathah), then it is read by making the letter located at the end of the sentence become silent with a little extended.
- When it stops at the end of the sentence, but the letter ends with the *fathah tanwin*, then only the *fathah* reads by two.
- If the last letter is *tasydid*, it is read by making it silent without removing the function of *tasydid*.
- *Hamzah* at the end of the word that is written on top of waw is *waqaf*, it will be silent and if it is *washal*, it will read shortly.

C. String Matching

String matching according to the Dictionary of Algorithms and Data Structures, the National Institute of Standards and Technology (NIST), is interpreted as a problem to find patterns of character strings in other strings or parts of text content [9]. Strings matching can be divided into two [10] that is:

1) Exact String Matching

Strings matching exactly to the order of characters in the matched string have the same number or sequence of characters. Examples of the word *step* will show a match with the word *step* only.

2) Inexact String Matching

String matching where the strings are matched have similarities but both have different character arrangements. Inexact String matching has two approaches namely matching based on the similarity of writing Approximate String Matching and matching based on the similarity of phonetic string matching speech.

a) Approximate String Matching

Strings matching based on similarity in terms of writing has the same number of characters but there are two different characters. If the difference between the two characters can be tolerated it is said to be suitable. Examples of the word *panitiea* with the *panitia*, have the same number of characters but there are different characters.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i3.935, Copyright ©2020
Submitted : 19 Juli 2020, Revised : 3 Agustus 2020, Accepted : 25 Agustus 2020, Published : 30 Agustus 2020

b) Phonetic String Matching

Strings matching based on similarity in terms of speech although in terms of writing is different. Examples of the word "*eye*" with the word "*ice*" have similarities in pronunciation but differ in writing, so the two strings are said to match.

D. N-gram

N-Gram is as many pieces of N characters taken from a longer string [11]. To make complete N-gram, "blanks*''* are needed at the beginning and end of a string as a marker, and usually, the blank that is often used is "underscore". The "LETTER" string example will consist of N-grams:

Unigram : L, E, T, T, E, R

Bigram : _L, LE, ET, TT, TE, ER, R_

Trigram : _LE, LET, ETT, TTE, TER, ER_, R_ _

It can be concluded that for strings of size n there will be n unigram, n + 1 bigram, n + 1 trigram, n + 1 quadgram, and so on. The advantage of N-Gram in matching strings can be resistant to textual errors due to the characteristics of N-Gram as part of a string, so errors in some strings will only result in differences in some N-Gram [7].

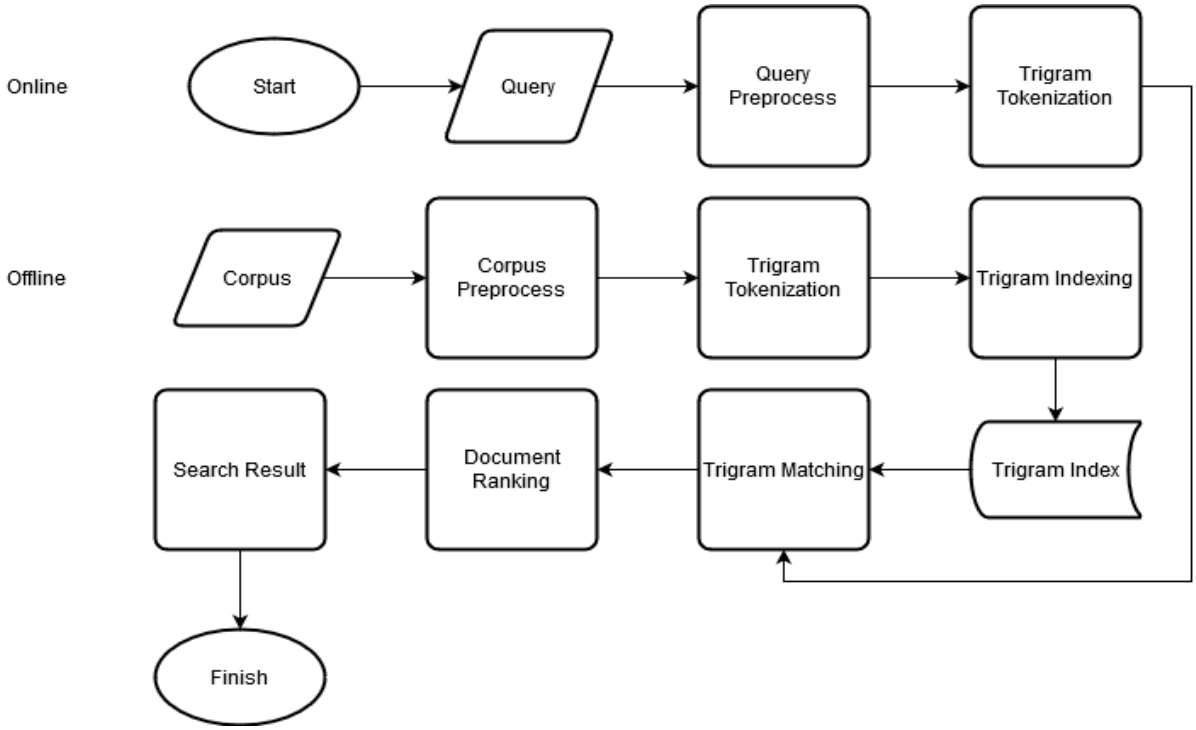

Fig. II. System Overview

E. MAP (Mean Average Precision)

MAP is the value obtained by calculating the average AP value of average precision [12]. The AP is counted for each relevant document taken and the relevant documents are counted in the results list. The following are examples of AP calculations listed in Table. II [3]

From the example table above, we get an AP score$\frac{\frac{1}{1}+\frac{2}{3}+\frac{3}{5}}{3} = 0{,}75$. A good MAP category can be seen from the AP results. If the AP value is high, it means that the precision of the system is good because the relevant information inputted by the system has an adjacent sequence and vice versa if the AP value is low, it means that the precision of the system is bad because the relevant information inputted does not have a contiguous sequence [13].

Table II. Example Table for Calculating AP

Output	True/False	Precision	Note
1	True	1/1	The sequence precision 1
2	False	-	Not precision
3	True	2/3	The sequence precision 3
4	False	-	Not precision
5	True	3/5	The sequence precision 5

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i3.935, Copyright ©2020
Submitted : 19 Juli 2020, Revised : 3 Agustus 2020, Accepted : 25 Agustus 2020, Published : 30 Agustus 2020

F. Recall

Recall is the number of relevant documents taken divided by the total amount of relevant information [14]. The maximum value for the recall is 1 and the drink value is 0. If the recall is 1, the system can successfully search according to information based on existing queries (Gold Standard). The following formula for calculating recall (1).

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

III. RESEARCH METHOD

A. System Overview

General description of the process on the system is illustrated in Figure I. In general, the process of the system is divided into two processes that are offline processing and online processing. Offline processing is done once to form an index, while online processing is done every time it is entered into the system and uses an index that has been formed to search [4].

B. Pre-Query Process

The preprocess query is input in the form of Latin text that is converted into phonetic code so that it can be matched with the results of preprocessing in the corpus. Some phonetic coding procedures for queries are explained in the following steps [4]:

1) Vowel Substitution

In the Arabic script, there are only three types of vowels, that are A, I, and U [15]. Whereas in the Latin alphabet there are other vowels, E and O. Therefore the vowel O is replaced by A, and for the vowel, E is replaced by I.

2) Multiple Character Elimination

The consonants and the same vowels that are next to each other become one letter.

3) Diphthong substitution

AI diphthong letters are changed to AY, while AU is changed to AW.

4) Idgham Reading Substitution

The rules on reading Idgham when *nun sukun* meets the *Idgham* letters, including *ya*, *nun*, *mim*, *wau*, *lam*, and *ra*. Then the letter N is omitted when meeting with the letter *idgham*.

5) Iqlab Reading Substitution

The rules on reading *iqlab* when *nun sukun* meet with the letter *ba* letter NB will change to MB.

6) Ikhfa Reading Substitution

The rules on reading *ikhfa* when the letters *nun sukun* meet with the letters *ikhfa*. *Ikhfa* reading is the sound of the letter N disguised, and sometimes written with "NG". The character G is omitted so it is equivalent to the coding of the Al-Qur'anic text.

7) Matching Letters

Matching letters from Latin text needs to consider Arabic letters that are represented in more than one consonant Latin letters. Created rules matching letters so that they can be adapted to the coding of the Al-Qur'an. Matching letters are written in Table III [4].

Table III. Latin Literacy Matching Rules

Latin Text	Equivalent
SH, TS, SY	S
KH, CH	H
ZH, DZ	Z
DH	D
TH	T
GH	G
NG (‘ain)	X
F, V, P	F
Q, K	K
J, Z	Z
‘,`	X

8) Elimination of Space

All spaces must be removed so that they are equivalent to the results in the procedure for coding the Al-Qur'anic text.

9) Addition Rule

Rules that can find words that can be *waqaf* [5]:

- Every query that ends in the letter *'h'* will be changed to the letter *'t'*. The word "*ghisyawah*" will be processed as "*ghisyawat*".
- Any query that ends in the letter *'h'* even if they are not located in a stop sign will still be displayed as output with the letter ending ' *h* '. For example, for the query "*alhamdulillah*" will still display results by the query.

C. Trigram Tokenization

The trigram tokenization process is the process of retrieving trigrams from phonetic code strings generated from the query and corpus of the al-Quran. The trigram formation does not require an initial or final marker on the string because the query must be part of the body of the Al-Qur'anic body [4]. The word to be tokenized must be changed into phonetic code, for example, the word "GHISYAWAH" is changed into phonetic code to "GISAWAH". Trigram of the word "GISAWAH" namely GIS, ISA, SAW, AWA, WAH.

D. Trigram Indexing

Trigram indexing is only done in the offline process. After tokenization on the corpus, the formation of the next inverted index. The inverted index uses trigrams as terms and verses of the Al-Qur'an as documents. One document in the index is one verse in the Al-Qur'an. The indexing process is done by 2

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i3.935, Copyright ©2020
Submitted : 19 Juli 2020, Revised : 3 Agustus 2020, Accepted : 25 Agustus 2020, Published : 30 Agustus 2020

methods, namely vocal and non-vocal [4]. In the vocal indexing method of the corpus of the Al-Qur'an, the phonetic code can be directly localized. However, in non-vocal indexing, all vowels (A, I, U) contained in the corpus of the Al-Qur'an are eliminated before they are localized.

E. Trigram Matching

Trigrams generated from queries are then compared with trigrams contained in the corpus of the Al-Qur'an then the same number of trigrams is calculated and multiplied by the threshold as a minimum score. The number of trigram matches more equal to the minimum score will be used as the output candidate.

F. Document Rating

A simple ranking of documents by scoring the same number of trigrams between the query and the corpus of the Al-Qur'an. The maximum value is as much as the number of trigrams and the drink value is one because the search process only takes one trigram of the same between the query and the corpus of the Al-Qur'an. Then the ranking of documents by giving a score on documents whose trigram positions appear ordered and dense compared to trigrams whose positions are randomized and fragmented. To search for candidates, the *Longest Increasing Subsequence* (LIS) is applied to the position where the trigram appears to give a sequence score [16]. The index which has the longest ranking will be the candidate results. LIS of an S sequence is a monoton subsequence up from S with a maximum length. To give a density score, the inverse average of the difference between the side of the elements by side is calculated [4] is calculated in equation2. The results from equation (2) are then multiplied by the length of the LIS that has been obtained to get an orderly score. For example, a line with an index [3, 10, 2, 1, 20] then the LIS [3, 10, 20] with a length of LIS 3 and a density score of 0.12 the index is given a score of 3 x 0.12 = 0.36

$$c = \frac{1}{n-1} \Sigma \frac{1}{s_{i+1} - s_i} \quad (2)$$

G. Search result

The output of the system is Al-Qur'an verses which contain the surah number, verse number, and the text of the Al-Qur'an verse in Arabic script. The verse that becomes the input query will be highlighted.

IV. RESULT AND DISCUSSION

Evaluation is carried out to determine the accuracy of the system based on the value of the recall and the value of MAP (*Mean Average Precision*).

A. Test Result

The tests are compared with previous research by adding 50 new test data taken from the Al-Qur'an murottal using the Tanzil application and testing the system of respondents. Test data is explained in the analysis of test results. Table IV shows the results of system testing in this research, where the results have a higher value than the previous system testing. The research results of Lafzi + get a recall value of 100% and MAP of 84% [5]. But when the system is tested by adding 50 new test data the system experiences a decrease in the value of recall and MAP. The decline in value occurs because the Lafzi + system cannot be implemented throughout the Al-Qur'an so there are still verses that cannot be displayed by the system.

Table IV. Table of Test Results

System	Sound Difference		Original Reading	
	Recall	MAP	Recall	MAP
Lafzi ++	100%	87%	99%	81%
Lafzi +	94%	73%	92%	74%
Lafzi	82%	65%	92%	74%

B. Analysis of Testing Results

System testing was conducted using 100 queries consisting of 50 previous research test data and 50 test data in this research. The first is testing queries that have different sounds from their writing. An example of the first scenario query can be seen in Table V. In the query "*Matsalanilqawm*" (مَثَلًالْقَوْم) it must be found in surah Al-A'raf verse 177 but in the Lafzi and Lafzi + query not found because the trigram differences generated between the input query and the corpus of the Al-Qur'an causes the input query not to be found by the system. By adding functions to the phonetic coding process and improving the process of index searching, the system can find the search results and increase the value of recall and MAP.

Table V. Example First Test Query

No	Arabic Text	Latin Text	Recall		
			Lafzi ++	Lafzi +	Lafzi
1	(فَنَبَذْوهُ)	Fanabazuh	1	0	1
2	(وَأَعَانَهُ)	Wa a'anah	1	0	0
3	(مَثَلًالْقَوْمُ)ا	Masalanil kaum	1	0	0

Second, testing the system using original reading in accordance with the Arabic script. Examples of queries for testing the second scenario can be seen in Table VI. The results of testing the second scenario get a recall value of 92% and MAP of 81% higher than the previous system. This is due to the addition of a function to the phonetic coding process and improving the process of searching indexes to produce search results where the Lafzi and Lafzi + systems are not found and the Lafzi ++ system is found.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i3.935, Copyright ©2020
Submitted : 19 Juli 2020, Revised : 3 Agustus 2020, Accepted : 25 Agustus 2020, Published : 30 Agustus 2020

Table VI. Example of Second Test Query

No	Arabic Text	Latin Text	Recall		
			Lafzi ++	Lafzi +	Lafzi
1	(لَهْوَاانفَضُّوا)	Lahwan infaddu	1	1	1
2	(وَأَعانَهُ)	Wa a'anahu	1	0	0
3	(مَثَلًالْقَوْمُ)	Masala al kaum	1	0	0

V. CONCLUSION

Based on the test results obtained and analyzed, it can be concluded that the verse search system in the Al-Qur'an for sound differences based on the phonetic similarity that has been developed does not reduce the recall value and the MAP of the previous system, wherein the previous system it gets a 100% recall value and MAP 84 Even though when tested with 50 new test data, the value decreases and after development it increases again with a recall value of 100% and MAP of 87%. Verse search based on original reading after the system was developed can increase the results of recall and MAP by 99% and 81%. This result is better than the previous Lafzi system and this research can already be implemented throughout the Al-Qur'an.

REFERENCES

[1] A. Sleit, and M. El-Haj B. Hammo, *Effectiveness of query expansion in searching the holy quran*., 2007.

[2] Tanzil. [Online]. http://tanzil.net

[3] IslamiCity. [Online]. http://www.islamicity.org

[4] M. A. Istiadi, "Sistem pencarian ayat al-qur'an berbasis kemiripan fonetis," 2012.

[5] M. A. Bijaksana, and K. M. Lhaksmana N. Rasyad, "Pencarian potongan ayat al-qur'an dengan perbedaan," *Jurnal Linguistik Komputasional*, vol. 1, 2019.

[6] Siti Nurhanifah, "Pencarian Informasi dengan Metode Trigram".

[7] K.-Y. Whang, J.-G. Lee, and M.-J. Lee M.-S. Kim, *n-gram/2l: A space and time efficient two-level n-gram*. In Proceedings of the 31st international conference on Very large data bases.

[8] Nadiazhr. 13 macam tanda waqaf yang wajib kamu ketahui.

[9] M. Syaroni and R. Munir, "Pencocokan string berdasarkan kemiripan ucapan (phonetic string matching) dalam bahasa inggris," *Islamic University of Indonesia*, 2005.

[10] A. Binstock and J. Rex, "Practical algorithms for programmers," *Addison-Wesley Longman Publishing Co.,Inc*, 1995.

[11] J. M. Trenkle, et al. W. B. Cavnar, "N-gram-based text categorization," *In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175.

[12] D. He, Z. Yue, and J. Jiang S. Han, "Contextual support for collaborative information retrieval," *In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, 2016.

[13] M. A. Bijaksana, and S. Al Faraby P. A. Arsaningtyas, "Sistem pencarian ayat al-quran berdasarkan kemiripan ucapan menggunakan algoritma soundex dan damerau-levenshtein distance," *Jurnnal Linguistik Komputasional*, 2018.

[14] D. Kelly, "Methods for evaluating interactive information retrieval systems with users. ," *Foundations and trends in Information Retrieval*, 2009.

[15] S. C. Soeratno, M. Ramlan, and I. D. P. Wijana S. Hadi, "Perubahan Fonologis Kata-kata Serapan dari Bahasa Arab dalam Bahasa Indonesia," *Gadjah Mada University*, 2003.

[16] D. Romik, "The surprising mathematics of longest increasing subsequences," *Cambridge University*, 2015.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i3.935, Copyright ©2020
Submitted : 19 Juli 2020, Revised : 3 Agustus 2020, Accepted : 25 Agustus 2020, Published : 30 Agustus 2020