Doa Search and Retrieval Using N-Gram

Nur Nabilah Abu Mangshor[1], Nurbaity Sabri[2], Zaidah Ibrahim[3], Zolidah Kasiran[4] and Anis Safura Ahmad

[1,2]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Kampus Jasin, 77300 Melaka Malaysia
[3,4]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor Malaysia

E-mail:[1]nurnabilah@melaka.uitm.edu.my,[2]nurbaity_sabri@melaka.uitm.edu.my,[3]zaidah@tmsk.uitm.edu.my,[4]zolidah@tmsk.uitm.edu.my

Received: 7 February 2017
Accepted: 9 October 2017

ABSTRACT

'Doa' is derived from Arabic word which means that one asks for the fulfillment or a need or the cure of sickness from him/her. Having to search and retrieve the relevant 'doa' for one needs at any particular time is beneficial. There are some search and retrieval applications that require using the exact match of the keyword search with the words stored in the database. This approach leads to the retrieval of insignificant results as users need to know the exact word to be searched. Therefore, this project allows for partial keyword search that utilises N-gram method for the search and retrieval process. Moreover, various words may have similar meaning thus to increase the accuracy of the retrieved result, this project compares the dice and overlap coefficient algorithms to find the synonyms of the searched word. The result produced indicates that overlap coefficient perform better than dice coefficient.

Keywords: *dice and overlap coefficient, n-gram algorithm, Malay keyword*

INTRODUCTION

Nowadays, science and technology has evolved tremendously. The incredible growth of the advanced technology in the new era makes the world's development more comprehensive and complex as to fulfil people's need and desires. One of these needs is search and retrieval of relevant *doa*.

Doa is about asking Allah for help or for the fulfilment of a particular need [1-4]. Information about *doa* can be retrieved from books, booklet or pamphlets. Although this approach is widely used, there exist some limitations. According to [4], searching using exact keyword retrieves some irrelevant information. Entering the exact keywords is also difficult [5]. One keyword can have different meanings or many keywords can have only one meaning [5]. Therefore, users may not know how to find relevant information based on keywords. Figure 1 shows an example when a user search for 'clean heart' at Dua Finder (*www.duaexplorer.com*) whereby the search results retrieves 66 counts of synonyms of 'clean heart' where most of them are irrelevant.

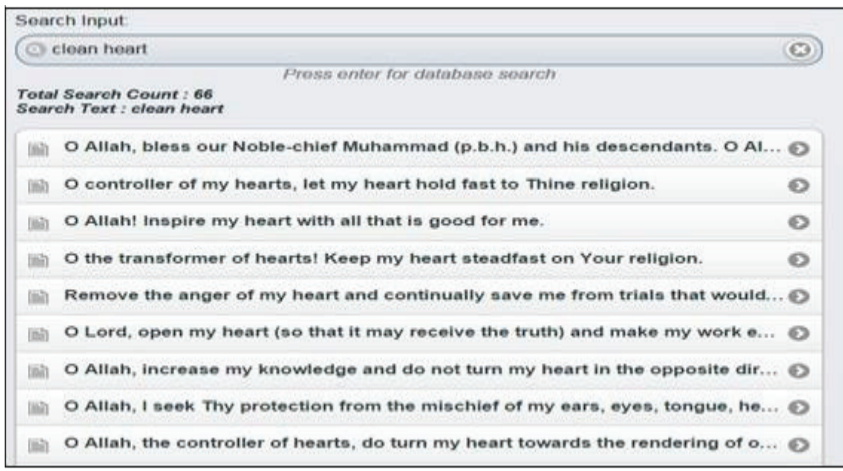

Figure 1: Sample Irrelevant Information Retrieved from Exact Keyword Search

Thus, this research investigates the use of N-grams for partial matching of the Malay keyword search and retrieval. The expert referred for the development of this prototype is Ustaz Hassan Abdullah from Darus Syifa Solok Gaung, Melaka.

METHOD

An N-gram is language independent [6-8] and can be looked at as a moving window. The incredible growth of the advanced technology in the new era makes the world's development more comprehensive and complex as to fulfil people's need and desires. One of these needs is search and retrieval of relevant *doa*.

Doa is about asking Allah for help or for the fulfilment of a particular need [1-4]. Information about *doa* can be looked at as a moving window on a word where the 'n' stands for the length of the sub-strings of the word. For example, if we were to N-gram the word *demam*, the results would depend on the length we have chosen. Table 1 shows the N-gram result for the word 'demam'.

Table 1 : Sample Result of N-gram for the Word *demam.*

Length	Result
Length 1(unigram):	[d,e,m,a,m]
Length 2(bigram):	[de,em, ma, am]
Length 3(trigram):	[dem, ema, mam]
Length 4(four-gram):	[dema, emam]
Length 5(five-gram):	[demam]

N-gram can also solve problems like partial matching or misspelling. In some Asian languages, different words are not separated by spaces, so a sentence is composed of many consecutives characters. Grammar knowledge is needed to separate those characters into words, which is a very difficult task to perform. Using N-grams, the system does not need to separate characters into words.

The N-gram model is a probabilistic model for predicting the next item in such a sequence by using the previous N-1 items [8]. The probability of a complete string sequence is as follows:

The equation above suggests that we could estimate the joint probability of an entire sequence of word by multiplying together a number of conditional probabilities. Table 2 shows an example of the calculation of the probability of a complete string by multiplying the appropriate bigram probabilities using the word 'This is Malaysia'.

The research chooses the trigram because it is a popular approach being applied by many researchers. Table 3 shows an example of the trigram for the keywords 'terang' and 'penerang' while Table 4 illustrates the result of Dice and Overlap coefficient for similarity matching [10]. The higher the result of the coefficient, the better is the similarity matching. N-gram had been applied in web based application in order to implement the advanced search function. Table 5 shows an overview of the searching process where for example, when user misspelled the entered keyword search 'blajar' and the result of the retrieved information is related to 'belajar'.

Table 2: Sample Calculation of the Probability for n-gram Method

<s>This is Malaysia</s>	
P(This|<s>) =2/3 = 0.67	0.67 * 0.67 * 0.5 * 0.5 =1.34
P(is|This) =2/3 = 0.67	
P(<s>|Malaysia) =1/2=0.50	
P(Malaysia|is) =1/2 = 0.50	

Table 3: Trigram Approach for *terang* and *penerang*

ter	pen
era	ene
ran	ner
ang	era
	ran
	ang

Table 4: Calculation for Dice and Overlap Coefficient for Similarity Matching

Dice Coefficient	Overlap Coefficient
(2M) / (P+Q)	M/min (P, Q)
2(3) / (8+6) = 0.42	3 / (8, 6) = 0.50

Table 5: Sample Misspelled Keyword or Partial Matching of Keyword Search

No.	User Input	Matched Word (Database)
1.	*erang*	*Penerang hati*
2.	*akhir*	*Pengakhiran baik*
3.	*blajar*	*Belajar*
4.	*sakit*	*Sakit kepala*
5.	*kerjaan*	*Pekerjaan*

RESULTS AND DISCUSSION

Collection of the *doa* and its uses have been gathered from Ustaz Hasan from Darus Syifa Solok Gaung, Melaka. He also verified the result of the *doa* and the uses of this prototype that consists of 15 different *doa*.Figure 3 to 4 illustrate sample interfaces of this prototype. Figure 2 is where user enters the keyword search. Then, user needs to click button 'Carian' and it displays the relevant information related to the keyword that user has entered as shown in Figure 3. According to Ustaz Hasan, there can be various *doa* that can be applied for a particular purpose. Thus, Figure 3 illustrates the various surah that user can select to retrieve the relevant *doa* as shown in Figure 4. By looking at Figure 4, it can be viewed the retrieved title of the *doa* entered by user, the verse from al-Quran and the Malay translation of the *doa*.

The experiments conducted between dice and overlap coefficient for similarity matching purpose between the entered keywords and the words in the database indicated that for all cases, overlap coefficient is better than dice coefficient. Table 5 illustrates the matching of 15 keywords that user entered and the words in the database.

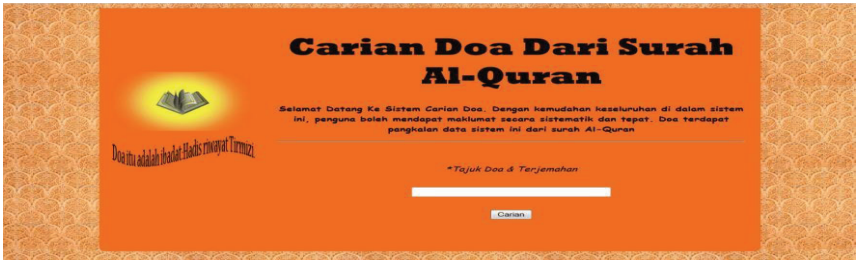

Figure 2: Interface to Enter Keyword

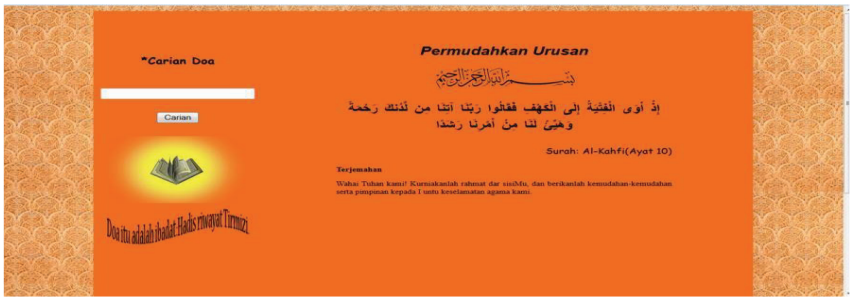

Figure 3: Retrieved Information

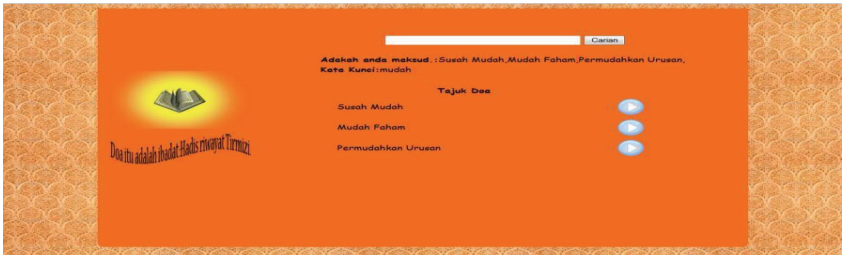

Figure 4: Final Result of the Retrieved Information

Table 5: Comparative Result between Dice and Overlap Coefficient for Similarity Matching

No.	Queries	*Tajuk **Doa***	Dice Coefficient (%)	Overlap Coefficient (%)
1.	*penerang*	*Penerang hati*	60	75
2.	*nerang*		44.44	66.66
3.	*masuk*	*Masuk dewan*	40	60
4.	*dewan*		40	60
5.	*hati*	*Penerang hati*	25	50
		Susah hati	30.76	50
		Ditetapkan hati	22.22	50
6.	*urus*	*Permudahkan urusan*	19.05	50
		Kebaikan urusan	22.22	50
7.	*urusan*	*Permudahkan urusan*	34.78	66.66
		Kebaikan urusan	40	66.66
8.	*ingat*	*Menguatkan ingatan*	27.27	27.27
9.	*faham*	*Mudah faham*	40	60
10.	*sabar*	*Diberi kesabaran*	20	60
11.	*teguh*	*Diteguhkan Iman*	31.57	50
12.	*iman*	*Diteguhkan Iman*	22.22	50
		Tetapbersama orang Iman	16	50
		Nabi sulaiman	25	50
13.	*dinding*	*Pendinding diri*	47.61	71.42
		Pendinding dari sihir	38.46	71.42
		Pendinding dari syaitan	35.71	71.42
14.	*nikmat*	*Mensyukuri nikmat*	36.36	66.66
15.	*rezeki*	*Limpahan rezeki*	40	66.66
		Keluasan rezeki	40	66.66
		Rezeki	66.66	66.66
		Rezeki Hari Raya	40	66.66

CONCLUSION

This prototype allows users to gain fruitful knowledge about *'doa'* and it uses whenever necessary without having to seek for the information from other sources such as books or meeting face-to-face with the experts. Instead, the development option in measuring keyword similarity is performed. In the searching process, it provides trigram that effectively conducts keyword search where partial matching or incorrect spelling of the entered keyword still produces relevant retrieved information. A comparative study between dice and overlap coefficient indicates that overlap coefficient is better than dice for similarity matching for retrieval purposes. Future work is to expand the database with *'doa'* from hadiths and includes audio processing where users can listen to the *'doa'* and its translation.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the help of Ministry of Education (MOE) and Universiti Teknologi MARA (UiTM) for sponsoring this research under the National Grant No. 600-RMI/FRGS 5/3 (165/2013).

REFERENCES

[1] S. Dhorat, 2011. The Importance of Dua, Islamic Da'wah Academy, pp. 1-3.

[2] N. Deuraseh, 2004. Al-Ruqyah with the Quran and the Du'a (the Prayer) in Islamic Medical Tradition. *JISHIM*, *Vol. 4*(April), pp 27-32.

[3] S. R. Ali, W. M. Liu, and M. Humedian, 2004. Islam 101: Understanding the Religion and Theraphy Implications, *Professional Psychology: Research and Practice, Vol. 3*(6), pp. 635-642. DOI: 10.1037/0735-7028.35.6.635.

[4] P. N. Reddy and Y. Swetha, 2013. Techniques for Efficient Keyword

Search in Cloud Computing, *International Journal of Computer Science and Information Technologies, Vol. 4*(1), pp. 66-68. DOI: 10.1.1.294.9906.

[5] J. Singthongchai and S. Niwattanakul, 2013. A Method for Measuring Keywords Similarity by Applying Jaccard's N-gram and Vector Space, *Lecture Notes on Information Technologies, Vol. 1*(4), pp. 159-169. DOI: 10.12720/lnit.1.4.159-164.

[6] R. Mahmoud and S. Majed, 2011. Improving Arabic Information Retrieval System using N-Gram Method, *World Scientific and Engineering Academy Society* (WSEAS), *10*(4), pp 1-9.

[7] M. Danesh, B. Minaei and O. Kashefi, 2013. A Distributed N-Gram Indexing System to Optimizing Persian Information Retrieval. In *International Journal of Computer Theory and Engineering, Vol. 5*(2), pp. 214-222. DOI: 10.7763/IJCTE.2013.V5.681.

[8] S. Yang, H. Zhu, A. Apostoli and P. Cao, 2007. N-Gram Statistics in English and Chinese: Similarities and Differences. In *International Conference on Semantic Computing*, pp. 454-460. DOI: 10.1109/ICSC.2007.46.

[9] Sreejith, Reghu & Indu, 2013. N-gram Based Algorithm fordistinguishing between Hindi and Sanskrit Texts, In *International Conference on Computing, Communication and Networking Technologies* (ICCCNT), 1-4. DOI: 10.1109/ICCCNT.2013.6726777.

[10] G. Recchia and M. Louwerse, 2013. A Comparison of String Similarity Measures for Toponym Matching. In *Comp 1/3 Proceeding of the First* ACM SIGSPATIAL *International Workshop on Computing Models of Place*, pp. 1-8. DOI: 10.1145/2534848.2534850.