How to cite this paper:

Ahmad Akmaluddin Mazlan, Norita Md Norwawi, Fauziah Abdul Wahid, Roesnita Ismail, & Ashraf Al Omoush. (2017). Query cost-reduction for quranic-arabic information retrieval using hexadecimal conversion algorithm in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference of Computing & Informatics (pp 91-98). Sintok: School of Computing.

QUERY COST-REDUCTION FOR QURANIC-ARABIC INFORMATION RETRIEVAL USING HEXADECIMAL CONVERSION ALGORITHM

Ahmad Akmaluddin Mazlan[1], Norita Md Norwawi[2], Fauziah Abdul Wahid[3], Roesnita Ismail[4], Ashraf Al Omoush[5]

[1, 2,3,4,5] *Universiti Sains Islam Malaysia, Malaysia*
{[1]*aamja03*, [5]*omoush2012*}@gmail.com, {[2]norita, [3]fauziah, [4]roesnita}@usim.edu.my

ABSTRACT. Digital Quran is a natural language document that use either Arabic font or images of the verses. In the Al-Quran there are 18994 unique words. Thus, the image approach uses a significant amount of memory space. However there is not much work has been done using machine translation (MT) technique for the Quranic representation. This paper will proposed Arabic information retrieval based on keywords search in Hexadecimal Representation using Al-Quran verses as the test case. All Quranic words will transliterate into machine language in the form of binary format after removing diacritic and duplication. This machine language approach in representing Digital Quran reduces the size of storage around 47-54% and retrieval time up to 20% hence reduce the query cost for Arabic information retrieval in general.

Keywords: hexadecimal conversion, arabic information retrieval, query cost reduction, digital quran representation

INTRODUCTION

Nowadays, Digital Quran can be found in every smart mobile device, computer, thousands of websites (Bakar, 2010), and tablet applications. However, users do not just recite the Quran only, but they also need to search, translate, and have its explication (Foda et al, 2013). Key term searches on Digital Quran involves query language based on non-Latin characters target document. Thus, the query search technique for the Arabic texts can be used to improve Arabic Information Retrieval (IR) (Elayeb et al, 2015).

Digital Quran is considered as a natural language text documents which suffers problem from retrieval time through machine translation (MT) technique (Elayeb et al, 2015). In this paper, an alternative approach for query processing using system programming will be presented. All language characters available for use in computers such as Arabic, Chinese, Thai etcetera are represented using an encoding method called Unicode that manipulate the binary representation in Hexadecimal form.

The Al-Quran is a corpus which consists of 30 Juzu', 60 Hizb, 114 Chapters, 6,236 Verses, 77,439 words, 320015 letters (Tabrizi, 2013, Foda et al, 2013). Due to the huge amount of words and lack of transliteration for machine understanding, this research proposed a technique to represent all words of the Al Quran by using Unicode. However, the Al-Quran has about 18,994 distinct word-non redundant words, or unique words. Thus, duplications of kalimah or words are removed hence improve machine readability at presentation layer (Foda et al, 2013).

Bilel mentioned there are 4 Arabic IR Translation techniques which are: (1) Dictionary-Based approaches, (2) Parallel Corpora-Based Approaches, (3) MT-Based approaches and, (4) Approaches Combing Arabic Translation Resources (Elayeb et al, 2015). Each IR processing techniques have its drawbacks. For example, the main problems reported in direct dictionary-based IR are: (i) the problem of inflection, (ii) translation ambiguity, (iii) compounds and phrases and their handling, (iv) proper name and other untranslatable words and their handling, and (v) lack of structuring (Pirkola, 2001).

However, this research focuses on the issue of lack of structuring in the application-based system through machine translation (MT). An improvement of the MT technique through encoding schemes will be able to provide alternative solution to the Arabic transliteration and spelling-related issues.

The rest of the paper is organize as follows: Section 2 considers the problems in query searching for Arabic information retrieval on MT technique, which are handling untranslatable words, and lack of structuring. Section 3 will present the proposed QuHex model consist of hexadecimal conversion algorithm. The hexadecimal conversion algorithm also will be discussed in section 3. Section 4 will discuss on the implementation strategy. Section 5 represents summarization of the results and suggestions for the future works.

RELATED WORKS

Hexadecimal Approach in Information Retrieval

Unicode representation enables every characters of many natural languages to be translated into a machine form to be implemented on any system. Every character has their own unique value or weightage. For example, Arabic character "ح" is valued as D8AD converted into hexadecimal value.

Every Arabic letters can be map to any ASCII characters. For example, LAM (ل)is denoted as D984 in hexadecimal. But the variation of the Arabic letters YEH (ي) is D98A and ALEF MAKSURA (ى) is D8A6, which both are different and has their own value [2].

The Al-Quran is a corpus which consists of 30 juzu, 60 hizb, 114 chapters, 6236 verses, 77439 words and 320015 letters (Basharat, 2015, Tabrizi, 2013, Foda et al, 2013). A research done by Mr. Khaled found that the Unicode representation is not the best way to represent the Holy Qur’an. However by using the Qur’anic code the research solved 5 out of 6 problems, and reached more than 65% reduction ratio, increase searching capability, and found standard way to store and present the Holy Qur’an on any electronic device (Foda et al, 2013).

Nevertheless, most studies revolves around Arabic Information researches (Alqahtani, et al, 2016, Amar et. Al, 2013, Basharat, 2013) actually based on the application layer of OSI.

Thus, an important component in this approach is the Hexadecimal conversion where a programming algorithm is required to converts every Arabic word into its own unique hexadecimal value. Table 1 lists the characters of Arabic and Latin and code points (in hex) (Foda et al, 2013).

Table 1. Hexadecimal Value for Each Characters

ل	ي	خ	ن

D984	D98A	D8AE	D986

THE QURANIC HEXADECIMAL (QUHEX) MODEL

The Quranic Hexadecimal (QuHex) model aims to identify the value of each natural keyword related to Quran which easily decomposes to characters. The model consists of the following functions:

- Q-IR will convert the user query to hexadecimal form to compare with the content in the index.
- The index stored hexadecimal form of keywords from 3 different languages related to the same verse of Quran.
- Reference on the contents of the Al-Quran by experts in 1 natural languages which is Arabic language.

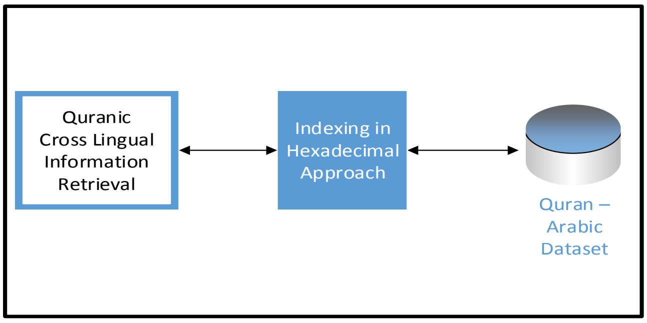

Quranic Cross Lingual Information Retrieval
Indexing in Hexadecimal Approach
Quran – Arabic Dataset

Figure 1. QuHex Model

Figure 1 illustrates the QuHex Model which integrate Arabic natural language to improve the readability on presentation layer of the Network OSI. The model is unique as the dataset is still in its natural form and the indexing link the information in each dataset. The indexing stored the keywords in hexadecimal form as shown in Table 1.

The general context of the orthographic Arabic word consists of a sequence of Arabic characters. The Arabic word is a finite set of Arabic characters that can be denoted as Eq. (1):

$$k = h_1 + h_2 + \cdots + h_n \tag{1}$$

Where k is the Arabic word and h each letter in the word. Each Arabic words has their own hexadecimal value and a zero or more diacritics with a finite length. The representation of an Arabic word kH in hexadecimal can be denoted as Eq. (2):

$$kH = \sum_i^n \boldsymbol{h_i} \tag{2}$$

Here, $\sum_i^n \boldsymbol{h_i}$ represents an Arabic word with a finite length where it combines each character value in hexadecimal form into a new hexadecimal value. For an example, the forth whitespace delimited token of verse 2:266, نخيل (Ashraf et. Al, (2015). This token is represented orthographically as in Table 2 where 4 characters are converted to 4 hexadecimal values.

Table 2. Conversion Arabic Word into Hexadecimal Form.

	Arabic

Word (k)	نخيل
Tokenization – Eq. (1)	ن + خ + ي + ل
Hex Value (H_i) – Eq. (2)	d986 + d8ae + d98a + d984
Sum Hex Value (*kH*)	36542

The Arabic character which is read from the right to the left have its own hexadecimal value based on the UTF-8. The sum of the 4 characters shows in Table 2 which is 36542.

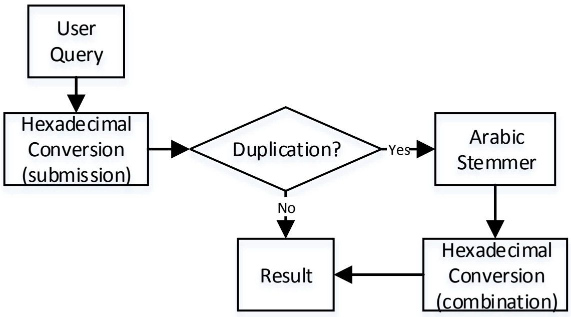

User Query
Hexadecimal Conversion (submission)
Duplication?
Yes
No
Arabic Stemmer
Result
Hexadecimal Conversion (combination)

Figure 2. The Hexadecimal Conversion Process Flow

Figure 2 shows when conflict happened when for two or more different words the Sum Hex Value (kH) has the same value after Hexadecimal Conversion based on Eq. (2). One alternative solution is, the conflicts was resolved through findings the root word. Then, the root word was converted to hexadecimal and the Sum Hex Value (kH) is the combination of letters of the root words. Finally, this resolved the equation of finding unique value for each words in the Arabic language by combining the actual word and the root words representation in Hexadecimal (refer Table 4.)

Figure 3 shows the algorithm translated from the Figure 1 for QuHex Model that will be implemented to the information retrieval for proof of concept.

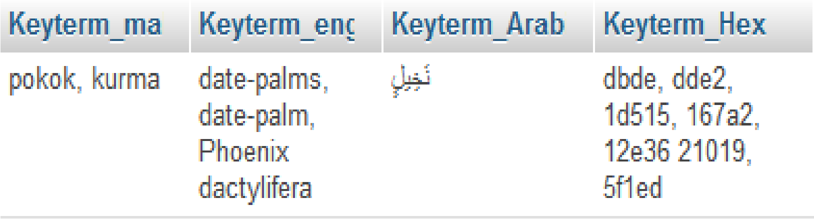
```
Function: Hexadecimal Conversion
Input: User query in Arabic
Output: Hexadecimal value of words
Start
User Query ← T
Tokenization word into characters, $T = \{c, c_2, \dots , c_u\}$
Check lookup table for Unicode
Character converted to Hexadecimal value
While (T! = ∅)
    While (Cn! = ∅)
        SumHex = $\sum$ (c, c2, … , cu)
    End Loop
    Read duplicateFile
    Let x = duplicateFile
    Identify conflict of duplication words
    If (SumHex = x)
        Find the root word
        Let R = RootWords (T)
        Combine the characters in Hexadecimal
```

```
                CombiHex = HexRootWord (R)
    End Loop
End
```

The results of hexadecimal conversion are shown in the Table 3, 4 and 5.

RESULTS AND FINDINGS

Table 3. Quran Arabic Keywords and Hexadecimal Value



Keyterm_ma	Keyterm_enç	Keyterm_Arab	Keyterm_Hex
pokok, kurma	date-palms, date-palm, Phoenix dactylifera	نَخِيلٍ	dbde, dde2, 1d515, 167a2, 12e36 21019, 5f1ed

Table 3 above shows that each language has its own respective keyword(s). At the last column, the multiple languages keywords were collected and converted into hexadecimal value per word in a single column. This hexadecimal can be matched to the user query. The comparison process uses the hexadecimal conversion algorithm. The result will produce the same relevance compared to the natural language of the user.

Table 4. Number of Words Converted through Hexadecimal Conversion

Surah	Total Kalimah (Without stopwords)	Repeated Words	Hexadecimal Conversion (Submission)	Hexadecimal Conversion (Combination)
Al-Fatihah	26	3	23	0
Al-Baqarah	2210	649	841	620

Table 4 above shows two data sets from the Digital Al-Quran (Tanzil © 2007–2017) consists of Surah Al-Fatihah and Surah Al-Baqarah. Surah Al-Baqarah is known as the longest datasets for Digital Quran. Both data sets are non-diacritic because of effective text similarity approach (basyarat, 2015, Abdelnasser, 2014). The actual total of kalimah for both Surah are 29 and 2266 respectively. After the pre-processing of Stopword removal, it left with 26 and 2210 respectively. Next, Hexadecimal Conversion (Submission) refer to the Eq.1 while Hexadecimal Conversion (Combination) was the combination of hex value of the first three letters of the root words.

Table 5. Size Reduction before and after Hexadecimal Conversion

Surah	Total Size of Arabic Text File (UTF-8) (Bytes)	Total Size of Hex Text File (Bytes)	Total Reduction in Size (%)
Al-Fatihah	332	159	47.89
Al-Baqarah	26,378	14,412	54.63

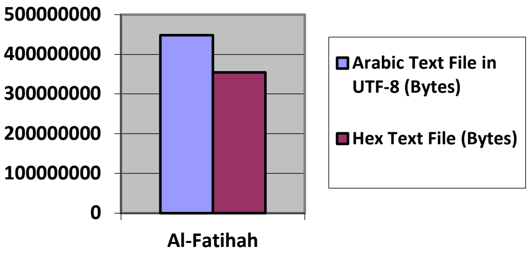


Figure 3. Size Comparison between Arabic Text File and Hex Text for Surah Al-Fatihah

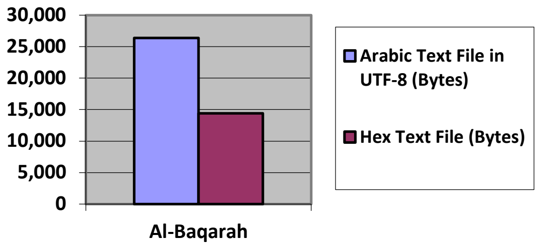


Figure 4. Size Comparison between Arabic Text File and Hex Text for Surah Al-Baqarah

From Figure 3 and 4, we can found out that the size for hex text are smaller than the Arabic Text (Surah Al-Fatihah and Al-Baqarah). From table 5, in summary, the reduction of file size is around 47-54 %. This can be imply that the processing speed could also be improve as the space needed for data transfer are much smaller. This prove that through system programming approach, the enhancement will improve the retrieval time especially for MT technique.

Table 6. Retrieval Time before and after using Hexadecimal Conversion

Surah	Average Time from 10 Runs without QuHex (nanosec)	Average Time from 10 Runs using QuHex (nanosec)	Retrieval Performance Time (%)
Al-Fatihah	448021762	354634733	20.84
Al-Baqarah	9409624071.9	9332212868.5	0.82

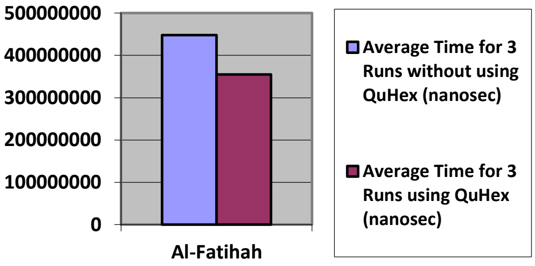


Figure 5. Average Time from 10 Runs with and without QuHex for Surah Al-Fatihah

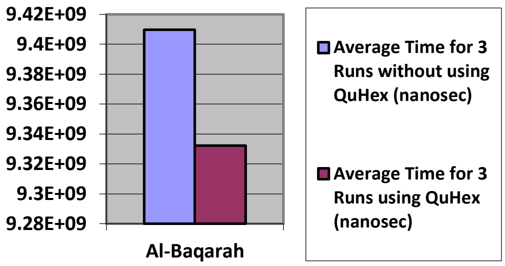


Figure 6. Average Time from 10 Runs with and without QuHex for Surah Al-Baqarah

Figure 5 and Figure 6, indicate that the size for hex text are smaller than the Arabic Text (Surah Al-Fatihah and Al-Baqarah). From table 6, in summary, the retrieval performance time is up to 20%. However, there are negligible result found for extra time because conflicts handling in hexadecimal conversion algorithm.

From this result, it proves that through exploiting the presentation layer, QuHex will decrease the processing complexity for every Arabic letters to a unique value (kH) of an Arabic word in hexadecimal form. Thus, this method will improve the MT technique for Arabic IR.

CONCLUSION

QuHex is presented as a potential solution to improve the readability of natural languages by using encoding approach. QuHex utilizes the hexadecimal conversion algorithm and the simulation of QuHex will give an early finding and result for further implementation. 3 different languages had been translated into hex value. Instead of 3 dictionaries, this research used only 1 dictionary in third party language (machine).

This research aim to improve the space usage as the reduction of space size is around 47-54 % thus will result a faster retrieval time of CLIR. The result found out that the retrieval time is improve up to 20%.

Currently, QuHex is still undergoing   a continuous prototyping development. Potential future research may focus on the application of the hexadecimal conversion algorithm to improve cryptographic algorithm as well as to provide security and reduce the size of the high quality digital images as well implementing 2nd phase in creating interpreter for any word in a hex value.

ACKNOWLEDGMENTS

The Ministry of Higher Education (MOHE), and Universiti Sains Islam Malaysia (USIM) under research grant USIM-NRGS-P/FST/8404/52113 are acknowledged for supports provided and facilities utilized during this study.

REFERENCES

A. Basharat, (2015). Comparative Study of Verse Similarity for Multi-lingual Representations of the Qur'an in Int'l Conf. *Artificial Intelligence*, ICAI'15.

Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., El-Makky, N., & Torki, M. (2014). *Al-Bayan: an arabic question answering system for the holy quran*. ANLP 2014, 57.

Alqahtani, M., & Atwell, E. (2016, June). Arabic Quranic Search Tool Based on Ontology. In *International Conference on Applications of Natural Language to Information Systems* (pp. 478-485). Springer International Publishing.

Amar Arbaoui, Alginahi, Y. M., & Menacer, M. (2013). Strategies for Collecting Electronic Resources on the Qur'anic Researches. *International Journal on Quranic Research (IJQR)*, 3(4), 57–78. http://Doi.Org/10.1017/CBO9781107415324.004 Accessed On 25th July 2016.

Ashraf Al_Omoush, Norita Md Norwawi, Roesnita Ismail, Fauziah Wahid & Ahmad Akmaludin Mazlan. On Unicode Hexadecimal Representation for Quranic Words. *The 4th FEIIC - International Conference on Engineering Education & Research 2015* (FICEER2015) (2015).

B Elayeb, I Bournas. Arabic Cross-Language Information Retrieval: A Review. *ACM Trans. Asian Low-Resou. Lang. Inf. Process, (2015)* Vol. 15, No. 3, Article 1.

Foda, K. M., Fahmy, A., Shehata, K., & Saleh, H. (2013, December). A Qur'anic Code for Representing the Holy Qur'an (Rasm Al-'Uthmani). *In Advances in Information Technology for the Holy Quran and Its Sciences (32519), 2013 Taibah University International Conference on* (pp. 304-309). IEEE.

K.M.S. Foda, A.Fahmy, K. Shehata, H. Saleh. A Qur'anic Code for Representing the Holly Qur'an (Rasm Al-'Uthmani), 2013 Taibah University *International Conference on Advances in Information Technology for the Al-Quran and Its Sciences*, 978-1-4799-2822-4/13 $31.00 © 2013 IEEE DOI 10.1109/NOORIC.2013.67

Pirkola A., Hedlund T., Keskustalo H., and Jarvelin K., *"Dictionary Based Information Retrieval: Problems, Methods and Research Findings."* Kluwer Academic Publishers. Information

Retrieval, 4, 209-230. J. Clerk Maxwell, a Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73, 2001.

Tabrizi, A.A., Mahmud, R., "*Issues of coherence analysis on English translations of Quran" 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013, pp. 1-6.

Tanzil © 2007–2017, website: http:// http://tanzil.net/wiki/Tanzil_Project. Accessed date: 29 Oktober 2016.