MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

USIM

Article

Quranic Cross-Lingual Information Retrieval Optimization Using Hexadecimal Conversion Algorithm

Ahmad Akmaluddin Mazlan[1,a], Norita Md Norwawi[1,b], Fauziah Abdul Wahid[1,c], Roesnita Ismail[1,d]

[1]Faculty of Science & Technology, University Sains Islam Malaysia (USIM), Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
E-mail: [a]aamja03@gmail.com[2], {[b]norita, [c]fauziah, roesnita[d]}@usim..edu.my

***Abstract*— The electronic Quran text has a huge amount of words and lack of transliteration for machine understanding which drop the query search performance of the Digital Quran. The Holy Quran consists of 30 juzu, 60 hizb, 114 chapters, 6236 verses, 77439 words and 320015 letters. Therefore, this research proposed the Quran Hexadecimal (QuHex) model as the solution to represent the words of the Holy Quran by using hexadecimal conversion algorithm to improve machine readability at the presentation layer. The result is 3 different languages into were translated into hex value. Instead of storing into 3 different dictionaries, we used only 1 dictionary in third party language (machine). This research also improves the space usage as the reduction of space size is around 47-54 %. This research hypothesizes that the higher the reduction of space usage, the faster the retrieval time of CLIR. This article highlights the proposed Quranic Cross-Lingual Information Retrieval (Q-CLIR) model and Hexadecimal Conversion Algorithm. Currently, QuHex is still undergoing a continuous prototyping development and will implement the 2nd phase in creating interpreter for any word in a hex value. The implementation strategy and suggestions for the future works are also given.**

***Keywords*— Hexadecimal Conversion; Quranic text; Quranic Cross-Language Information Retrieval; Quran Hexadecimal Model; Unicode**

I. INTRODUCTION

Cross Lingual Information Retrieval involves query language which is different from the target document. The classical Arabic text is an important test set that use to improve Arabic Information Retrieval (IR) and Cross-Lingual Information Retrieval (CLIR), discussed in [6]. This study exploits a standard text reference for Muslim's called The Holy Quran. Which remain valid throughout time. The Holy Quran is a corpus which consists of 30 juzu, 60 hizb, 114 chapters, 6236 verses, 77439 words and 320015 letters.

Digital Qurans that are available on the internet are considered as natural language text documents which also present serious problems for achieving machine interoperability. Due to this huge amount of words and lack of transliteration for machine understanding, this research proposed a technique to represent all words of the Holy Quran by using Unicode to remove the duplication of kalimah or words and the improve retrieval time at the presentation layer in, discussed in [6] and [8].

The Muslim population has reached billions with different cultures and natives. Besides arabs, those who called 'ajam are people who are non-native arab speaker. Thus, cross-lingual study for understanding in the Quran is needed, discussed in [8]. Cross-Language Information Retrieval (CLIR) is developed to allow users to retrieve text documents and acquire relevant information in a language different from the language of the user's query. Elayeb [9] discusses four Arabic CLIR Translation techniques which are: (1) Dictionary-Based approaches, (2) Parallel Corpora-Based Approaches, (3) MT-Based approaches and, (4) Approaches Combing Arabic Translation Resources. This research will focus to improve dictionary and parallel-corpora based translation system.

The main problems reported in direct dictionary-based CLIR are: (1) the problem of inflection, (2) translation ambiguity, (3) compounds and phrases and their handling, (4) proper names and other untranslatable words and their handling, and (5) lack of structuring discussed in [1]. However, this research will focus on the issue of lack of structuring in the application-based system through machine translation.

The main problems faced by parallel-corpora translation system is in translating more than two (2) languages. For example, if we are translating three (3) languages simultaneously means we need three (3) dictionaries for each language. However, through machine translation, we can have only 1 dictionary based on its value, discussed in [10].

The structure of this paper is as follows. Section II considers the problems in query translation using a cross-lingual dictionary, which are untranslatable words and their handling, and lack of structuring. Section III will show the proposed QuHex model consist of hexadecimal conversion algorithm. Section IV shows the results of this research. Section V will discuss the implementation strategy. Section VI represents summarization and suggestions for the future works.

II. Related Works

These are the works that related to this research.

A. CLIR for Arabic, English, Malay.

According to Elayeb [9], there are three main approaches to translation in CLIR which are: (i) translation by a bilingual machine-readable dictionary (MRD), (ii) parallel or comparable corpora-based methods, and (iii) Machine-Translation (MT). Approach (i) problem is related to the translation ambiguity associated with these resources. However, query translation for Arabic-English CLIR through MRD is cost effective compared to the other methods. Additionally, as discussed in [14] said it is tough to find this kind of wide-ranging dictionary. Next is every-match (EM) method, in which they evaluated the impact of simple word-by-word translation of Arabic-English retrieval performance. However, this method suffers from ambiguous translations, as many extraneous terms are added to the original query, discussed in [12].

Approach (ii) problem is it needed bigger usage of memory space as it requires a large dictionary for all possible words for different-pairs of language. (Mult)iSearcher, supported by an automatic translation process, benefited from parallel corpora that are more available and with free access, discussed in [2]. Approach (ii) based on statistical/probabilistic methods applied on parallel text with the aim of selecting correct translations provide good performance, but they are challenged by accuracy based on domain and absence of parallel texts in different pairs of languages.

Approach (iii) currently uses a Natural language processing technique to find the match of the query. For example, Lavie [13] developed a basic Hindi-to-English MT system, by enhancing the performance of a syntactically transfer-based approach, using strong statistical methods. Other than that, Hatem and Omar [14] proposed a transfer-based approach in Arabic to English MT, in order to solve the word ordering problem. Both had improved the performance of information retrieval.

Recently, however, it is hard to find research which uses system programming approach on information retrieval which benefiting the hexadecimal representation of the Arabic texts where query processing will take place at the presentation layer of the OSI.

B. Hexadecimal Approach for Encoding.

Numerous researches have led to an enhanced Arabic model that make the i'jam patterns of Arabic letters normative by splitting Joining_Group classes related to the UTF-8 Unicode Standard. By meeting the requirements of byte-oriented, ASCII-based systems, the Unicode Standard defines UTF-8. UTF-8 is a variable-length, byte-based encoding that preserves ASCII transparency. UTF-8 maintains transparency for all of the ASCII code values (0...127). These values in Table 2.4 do not appear in any byte of a transformed result except as the direct representation of the ASCII values, discussed in [11].

Unicode representation enables the machine to translate every character of many natural languages to be implemented on any system. Every character has their own unique value or weightage. For example, the Arabic character "ح" is valued as d8ad converted into hexadecimal value.

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, discussed in [8].

Most studies revolve around Arabic Information researches discussed in [1], [3] and [5] based on the application layer of the OSI.

Thus, hexadecimal conversion is required, which is a programming algorithm to convert every Arabic word into its own unique hexadecimal value. Table I lists the characters of Arabic and Latin and code points (in hex), discussed in [10].

TABLE 1
HEXADECIMAL VALUE FOR EACH CHARACTER.

ل	ي	خ	ن
d984	d98a	d8ae	d986

Thus, hexadecimal conversion is required which is a programming algorithm to convert every Arabic word into its own unique hexadecimal value. Table II lists reference table (Unicode) for characters of Arabic and Latin and its code points (in hex), discussed in [10].

TABLE II
UNICODE FOR ARABIC AND LATIN CHARACTERS

Name	Character	Code Point	UTF-8 (Hex.)
Small letter a	a	U+0061	61
Small letter b	b	U+0062	62
Small letter c	c	U+0063	63
Small letter d	d	U+0064	64
Small letter e	e	U+0065	65
Small letter f	f	U+0066	66
Small letter g	g	U+0067	67
Small letter h	h	U+0068	68
Small letter i	i	U+0069	69
Small letter j	j	U+006A	6a
Small letter k	k	U+006B	6b
Small letter l	l	U+006C	6c
Small letter m	m	U+006D	6d
Small letter n	n	U+006E	6e
Small letter o	o	U+006F	6f
Small letter p	p	U+0070	70
Small letter q	q	U+0071	71
Small letter r	r	U+0072	72
Small letter s	s	U+0073	73
Small letter t	t	U+0074	74
Small letter u	u	U+0075	75
Small letter v	v	U+0076	76
Small letter x	x	U+0078	78
Small letter y	y	U+0079	79
Small letter z	z	U+007A	7a
Alef	ا	U+0627	d8 a7
Beh	ب	U+0628	d8 a8
Teh marbuta	ة	U+0629	d8 a9
Teh	ت	U+062A	d8 aa
Theh	ث	U+062B	d8 ab
Jeem	ج	U+062C	d8 ac
Hah	ح	U+062D	d8 ad
Khah	خ	U+062E	d8 ae

Dal	د	U+062F	d8 af
Thal	ذ	U+0630	d8 b0
Reh	ر	U+0631	d8 b1
Zain	ز	U+0632	d8 b2
Seen	س	U+0633	d8 b3
Sheen	ش	U+0634	d8 b4
Sad	ص	U+0635	d8 b5
Dad	ض	U+0636	d8 b6
Tah	ط	U+0637	d8 b7
Zah	ظ	U+0638	d8 b8
Ain	ع	U+0639	d8 b9
Feh	ف	U+0641	d9 81
Qaf	ق	U+0642	d9 82
Kaf	ك	U+0643	d9 83
Lam	ل	U+0644	d9 84
Meem	م	U+0645	d9 85
Noon	ن	U+0646	d9 86
Heh	ه	U+0647	d9 87
Waw	و	U+0648	d9 88
Alef maksura	ى	U+0649	d9 89
Yeh	ي	U+064A	d9 8a

III. The Quranic Hexadecimal (QuHex) Model

The Quranic Hexadecimal model aims to identify the value of each natural keyword related to the Quran, which could easily be decomposed to characters. The model consists of the following functions:

- QuHex will convert the user query to hexadecimal form to compare with the content in the index.
- The index will store the hexadecimal form of keywords from the Arabic language related to the same verse in the Al-Quran.
- The contents of the Al-Quran will be referenced with the help of experts in the Arabic language.

Fig. 1 illustrates the QuHex model.

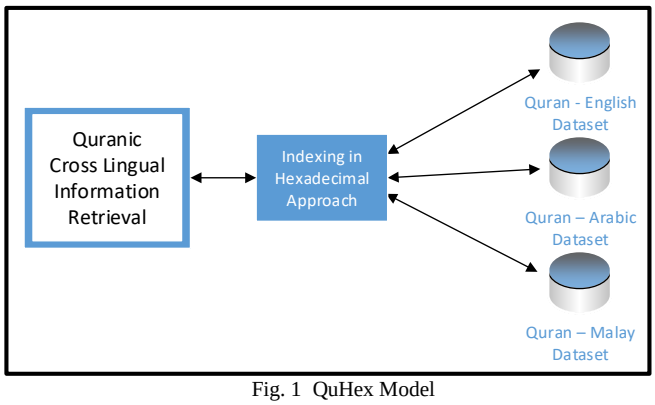

Fig. 1 QuHex Model

Fig. 1 shows that the Quran Hexadecimal (QuHex) Model which integrate different natural language to improve the readability on representation layer for the machine. The model is unique as the dataset is still in its natural form and the indexing gives the relation to the information in each dataset.

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:

$$k = h_1 + h_2 + \cdots + h_n \quad (1)$$

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

$$\boldsymbol{k} = \sum_{\mathbf{i}}^{\mathbf{n}} \boldsymbol{h_i} \quad (2)$$

Here, 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, الثمرات.

IV. Results

The sum of the four characters is shown in Table III with its translation in Malay and English.

Table III
Arabic/Malay/English Word in Hexadecimal Form

	Arabic	**Malay**	**English**
Word	نخيل	Kurma	date
Hex Value	36542	220	19E

Keyterm_ma	Keyterm_enc	Keyterm_Arab	Keyterm_Hex
pokok, kurma	date-palms, date-palm, Phoenix dactylifera	نَخِيلٍ	dbde, dde2, 1d515, 167a2, 12e36 21019, 5f1ed
benang, biji kurma	thread, date seed	فَتِيلًا	13f2d, ccd2 dde2, 14831, d8c6 d8c9, 6ca89
titik, biji kurma	speck, date seed	نَقِيرًا	e8d2, ccd2 dde2, d8d3, d8c6 d8c9, 6ca93

Fig. 2 Quran – Keywords

Fig. 2 shows that each language has its own respective keyword(s). At most-right column, the multiple languages keywords were 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. The hexadecimal conversion algorithm will be discussed in the next section.

The algorithm shown in Fig. 3 is used to improve the readability of natural language for machine based on hexadecimal values of a word.

Input: User query in Arabic/English/Malay
Output: Hexadecimal value of words

Start
User Query ← T
Tokenization word into characters, T = {c, c_2, … , c_u}
Check lookup table for Unicode
Character converted to Hexadecimal value
While (T != Ø)
  While (C_n != Ø)
    SumHex = ∑ (c, c_2, … , c_u)
  End Loop
End Loop
End

Fig. 3 The Hexadecimal Conversion Algorithm

TABLE 2
THE SIZE OF TEXT FILES BEFORE AND AFTER HEXADECIMAL

Surah	**Total Kalimah** (Without Stopwords)	**Arabic Text File (UTF-8)** (Bytes)	**Hex Text File** (Bytes)	**Total Reduction in Size (%)**
Al-Fatihah	26	332	159	47.89
Al-Baqarah	2210	26,378	14,412	54.63

From Table 4, we can find out that the size for hex value is smaller than the Arabic Text (surah al-fatihah and al-baqarah). The reduction of file size is around 47-54 %. This can be implied that the processing speed could also be improved as the space needed for data transfer are much smaller. This proves that system programming enhancement will improve the retrieval time, especially for MT technique.

V. DISCUSSION AND IMPLEMENTATION STRATEGY

The strategy of the implementation is divided into 2 phases:

A. Phase 1:

This phase will be implemented as an early experiment using only keywords. The proposed model, also known as QuHex, will be given the Quranic dataset in 3 different languages. Then, QuHex converts the keywords into a hexadecimal form which converts 3 different languages into the same column.

B. Phase 2:

The general model for the Quranic Cross-Lingual Information Retrieval (Q-CLIR) is parallel and intersects. This Q-CLIR will embed QuHex as an interpreter at indexing phase and converting a query to hexadecimal form in order to improve machine readability and reduce data requirements at query matching while maintaining original data sources. Figure 4 illustrates the immutable orthographic model of the Q-CLIR where the model can be read, searched and the dataset will be unchangeable. This research aims to improve data processing speed with word interpreter called QuHex.

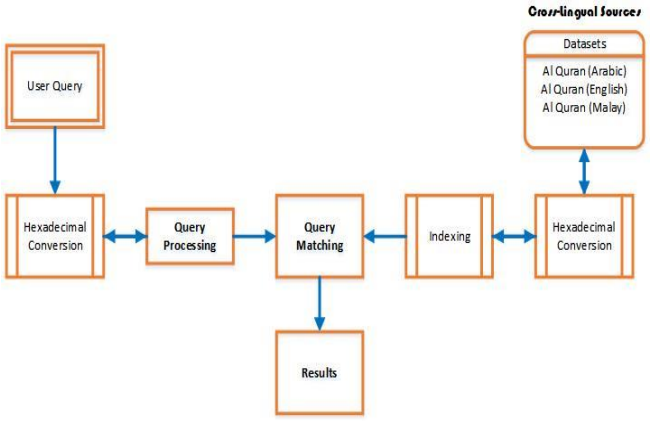

Fig. 4 Quranic Cross-Lingual Information Retrieval (Q-CLIR)

The Quranic Cross-Lingual Information Retrieval (Q-CLIR) is composed of left to right then to middle sections which comprise the following definition:

- The user query is a text-based query from the user in a form of words or sentences which can be in many forms or languages. The computer does not differentiate natural language.
- Hexadecimal Conversion is a programming algorithm which converts every word into its own unique hexadecimal value.
- Datasets are the core part which holds the information of the systems, which is the cross-lingual source like Arabic, English, and Malay languages.
- Indexing is structured information which holds the tagging of each data from the datasets.
- Query Matching is the techniques which match the query to the datasets to produce relevant results for the user.
-

VI. CONCLUSIONS AND FUTURE WORKS

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

Next, we also improve the space usage as the reduction of space size is around 47-54 %. This research hypothesizes that the higher the reduction of space usage, the faster the retrieval time of CLIR.

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 as implementing the second phase in creating interpreter for any word in a hex value.

ACKNOWLEDGMENT

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

[1] A. Basharat et al, "Comparative Study of Verse Similarity for Multi-lingual Representations of the Qur'an" in *Int'l Conf. Artificial Intelligence, ICAI'15*, 2015.

[2] A. Farag and A. Nurnberger, "Translation ambiguity resolution using interactive contextual information," *Computational Linguistics. Studies in Computational Intelligence*, vol. 458, pp. 219–240, 2013.

[3] M. Alqahtani and E. Atwell, "Arabic Quranic Search Tool Based on Ontology" in *21st International Conference on Applications of Natural Language to Information Systems*, 2016, pp. 478-485.

[4] A. Alshareef, A. E. Saddik, "A Quranic Quote Verification Algorithm For Verses Authentication," in *Innovations in Information Technology (IIT), 2012 International Conference*, 2012, pp. 339 – 343

[5] A. Arbaoui, Y. M. Alginahi and M. Menacer, "Strategies For Collecting Electronic Resources On The Qur'anic Researches," *International Journal On Quranic Research (IJQR)*, 3(4), pp. 57–78, 2013.

[6] A. A. Omoush, N. M. Norwawi, R. Ismail, F. Wahid & A. A. Mazlan. (2015). "Unicode Hexadecimal Representation for Quranic Words," in *The 4th FEIIC - International Conference on Engineering Education & Research*, 2015.

[7] B. Elayeb, "Arabic Cross-Language Information Retrieval: A Review," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 15(3), 18, Dec 2015.

[8] A. B. A. Bakar, "Evaluating the accessibility and visibility of Quran websites," in *International Symposium in Information Technology (ITSim)*, 2010, pp. 1-4.

[9] B. Elayeb and I. Bounhas, "Arabic cross-language information retrieval: a review," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol.15(3), 18, 2016.

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

[11] A. Pirkola, T. Hedlund, H. Keskustalo and K. Jarvelin, "Dictionary-Based Information Retrieval: Problems, Methods and Research Findings," Kluwer Academic Publishers. Information Retrieval, vol. 4, pp. 209-230, 2001.

[12] M. Aljlayl and O. Frieder, "Effective Arabic-English cross-language information retrieval via machine readable dictionaries and machine translation," in *Proceedings of the 2001 ACM International Conference on Information and Knowledge Management (CIKM'01)*, 2001, pp. 295–302.

[13] A. Lavie, K. Probst, E. Peterson, S. Vogel, L. Levin, A. Font-Llitjos and J. Carbonell, "A trainable transfer-based machine translation approach for languages with limited resources," in *Proceedings of Workshop of the European Association for Machine Translation*, 2004.

[14] A. Hatem and N. Omar, "Syntactic reordering for Arabic-English phrase-based machine translation." *Database Theory and Application, Bio-Science and Bio-Technology. Communications in Computer and Information Science*, vol. 118, pp. 198-206, 2010.