Journal of Sustainability Science and Management
Volume 20 Number 9, September 2025: 1792-1806
eISSN: 2672-7226

TRACKING CONCEPT-BASED STRUCTURES THROUGH REPETITIVE TEXT PATTERNS IN THE QURAN REVEALING POSSIBLE ONTOLOGICAL RELATIONSHIPS

NAZIA NISHAT[1], ROSALINA ABDUL SALAM[1]*, YUSUF MAHBUBUL ISLAM[2] AND ZULKIFLY MOHD ZAKI[1]

[1]Faculty of Sciences and Technology, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia. [2]School of Engineering, Technology and Sciences, Independent University, Bangladesh, 1229, Dhaka, Bangladesh.

**Corresponding author: rosalina@usim.edu.my*
http://doi.org/10.46754/jssm.2025.09.003
Submitted: 2 September 2024 Revised: 30 December 2024 Accepted: 24 February 2025 Published: 15 September 2025

Abstract: The Quran is a globally accessed sacred scripture from which many seek guidance. It has been asserted that the Quran explores comparable themes dispersed across its chapters, addressing multiple contexts. Prior studies have found that approximately 75% of its chapters contain recurring textual elements. However, apparent repetitions in the Quran have not been studied for potential interlinks such as creating an index of recurring text patterns or concepts to support contextual information retrieval. This study aims to advance the analysis of textual pattern repetitions in the Quran by devising a methodology to group and interlink recurring patterns, thereby exploring their relationship with the structure and presentation of the text. A consistent ratio of approximately 0.6 was identified between the number of repeated *n*-grams and the subsequent (n+1) grams, revealing a consistently converging "ring structure" among repeating text patterns. Findings show that repeating patterns enable the retrieval of 5WH ontological information on a basic repeating pattern. The proposed unsupervised information retrieval approach is independent of expert input and therefore offers greater sustainability.

Keywords: Concept-based structure, repetitive text patterns, ontological relationships, information retrieval, indexing.

Introduction

The Quran is acknowledged to possess an inherent repetitive structure (Oktaviani *et al*., 2019). Mathematical analyses specifically highlighted ring-like structures (Farrin, 2010; Ishak *et al*., 2020) and various textual patterns (Ebrahimi *et al*., 2012; Bentrcia *et al*., 2018), all of which point to the need for further exploration of possible relationships and interlinks among these patterns.

The relationship between such structures and the overall presentation of the Quran, as well as their role in providing additional guidance, remains largely unexplored. This gap in exploration limits understanding of how these structures may facilitate information retrieval, especially in extracting thematically related topics from this closed-domain text.

Existing studies demonstrate the statistical significance of patterns in the Quran (Bentrcia *et al*., 2018), employ word clouds to highlight frequently occurring words and patterns (Alhawarat *et al*., 2015), and propose a dataset of repeating patterns of up to 40 sequential words across give chapters chapters for further analysis (Oktaviani *et al*., 2019). Recent research (Bashir *et al*., 2023) has yet to thoroughly examine the possibility of an overarching organic structure, a critical direction previously proposed by El-Awa (2006). Both El-Awa (2006) and Oktaviani *et al*. (2019) stress the importance of understanding the Quran's structure for identifying connections between similar verses.

In classical exegesis such as *Tafsir Ibn Kathir*, clarification of specific verses is often achieved by referencing others. However, despite this traditional approach, the patterns identified in the Quran have not been applied in modern machine-based methods to identify similar verses to improve context-based information retrieval.

The Quran Offers Its Own Best Explanation

The science of the Quran interpretation is categorised into two types (Basyony, 2023): *Tafsir bil Riwaya*, which employs the Quran along with hadith to explain verses and *Tafsir bil Al-Ra'y*, also known as "ijtihad", which relies solely on other Quranic verses for interpretive support. Both approaches face the challenge of identifying verses that elucidate or provide additional information about a given concept. The presents a technological task: To facilitate the discovery, linkage, or indexing of such verses to enable context-based retrieval.

Sharaf and Atwell (2012b) attributed the complexity of the structure of Quranic verses to the frequent summarisation of concepts in one verse that are elaborated in others. Their earlier work (2012a) mapped approximately 1,050 concepts and 2,700 relations in *Tafsir Ibn Kathir* by assigning antecedents of pronouns. The objective of seeking explanatory verses and the challenge in using statistical similarity measures to find such verses can be illustrated by comparing the verses of *Surah* 44, Verse 3 (Ad-Dukhan 44:3) and *Surah* 97, Verse 1 (Al-Qadr 97:1). Table 1 presents both the original Arabic and the English translation.

As shown in Table 1, English translations yield only five common words of 15 words in *Surah* 44, verse 3 and five of 11 words in *Surah* 97, verse 1, making similarities difficult to detect. However, the Arabic reveals a stronger connection, with the longest common string (إنا أنزلناه في ليلة) forming a four-word pattern (four-grams). Four of the five words in *Surah* 97, verse 1, match with the additional word القدر or "power", enabling comparison for the purpose of *ijtihad* to determine which night is referenced and why it is considered blessed.

Table 1: Finding explanatory verses in Arabic and translation

Verses that Support Each Other	English Translation (Sahih English Translation)	Further Explanation Needed	Quranic Arabic Tanzil.net (Tanzil.net, 2022)	Longest Common Substring (LCS) Analysis
44.3	Indeed, We sent it down during a blessed night. Indeed, We were to warn [mankind].	Which is the blessed night? Why is the night blessed? These answers are not given in 44.3.	إنا أنزلناه في ليلة مباركة إنا كنا منذرين	إنا أنزلناه في ليلة
97.1	Indeed, We sent the Quran down during the Night of Decree.	The answer to the questions is found in this verse: Night of Decree. The night is blessed as Quran revelation starts. The answers required are found by comparing 97.1 with 44.3.	إنا أنزلناه في ليلة القدر	إنا أنزلناه في ليلة
What matches in the two verses?	"We sent… night" matches. The remaining words in the verses do not match, resulting in a poor match when employing a similarity measure.	Answers are identified through a comparison with a supporting verse in Arabic. In translation, matching the verses would pose a challenge.		From both Arabic verses the Longest Common Substring (text pattern) that matches is provided. The Arabic LCS or four-gram (four consecutive words) matches.

In such cases, Arabic text pattern matching —using bi-grams, tri-grams, or four-grams— proves more effective than vector-based word matching. In the case of these two verses, similarity matching using Arabic text patterns would be more effective. For example, in *Ayatul Kursi* (Al-Baqarah 2:255), which contains multiple concepts (as evidenced by different text patterns), Sharaf and Atwell (2012b) show that matching based on common patterns, rather than entire verses, enhances concept identification. Specifically, using cosine similarity, determining which concept or text pattern of a long verse should be matched presents a significant challenge.

As an initial step, this research draws on studies by Akour *et al*. (2014), Putra *et al*. (2018), and Oktaviani *et al*. (2019), which identified common text patterns or Longest Common Substrings (LCS) in Arabic. Although statistical measures such as Term Frequency-Inverse Document Frequency (TF-IDF) (Qaiser & Ali, 2018) is useful in demonstrating the frequent use of individual terms like "Allah", they do not help in identifying explanatory verses. Such statistics reflect term importance but not conceptual connections. Recognising that each repetition of a text pattern may contributes contextual or ontological information, identifying similar verses is crucial for a comprehensive understanding of a topic. This consideration is essential for effective information retrieval from the Quran.

Materials and Methods

Collecting Recurring Patterns in Verses of the Quran

The initial step involves tokenising all individual words in the Quran, followed by the identification of repetitive patterns. In the study by Oktaviani *et al*. (2019), Ukkonen's algorithm is employed to detect the LCS and count their occurrences, as illustrated in Figure 1. While this approach isolates and counts repeated LCS, it does not separately identify overlaps with common shorter patterns within longer substrings (matching text patterns).

For instance, as show in Figure 1, when comparing rows 5 and 7, the identified LCS appears nine times while a shorter pattern repeats 15 times. The smaller string "EalaY himo wa", which occurs 15 times is part of the larger string pattern "EalaY himo wa laA", which occurs nine times. This suggests a possible semantic structural relationship between the shorter and longer text patterns. The shorter string must therefore appear in six other verses. Recognising this shorter substring as a component of the longer one supports the establishment of a <part of> relationship, indicating a semantic connection between patterns, as illustrated in Figure 2. This <part of> relationship reflects how individual words are sequentially added to convey a complete meaning, independent of the Parts of Speech (POS) involved.

	A	B	C	D	E
1	jumlah_kata/ word count	lcs	first_verse	verses	verse_count
2	4	{l r~aHoma`n {l r~aHiym	1:1	1:1, 1:3, 2:163	3
3	3	rab~ {lo Ea`lamiyn	1:2	1:2, 2:131, 5:28	3
4	4	{ll~ah rab~ {lo Ea`lamiyn	1:2	1:2, 5:28	2
5	4	EalaY` himo wa laA	1:7	1:7, 2:38, 2:62, 2:112, 2:262, 2:274, 2:277, 3:170, 5:69	9
6	3	wa laA {l	1:7	1:7, 2:120, 5:2	3
7	3	EalaY` himo wa	1:7	1:7, 2:38, 2:62, 2:112, 2:160, 2:167, 2:262, 2:274, 2:277, 3:170, 4:6, 4:17, 5:69, 5:80, 5:117	15
8	3	l~i lo mut~aqiyn	2:2	2:2, 2:66, 3:138, 5:46	4

Figure 1: Repeating LCS text patterns identified using Ukkenon's algorithm (Oktaviani *et al*., 2019)

Ismail *et al*. (2017) attempted to identify contextual relations using the specific term "Allah" in the English translation of the Quran. However, keyword-based indexing fails to capture contextual meaning (Masri, 2020), as it does not account for how keywords are used within specific contexts during retrieval. As a result, such indexing often yields limited or irrelevant results.

Sedek *et al.* (2020) indexed verses based on chapter, *ayat*, and expert themes. Their main technical contribution was migrating the traditional database into a state-of-the-art serverless index. However, their study focused on the Malay translation and did not address the repetitive text-pattern structure of the Quran in Arabic. This presents an opportunity to index recurring meaningful phrases or text patterns in individual *ayats* in the original Arabic.

Similarly, Putra *et al*. (2018) enhanced the digital index of the Indonesian translation by incorporating 6,794 unigram words and 60,323 bi-grams, using TF-IDF to identify meaningful terms. However, merely calculating term frequency in the Quran does not provide access to the additional information contained in verses sharing common *n*-grams. The LCS patterns identified by Oktaviani *et al*. (2019) also do not account for all the smaller matching text patterns. Since both smaller and larger patterns recur independently throughout the Quran, this study proposes establishing a <part of> relationship connecting all repeating *n*-gram and (*n*+1)-gram patterns, as illustrated in Figure 2.

Use of n-grams

To enhance the available keyword index, Putra *et al*. (2018) used bi-grams, i.e., two adjacent word sequences. Similarly, Akour *et al*. (2014) employed four-gram patterns to differentiate between *ayats* revealed in Mecca and/or Medina. These studies suggest that the higher *n*-grams produce better classification and matching results. Akour *et al*. (2014) used Support Vector Machines (SVM) to classify theme and location, creating vectors for each four-gram and matching them with vectors representing entire *ayats*. However, longer *ayats* showed a poorer match. Therefore, matching only the recurring *n*-gram may be more effective, as the remainder of the *ayat* may merely elaborate on the concept expressed in the core text pattern.

Illustration of Possible Relationship between Recurring Patterns

The potential structural relationship between recurring patterns is depicted in Figure 2, where a shorter pattern, (e.g., *XY*) is considered a <part of> a longer pattern (e.g., *XYZ*, *XYA*, ...*XYB*). If the shorter pattern *XY* appears 15 times, all (*n*+1)-gram occurrences of containing *XY*, unless *XY* is found in a verse with only two words or occurs at the end of a verse, should collectively total 15, as illustrated in Figure 2. Similarly, each sequential pattern (e.g., *XYZ*) may appear within larger (*n*+1)-gram patterns (e.g., *XYZA*, *XYZB*, …*XYZm*) and their total instances should match the frequency of *XYZ*, except where *XYZ* occurs in a three-word verse on ends a verse.

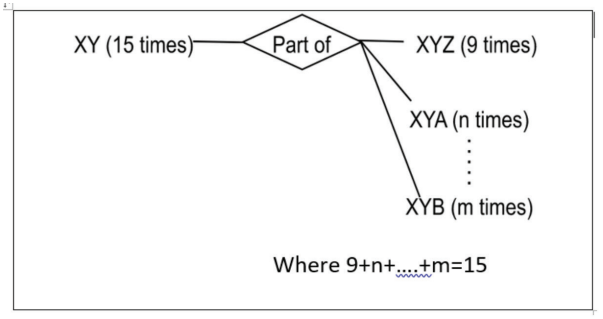

Figure 2: An example of <part of> links between patterns

Since this research addresses text patterns at the word level (rather than the character level), each word is referred to as a "gram". While the term *n*-gram is commonly associated with Markov-based next-word prediction (Patel, 2022), it is used here simply to denore the number of sequential words in a text pattern. Drawing on the concept of (*n*+1)-gram prediction, the <part of> relationship in this study refers to the link between an *n*-gram and (*n*+1)-gram words. Rather than predicting the probability of the next word or gram, the <part of> relationship aims to link all possible next words, showcasing how and which next words are utilised to string together each subsequent word, revealing the presentation of a verse in the closed-domain text of the Quran.

This research proposes to track the incremental development of repeating bi-gram concepts with each additional word by connecting repeating text patterns found throughout the Quran, intending to capture its underlying concept-based structure. The <part of> relationship, as depicted in Figure 2, signifies that a bi-gram is a <part of> all corresponding tri-gram patterns, whether repeated or not.

This relationship differs from that proposed by Saad *et al*. (2013), which is based on POS. For example, if the string "i.e.," appears, it precedes an example such as "pray i.e., funeral prayer", where "funeral prayer" is part of "pray". In the present research, however, the <part of> links to the subsequent text without consideration of the POS i.e., simply a sequential text pattern. Each subsequent word is added to the bi-gram concept phrase to present a complete narrative in each verse. Thus, the initial bi-gram becomes a <part of> the tri-gram phrase, which in turn becomes a <part of> the four-gram phrase and so on. Since each n-gram (bi-grams, tri-grams, four-grams, etc.) is repeated in the entirety of the closed domain text of the Quran, they are therefore treated as repeating text patterns.

Bi-grams serve as a starting point for matching text patterns given their increased semantic specificity compared with unigrams as demonstrated by Putra *et al*. (2018), making them more accurate in feature representation for classification (Elghannam, 2021) by providing a richer context. Utilising POS to identify grammatical patterns requires an in-depth study of various language structures (Saad *et al*., 2013). Extracting individual repeated terms such as "Allah" excludes the surrounding text (Ismail *et al*., 2017) that contributes to specific meaning. For example, the term "fear Allah" forms a collocation (Ebrahimi *et al*., 2012), aiding in concept extraction. However, for phrases like "purify Allah", the concept may become ambiguous. Hence, enhancing term combinations for context clarification and incrementally expanding them is essential to reveal the complete meaning.

To explore the repetitive text structure of the Quran, this study proposes analysing the <part of> relationships between smaller and larger patterns across the entire Quran. Documenting the <part of> relationships between patterns may help elucidate the Quran's structural organisation.

Use of Standard n-gram Extraction Algorithms

Oktaviani *et al*. (2019) applied Ukkonen's algorithm to analyse five *surahs* of the Quran for identifying the LCS. However, because Ukkonen's algorithm operates at the character level, extending the analysis to the entire Quran with its 114 *surahs* proved excessively time-consuming. Moreover, smaller common text patterns within the LCS were not separately identified, as illustrated in Figure 1.

To improve processing efficiency and to identify all repeating or common text patterns, not just the LCS, standard algorithms (Hashmi, n.d.) were employed. To optimise processing time, each *n*-gram extraction routine was executed separately for each *n*-value (n = 2, 3, 4, 5, 6, 7, 8, 9). A dedicated routine was subsequently developed to identify all matching *n*-grams and calculate the total number of repeating sets of *n*-grams. The results for the number of repeating *n*-grams are presented in Figure 3.

Analysis and Discussion

Exploration of Relationships between Text Patterns

Figure 3 reveals a consistent convergence in the number of repeating *n*-gram patterns as *n* increases. The converging ratio remains consistently around 0.653, suggesting that repeating text patterns may be an intentional feature of the Quran. To demonstrate the consistent reduction, a Geometric Progression (GP) curve fit was employed, yielding the following formula:

$$R_n = 9019*0.653\text{^}(n\text{-}2) \quad \text{(i)}$$

In the GP formula (i), *n* is the number of sequential words being considered, R_n is the number of repeating text patterns found, and 0.653 is the average ratio between successive *n*- and (*n*+1)-grams. A total of 9,019 repeating sets of unique bi-gram word patterns were found, forming the first term in the geometric progression. As this is the first term, the index of the ratio has to be *n*-2. The maximum deviation of the GP formula is 8.9%, as shown superimposed on the bar graph in Figure 3. For example, the bi-gram السماوات والأرض (heavens and the earth) appears 133 times and the bi–gram ان الله (God is) appears 257 times—each counted as one of the 9,019 repeating bi-grams in the Quran.

Furthermore, to explore possible relationships between the *n*-gram repeating patterns, various ratios between the *n*-grams, as presented in Table 2 were calculated. Considering each group of repeating *n*-grams as a unit, the ratio of total repeating (*n*+1)-grams and *n*-grams was first calculated as serial (i) in Table 2.

- Serial (i) in Table 2 presents the ratio of total repeating (*n*+1)-grams to *n*-grams. Each successive ratio approximates a consistent value of around 0.6. Collectively, successive numbers of repeating *n*-grams reduce with a consistent ratio.
- Serial (ii) presents the ratio of [(*n*+1)-gram + *n*-gram] to *n*-gram, averaging 1.64 for all *n*-grams from 2 to 6. Hence, a consistent ratio approximating 1.6, popularly known as the golden ratio or divine ratio (Cuemath, 2022) is found between the number of common or repeated *n*-grams and (*n*+1)-grams.

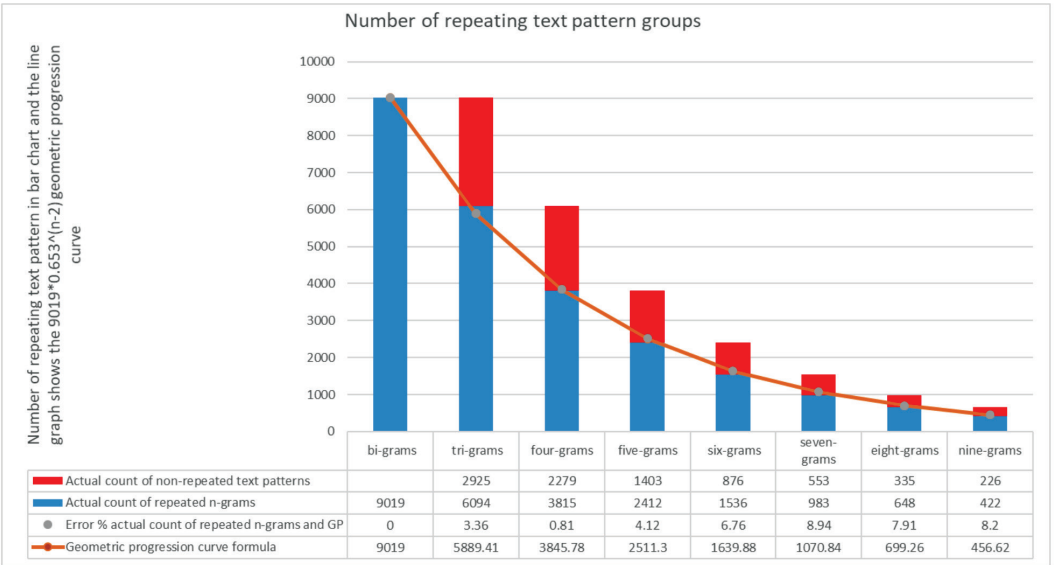

Figure 3: Number of repeating text pattern groups

Table 2: Ratios calculated between the repeating and increasing *n*-grams

SI.	Ratio Formula with Repeating *n*-grams	*n* = 2	*n* = 3	*n* = 4	*n* = 5	*n* = 6
i)	$\frac{(n+1)-gram}{n-gram}$	$\frac{6094}{9019} = 0.68$	$\frac{3815}{6094} = 0.63$	$\frac{2412}{3815} = 0.63$	$\frac{1536}{2412} = 0.64$	$\frac{983}{1536} = 0.64$
ii)	$\frac{[(n+1)-gram+n-gram]}{n-gram}$	$\frac{6094+9019}{9019}$ = 1.68	$\frac{3815+6094}{6094}$ = 1.63	$\frac{2412+3815}{3815}$ = 1.63	$\frac{1536+2412}{2412}$ = 1.64	$\frac{983+1536}{1536}$ = 1.64
iii)	$\frac{n-gram}{bi-gram}$	$\frac{9019}{9019} = 1.0$	$\frac{6094}{9019} = 0.68$	$\frac{3815}{9019} = 0.42$	$\frac{2412}{9019} = 0.27$	$\frac{1536}{9019} = 0.17$

- Serial (iii) shows the ratio of bi-gram to a higher specific *n*-gram, revealing that the number of matching clusters diminishes with increasing *n*-grams. As the ratio of bi-gram to six-gram falls to less than 20%, the study is limited to patterns from bi-grams to five-grams. However, similar convergence trends are observed even at *n* = 9.

Unique Text Pattern Structure of Al-Quran

To verify the uniqueness of the consistent convergence of the repeating *n*-gram structure found in the original Arabic text of the Quran, a comparison was made with other closed-domain data sources, as shown in Table 3. The first dataset used for comparison was the Sahih International English translation of the Quran, sourced from Tanzil.net (Tanzil.net, 2022). In the preprocessing stage, punctuation was removed and all uppercase texts were converted to lowercase to ensure uniformity in *n*-gram representations, considering the fact that the Arabic text does not distinguish between uppercase and lowercase.

The second comparison dataset was the King James Version (KJV) of the Bible, obtained from the Kaggle.com corpus uploaded by Hartono (2017). Table 3 presents the ratios of successive repeating *n*-grams across the three datasets.

The data show that, even in the English translation of the Quran, the consistency observed in the Arabic original is lost. This discrepancy is likely due to structural differences between Arabic and English. In the Arabic Quran, the successive ratio of repeating *n*-grams to (*n*+1)-grams consistently remains around 0.6. In contrast, the English translation of the Quran exhibits more variation, with ratios ranging from > 1 to approximately 0.7.

Similarly, the KJV Bible also displays fluctuating ratios from > 1, dropping to 0.6 and then rising to 0.7. In contrast, the Arabic Quran maintains a steady decrease in successive *n*-gram ratios, a pattern not mirrored in the translated Quran and the Bible, as indicated in Table 3.

Table 3: Comparison of successive repeating *n*-gram ratios in the Arabic Quran, its English translation, and the KJV Bible

Repeating *n*-gram Ratio	Arabic Quran	Sahih International English Translation of the Quran	King James Version of the Bible	The value of *n* in *n*-grams
$\frac{(n+1)-gram}{n-gram}$	0.68	1.29	1.49	n = 2
	0.63	0.81	0.76	n = 3
	0.63	0.72	0.65	n = 4
	0.64	0.72	0.67	n = 5
	0.64	0.73	0.71	n = 6

These findings strongly support the assertion that the repeating pattern structure of the Quran in Arabic is unique. El-Awa (2006) characterised the structure as "organic" and suggested that this inherent structure must be considered for successful information retrieval.

Unique Linguistic Structure of the Arabic Quran

The variation in repeating *n*-grams not only differs in the translation of the Quran, but is also evident in the count of unique repeating groups of *n*-grams. For instance, the number of repeating bi-grams in the Arabic Quran is 9,019, whereas in the English translation, it is 14,656. An analysis of bi-grams on specific topic sheds light on this discrepancy. The difference primarily stems from the distinct characteristics of the English language, where prepositions and articles are separate and may form bi-grams.

For example, the Arabic unigram الصلاة is translated as the English bi-gram "the prayer", increasing the overall bi-gram count. Examining specific bi-gram words further illustrates this point. In the Sahih translation, the bi-gram "of the" appears 969 times while in the KJV Bible (Hartono, 2017), it appears 11,528 times. The prevalence of the phrase "of the" in English, forming a large number of unconnected tri-gram words, contributes to a comparatively lower count of unique repeating bi-grams.

Moreover, the likelihood of bi-grams being common across the English translation of the Quran is higher due to linguistic structure. Common English bi-grams such as "of the", "those who", "and the", "do not", "will be", "is the", and "in the", occur repeatedly, forming a variety of repeated tri-grams with the same bi-grams. As these common bi-grams are grouped together, the count of bi-grams is lower than that of tri-grams. For example, the bi-gram "of the" may appear in "of the heaven", "of the people", or "of the worlds", resulting in multiple tri-grams unrelated to one another. This results in the count of repeated bi-grams being higher than that of repeated tri-grams. Consequently, the number of repeated *n*-grams initially rises before decreasing, leading to the apparent anomaly of the ratio of repeated bi-grams to repeated tri-grams being greater than 1, as shown in Table 3.

Expressing the Repeating Text Pattern Structure of the Quran

For each repeating *n*-gram, the (*n*+1) word may either repeat or be unique. Taking $n = 2$ and $n+1 = 3$ as an example, the third word following a repeating bi-gram may or may not match with other third words for the same bi-gram. Once a repeating bi-gram word is identified (of which there are 9,019, consisting of 4,110 unique single words), it is treated as a repeating text pattern, which may or may not be semantically complete. For example, the Arabic bi-gram ان الله ("God is") appears 257 times, but is semantically incomplete while السماوات والأرض ("heavens and the earth"), which appears 133 times, conveys a complete concept of creation.

As a direction for future research, the 9,019 repeating bi-grams (or their combined root word equivalents) may be compared with the 1,050 concepts identified in *Tafsir Ibn Kathir*, as counted by Sharaf and Atwell (2012a).

Each repeating *n*-gram, along with the next (*n*+1) word is considered as a family of concepts of the bi-gram cluster, as illustrated in Figure 4. Within a repeating *n*-gram cluster, the (*n*+1) word may either repeat or be unique. All repeating (*n*+1) words are collected and classified together as a sub-cluster, each with a matching (*n*+1) word as shown in Figure 4. For the scope of this research, the sub-cluster with all singular (*n*+1) word is not considered, as the focus is on what retrieved verses with repeating patterns may reveal.

By tracking the bi-gram pattern through incremental *n*-grams, each *n*-gram is linked to the (*n*+1) word to create the subsequent (*n*+1)-gram. Figure 4 visually depicts the overall scenario of repeating (matching) and unique (non-matching) (*n*+1) words for repeating *n*-gram clusters. The figure also illustrates how a main *n*-gram cluster (denoted as *x*) has two sub-clusters with different repeating (*n*+1) words. When the (*n*+1) word matches, it becomes part of the subsequent cluster for the (*n*+1)-gram.

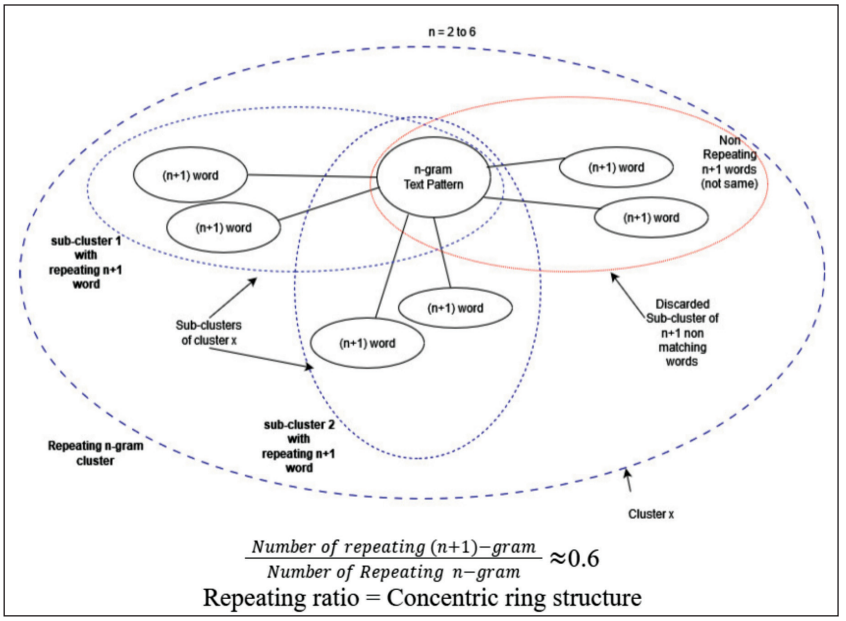

Figure 4: Reducing concentric ring structure of repeating text patterns, checked for $n = 2$ to $n = 9$

To explore how the overall converging structure (Figure 3) applies to the repeating 9,019 bi-grams, it is noteworthy that this same converging pattern is reflected in the larger bi-grams tested. For example, both individual bi-grams as shown in Figures 5 and 6, السماوات والأرض ("heavens and the earth") and ان الله ("God is") demonstrate similar converging

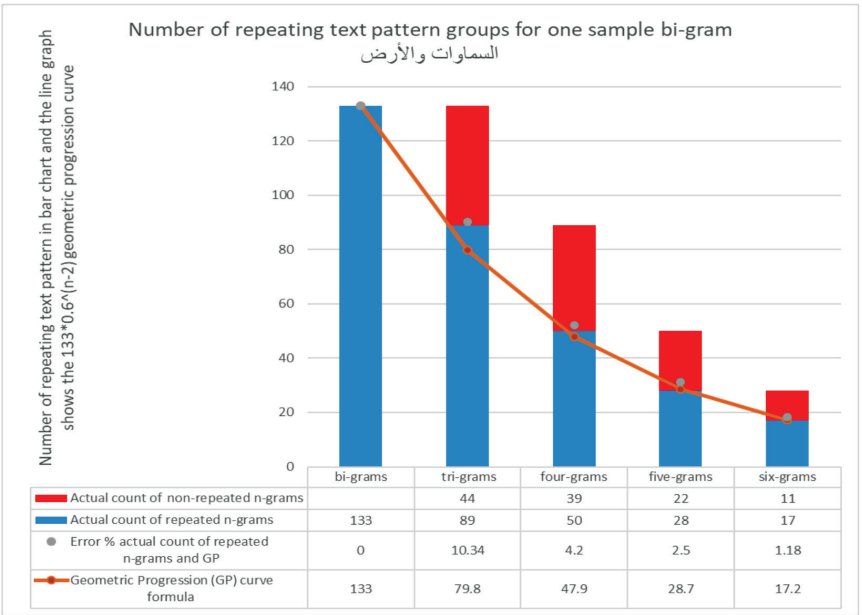

Figure 5: Number of repeating and non-repeating successive text pattern groups for the sample bi-gram السماوات والأرض

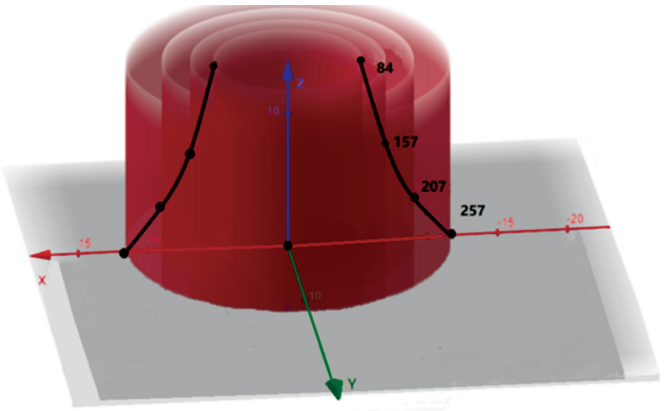

Figure 6: Converging common higher repeating patterns for "God is" (ان الله) in the Quran

behaviour in different contexts. The repeating *n*-grams are again treated collectively as a unit, supporting the idea that *ayats* sharing the same repeating text patterns may provide ontological information in a converging basis on the bi-gram topic. Calculating the successive ratios between repeating *n*-grams to (*n*+1)-grams reveals a consistent value of approximately 0.6, as aligning with the overall pattern detailed in Table 2 (i). Each *n*-gram has its own set of sub-clusters and all the sub-clusters together form an *n*-gram family of concepts. The repeating *n*-gram ratio was examined for $n = 2$ to $n = 9$, the structure is generalised in Figure 4 to represent the repeating text patterns observed across the Quran for the tested range.

Results

Using the Repeating Structure to Find Verses that Provide Contextual Information

As an example, the bi-gram السماوات والأرض ("heavens and the earth") from *Ayatul Kursi* (Al-Baqarah 2:255) is used to examine whether the Quran's repeating structure can help retrieve verses that provide additional information. Using the unique pattern-based approach, 25 different repeating-pattern clusters of related verses, ranging from bi-grams to five-grams were identified. These are detailed in Table 4.

Each cluster provides insight into a distinct aspect of the multiword expression "heavens and the earth", offering guidance and ontological information about the creation of the heavens and the earth. Sharaf and Atwell (2012b) noted that verses containing important multiword expressions are difficult to match when the full verse is compared using statistical techniques such as cosine similarity or SVM, potentially overlooking explanatory verses. By contrast, Bentrcia *et al*. (2018) directly extracted all 133 verses containing the conjunctive phrase "heavens and the earth" to analyse the relationship between the nouns of the phrase.

Table 4 captures the repeated phrases with السماوات والأرض ("heavens and the earth") within the Quran and follows the repeated patterns based on the structure. It demonstrates how a relevant phrase from *Ayatul Kursi* recurs in other verses, each presenting different contexts or additional ontological insights.

The bi-gram السماوات والأرض ("heavens and the earth") appears in the repeated tri-gram السماوات والأرض وما (23 repetitions), which itself extends to the four-gram السماوات والأرض وما بينهما (17 repetitions), and further into the five-gram السماوات والأرض وما بينهما في (three repetitions). These five-grams occur *Surah* 25, Verse 59; *Surah* 32, Verse 4; and *Surah* 50, Verse 38. Collectively, these three verses address fundamental questions such as who created the heavens and the earth and how?

In Cluster 1 of Table 4, *Surah* 25, Verse 59; *Surah* 32, Verse 4; and *Surah* 50, Verse 38 describe the creation of the heavens and the

Table 4: How the structure of the Quran handles a random phrase السماوات والأرض found in Ayatul Kursi

Bi-Gram Concept Taken from 2:255	Repeated Tri-Grams on The Concept Found	Repeated Four-Grams on The Concept Found	Repeated Five-Gram on The Concept Found	The Resulting Verse Cluster Found	Verse Cluster Serial Number	Basic Questions that the Clusters Appear to Answer
السماوات والأرض	السماوات والأرض وما	السماوات والأرض وما بينهما	السماوات والأرض وما بينهما في	25:59, 32:4, 50:38	1	How did Allah create heaven and earth?
	السماوات والأرض وما	السماوات والأرض وما بينهما	السماوات والأرض وما بينهما إلا	15:85, 30:8, 46:3	2	Why did Allah create heaven and earth? What was the purpose of creation?
	السماوات والأرض وما	السماوات والأرض وما بينهما	السماوات والأرض وما بينهما إن	26:24, 44:7	3	Who is Allah in relation to the heavens and the earth?
	السماوات والأرض بالحق	السماوات والأرض بالحق إن	No repeated five-gram concept found	14:19, 29:44	4	Why did Allah create heaven and earth?
	السماوات والأرض وهو	السماوات والأرض وهو العزيز	السماوات والأرض وهو العزيز الحكيم	30:27, 45:37, 57:1, 59:24	5	What is the dominion of Allah?
	السماوات والأرض في	السماوات والأرض في ستة	السماوات والأرض في ستة أيام	7:54, 10:3, 11:7, 57:4	6	How long did it take to create heavens and the earth?
	السماوات والأرض والله	السماوات والأرض والله على	السماوات والأرض والله على كل	3:189, 85.9	7	Who controls everything?
	السماوات والأرض وكان	السماوات والأرض وكان الله	السماوات والأرض وكان الله عليما	4:170, 48:4	8	What are the attributes of Allah?
	السماوات والأرض ليقولن	السماوات والأرض ليقولن الله	السماوات والأرض ليقولن الله قل	31:25, 39:38	9	What do disbelievers say about the creator?
	السماوات والأرض كل	السماوات والأرض كل له	السماوات والأرض كل له قانتون	2:116, 30:26	10	How everything is in Allah's control?

Who is the owner of the heavens and the earth?	11	7:158, 7:187, 57:10	No repeated five-gram concept found	No repeated four-gram concept found	السماوات والأرض لا
What are the signs? Who can understand signs?	12	2:164, 3:190	السماوات والأرض واختلاف الليل والنهار	السماوات والأرض واختلاف الليل	السماوات والأرض واختلاف
How long would people live in paradise?	13	11:107, 11:108	السماوات والأرض إلا ما شاء	السماوات والأرض إلا ما	السماوات والأرض إلا
Is there any other deity?	14	13:16, 34:24	No repeated five-gram concept found	السماوات والأرض قل الله	السماوات والأرض قل
To whom belongs to the dominion of heaven and earth?	15	24:42, 57:5	No repeated five-gram concept found	السماوات والأرض وإلى الله	السماوات والأرض وإلى
How does Allah nurture us?	16	14:32, 27:60	No repeated five-gram concept found	No repeated four-gram concept found	السماوات والأرض وأنزل
Who has power over life and death?	17	9:116, 57:2	No repeated five-gram concept found	السماوات والأرض يحيي ويميت	السماوات والأرض يحيي
Who knows the secrets of heaven and the earth?	18	25:6, 35:38	No repeated five-gram concept found	No repeated four-gram concept found	السماوات والأرض إنه
What is the relation between Allah and believers?	19	21:19, 23:71	No repeated five-gram concept found	No repeated four-gram concept found	السماوات والأرض ومن
Can anything be hidden from Allah?	20	27:25, 64:4	No repeated five-gram concept found	السماوات والأرض ويعلم ما	السماوات والأرض ويعلم

السماوات والأرض والذين	No repeated four-gram concept found	No repeated five-gram concept found	29:52, 39:63	21	Who are the losers?
السماوات والأرض طوعا	السماوات والأرض طوعا وكرها	No repeated five-gram concept found	3:83, 13:15	22	How does the entire creation submit to Allah?
السماوات والأرض ولا	No repeated four-gram concept found	No repeated five-gram concept found	2:255, 18:51	23	Does Allah need any helper?
السماوات والأرض لآيات	No repeated four-gram concept found	No repeated five-gram concept found	10:6, 45:3	24	Why are the signs given?
السماوات والأرض ولم	No repeated four-gram concept found	No repeated five-gram concept found	25:2, 46:33	25	What is the power of Allah?

earth in six days. Cluster 2, which contains *Surah* 15, Verse 85; *Surah* 30, Verse 8; and *Surah* 46, Verse 3, discusses the purpose of creation, i.e.,. why did Allah create? The answer that can be extracted is to "establish the truth about Himself". These examples illustrate how clusters generated using the repeating structure can effectively lead to verse groupings that answer basic ontological questions.

By contrast, cosine similarity, which statistically assesses relatedness between verse pairs, fails to deliver this structural insight. The repeating-pattern approach, therefore, offers a more meaningful way to match semantically rich multiword expressions, which traditional statistical methods struggle to align.

Analysis of each five-gram sub-cluster (right-most column of Table 4) suggests that basic 5WH questions concerning السماوات والأرض (heavens and the earth) may be answered. Collectively, the retrieved answers indicate the potential to construct a TBox ontology (Nasution, 2018) based on each bi-gram topic.

Conclusions

Extending *n*-grams from $n = 2$ to $n = 6$ reveals a consistent ratio of approximately 1.6 (commonly referred to as the golden ratio or divine ratio) between the number of repeated *n*-grams and (*n*-1)-grams, suggesting a structural regularity not observed in other texts. Upon further examination, repeating *n*-grams and their associated (*n*+1)-grams form a reducing concentric "ring structure", where successively smaller clusters emerge.

By tracing this structural pattern, it is found that specific information on the repeating bi-grams can be found in retrieved verses. The verses uncovered reveal various details about the important bi-gram phrase, which may support the potential development of a 5WH TBox ontology for each bi-gram topic. The structure offers a method to extracting meaningful verses that may aid in *ijtihad*, i.e., narrowing down the search to understand a topic using verse clusters related to it.

Further analysis may also help to investigate ontological information. Convergence, i.e. the reduction of repetitions in the number of repeated concepts, indicates that some concepts have no more repetitions. The narrowed-down common patterns may highlight verses suitable for *ijtihad*.

To utilise and allow the study of *ayats* in converging sub-clusters, it is recommended that an algorithm be developed to identify and link or index common LCS word patterns to support information retrieval from the entire Quran. The output could be compared with various ontological research conducted on the Quran.

Acknowledgements

This work is funded by the Universiti Sains Islam Malaysia (USIM).

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

Akour, M., Alsmadi, I. M., & Alazzam, I. (2014). MQVC: Measuring Quranic verses similarity and sura classification using N-gram.

Alhawarat, M., Hegazi, M., & Hilal, A. (2015). Processing the text of the Holy Quran: A text mining study. *International Journal of Advanced Computer Science and Applications*, *6*(2), 262-267.

Bashir, M. H., Azmi, A. M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., & Qadir, J. (2023). Arabic natural language processing for Qur'anic research: A systematic review. *Artificial Intelligence Review*, *56*(7), 6801-6854.

Basyony, S. (2023). *What is Tafsir? – Meaning, history, importance, types, books, and more!* https://bayanulquran-academy.com/what-is-tafsir-in-islam/

Bentrcia, R., Zidat, S., & Marir, F. (2018). Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns. *Journal of King Saud University-Computer and Information Sciences*, *30*(3), 382-390.

Cuemath. (2022). *Golden ratio*. https://www.cuemath.com/commercial-math/golden-ratio/

Ebrahimi, S., Pahlavannezhad, M. R., & Nadernezhad, G. (2012). The analysis of Quranic collocations in the Orchard" Boostan of Sa'di". *International Journal of Linguistics*, *4*(3), 274.

El-Awa, S. M. (2006). *Textual relations in the Qur'an: Relevance, coherence and structure.* Routledge.

Elghannam, F. (2021). Text representation and classification based on the bi-gram alphabet. *Journal of King Saud University-Computer and Information Sciences*, *33*(2), 235-242.

Farrin, R. K. (2010). Surat al-Baqara: A Structural Analysis* muwo_1299 17. 32. The Muslim World, 100.

Hartono, O. R. (2017). *Bible Corpus*. https://www.kaggle.com/datasets/ oswinrh/bible

Hashmi, F. (n.d.). *Thinking Neuron*. https://thinkingneuron.com/how-to-generate-n-grams-in-python/

Ishak, N. D., Kilicman, A., Husain, S. K. S., & Din, R. (2020). Mathematical Wondrous in the Al-Quran through Surah Al-Alaq. *Journal of Personalised Learning*, *3*(1), 31-39.

Ismail, R., Abd Rahman, N., & Bakar, Z. A. (2017). A pattern for concept identification from English translated Quran. In *MATEC Web of Conferences* (Vol. 135, p. 00067). EDP Sciences.

Masri, N. (2020). *An innovative automatic indexing method for Arabic text* [Master's Thesis, Lebanese American University].

Nasution, M. K. (2018). Ontology. *Journal of Physics: Conference Series*, *1116*, 022030.

Oktaviani, D., Bijaksana, M. A., & Asror, I. (2019). Building a database of recurring text in the Quran and its translation. *Procedia Computer Science*, *157*, 125-133.

Patel, V. (2022). *Markov Chain Explained*. https://builtin.com/machine-learning/markov-chain

Putra, S. J., Gunawan, M. N., & Suryatno, A. (2018). Tokenisation and N-Gram for Indexing Indonesian translation of the Quran. In *2018 6th International Conference on Information and Communication Technology (ICoICT)* (pp. 158-161). IEEE.

Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29.

Saad, S., Noah, S. A. M., Salim, N., & Zainal, H. (2013). Rules and natural language patterns in extracting Quranic knowledge. In *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences* (pp. 381-386). IEEE.

Sedek, K. A., & Osman, M. N. (2020). Quran indexing using Cloud Database. In *Charting the sustainable future of ASEAN in science and technology* (pp. 37-47). Singapore: Springer.

Sharaf, A. B. M., & Atwell, E. (2012a). QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *LREC* (pp. 130-137).

Sharaf, A. B. M., & Atwell, E. (2012b). QurSim: A corpus for evaluation of relatedness in short texts. In *LREC* (pp. 2295-2302).

Tanzil.net (2022). Download the Quran Text. https://tanzil.net/download/