Understanding the thematic structure of the Qur’an: an exploratory multivariate approach

Naglaa Thabet
School of English Literature, Language and Linguistics
University of Newcastle
Newcastle upon Tyne, UK, NE1 7RU
n.a.thabet@ncl.ac.uk

Abstract

In this paper, we develop a methodology for discovering the thematic structure of the Qur’an based on a fundamental idea in data mining and related disciplines: that, with respect to some collection of texts, the lexical frequency profiles of the individual texts are a good indicator of their conceptual content, and thus provide a reliable criterion for their classification relative to one another. This idea is applied to the discovery of thematic interrelationships among the suras (chapters) of the Qur’an by abstracting lexical frequency data from them and then applying hierarchical cluster analysis to that data. The results reported here indicate that the proposed methodology yields usable results in understanding the Qur’an on the basis of its lexical semantics.

1 Introduction

The Qur’an is one of the great religious books of the world, and is at the heart of Islamic culture. Careful, well-informed interpretation of the Qur’an is fundamental both to the faith of millions of Muslims throughout the world, and to the non-Islamic world’s understanding of their religion. There is a long tradition of scholarly quranic interpretation, and it has necessarily been based on traditional literary-historical methods of manual textual exegesis. However, developments in electronic text representation and analysis now offer the opportunity of applying the technologies of newly-emerging research areas such as data mining (Hand et al., 2001) to the interpretation of the Qur’an. Studies on computational analyses of the Qur’an are almost lacking. Contributions to this field include the development of a morphological analyser for the Qur’an (Dror et al., 2004).

The Qur’an consists of 114 chapters called suras which range in length from the shortest, Al-Kawthar, consisting of 4 ayat (verses) to the longest, Al-Baqarah, consisting of 286 ayat. There is no obvious reason why the suras are sequenced as they are in the text. They are not in chronological order, and seem, in fact, to be ordered roughly by length, from longest at the beginning of the text to shortest at the end. Given this, apparently arbitrary sequencing, one of the first steps in interpreting the Qur’an as a whole must be to discover thematic interrelationships among the suras. The present paper proposes a methodology for doing this using exploratory multivariate analysis.

The paper is in five parts; the first part is the introduction. The second presents the quranic text and the data preparation prior to the analysis. The third part deals with the application of cluster analysis techniques to the Qur’an and the interpretation of the results. The fourth part draws the conclusion and suggests future research to be undertaken.

2 Data

The data for this study is based on an electronic version of the Qur’an produced by Muslimnet[1]. This version is a Western alphabetic transliteration of the Arabic orthography. The data is transliterated into Latin based ASCII characters, mostly with single-symbol equivalents of the Arabic phonemes and by replacing diacritics and

[1] http://www.usc.edu/dept/MSA/quran/transliteration/

Proceedings of the ACL Student Research Workshop, pages 7–12,
Ann Arbor, Michigan, June 2005. ©2005 Association for Computational Linguistics

glyphs which represent short vowels in Arabic orthography with appropriate Roman letters.

A frequency matrix F is constructed in which the rows are the suras, the columns are lexical items, and every cell F_{ij} contains an integer that represents the number of times lexical item j occurs in sura i. Construction of such a matrix is straightforward in principle, but in practice some well known issues arise.

2.1 Tokenization

Given that one wants to count words, what is a word? The answer is surprisingly difficult, and is a traditional problem both in linguistics and in natural language processing (Manning and Shütze, 1999). Even confined to written language, as here, two issues arise:

- Word segmentation: In English text, the commonsensical view that a word is the string of letters bounded by punctuation and/or white space is quite robust, but it is less so for other languages.
- Stemming: In languages with a significant element of morphological elaboration of words stems, do the various morphological variants of a given stem count as different words?

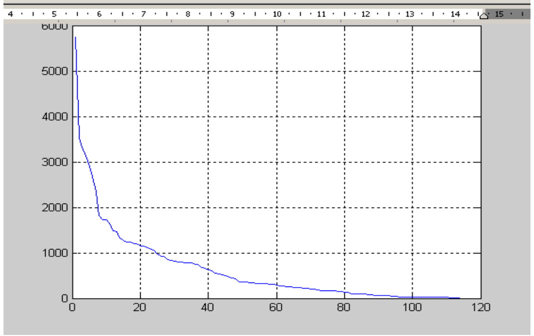
For present purposes, the words segmentation problem is easily resolved in that the Qur'an's orthography is such that words can be reliably identified using the 'string of letters between punctuation and/or white space' criterion. With regard to stemming, morphological variants are treated as single word types, and to achieve this, the electronic text of the Qur'an was processed using a purpose-built stemmer whose characteristics and performance are described in Thabet (2004).

2.2 Keyword selection

Function words like determiners and prepositions were removed from the text, and only content words were retained. In addition, the (many) words with frequency 1 were removed, since these cannot contribute to determination of relationship among suras.

2.3 Standardization for text length

The introduction noted that the suras vary in length from fewer than a dozen to several thousand words. The following plot of number of content words per sura, sorted in order of descending magnitude.



"Figure 1. Plot of number of words per sura"

Clearly, given a word with some probability of occurrence, it is more likely to occur in a long text than a short one. In order to compare the suras meaningfully on the basis of their word frequency profiles, the raw frequencies have to be adjusted to compensate for sura length variation. This was done on the following basis:

$$freq'(F_{ij}) = freq(F_{ij}) \times \left(\frac{\mu}{l}\right)$$

where *freq*' is the adjusted frequency, F_{ij} is the value at the (i,j) coordinates of the data matrix F, *freq* is the raw frequency, μ is the mean number of words per sura across all 114 suras, and l is the number of words in sura *i*.

That said, it has also to be observed that, as text length decreases, so does the probability that any given word will occur even once in it, and its frequency vector will therefore become increasingly sparse, consisting mainly of zeros. Because 0 is non-adjustable, functions that compensate for variable text length generate increasingly unreliable results as length decreases. In the present application, therefore, only relatively long suras are considered for analysis, and more specifically those 24 containing 1000 or more content words.

Sura name	Words	Sura name	Words
Al-Baqarah	5739	Al-Israa	1464
Al-Imran	3316	Al-Kahf	1489
Al-Nisa	3543	Ta-Ha	1265
Al-Maidah	2681	Al-Anbiyaa	1077
Al-An'am	2895	Al-Hajj	1195
Al-A'raf	3127	Al-Nur	1236
Al-Anfal	1156	Al-Shu'araa	1208

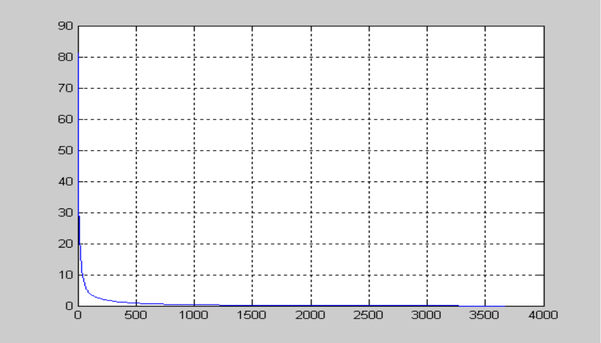
Al-Tawba	2345	Al-Naml	1069
Yunus	1732	Al-Qasas	1332
Hud	1809	Al-Ahzab	1239
Yusuf	1665	Al-Zumr	1107
Al-Nahl	1729	Ghafir	1156

“Table 1. Suras with more than 1000 words”

The choice of 1000 as the length threshold is arbitrary. Arbitrariness does no harm in a methodological paper such as this one. Clearly, however, any legitimate analysis of the Qur’an using this methodology will have to face the problem of which suras, if any, to exclude on length grounds in a principled way.

2.4 Dimensionality reduction

After function words and words with frequency 1 were eliminated and morphological variants stemmed, 3672 significant ‘content’ words remained, requiring a matrix with 3672 columns. Given only 24 data points, this results in an extremely sparsely populated data space whose dimensionality should be reduced as much as possible consistent with the need to represent the data domain adequately. For a discussion of dimensionality issues in data analysis see Verleysen (2003). To do this, the variances for all 3672 columns of the frequency matrix F were calculated, sorted in decreasing order of magnitude, and plotted:



“Figure 2. Plot of variances for 3762 columns”

This is what one would expect from the typical word frequency distribution in natural language text articulated by Zipf’s Law (Manning and Shütze, 1999; 20-29): almost all the variance in the data is concentrated in a small number of variables –the 500 or so on the left. The variance in the remainder is so small that it cannot contribute significantly to differentiating the data matrix rows and, therefore, can be disregarded. The matrix is thus truncated to 500 variables / columns, resulting in a 24 x 500 matrix for cluster analysis.

3 Analysis

3.1 Hierarchical cluster analysis

Cluster analysis aims to identify and graphically represent nonrandomness in the distribution of vectors in a data space such that intra-group distance is small relative to the dimensions of the space, and inter-group distance is relatively large. Detailed accounts of hierarchical cluster analysis are in Everitt (2001), Gordon (1999; 69-109), and Gore (2000). For briefer discussions see Dunn and Everitt (2001; 125-160), Hair et al. (1998; 469-518), Flynn et al. (1999; 275-9), Kachigan (1991; 261-70), Oakes (1998; 110-120). There are two main varieties: hierarchical and nonhierarchical. The former aims not only to discover and graphically represent clusters, but also to show constituency relations among data items and data item clusters as ‘dendrograms’ or trees.

Hierarchical analysis uses relative proximity among vectors as the basis for clustering, where proximity can be measured in terms either of similarity or of distance; distance is most often used, and is adopted here. Assuming the existence of a data matrix containing numerical values such as the one described above, construction of a distance-based cluster tree is a two-stage procedure. In the first step, a table of distances between data items, that is, between row vectors of the data matrix, is generated. A frequently used measure is the Euclidean; there the distance between vectors A and B is calculated using the well known formula: $length(z) = \sqrt{(length(x))^2 + (length(y))^2}$, but there are many others as in Gordon (1999; 15-3) and Flynn et al. (1999; 271-4).

The second step then uses the distance table to build clusters with the following generic algorithm:

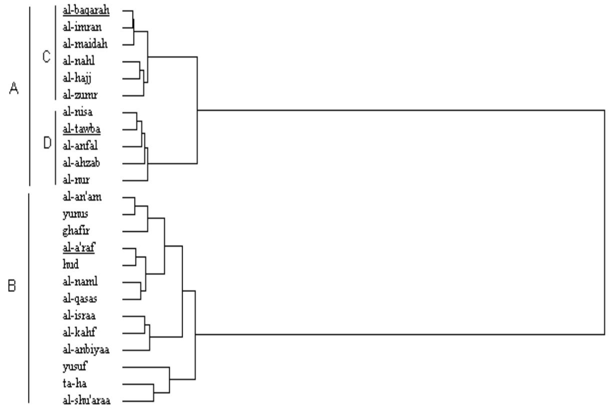
- Initially, every data vector is its own cluster.
- Using as many steps as necessary, at each step combine the two nearest clusters to form a new, composite cluster, thus reducing the number of clusters by 1.

- When only one cluster remains, incorporating all the cases in the distance matrix, stop.

An example of a tree generated by this procedure follows in the next section.

3.2 Cluster analysis of the quranic data

The above generic clustering algorithm glosses over an important point: determination of distances between data items is given by the distance table, but the distances between composite clusters is not, and needs to be calculated at each step. How are these distances calculated? There is no single answer. Various definitions of what constitutes a cluster exist, and, in any given application, one is free to choose among them. The problem is that the numerous combinations of distance measure and cluster definition available to the researcher typically generate different analyses of the same data, and there is currently no objective criterion for choosing among them. This indeterminacy is, in fact, the main drawback in using hierarchical clustering for data analysis. The present discussion sidesteps this important issue on the grounds that its aim is methodological: the intention at this stage of research is not to present a definitive cluster analysis of the Qur'an, but to develop an approach to doing so. One particular combination of distance measure and cluster definition was therefore chosen at random and applied to the data: squared Euclidean distance and Ward's Method. The result was as follows (the A - D labels on the left are for later reference):



"Figure 3. Tree generated by cluster analysis"

3.3 Interpretation

Given that the lengths of the vertical lines in the above tree represent relative distance between subclusters, interpretation of the tree in terms of the constituency relations among suras is obvious: there are two main subclusters A and B; A consists of two subclusters C and D, and so on. Knowing the constituency structure of the suras is a necessary precondition for understanding their thematic interrelationships –the object of this exercise—but it is not sufficient because it provides no information about the thematic characteristics of the clusters and the thematic differences between and among them. This information can be derived from the lexical semantics of the column labels in the data matrix, as follows.

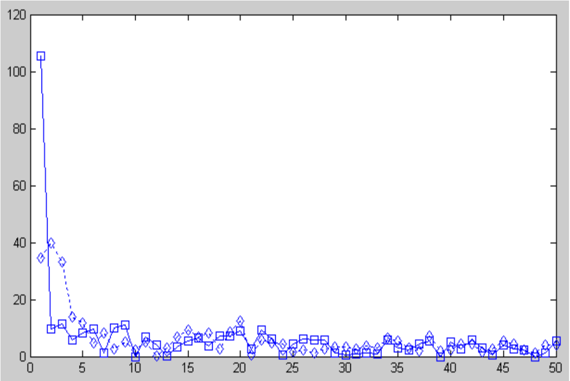
Each row in the data matrix is a lexical frequency profile of the corresponding sura. Since hierarchical analysis clusters the rows of the data matrix in terms of their relative distance from one another in the data space, it follows that the lexical frequency profiles in a given cluster G are closer to one another than to any other profile in the data set. The profiles of G can be summarized by a vector *s* whose dimensionality is that of the data, and each of whose elements contains the mean of the frequencies for the corresponding data matrix column:

$$s_j = \left(\sum_{i=1..n} F_{i,j}\right) / n$$

where j is the index to the jth element of *s*, i indexes the rows of the data matrix F, and n is the total number of rows in cluster G. If *s* is interpreted in terms of the semantics of the matrix column labels, it becomes a thematic profile for G: relative to the frequency range of *s*, a high-frequency word indicates that the suras which constitute G are concerned with the denotation of that word, and the indication for a low-frequency one is the obverse. Such a thematic profile can be constructed for each subcluster, and thematic differences between subclusters can be derived by comparing the profiles.

The general procedure for thematic interpretation of the cluster tree, therefore, is to work through the levels of the tree from the top, constructing and comparing thematic profiles for the subclusters at each level as far down the tree as is felt to be useful.

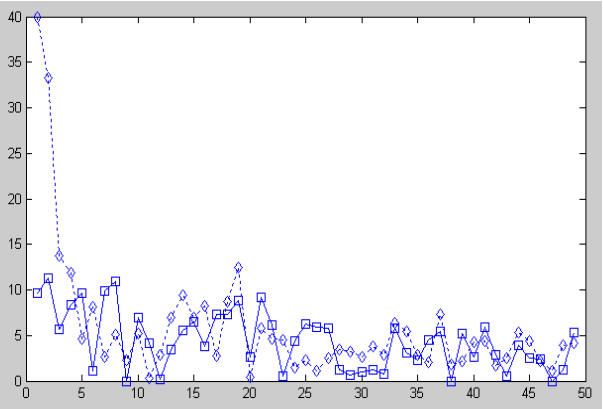
By way of example, consider the application of this general procedure to subtrees A and B in the above cluster tree. Two mean frequency vectors were generated, one for the component suras of cluster A, and one for those of cluster B. These were then plotted relative to one another; the solid line with square nodes represents cluster A, and the dotted line with diamond nodes cluster B; for clarity, only the 50 highest-variance variables are shown, in descending order of magnitude from the left:



“Figure 4. Initial plot of groups A and B”

The suras of cluster A are strikingly more concerned with the denotation of variable 1, the highest-variance variable in the Qur’an, than the suras of cluster B. This variable is the lexical item ‘Allah’, which is central in Islam; the disparity in the frequency of its occurrence in A and B is the first significant finding of the proposed methodology.

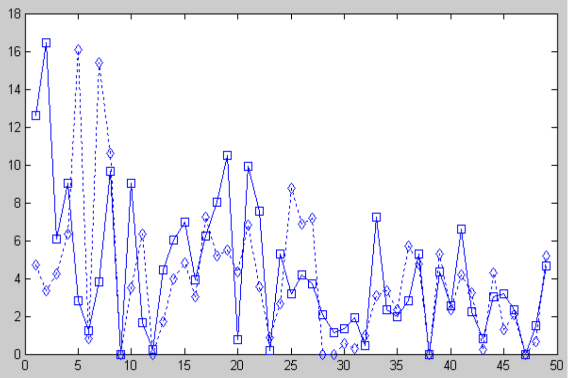
The scaling of the ‘Allah’ variable dominates all the other variables. To gain resolution for the others, ‘Allah’ was eliminated from the lexical frequency vectors, and the vectors were re-plotted:



“Figure 5. Re-plotting of groups A and B”

Awareness of the historical background of the Qur’an’s revelation to Mohamed is crucial at this point of interpretation. The suras revealed to Mohamed before his migration to Madinah are called Makkan suras, whereas those sent down after the migration are called Madinan. Makkan suras stress the unity and majesty of Allah, promise paradise for the righteous and warn wrongdoers of their punishment, confirm the prophethood of Mohamed and the coming resurrection, and remind humanity of the past prophets and events of their times. On the other hand, the Madinan suras outline ritualistic aspects of Islam, lay down moral and ethical codes, criminal laws, social, economic and state policies, and give guidelines for foreign relations and regulations for battles and captives of war. The results emerging from the initial clustering classification in figure 3 highlighted such thematic distiction. All the suras in cluster A are Madinan suras (apart from ‘Al-Nahl’ and ‘Al-Zumr’ which are Makkan suras; yet they do contain some verses that were revealed in Madina). The 13 suras which compose cluster B are all Madinan suras. The distribution of the variables (keywords) in figure 5 is also highly significant, e.g. variable 1 ‘qAl’ (said) is prevalent in the suras of cluster B. The suras of this group contain many narratives which illustrate important aspects of the quranic message, remind of the earlier prophets and their struggle and strengthen Prophet Mohamed’s message of Islam. This signifies the use of the verb ‘qAl’ as a keyword in narrative style. Variable 4 ‘qul’ (say, imperative) is more frequent in group B than group A. Most of the passages of these Makkan suras start with the word ‘qul’, which is an instruction to Prophet Mohamed to address the words following this introduction to his audience in a particular situation, such as in reply to a question that has been raised, or an assertion of a matter of belief. The use of this word was appropriate with Mohamed’s invitation to belief in God and Islam in Makkan suras. Variable 5 ‘mu/min’ (believers), variable 8 ‘Aman’ (believe) and variable 24 ‘ittaq’ (have faith) highly occur in group A. These are the Madinan suras in which prophet Mohamed addresses those who already believed in his message and hence focusing on introducing them to the other social and ethical aspects of Islam. Other variables prevelant in group B are variables 14 and 28 ‘AyAt , Ayat’ (signs/sign). The use of the two words was very important for Prophet Mohamed in the early phase of Islam in Makkah. He had to provide evidence and signs to people to support his invitation to belief in Allah and Islam.

The same procedure of clustering can be applied to the subclusters of A and B. Again, the scaling of Allah' dominates, and removing it from the mean frequency vectors gives better resolution for the remaining variables. Plotting the lexical frequency vectors for C and D, for example, yields the following:



"Figure 6. Plot of groups C and D"

Results from figure 6 are also supportive of the thematic structure of each group. Suras of group C are more abundant in the use of narratives and addressing Mohamed to provide evidence of his message to people. Suras of group B are more concerned with addressing believers about the reward for their righteous conduct. Occurrences of relative variables to those themes are indicative of such distinction.

4 Conclusion and future directions

The above preliminary results indicate that construction and semantic interpretation of cluster trees based on lexical frequency is a useful approach to discovering thematic interrelationships among the suras that constitute the Qur'an. Usable results can, however, only be generated when two main issues have been resolved:

- Standardization of the data for variation in sura length, as discussed in section (2.3)
- Variation in tree structure with different combinations of distance measure and cluster definition, as discussed in section (3.2)

Work on these is ongoing.

To conclude, hierarchical cluster analysis is known to give different results for different distance measure / clustering rule combinations, and consequently cannot be relied on to provide a definitive analysis. The next step is to see if interpretation of the principal components of a principal component analysis of the frequency matrix yields results consistent with those described above. Another multivariate method to be applied to the data is multidimensional scaling. In the longer term, the aim is to use nonlinear methods such as the self organizing map in order to take account of any nonlinearities in the data.

References

Dror, J., Shaharabani, D., Talmon, R., Wintner, S. 2004. *Morphological Analysis of the Qur'an*. Literary and Linguistic Computing, 19(4):431-452.

Dunn, G. and Everitt, B. (2001). *Applied Multivariate Data Analysis*, 2nd ed. Arnold, London.

Everitt, B. (2001). *Cluster Analysis*, 4th ed. Arnold, London.

Flynn, P., Jain, A., and Murty, M. (1999). *Data clustering: A review*. In: ACM Computing Surveys 31, 264–323.

Gordon, A. (1999). *Classification*, 2nd ed. Chapman & Hall, London.

Gore, P. (2000). *Cluster Analysis*. In H. E. A. Tinsley & S. D. Brown (Eds.), Handbook of applied multivariate statistics and mathematical modeling (pp. 297-321). Academic Press, San Diego, CA

Hair, H., Anderson, J., Black, W. and Tatham, R. (1998). *Multivariate Data Analysis*, 5th ed. Prentice-Hall International, London.

Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*, MIT Press.

Kachigan, S. (1991). *Multivariate Statistical Analysis. A conceptual introduction*. Radius Press, New York

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass, MIT Press.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh

Thabet, N. (2004). "*Stemming the Qur'an*". In Proceedings of Arabic Script-Based Languages Workshop, COLING-04, Switzerland, August 2004.

Verleysen, M. (2003). *Learning high-dimensional data*. In: *Limitations and future trends in neural computation*. IOS Press, Amesterdam, pp141-162.