Lexical Analysis of The Quran using Frequent Itemset Mining

Syed Zubair Ahmad Shah* and Mohammad Amjad
Department of Computer Engineering
Jamia Millia Islamia, New Delhi
*Email: zub12345@gmail.com

ABSTRACT

This paper presents an analysis of the frequency of presence of words in different chapters of the Quran. It also presents an analysis of the repetition of verses in the Quran. For this purpose we have introduced a novel algorithm. Our research is based on a fundamental concept of frequent pattern mining: that, with regard to some collection of documents, the lexical frequency profiles of individual documents are a good indicator of their conceptual content. The results obtained here show that the proposed approach produces results that are useful in getting onto the fundamental concepts of the Quran and can be of much help to the people of theology and to those who are interested in objective study of religious scriptures.

Keywords: Frequent Itemset Mining, Quran, Lexical Analysis.

1. INTRODUCTION

Nearly 1.6 Billion people around the world believe Quran to be the word of God. This makes Quran an important subject of research, for Muslims and Non-Muslims alike. There are many approaches to analyze the text of any sacred book. One approach is to have a manual reading of it and write the gist and the commonalities/ contradictions (if any) between its various chapters/parts or the concepts it presents. This approach is subject to human prejudice, level of intellect and memory power. Another approach is to use data mining for analysis. The above mentioned three limitations of human mind do not affect this approach in any way. This approach is based purely on the analysis of words in the text of the scripture.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It is also defined as the extraction of hidden predictive information from large databases or data-files. One of the fields of data mining is frequent itemset mining. Frequent itemset mining is a method used for market basket analysis. It aims at finding sets of products that are frequently bought together by customers from a super-market or online shops. This field of data mining has also been used to find out terms that are frequently used in a document set and the terms that are frequently found together in documents.

Some researches on the text of the Quran have already been done. [1] attempts to discover thematic interrelationships among the chapters of the Quran by abstracting lexical frequency data from them and then applying hierarchical cluster analysis to that data, [2] uses association rules for ontology extraction from a Quran corpus, [3] presents an approach to capture the common relationships between the words' tags that highly show up in the Quran, [4] proposes a method of representing the Quranic text corpus as a graph and then applies a frequent sub-path mining algorithm on it to generate frequent patterns, [5] uses frequent pattern mining to extract non-trivial patterns and interesting relations in the chapters of the exegesis of the Quran.

In this paper, we have used frequent itemset mining on a Quran document corpus. Instead of itemsets, we find frequent termsets in the corpus. Of the frequent itemset/termset mining algorithms, we have used algorithm Apriori [6].

The rest of this paper is organized as follows. Section 2 briefly defines some of the terms related to our work. In section 3, we introduce algorithms that we used for analyzing the Quran. An experimental evaluation on a translation of the Quran was conducted, and section 4 reports its major results. Section 5 summarizes the paper and outlines some interesting directions for future research.

2. BASIC PRELIMINARIES

We quickly review some definitions related to our work. Let $D=\{C_1,C_2,C_3,\ldots\}$ represent the *document set* or document corpus which in our case is a set of all the chapters of the Quran and let $W=\{w_1,w_2,w_3,\ldots)$ represent set of all the words in D (i.e. $w_i \in D\ \forall\ i$), each term mentioned only once in the set. The *support of a word* w, denoted by supp(w), is the fraction of chapters containing w. A word whose support is equal to or greater than user-specified minimum is a frequent word i.e. if $supp(w) \geq min\_sup$, then w is a *frequent word*. Let $V=\{v_1,v_2,v_3,\ldots\}$ represent the set of all the verses in the Quran. Two verses that have more than user specified percentage of common words are *similar verses*. The support of a verse v, denoted by sup(v), is the fraction of verses that are similar to v (including v). A verse whose support is equal to or greater than a user-specified minimum verse threshold is a *frequent verse*.

3. ALGORITHMS

This section introduces the algorithms used for the analysis of the document set. All of the implementation and evaluation has been done in R. Before the application of algorithms, the document set is pre-processed. In pre-processing of document set, following operations are performed.

1. Extra white-spaces are removed.
2. All alphabets are converted to lower case.
3. All (verse) numbers are removed.
4. All punctuation symbols are removed.
5. All stop words are removed.
6. Stemming is done on the remaining words of the document set.

For analysis of the pre-processed document set, two algorithms were used. One is Apriori [6] that finds out the frequent words. This algorithm finds out all those words from the pre-processed document set that are there in at least or more than the user specified number of chapters i.e. those words whose support is equal to or greater than a user-specified minimum support threshold.

The second algorithm does not apply on individual chapters of the Quran but to the scripture as a whole. Let us call the single-file pre-processed scripture as S_p. Let NS be the number of verses in S_p.

Algorithm Verse_Analyzer:

1. Split S_p using line-break as sentinel and store the result in an array A of size NS.
2. Put I=1.
3. Split A[I] using blank-space as sentinel and put result in array X.
4. If Size(X) = 0 or 1 or 2 goto step 13.
5. Put J=1 and N=0.
6. If Size(X) = 3 or 4 then min_word_threshold=1, Else min_word_threshold=0.8
7. Split A[J] using space as sentinel and put result in array Y.
8. Find out how many words do X and Y have in common. Store the result in NCW.
9. If NCW>=min_word_threshold*Size(X), increment N by 1. (The truthfulness of this condition implies A[J] is a match of A[I] or at least a part of A[I]).
10. Increment J by 1.
11. If J<=Size(A) goto step 7, Else goto step 12.
12. If N>=min_verse_threshold*NS, store A[I] as an element in resultant array R (preventing redundancy).
13. Increment I by 1.
14. If I<Size(A) goto step 3, Else exit.

(min_verse_threshold is the minimum fraction of verses that must be a match of or a superset of the verse under consideration for it to be considered a frequent verse)

This algorithm finds out frequent verses in the Quran. In other words it finds out those verses that are of much importance and have been repeated very often.

This algorithm initially stores the whole (pre-processed) scripture in an array A with each element of A storing exactly one verse of the Quran. Then, one by one, each verse is matched with every other verse to find its match. This is done by splitting each verse into words and comparing how many words do two verses have in common. Condition for a match is kept such that at least 80 percent of the words of the two verses should be the same. This condition is meant for verses with word length greater than or equal to five. Verses with word length (after preprocessing) less than three are not considered as such small verses have a high probability that their words will be present in some other verse. For example verse “The Merciful” after preprocessing becomes “merciful”, and this word “merciful” is present in hundreds of other verses. Considering such verses won’t be of much benefit.

For verses with word length three or four the minimum word threshold has been kept to 1.0 as for verses of such small length all the words of one verse should be present in another verse so as to consider a match. If the total number of matches for a particular verse cross the user specified minimum verse threshold, then only is the verse considered to be a frequent verse.

4. RESULTS AND DISCUSSION

We downloaded a copy of English translation of the Quran from [7]. The translation is by Wahiduddin Khan [8]. This document set is of size 114 which implies that the number of chapters in the Quran is 114. After pre-processing, we executed Apriori [6] on this copy of the Quran. The results are shown in table 1.

Table 1 (Results of Apriori execution)

Minimum Support Threshold	**Number of Words Found**	**Words Found (in alphabetical order)**
0.80	2	"god" "lord"
0.75	5	"day" "god" "lord" "one" "truth"
0.70	9	"day" "deni" "earth" "god" "know" "lord" "one" "peopl" "truth"
0.65	13	"believ" "day" "deni" "earth" "god" "inde" "know" "lord" "one" "peopl" "say" "sure" "truth"
0.60	21	In addition to above: "come" "creat" "give" "good" "heaven" "make" "see" "turn"
0.55	29	In addition to above: "away" "fire" "heart" "made" "man" "messeng" "punish" "sent"
0.50	55	In addition to above: "among" "ask" "back" "bring" "call" "can" "deed" "evil" "fear" "follow" "garden" "given" "guid" "knowledg" "let" "life" "like" "may" "power" "reward" "right" "thing" "true" "truli" "upon" "yet"

The results are very interesting. A minimum support threshold of 0.80 yielded only two words – “God” and “Lord”. This means the words “God” and “Lord” are in more than 80% of the chapters of the Quran. No other word crosses this threshold of 80% which obviously implies that the central figure in the Quran is God.

The word “Lord” in the Quran is the English translation of the Arabic word "Rabb" which also means Creator, Sustainer, Master, Nourisher and Regulator (of the Universe), the word "Day" in the Quran repeatedly has been used for Day of Judgement and the word "Truth" is the English translation of the Arabic word "Al-Haq". Both Rabb and Al-Haq, in the Quran, are used as attributes of God. Since no other word except Day, God, Lord, One and Truth crosses the minimum support threshold of 0.75, this implies that these five words represent the most basic concepts that the Quran presents i.e. the concept of oneness of God, the concept that He is the Lord and He is the Truth and the concept of Day of Judgement. Another important observation that we came across was that 113 of the total 114 chapters of the Quran contain atleast one of these terms.

With a minimum support threshold of 0.70, again only the most important concepts of the Quran are highlighted by these few words. A very meaningful set of sentences can be constructed using these words – “People of Earth Know that your God is one God. He is the Truth. He is your Lord. Do not Deny Him. Know that the coming of the Day of Judgement is sure.”

Similar deductions can be made from other table entries also. Even at a minimum support threshold of 0.50, the words that come out are very meaningful in terms of the important concepts that the Quran presents and its overall message.

The results obtained after executing our algorithm Verse_Analyzer on the (pre-processed) Quran are shown in table 2.

Table 2 (Results of Verse_Analyzer execution)

Min. Word Thres-hold	**Min. Verse Thres-hold**	**Verses Found**	**Times of Occur-rence**
1.00 for verses of length = 3 or 4 0.80 for verses of length > 4	0.005	1. "god will turn merci forgiv"	33
		2. "will say believ"	41
		3. "god protector believ deni truth"	40
		4. "lord wonder deni"	31
		5. "say deni truth"	61
		6. "say god one"	56

The verses of the result (without preprocessing) are shown below:

1. "then after that, God will turn in His mercy to whom He wills: God is forgiving and merciful." (9:27)
2. "They will say, "No! It was you who would not believe --" (37:29)
3. "That is because God is the protector of the believers, and those who deny the truth have no protector at all." (47:11)
4. "Which of your Lord's wonders would you deny?" (55:16)
5. "Say, "You who deny the Truth," (109:1)
6. "Say, "He is God, the One," (112:1)

(The numbers in parenthesis represent chapter number and verse number, respectively)

These results again carry with them the very fundamental concepts of Quran. They could be very useful in understanding the areas that Quran puts a lot of stress on. God's oneness, His forgiving and merciful nature, His being the only true protector, His majesty, His wonders in creation – these concepts are the highlights of these verses.

We leave the work of further interpretation to the people of the field of theology, the students of religious studies and those interested in objective study of religious scriptures.

5. CONCLUSIONS

This paper presented an analysis of the English translation of the Quran. The analysis was done using frequent itemset mining. An algorithm was also introduced for the analysis. The results were astonishing and can be of much help to the people of theology.

Our algorithms are not limited to the analysis of Quran only. They can be used for analyzing any religious or non-religious scripture. Our future plans are to analyze the Old and New Testament, and the writings of poets of English language.

REFERENCES

[1] N. Thabet, "Understanding the thematic structure of the Qur'an: an exploratory multivariate approach", in **Proc. ACL student research workshop**, Michigan, USA, 2005, pp. 7–12.

[2] F. Harrag, A. Al-Nasser, A. Al-Musnad, R. Al-Shaya and A. S. Al-Salman, "Using association rules for ontology extraction from a Quran corpus", in **Proc. 5th int. conf. Arabic language processing**, Oujda, Morocco, 2014.

[3] D. E. M. A. AbuZeina and M. H. Alsaheb, "Capturing the common syntactical rules for the Holy Quran: A data mining approach", in **Proc. int. conf. advances in information technology for the Holy Quran and Its Sciences**, Madina, Saudi Arabia, 2013, pp. 670–680.

[4] I. Ali, "Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of the Arabic", **International journal of software engineering and its applications**, vol. 6(3), 2012, pp. 127–134.

[5] S. Chua and P. N. E. B. Nohuddin, "Frequent pattern extraction in the Tafseer of Al-Quran", in **Proc. 5th int. conf. information and communication technology for the Muslim world**, Kuching, Malaysia, 2014, pp. 1–5.

[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", in **Proc. 20th int. conf. very large data bases**, 1994, pp. 487–499.

[7] tanzil.net

[8] http://www.cpsglobal.org/mwk