A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran

Mohamadou Nassourou
Department of Computer Philology & Modern German Literature
University of Würzburg Am Hubland D - 97074 Würzburg
mohamadou.nassourou@uni-wuerzburg.de

Abstract

Computationally categorizing Quran's chapters has been mainly confined to the determination of chapters' revelation places. However this broad classification is not sufficient to effectively and thoroughly understand and interpret the Quran. The chronology of revelation would not only improve comprehending the philosophy of Islam, but also the easiness of implementing and memorizing its laws and recommendations.
This paper attempts estimating possible chapters' dates of revelation through their lexical frequency profiles. A hybrid statistical classifier consisting of stemming and clustering algorithms for comparing lexical frequency profiles of chapters, and deriving dates of revelation has been developed. The classifier is trained using some chapters with known dates of revelation. Then it classifies chapters with uncertain dates of revelation by computing their proximity to the training ones. The results reported here indicate that the proposed methodology yields usable results in estimating dates of revelation of the Quran's chapters based on their lexical contents.

Keywords: text categorization, Bayesian classifier, distance-based classifier

1. Introduction

Computationally classifying text documents refers to the process of developing software capable of automatically assigning unseen text documents to pre-defined classes. This study has used supervised learning approach to build a text classifier for estimating dates of revelation of Quran's chapters based on their lexical frequency profiles.
It is important to mention that dates of revelation refer to the revelation of the first verses of a given chapter.
The Qur'an consists of 114 chapters having unequal length. They are not arranged in chronological order, and appear to be roughly compiled by descending order of length, with long chapters at the beginning and short ones at the end.
Based on traditional Quran scholars' classification as well as western orientalists ones such as [12], 13 out of 114 Quran's chapters have been reasonably dated. Some of those dated chapters are the ones that contain verses pertaining to historical facts and events such as battles, migration to Medina.
Out of the 13 dated chapters 5 belong to the period between 610 AD to 622 AD corresponding to the Meccan[1] phase, and the remaining 8 correspond to the Medinan[2] phase between 622 AD and 632 AD.
These chapters with known dates of revelation have been labeled, and then supplied to the classifier to learn and generate categories according to dates of revelation of the chapters. Using the generated categories the classifier will classify each chapter with unknown date of revelation by computing its proximity to the categorized chapters.

The Qur'an consists of 114 chapters called suras having unequal length. They are not arranged in chronological order, and appear to be roughly compiled by descending order of length, with long chapters at the beginning and short ones at the end. However traditional Islamic scholars through narrations (hadiths) of the prophet Mohammad have shown that the actual arrangement of the chapters has been made by the prophet himself under the recommendation of God.
As no one has recorded the chronology of revelation, the present study attempts to reconstruct it with the help of machine learning methods. A hybrid method combining Bayesian and distance-based classifiers has been used.
First the Bayesian classifier is applied to assign each testing chapter to one of the categories of the 13 chapters, then a distance classifier is utilized to derive plausible dates of revelation of each chapter.

2. Data

The main objective of the present study is to attempt reconstructing the Quran chronology with the help of computational techniques. Knowing the difficulties of developing or finding a software program capable of producing unambiguously roots of Arabic words of the Quran, I found it easier and more effective to use the transliteration version. The transliteration resolves significantly the problem of diacritics in the Quran. In fact [2] has already used this approach in order to develop a light stemmer for the Quran.
An electronic version of the transliteration formatted with HTML was downloaded from the Muslimnet [1] website and preprocessed in order to clean and produced plain text.

3. Chapter Representation

It consists of representing each chapter with vectors having keywords as entries.
Preprocessing downloaded chapters in order to produce vectors of words is the main activity of this section.

3.1 Preprocessing

The preprocessing consists of the following tasks:

a. Generation of plain text file

Plain text files equivalent to the downloaded HTML ones were created by filtering out HTML tags and replacing underlined letters with others consisting of combining two or three ASCII characters.

b. Removal of functional words such as determiners and prepositions.

c. Segmentation.

The segmentation involves tokenization of the text files based on the white space criterion in order to create vectors of words known as features or keywords.

d. Stemming

Stemming helps reduce size of vectors by eliminating redundancies. It is the process whereby features are grouped based on semantic similarity. Usually it involves removal of affixes from words. Affix comprises prefixes, infixes and suffixes. However stemmers used in text mining are usually called light-stemmers, because they fail to remove infixes.

In this study a light stemmer similar to the one described in [2] was developed. It removes duals and plurals for masculine and feminine, possessive forms, definite articles, and pronouns.

Before stemming is applied, a stopword list consisting of words to be excluded from the stemming process was compiled. Then the following algorithm was used.

For every word in the text document:

1. IF the word length < 3 characters then delete it.
2. Remove prefixes if the word is not a stop word.
3. If the word is not a stop word, then remove suffixes whose length > 2.

4. Dimensionality Reduction

This is a kind of global applied to all the categories.

Each chapter is represented as a vector of length equal to the total number of keywords it contains. Computational time needed to weigh features of long chapters, as well as during the classification phase is usually long. Additionally, unnecessary long vectors are simply waste of storage. To cope with this situation the size of the vectors is practically reduced through the elimination of less important keywords.

In fact the number of keywords needed for classification is usually less than the actual size of the vectors. Based on this fact several techniques have been devised to reduce the dimension of the vectors [10].

In this study, the following learning algorithm was used to select the most important features for categorizing the chapters.

For each chapter:

a. compute weight of each feature using the formula TF / (Document length), where TF is term frequency (number of occurrences of a keyword)
b. compute standard deviation of the mean of weights
c. select features whose weights are above the mean of weights minus the standard deviation

5. Selection of Number of Keywords per Category

This could be viewed as a kind of local policy for selecting features in classification step.

Because the chapters vary enormously in size, and short chapters might even see their entire keywords present in the long ones, the number of features per category has been limited to some most frequent features plus some uncommon ones to each other.

Based on my personal observation and understanding of the Quran's text, I found that two successive chapters do not have a lot of common terms. Based on traditional Quran scholars' classification one can see for instance chapters 93, 94, or 2, 8 do not share many common terms even though they were revealed successively. To the contrary only one or very few terms are linking them together. This is in fact where the Quran's philosophy lies. Hence I do not need to compare most or least frequent features of classifying and testing chapters. But for the sake of verifying my observations I did compare all the possible combinations which are described below.

For the Bayesian classifier approach I distinguished 4 cases based on these observations. The 4 cases are formulated as rules for selecting features for classes and testing chapters.

The rules are explained below.

a. Compare most frequent features of classifying and testing chapters.
b. Compare least frequent features of classifying and testing chapters.
c. Compare most frequent features of classifying chapters with least frequent features of testing chapters.
d. Compare least frequent features of classifying chapters with most frequent features of testing chapters

The result that was closer to the traditional classification was obtained with case (b)

5.1 Categories Definition

The Quran was revealed in 23 years starting from 610 to 632 AD. Each year is considered as a class. The features of a class are the keywords of chapters with known dates of revelation that were revealed during that year. Initially only 9 classes do have features.

It is important to mention that a year of revelation refers to the revelation of the verses of a chapter.

The classes and their chapters are shown in table 1.

Class	Chapter
610	96
613	74
620	17
622	22 , 23, 2
624	3, 8
627	33
628	48
630	9
632	110, 5

Table1. Classes with contained chapters

From now on I will refer to these classes as non-empty classes, and the remaining ones as empty classes. It is possible to view the non-empty classes as training classes and the empty ones as testing ones.

It is important to notice that we are dealing with a semi-supervised learning problem whereby initially we have some training sets from which new classes will be induced. In fact some chapters will belong to some of the non-empty classes, while others will form new classes on their own.

6. Classification Methodology

This is the section that deals with the actual categorization of chapters.
The methodology is described as follows:

1. Classify chapters according to place of revelation i.e revealed in Mecca or Medina.
2. 5 chapters among the Meccan chapters have known dates of revelation, use them to construct features of categories that Meccan chapters (testing chapters) with unknown dates of revelation will be assigned to.
3. 8 chapters among the Medinan chapters have known dates of revelation, use them to construct features of categories that Medinan chapters (testing chapters) with unknown dates of revelation will be assigned to.
4. Use Bayesian classifier to assign each testing chapter to a category by separating Meccan categories(610-622 AD) with Medinan categories(622-632 AD)
5. To determine exact date of revelation of each chapter the following procedure is used:
   a. Compute similarity and difference between every two successive categories by considering the initial features only.
   b. Divide the similarity and difference values by the number of years between them in order to create intervals for allocating testing chapters.
   c. The date of revelation of each testing chapter depends on the interval where its similarity and difference values will lie.

7. Classifiers

There are several types of classifiers with each one having its own method of categorizing documents or objects. Among them we find statistical, functional, neural, decision trees, and fuzzy classifiers. The most widely used classifiers are the statistical ones such as the Bayesian and distance-based classifiers. However some functional classifiers such as K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) are also intensively researched.
For this study, Bayesian and distance-based classifiers have been selected because of their simplicity and effectiveness.

7.1 Naive Bayesian Classifier

Bayesian classifiers assign a document X to a class C_i by applying the following Bayes' theorem:

$P(C_i|X) = P(X|C_i)\, P(C_i) / P(X)$

In traditional text classification, a document X is represented by an n-dimensional vector X = {x1, x2, . . . , xn}.
If the vector X is huge, processing time would be long to compute $P(X|C_i)$. Hence the naïve Bayesian classifier assumes that classes are conditionally independent.
In other words for a given vector X its entries x1, x2,…xn are conditionally independent of one another.

Mathematically this implies that:

$$P(\mathbf{X}|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i).$$

Given a document X, the classifier will predict that X belongs to the class Ci if and only if

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \qquad \text{for } 1 \leq j \leq m,\ j \neq i.$$

$P(C_i|X) = P(X|C_i)\, P(C_i) / P(X)$

For the current research P(X) is same for all classes, therefore only $P(X|C_i)P(C_i)$ needs to be maximized.
Moreover the class a priori probabilities $P(C_i)$ are not known, therefore I supposed they are all equal,
$P(C_1) = P(C_2) = \ldots = P(C_k)$.
Hence only P(X|Ci) needs to be maximized.

$P(X|C_i)$ is computed as follows:

$P(X|C_i) = w_f / d_f$

Where w_f is the frequency of a word x_i in the testing chapter, d_f is the number of classifying chapters per category that contain w_f.

w_f = number of occurrences of word / chapter length

If a word does appear in any of the classifying chapters, d_f will be zero, for this reason one is added to d_f.

Hence the formula is rewritten as:

$P(X|C_i) = w_f / (d_f + 1)$

7.2 Distance-based Classifier

It measures similarities and differences between chapters of a class to the base chapter (classifying chapter) of a neighboring class. Depending on these measurements it will assign a chapter as either pre-base chapter (revealed before the base chapter) or post-base chapter (revealed after the base chapter). Then the distance between base chapters of neighboring classes is divided by the number of years that separate them, in order to create intervals to which each testing chapter will be assigned to.

If the keywords of a base chapter X of class C_i are {x1, x2,…xn}, and those of a neighboring base chapter Y of class C_{i+1} are {y1, y2,…yn}, then the summation of keywords weights that they share is the similarity measure (SIM), and the summation of uncommon keywords weights is the difference measure (DIFF).

Now the degree of similarity and difference between each chapter of class C_i with the base chapter of class C_{i+1} will be estimated, and vice versa. Each chapter will be assigned a date according to its location on the partitioned distance between the two base chapters.

8. Implementation

The classification algorithm described above has been implemented in PHP.

9. Preliminary Results and Interpretation

Currently I have computed dates of revelation of Medinan chapters. The result is shown in table 2.

Date of revelation	Chapter
622	2, 23
623	22, 76
624	8, 3
625	4, 57, 61
626	24, 13, 47
627	33, 62, 55
628	99, 66, 98, 59
629	65, 63, 9, 49
630	48, 58, 64, 63
632	110, 5

Table 2. Chapters with their estimated dates of revelation

As mentioned previously two successive chapters have few words in common. The shared words most of the time are an indication about the main theme discussed in the preceding chapter.

The above classification is for the time being a statistical approach of proving my observations.

This classification does neither correspond to the traditional Quran's scholars nor to the western orientalists classifications. However it is closer to the traditional scholars' one in the sense that the ideology of linking the chapters seems to be similar. For instance chapter 2 being the first to be revealed after emigration discusses the contact between the prophet and Jewish community of Medina. Then followed by 76 reminding them of their Lord who created them so that they convert to Islam. Chapter 8 suggest that a battle took place (the first battle of Badr), and the Muslims inquire about the way to share the booty. Chapter 3 reminds them not to forget the victory that God gave them against the non-believers; they should stay fast and continue the battle. Following chapters introduces the rules for treating female slaves and women in general. And remaining ones discuss difficulties and oppositions that Muslims have been facing till ultimately their religion Islam prevailed.

I find this classification plausible to the real chronology of the Quran, because after migration it seems to me that battling the pagans of Mecca, where the mosque Kaaba is located and the hometown of the prophet himself, should have been more important than delivering chapters for condemning the Jews of Medina.

However this study is just the beginning of a process that needs to be carefully re-examined and validated with the help other computational methods. I am currently working on that.

10. Conclusion and Future Work

In this study I have presented a methodology of reconstructing Quran's chronology based on machine learning techniques. A hybrid statistical classifier consisting of Bayesian and distance-based classifiers have been employed, in order to arrive at plausible dates of revelation in accordance with the traditional Islamic scholars and western orientalists chronologies. A transliteration version of the Quran was used because of its advantages in stemming process.

The results obtained so far are encouraging since they agree mostly with those of the traditional Quran's scholars.

The next activity will be to complete the classification for the Meccan chapters as well.

The task done in this research could be extended by applying other methods such as functional classifiers namely SVM and KNN classifiers, in order to further validate the obtained results.

Widening the number of base chapters per class might improve the accuracy of the result found here. This is actually my current task. I am currently consulting more historical documents about Islam in the 7th and 8th century AD, and this include the hadiths (traditions) of the prophet, as well as writings of his companions and their successors.

References

[1]http://www.usc.edu/schools/college/crcc/engagement/resources/texts/muslim/quran/transliteration/

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

[3] Kanaan, G., AL-Kabi, M. N., and AL-Shalabi, R. ,2005, "Statistical Classifier of the Holy Quran Verses (Fatiha and YaSeen Chapters)", Journal of Applied Science, 5(3), pp.580-583.

[4] HERMANN MOISL, University of Newcastle, UK "Sura Length and Lexical Probability Estimation in Cluster Analysis of the Qur'an", ACMTransactions on Asian Language Information Processing, Vol. 8, No. 4, Article 19, Pub. date: December 2009

[5] Marwan Ali.H. Omer Ma shi long, School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics Beijing, China "STEMMING ALGORITHM TO CLASSIFY ARABIC DOCUMENTS"

[6] Khitam Jbara Department of Computer Science, King Abdullah II School for Information Technology, University Of Jordan, "Knowledge Discovery in Al-Hadith Using Text Classification Algorithm", Journal of American Science, 2010

[7] AL-Kabi, M. N., and AL-Shalabi (2005),"ALHadith Text Classifier", Journal of applied sciences, 5(3), pp.548-587.

[8] Zhang J. (2004)."The Optimality of Naïve Bayes ", Proceedings of the 17th International Florida Artificial Intelligence Research Society Conferences, Florida, pp.562-567

[9] Manar Alkhatib Zayed, "Classification of Al-Hadith Al-Shareef Using Data Mining Algorithm", University College of Information Technology. Abu Dhabi, UAE

[10] Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R. (2008),"Text Document pre-processing with the Bayes formula for Classification using the Support Vector Machine", IEEE Transactions on Knowledge and Data engineering, Volume 20, Issue 9, pp. 1264 – 1272

[11] Hirotoshi Taira "Text Categorization using Machine Learning", Doctor's Thesis, Nara Institute of Science and Technology, 2002

[12] http://en.wikipedia.org/wiki/Theodor_Nöldeke