Using Machine Learning Algorithms for Categorizing Quranic Chapters by Major Phases of Prophet Mohammad's Messengership

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

Abstract

This paper discusses the categorization of Quranic chapters by major phases of Prophet Mohammad's messengership using machine learning algorithms. First, the chapters were categorized by places of revelation using Support Vector Machine and naïve Bayesian classifiers separately, and their results were compared to each other, as well as to the existing traditional Islamic and western orientalists classifications. The chapters were categorized into Meccan (revealed in Mecca) and Medinan (revealed in Medina).

After that, chapters of each category were clustered using a kind of fuzzy-single linkage clustering approach, in order to correspond to the major phases of Prophet Mohammad's life.

The major phases of the Prophet's life were manually derived from the Quranic text, as well as from the secondary Islamic literature e.g hadiths, exegesis.

Previous studies on computing the places of revelation of Quranic chapters relied heavily on features extracted from existing background knowledge of the chapters. For instance, it is known that Meccan chapters contain mostly verses about faith and related problems, while Medinan ones encompass verses dealing with social issues, battles…etc. These features are by themselves insufficient as a basis for assigning the chapters to their respective places of revelation. In fact, there are exceptions, since some chapters do contain both Meccan and Medinan features.

In this study, features of each category were automatically created from very few chapters, whose places of revelation have been determined through identification of historical facts and events such as battles, migration to Medina…etc.

Chapters having unanimously agreed places of revelation were used as the initial training set, while the remaining chapters formed the testing set. The classification process was made recursive by regularly augmenting the training set with correctly classified chapters, in order to classify the whole testing set.

Each chapter was preprocessed by removing unimportant words, stemming, and representation with vector space model.

The result of this study shows that, the two classifiers have produced useable results, with an outperformance of the support vector machine classifier.

This study indicates that, the proposed methodology yields encouraging results for arranging Quranic chapters by phases of Prophet Mohammad's messengership.

Keywords: Text categorization, Support Vector Machine, Naïve Bayesian, Clustering, Place of revelation, Stages of Prophet Mohammad's messengership, Quran

1. Introduction

The presence of huge amount of data on the web, and on some intranets has made finding relevant information very tedious and a complicated task. One of the solutions to tackle this issue has been to develop methods for automatically grouping information by relevant criteria. Text categorization is one of those methods used for grouping information according to some pre-defined classes.

Computationally classifying or categorizing text documents refers to the process of developing software capable of automatically assigning unseen text documents to pre-defined classes.

Several text categorization algorithms have been developed among them naïve Bayes, support vector machines, k-nearest neighbors, and neural networks.

Previous studies on computing the places of revelation of Quranic chapters relied heavily on features extracted from existing background knowledge of the chapters. For instance, it is known that Meccan chapters contain mostly verses about faith and related problems, while Medinan ones encompass verses dealing with social issues, battles…etc. However these features are by themselves insufficient as a basis for assigning chapters to their respective places of revelation. In fact, there are exceptions, since some chapters do contain both Meccan and Medinan features.

In this study, naïve Bayes and Support Vector Machine (SVM) classifiers have been used for estimating places of revelation of the Quranic chapters based on their lexical frequency profiles. It is important to mention that, place of revelation refers to the revelation of the first verses of a given chapter.

The Quran 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 the classifications of traditional Islamic scholars and western orientalist researchers such as [13, 14, 15], 13 out of 114 Quranic 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 phase, and the remaining 8 correspond to the Medinan phase between 622 AD and 632 AD.

These chapters with known places of revelation have been combined according to places of revelation, and then supplied to the classifiers to learn. Two categories namely Meccan (revealed in Mecca) and Medinan (revealed in Medina) were generated. Using the generated categories the classifiers assign each chapter with unknown place of revelation to the appropriate category.

The classification process was made recursive by regularly augmenting the training set with correctly classified chapters, in order to classify the whole testing set. In fact due to the very few number of initial training set (13 chapters out of 114), it would be unrealistic to expect the testing set to be correctly classified at once.

Each chapter was preprocessed by removing unimportant words, stemming, and representation with vector space model.

For the naïve Bayesian classifier two iterations were needed to achieve complete classification of the chapters, while the SVM classifier produced the same result after a single iteration.

After the classification process, results of SVM and naïve Bayesian classifiers were compared, and then each result was validated with the existing classifications of traditional Islamic scholars and western orientalists researchers. The result of this study shows that the two classifiers have produced similar results, with an outperformance of the support vector machine classifier.

Having classified the chapters into Meccan (revealed in Mecca) and Medinan (revealed in Medina), the process of clustering chapters of each category was initiated. Chapters of each category were initially automatically clustered through similarity of their features (a kind of fuzzy-connectivity based clustering approach), then manually adjusted in order to correctly correspond to the major phases of Prophet Mohammad’s messengership.

The major phases of the Prophet’s messengership were manually derived from the Quranic text, as well as from the secondary Islamic literature e.g hadiths, exegesis.

Ten phases have been identified, whereby seven are considered to be Meccan, and three belong to Medinan period. Extending the categorization by places of revelation to the phases of Prophet Mohammad’s messenegership, suggests a better way of effectively understanding the message of the Quran, and thereby optimizing the reconstruction of its chronology as explained in [1].

2. Objectives

This study has got three objectives.

a. Measuring the performance of SVM and naïve Bayesian classifiers for categorizing the Quranic chapters by places of revelation.
b. Comparing the results of this research with the classifications of traditional Islamic scholars, and western orientalists researchers.
c. Grouping Quranic chapters according to different phases of Prophet Mohammad’s messengership.

This research could help understand or figure out the logic behind the existing classifications in general.

It might be important to mention that, the classifiers’ training chapters have been selected based on some internal evidences that, both Islamic scholars and western orientalist researchers have agreed upon, as far as their places of revelation are concerned.

3. Generic Categorization Model based on Phases of Prophet Mohammad’s Messengership

Simply analyzing the text of sacred scriptures without good knowledge of the origin of the text could lead to erroneous conclusions.

According to traditional Islamic scholars point of view, there are mostly a single Meccan phase and a single Medinan phase, whereby each phase has got its own purpose, and cannot be explicitly dissociated from each other. From the angle of western orientalists researchers such as [13, 14, 15] , the Quranic chapters are explicitly separable based on the critical analysis of the chapters. Nöldeke identified three Meccan phases and one Medinan phase, while [16] determined five Meccan phases and one Medinan phase.

In this study, I am proposing a new categorization based on the life of messengers of God in general, and on the phases of Prophet Mohammad’s messengership in particular. Categorizing Quranic chapters according to major phases of Prophet Mohammad’s life is more objective than using the styles of the verses along with personal judgment.

In fact the phases of the Prophet’s life are quite well known, and reasonably imaginable by any ordinary human being.

After having considerably analyzed religious texts by mainly concentrating on the development stages of the tasks of the messengers, I found that almost every messenger undergoes the following stages:

1. Notification of messengership: God informs the person that he is selected to be a messenger of God.
2. Secret announcement / call: the messenger informs others (individuals) secretly about his messengership and the message.
3. Public announcement: the messenger publicly declares his messengership and the message to the people.
4. Rejection / evidences: the people reject the messenger and his message, and the messenger brings arguments, proofs…etc.
5. Negotiation / promises: the people start negotiating with the messenger so that he abandons his message, the messenger replies by giving glad tidings to his followers, and horrible torment to disbelievers, and he avoids too much debates.
6. Threatening: the people start threatening the messenger after negotiations have failed, the messenger persists on his message by mostly strengthening the believers among themselves, showing confidence and calmness, and warning disbelievers through punishments and analogies with anterior disbelievers in the message of God.
7. Persecution: the people start attacking the messenger with his followers physically or violent chidings and speeches.
8. Migration, evidences and hope: the messenger with his followers finally leaves the city, and hope for victory.
9. Resistance and counter-attack: the messenger with his followers start retaliating to the attacks of the people.
10. Victory of the messenger with his message: finally the messenger with his message prevail.

For the messengership of Prophet Mohammad, the Meccan period covers steps 1 to 7, while the Medinan one goes from step 8 to 10.

This method of categorization suggests a better way of effectively understanding the message of the Quran, and thereby optimizing the reconstruction of its chronology as explained in [1].

4. Problem statement

The problem to be solved could be stated as follows:
Given places of revelation of very few chapters, would it be possible with the help of machine learning algorithms to determine the places of revelation of the remaining chapters, and cluster them according to major phases of Prophet Mohammad's messengership?

5. Proposed solution

The basic idea is to organize the Quranic chapters according to major stages of Prophet Mohammad's life using the following steps:

a. First, two groups have to be created using SVM, naïve Bayesian classifiers: chapters revealed in Mecca, and chapters revealed in Medina.

   Some Quranic chapters were revealed in Mecca while others in Medina. Therefore there are only two categories to which the Quranic chapters have to be assigned to. Based on this fact, it is advantageous to solve this problem using a machine learning algorithm that separates data into two distinct groups, with a wide margin between the groups. One of the best algorithms for dividing data into two groups with maximum margin possible is the support vector machine (SVM). For this reason SVM has been used for categorizing the chapters.

   In order to leverage the results obtained from the SVM classification, a naïve Bayesian classifier was used afterwards.

   The categorization process for each classifier was made recursive by regularly augmenting the training set with correctly classified chapters, in order to classify the whole testing set. In fact due to the very few number of initial training set (13 chapters out of 114), it would be unrealistic to expect the testing set to be correctly classified at once.

b. Then, within each group, chapters for each phase of Prophet Mohammad's messengership have to be determined using clustering by similarity of features (a kind of fuzzy-single linkage clustering approach), and the clusters could be manually adjusted if necessary.

6. Data

Knowing the difficulties of developing or finding a software program capable of producing unambiguously roots of Arabic words of the holy Quran, I found it easier and more effective to use the transliteration version. The transliteration resolves significantly the problem of diacritics in the Quran. An electronic version of the transliteration formatted with HTML was downloaded from the Muslimnet [2] website, and preprocessed in order to clean and produce plain text. Practically a wrapper as explained in [5] was used for this task.

7. Prep-processing

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 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.
```

e. Categories creation

There are two categories:

(i) Meccan category holding features of chapters revealed in Mecca, (Meccan chapters with known dates of revelation: 96, 68, 74, 73, 17, 23, 22)

(ii) Medinan category containing features of chapters revealed in Medina, (Medinan chapters with known dates of revelation: 8, 33, 24, 48, 9, 5, 110).
Chapters 2 and 3 were excluded from the training set in order to minimize overfitting problem. They contain several keywords belonging to both Medinan and Meccan classes.

Chapters of each category were combined to produce a single vector for that category. Then the two categories were cross checked, so that common features are deleted. Finally two distinct classes were obtained, which were transmitted to the dimensionality reduction step as explained in the following section.

8. Dimensionality Reduction

This is a kind of global policy 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 [3].
Contrary to the research carried out in [1], in this study, for the training set, the number of features per category was simply adjusted to the number of features of the category having less number of features. The reason for doing that was the fact that, SVM requires more training features in order to produce satisfactory results.

Concerning the testing chapters, the following learning algorithm was used to select the most important keywords.
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

Now that the data is preprocessed, there is need to distinguish between the inputs for each classifier. As for the Bayesian classifier, I wrote a PHP script from scratch that processes the vectors directly and produces the classification result. Concerning the SVM classifier, the LibSVM library was used. The library requires the conversion of input vectors into numerical values. This conversion was carried out as described in the following section.

9. Converting Data into The LibSVM Format

In this step, all the created vectors have to be converted to libSVM format. First the features of the categories were concatenated, and used as columns of a matrix, whose first column was filled as follows:

a. +1 for a training Medinan chapter
b. -1 for a training Meccan chapter
c. 0 for a testing chapter

Each entry of every chapter's vector was checked whether corresponding features (i.e column names) occur in it. If so, the position's value for each corresponding feature was set to one, otherwise it was set to zero.

During the training phase, the matrix contain only corresponding values of training chapters. While during the testing, it consists of the testing chapters.

10. 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, naive Bayesian and SVM classifiers are used, in order to compare their results individually to the existing categorization of Quranic chapters by places of revelation.

i. Support Vector Machine

SVM is basically a linear classifier for finding a hyperplane with maximum Euclidean distance (i.e gap) to data points on the boundaries (support vectors) as shown in fig 1.

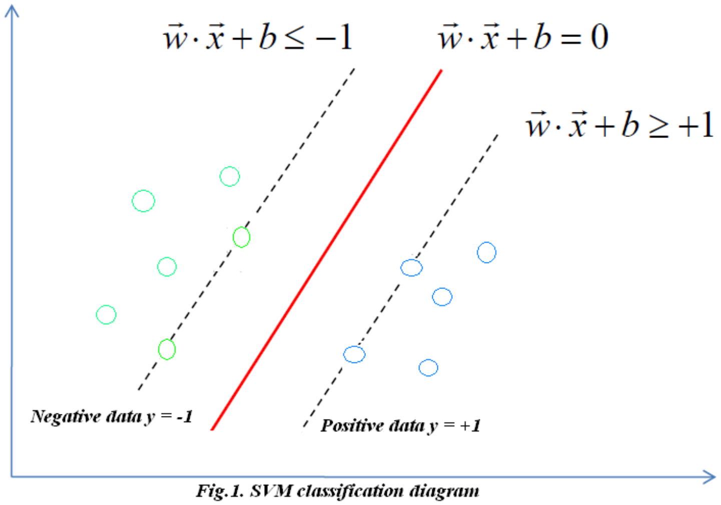
$\vec{w}\cdot\vec{x}+b\le -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\ge +1$
Negative data y = -1
Positive data y = +1

Fig.1. SVM classification diagram

The gap is the distance between parallel hyperplanes:
$\vec{w}\bullet\vec{x}+b=-1$ and $\vec{w}\bullet\vec{x}+b=+1$
Solving the two equations simultaneously yields:
The distance D = $2/\|\vec{w}\|$
Since the aim is to maximize D, then $\|\vec{w}\|$ or equivalently $1/2(\|\vec{w}\|^2)$ has to be minimized.
However when minimizing the denominator of D, we need to make sure that any given point having y=-1 or y=+1 is correctly classified accordingly. This condition results in a constraint that could be formulated as follows:
$y_i(\vec{w}\bullet\vec{x}+b)\ge +1$ for i= 1,…,N
With this constraint at hand, the classifier needs simply to learn this function:
$f(\vec{x}) = sign(\vec{w}\bullet\vec{x}+b)$
SVM is based on the principle of structural risk minimization. In fact, it tries to find the best plane possible so that the lowest true error is guaranteed.
Minimizing the gap D under the defined constraint, suggests that we are dealing with a quadratic programming (QP) optimization problem with *n* variables (w_i, i= 1,…,n), where *n* is the number of

features in the dataset. This form of expressing the optimization problem is called “primal formulation of linear SVMs”.
There is also “dual formulation of linear SVMs“, whereby the n is the number of samples, and the $\vec{w}$ vector is expressed in terms of α_i : $\vec{w} = \sum \alpha_i y_i \vec{x}_i$

For non-linear problems, one could either use soft-margin approach, or map the data into a higher dimensional space called feature space. Then in the feature space, the linear SVM algorithm could be easily applied. Usually it is done with the help of a kernel function, which is a dot product in the feature space.
There are several kernel functions [12] among them polynomials, Gaussian Radial Basis, hyperbolic tangent, exponential functions.

ii. 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 C_i 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,
$P(C_1) = P(C_2) = . . . = P(C_k)$.
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 not 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:
$P(X|C_i) = w_f / (d_f + 1)$

11. Clustering

After having categorized the chapters by cities of revelation, the task of identifying chapters for each phase of Prophet Mohammad’s messengership was tackled.

This process was achieved in two steps:

i. Clustering the chapters of each category using a kind of single linkage clustering algorithm.
The Meccan chapters have to be clustered into 7 groups, and the Medinan ones into 3 groups.

ii. Manually, analyze the generated clusters, and adjust them if necessary.

12. Implementation

a. SVM classification

The LibSVM library is an open source SVM package [11], which contains all the tools required for performing SVM classification.
It supports several basic kernel functions.
However knowing that several Quranic chapters contain both Meccan and Medinan features, the Gaussian radial basis function (RBF) was selected as the kernel function.

Gaussian radial basis function:
$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2)$, for $\gamma > 0$.
Sometimes parameterized using $\gamma = 1 / 2\sigma^2$

b. Naïve Bayesian classifier

For the Bayesian classification, a PHP script was written to directly process the vectors of keywords, and assigns chapters to their corresponding categories.

13. Results and Discussion

a. Using SVM classifier

For the SVM classification, 2212 features were selected.
There were total 103 chapters to be classified. Based on the existing classifications, among the 103 testing chapters, 21 are Medinan, and 82 Meccan.

Chapters 2 and 3 were used among the testing set for cross-validation purpose.

For Medinan category:
During the initial classification, 11 out of the 103 chapters have been correctly classified as Medinan chapters, and 10 misclassified, and they are: 99, 13, 55, 76, 98, 59, 63, 49, 66, and 62.

The precision is: 11/103 = 10.67%
The recall is: 11/21 = 52.38%

These misclassified chapters have been classified through iterative updating of the training set with so far correctly classified chapters. After one iteration the chapters were correctly assigned to the correct Medinan category.

For Meccan category:
82 out of 103 chapters have been correctly classified as Meccan chapters, therefore a 100% success.

The precision is: 82/103 = 79.61%
The recall is: 100%

So the SVM classifier is able to categorize Meccan chapters better than Medinan ones. Why is it so? This is part of the task currently under investigation.

Remarks:
Reducing Meccan features increases classification of Medinan chapters, but reduces classification of Meccan chapters.
Reducing features of Medinan ones does not affect classification of Meccan chapters.

b. Using naïve Bayesian classifier

Using keywords with frequency > 2, and 55 features for each class (because Meccan chapters got 55 keywords remaining), the following results were obtained.

For Medinan category:

During the initial classification, 18 out of 103 testing chapters have been correctly classified as Medinan chapters.
Therefore, 3 Medinan chapters have been misclassified, and they are: 99, 76, 55.

The precision is: 18/103 = 14.47%
The recall is: 18/21 = 85.71%

Even by updating the training set with the classified chapters, these three chapters have always been assigned to Meccan category. After having examined these chapters, I found that, in fact they have got more Meccan features than Medinan ones. So the classifiers (both SVM and naïve Bayes) are quite right to assign them to Meccan category.

Now the question is, why some traditional Islamic scholars have classified them as Medinan chapters?
Even among the Islamic scholars there is disagreement about the places of revelation of these three chapters.
So this study supports the view of those who think they were revealed in Mecca. It might be important to mention that western orientalists researchers such as [13, 14, 15] also classified these chapters as Meccan.
Maybe this study gives us an idea of the logic and procedure that, some Islamic scholars and western orientalists might have followed to derive places of revelation of the Quranic chapters.

Here, it is important to emphasize that, this whole study has not considered background information of the Quranic text reported in some hadiths or some exegesis, which might be in this instance necessary for correctly classifying these three chapters

Therefore, their correct classification can only be achieved with the help of background information reported in the hadiths (narrations of Prophet Mohammad) and tafisr (exegesis)

For Meccan category:

During the initial classification, 53 out of 103 testing chapters have been correctly classified as Meccan chapters, and therefore 29 misclassified.

The precision is: 53/103 = 51.45%
The recall is: 53/82 = 64.63%

The classification of these 29 chapters was achieved through iterative updating of the training set with correctly classified chapters. It took two iterations two completely classify them.

This automatic updating of the training set is a kind of improving the classification process using the unlabeled data. The unlabeled data being the initial testing set from which some features have been extracted, and added to the initial training set after they have been properly classified..

Comparing the classifiers results:

The precision does not tell us a lot since all the 103 chapters are every time classified.
Therefore, it is the recall that makes the difference.
For Medinan chapters, naïve Bayes has got the higher recall 85.71%, while SVM performs 52.38%.
For Meccan chapters, SVM outperformed the naïve Bayes, with 35.37% more.

The overall performance for both classifiers is measured using precision.
For SVM, precision is: (82+11) /103 = 90.29%.
For naïve Bayes precision is: (53+18) = 68.93%

Hence SVM outperforms naïve Bayes with 21.36%.

c. Clustering

After categorizing the chapters by cities of revelation, and using a fuzzy-single linkage clustering algorithm that checks whether clusters share at least a chapter in common, the following clusters were generated.

For the Meccan period:

Cluster 1:
92,36,43,37,75,80,89,93,46,69,102,107,106,56,68,77,94,100,20,35,41,74,32,45,73,108,11,23,1,70,83,84,105,113,114,22,30,39,42,85,88,95,96,103,112,97,26,54,78,82,79,7,
27,71
Cluster 2: 111, 86, 18, 101, 81, 29, 40, 53
Cluster 3: 31, 72, 87, 91, 104, 109, 56
Cluster 4: 50, 38, 15, 51
Cluster 5: 21, 34, 90, 17
Cluster 6: 6, 25, 52, 10, 67, 16
Cluster 7: 14, 19, 28, 44

For the Medinan Period:

Cluster 1: 2, 22, 64, 60, 99, 57, 47, 13, 55, 76, 98
Cluster 2: 8, 3, 33, 9, 66, 62, 63, 58, 59
Cluster 3: 4, 65, 24, 48, 49, 5, 110

These clusters have to be mapped to the phases of Prophet Mohammad's messengership. The above listing of the clusters does not necessarily imply that, they respectively correspond to the stages of the messengership.

The mapping to the phases of messengership is currently being manually carried out, and the result will be published as soon as possible.

14. Conclusion and Future Work

In this study, the performance of SVM, and naïve Bayesian classifiers for categorizing Quranic chapters has been discussed.
The chapters have been classified into Meccan and Medinan categories, and the result has been validated with the existing traditional classification.
The results of the classifiers were compared, and it is found that the SVM outperforms the naïve Bayesian classifier.

A generic categorization model of the chapters based on the phases of the messengership of Prophet Mohammad has also been presented.
Using a fuzzy-single linkage clustering algorithm, the Quranic chapters in each category (i.e revealed in Meccan or in Medina) have been clustered. Chapters of the Meccan category have been clustered into 7 clusters, while the Medinan ones got grouped into 3 clusters. The mapping of the clusters to the phases of the messengership is currently being manually dealt with.

This mapping will be supplemented with the result of [1], whereby the reconstruction of the yearly chronology of the Quran was done.

Reference

[1] Mohamadou Nassourou, "A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran", accepted in WEBIST/WTM 2011, The Netherlands http://nbn-resolving.de/urn:nbn:de:bvb:20-opus-54712

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

[3] 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

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

[5] Mohamadou Nassourou, "Empirical Study on Screen Scraping Web Service Creation: Case of Rhein-Main-Verkehrsverbund (RMV)**",** http:// nbn-resolving.de/urn:nbn:de:bvb:20-opus-49396

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

[7] 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.

[8]Vapnik, V. Statistical Learning Theory. Wiley, Chichester, GB, 1998.

[9] Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, 1997.

[10] Fatimah Wulandini, Anto Satriyo Nugroho, "Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases", International Conference on Rural Information and Communication Technology 2009

[11] C. Chang, C. Lin „LIBSVM -- A Library for Support Vector Machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] Support vector machine, http://en.wikipedia.org/w/index.php?title=Support_vector_machine &oldid=465198758 (last visited Dec. 20, 2011).

[13] Theodor Nöldeke, http://en.wikipedia.org/w/index.php?title=Theodor_N %C3%B6ldeke&oldid=456389258 (last visited Dec. 21, 2011).

[14] Régis Blachère, Le Coran (traduit de l'arabe) (Paris, 1957)

[15] Gustav Weil, http://en.wikipedia.org/w/index.php?title=Gustav_Weil &oldid=460254815 (last visited Dec. 21, 2011).

[16] William Muir, http://en.wikipedia.org/w/index.php?title=William_Muir &oldid=466518722 (last visited Dec. 21, 2011).