Visual Analytics Based Authorship discrimination Using Gaussian Mixture Models and Self Organising Maps: Application on Quran and Hadith

H. Sayoud

http://sayoud.net

Abstract. An interesting way, to analyse the authorship authenticity of a document, is the use of stylometry. However the use of conventional features and classifiers has some disadvantages such as the automatic authorship decision, which usually gives us a speechless authorship classification without (often) any way to measure or interpret the consistency of the results.
In this paper, we present a visual analytics based approach for the task of authorship discrimination. A specific application is dedicated to the authorship comparison between two ancient religious books: the Quran and Hadith. In fact, an important raising question is: could these ancient books be written by the same Author?
Thus, seven types of features are combined and normalized by PCA reduction and three visual analytical clustering methods are employed and commented, namely: Principal Component Analysis, Gaussian Mixture Models and Self Organizing Maps.
The new visual analytical approach appears interesting, since it does not only show the distinction between the author styles, but also sheds light on how consistent was that distinction (i.e. visually).
Concerning the discrimination application on the ancient religious books, the results have shown the appearance of two main clusters: namely a Quran cluster and Hadith cluster. The clusters distinction corresponds to a clear authorship difference between the two investigated books, which implies that the two authors of these two books should be different.

Keywords: Artificial Intelligence, Data mining, Visual analytics, Natural Language Processing, Authorship attribution, Quran.

1 Introduction

Visual Analytics (VA) is defined as the graphical visualisation of the information resulting from an Analytical Modelling (AM). This graphical visualisation represents a bridge between the human and the mathematical results, and helps the experts extracting the important information for taking a decision [Wijk 2013] [Endert 2014] [Blascheck 2016].

It is impossible to dissociate the VA from AM, but in the contrary the two entities have to be associated to help the experts getting clear information from the analysed data.

Authorship Discrimination (AD) [Sayoud 2015-a], which represents a sub-field of stylometry, consists in checking whether two text documents belong to the same author or not. This research field can efficiently respond to some literary disputes with regards to the authentic writer of a document [Sayoud 2015-b].

Mostly, stylometry (or authorship attribution) uses AM computations to evaluate the probability that a specific author could have written a given piece of text. This manner, the user or expert can difficultly manage to make a decision with regards to the real author supposed to be the writer of that document.

The originality of this research work is that we propose a new way of authorship analysis by using the VA approach. Furthermore we propose a new set of linguistic features that are also original in stylometry.

The principal application of our work is the analysis of the authorship authenticity of the Quran. This task is made by applying an authorship discrimination between the Quran, claimed to be from God [Wi1 2012] [Nasr 2013] [Ibrahim 1996], and the Hadith (i.e. statements of the Prophet).

It can be seen why Holmes [Mills 2003] explained that the area of stylistic analysis is the main contribution of statistics to religious studies. For example, early in the nineteenth century, Schleiermacher discussed the authorship of the Pauline Pastoral Epistle 1 Timothy [Mills 2003]. Consequently, other German speaking theologians, namely, F.C. Baur and H.J. Holtzmann, initiated similar studies of New Testament books [Mills 2003].

In such problems, it is crucial to use rigorous scientific tools and it is important to interpret the results very carefully. That is, knowing that authors possess specific stylistic features that make them differentiable [Jiexun 2006], we tried to make some experiments of authorship discrimination [Tambouratzis 2003] [Tambouratzis 2004-a,b] [Tambouratzis 2005] [Tambouratzis 2007] between the Quran (*i.e. words of God*) and the Hadith (*i.e. Prophet's statements*) by using the VA approach. Our corpus consists of the two cited ancient books, which are segmented into text segments of the same size: 14 different text segments for the Quran and 11 different text segments for the Hadith. The segments have a medium size of about 2076 words per text.

2 Stylometric features

Several linguistic features are proposed in the field of authorship attribution [Ranatunga 2013]. We can quote four main types:

Vocabulary based Features: In general, the typical words, an author is used to write, can reveal his or her identity. The problem with such features is that the data can be faked easily. A more reliable method would be able to take into account a large fraction of the words in the document [Juola, 2006] as the average sentence length.

Syntax based Features: One reason that function words perform well is because they are topic-independent [Juola, 2006]. A person's preferred syntactic constructions can be cues to his authorship. One simple way to capture this is to tag the relevant documents for part of speech or other syntactic constructions [Stamatatos, 2001] using a tagger.

Orthographic based features: One weakness of vocabulary-based approaches is that they do not take advantage of morphologically related words. A person who writes of “work” is also likely to write of “working”, “worker”, etc. [Juola, 2006].

Characters based features: Some researchers [Peng, 2003] have proposed to analyze documents as sequences of characters. This type of parameter can replace several other high-level linguistic features. Furthermore, several experiments showed that character n-gram is quite reliable in authorship attribution [Stamatatos, 2009].

In our investigation, a mixture of different features is proposed: Author Related Pronouns (ARP), Father Based Surname (FBS), Discriminative Words (DisW), COST value, Word Length Frequency (WLF), Coordination Conjunction (CC) and Starting Coordination conjunction (SCC). All those features are original and some of them are used for the first time in stylometry (*during the preparation of this work*). The features are collected from the two books to be discriminated and normalized by the maximum so that the different numerical values will range approximately between 0 and 1.
Those seven features are described as follows:

2.1 Author’s Pronoun Based Feature

In Arabic, the pronoun I (إني - أنا) is the most used one for representing the speaker person (i.e. myself). However, in some few cases, the author’s pronouns He (هو) and We (نحن - إنا) are also employed, instead of I, at least in special circumstances.

In fact, most speakers use the pronoun “I”, which is normal, when speaking or writing, like in the following sentence:
"انا سعيد لرؤيتك", meaning « I am happy to see you ».

Sometimes the use of He is also possible, such as:
"سألتني إذا كنت لا أزال سعيد ... لم يعد كذلك", meaning “you asked me if I am still happy… He is no longer so!”.

And sometimes, even the use of We is preferable than I, such as in royal discourses for expressing the splendor of the speaker, as we can notice in the following sentence:
"قال الملك: نحن لا نقبل أي عصيان في المملكة", meaning “the king said: We do not accept any disobedience in the kingdom”.
This great variety of speaker’s pronoun in Arabic makes a great challenge in trying using them in stylometry.

2.2 On the use of “ أبا “ (*father of*) for naming people

In the Arabic language, it is usual to call a person using the name of his oldest child (often the son). That is, if somebody has a son called Youssof for instance, then it is possible to call him *Aba*-Youssof, which can be translated in English into *Father-of*-Youssof. This fact is often noticed in verbal communications, when somebody talks with his companions. Nowadays, although this means of appellation has become less frequently used, it is still widely employed in some countries of the Arabic gulf region and middle-east.

2.3 Frequency of some discriminative words

The key idea is to investigate the use of some words that are very discriminative. In practice, we remarked that such words, for instance: الذين (*in English: THOSE or WHO in a plural form*), are very commonly used by certain speakers. In fact, some authors like the use of the plural form even if the matter does not concern a plural situation, such as in the following sentence: "أين هم الذين يفهمون ما نقوم به؟ ", meaning "Where are those who understand what we do?", instead of "Who understand what we do?". As another example one can cite the word الأرض (*in English: EARTH*), which is frequently used in several Arabic religious books.

2.4 COST parameter Based feature

Usually, when poets write a series of poems, they make a termination similarity between the neighboring sentences of the poem, such as a same final syllable or letter. To evaluate that termination similarity, a new parameter estimating the degree of text chain (*in a text of several sentences*) has been used: the COST parameter [Sayoud 2012].

Thus, the COST parameter for a sentence "j" is computed by adding all the occurrence marks (*values*) between sentence "j" and its neighboring sentences (*sentence "j-1" and sentence "j+1"*). In our case, the occurrence marks concern only the two last letters of the sentence.

The COST parameter, in this case, can give some information on the structure of the text (*ending structure*). In this investigation, it has been employed to see if the analysed documents respect certain regularities in the text structure or not and, if so, to assess the corresponding regularity ratio.

For instance, let us observe the following English poem:

- Life is so short to let things kill our mi**nd** → COST = 2
- What to do in such situations dear frie**nd** → COST = 4
- It is true that it is hard but victory will be in ha**nd** → COST = 3
- This is only a text, but thank you for listeni**ng** → COST = 1

If we consider the 2nd sentence (ending with "**nd**"), we notice that the previous and next sentences (sentence 1 and 3) are ended with the same last 2 characters (i.e. "**nd**"). So by counting the number of similar characters (i.e. (1+1) + (1+1) = 4), we get a COST value of 4. The same procedure is repeated for each sentence until the last one.

2.5 Word length frequency

The fifth feature is the word length frequency. Herein, we must define some technical terms employed in our experiments:

-The word length is the number of letters composing that word.

-The word length frequency F(n) for a specific length ‘n’, represents the number (*in percent*) of words composed of n letters each, present in the text (*In practice we chose n < 11*).

So, if we denote by $F^{i}_{Q}(j)$ the frequency of the words with “j” letters in the i^{th} Quran segment and by $F^{i}_{H}(j)$ the frequency of the words with “j” letters in the i^{th} Hadith segment. Then, the i^{th} Quran segment will be represented by the vector $F^{i}_{Q}(j)$, j=1..10 and the i^{th} Hadith segments will be represented by the vector $F^{i}_{H}(j)$, j=1..10.

2.6 Frequency of the coordination conjunction «و» (meaning AND in English)

The coordination conjunctions represent an interesting type of features, which are widely used in the Arabic literature. In this study, we have limited our investigation to one of the most interesting conjunction, it is the conjunction ” و” , which corresponds to the coordination conjunction AND (*in English*) and which is widely used in Arabic. In fact some previous studies in the matter have shown that it often represents the most frequent word in Arabic text documents, which gives a special importance to that conjunction.

2.7 Frequency of the coordination conjunction «و» at the beginning of sentence

Similarly to the previous feature, herein we are still interested in the frequency of the coordination conjunction ” و”. However, in this case we only keep the conjunctions that are localized at the beginning of sentences, such as in the following sentence:

"والآن، ماذا يجب أن نفعل؟ ", “And now, what should we do?”.

If we carefully look at this sentence, we notice that this coordination conjunction has lost its initial meaning (*i.e. AND meaning*) for another meaning close to “Thus” or “Hence”. So, it could be interesting to try using this type of feature for characterizing some author styles.

3 Visual Analytics based Clustering methods

By definition, the term clustering corresponds to the fact of grouping some things together; which can be physical objects, numerical data, concepts or any sort of elements.

In pattern recognition, cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (*ie. cluster*) are more similar to each other than to those in other groups [Wi2 2014] [Norusis 2008]. This task is commonly used in data mining, statistical data analysis, machine learning and information retrieval.

On the other hand, visual analytics [Blascheck 2016] [Wi3 2014] [Ellis 2010], which is a combination of several fields (*ie. computer science, information visualization and* *graphic design*) is often used in cluster analysis to make the analyst's judgment easier to develop and more objective.

That is, the combination of those two research fields can lead to a strong and efficient analysis tool for handling some classification tasks that could be extremely difficult to perform with conventional analytic tools.

Furthermore, a great advantage of clustering over conventional classification tools is its non-supervised property (*for several clustering techniques*).

Consequently, it appears that the association of visual analytics with clustering analysis may be interesting for solving some stylometric problems, for which we do not possess any training possibility or information to make a supervised classification task. So, it should be extremely motivating to apply them in our application of authorship discrimination (*ie. Quran vs Hadith*).

Concerning the methods using the association of visual analytics with clustering analysis, there exist several approaches that have been proposed during the last five decades, such as: K-mean Clustering, Hierarchical Clustering, Sammon Mapping, Fuzzy C-mean Clustering, Principal Component Analysis, etc.

In our survey, we propose to use the Gaussian Mixtures Models and Self Organizing Maps, separately in order to find out the possible clusters related to the different investigated text segments.

Our corpus consists of two ancient Arabic religious books, namely; the holy Quran and Hadith. The first one represents the words of God, while the second one represents the statements of the Prophet. However, since the sizes of the two books are different, we were obliged to segment them into segments of the same size: there are 14 different text segments for the Quran and 11 different text segments for the Hadith. The segments have more or less the same size in terms of words and the medium size is about 2076 words per text.

3.1 Principal Components Analysis

3.1.1 Definition

Principal component analysis (PCA) can be considered as one of the most interesting results of applied linear algebra. PCA is used abundantly in all forms of analysis - from neuroscience to computer graphics, because it is a simple and non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it [Shlens 2003].

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (*i.e. the first principal component*), the second greatest variance on the second coordinate, and so on.

Consider a data matrix, X, where the sample mean of each column has been shifted to zero and where each of the *n* rows represents a different repetition of the experiment, and each of the *p* columns gives a particular kind of datum (e.g. results of a sensor).

Mathematically, the transformation is defined by a set of *p*-dimensional vectors of weights or *loadings* $W_{(k)}=(w_1,\dots,w_p)_{(k)}$ that map each row vector $X_{(i)}$ of **X** to a new vector of principal component *scores* $t_{(i)}=(t_1,\dots,t_p)_{(i)}$, given by $t_{k(i)}= X_{(i)} . W_{(k)}$ (1)

This is made in such a way that the individual variables of **t** considered over the data set successively inherit the maximum possible variance from **x**, with each loading vector **w** constrained to be a unit vector.

PCA is quite interesting in complex data analysis, when the most important features are not known in advance. Furthermore, by reducing the dimensionality to a lower more consistent one, the data analysis becomes usually easier and more pertinent.

3.1.2 Application of the PCA analysis on the two ancient books

A PCA representation of the data, using the 3 most important eigenvectors, is given in the following figure. The Quran texts are symbolized by red circles, while the Hadith texts are symbolized by blue crosses. In that figure, we can notice that all the Quran documents are grouped together in the right side, while all the Hadith ones are grouped in the left side. That is, the stylistic discrimination can be easily noticeable in the 3D based PCA representation.

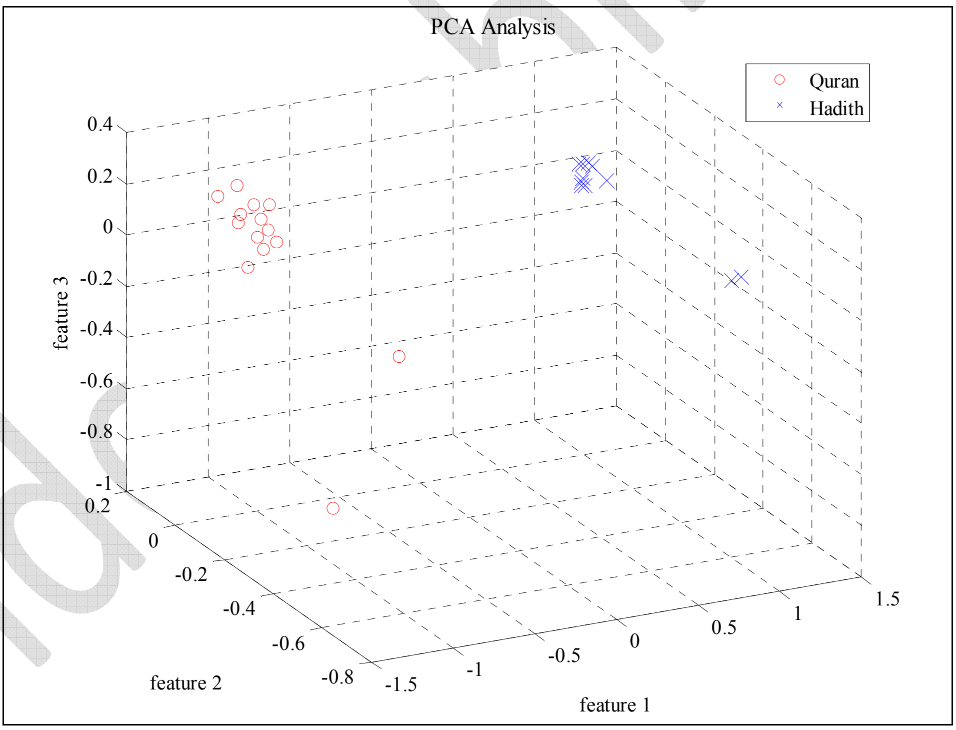

Fig. 1: PCA representation of the data using the 3 most important eigenvectors. Quran texts are represented by red circles while Hadith texts are represented by blue crosses.

3.2 Gaussian Mixture Model based clustering

3.2.1 Definition

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modelling of a wide variety of random phenomena [McLachlan 2003]. Because of their usefulness as an extremely flexible method of modelling, finite mixture models have continued to receive increasing attention over the years, both from a practical and theoretical point of view. For multivariate data of a continuous nature, attention has focused on the use of multi-variate normal components because of their wide applicability and computational convenience. They can be easily fitted iteratively by maximum likelihood via the expectation maximization algorithm.

With a normal mixture model-based approach, it is assumed that the data to be clustered are from a mixture of an initially specified number g of multivariate normal densities in some unknown proportions pi$_1$, …, pi$_g$, That is, each data point is taken to be at realization of the mixture probability density function.

$$f(u;\Psi) = \sum_{i=1}^{g} \pi_i \phi(y;\ \mu_i, \Sigma_i) \qquad (2)$$

where $\phi(y;\ \mu_i, \Sigma_i)$ denotes the p-variate normal density probability function with mean μ_i, and covariance Σ_i.

Here, the vector Ψ of unknown parameters consists of the mixing proportions π_i, the elements of the component means μ_i and the distinct elements of the component covariance matrices Σ_i.

Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. An outright assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging [McLachlan 2001].

3.2.2 Application of the GMM based clustering on the two ancient books

A GMM based clustering is performed after PCA reduction into the 2 most important components. That is, two types of visualizations are provided: a 2D representation (*with those two components*) and a 3D representation including the probability density function as third component (*see figures 2.a and 2.b*).

In both figures, we notice that the different text samples have been clustered into 2 main groups: Quran cluster, at the bottom left side, gathering all the Quran texts and a Hadith cluster at top right, gathering all Hadith texts.

In the 2D representation, the Gaussian mixtures are represented by different ellipsoids surrounding the two clusters, while in the 3D representation, the Gaussians are more visible since they are represented in form of 3D Gaussians surrounding the different clusters.

While, the first representation is sharper, the two representations are similar in terms of clustering information: so, we easily notice that all Quran texts are closely clustered together and all Hadith ones are closely grouped together too. This fact confirms, once again, that the two writing styles of the 2 books are probably different.

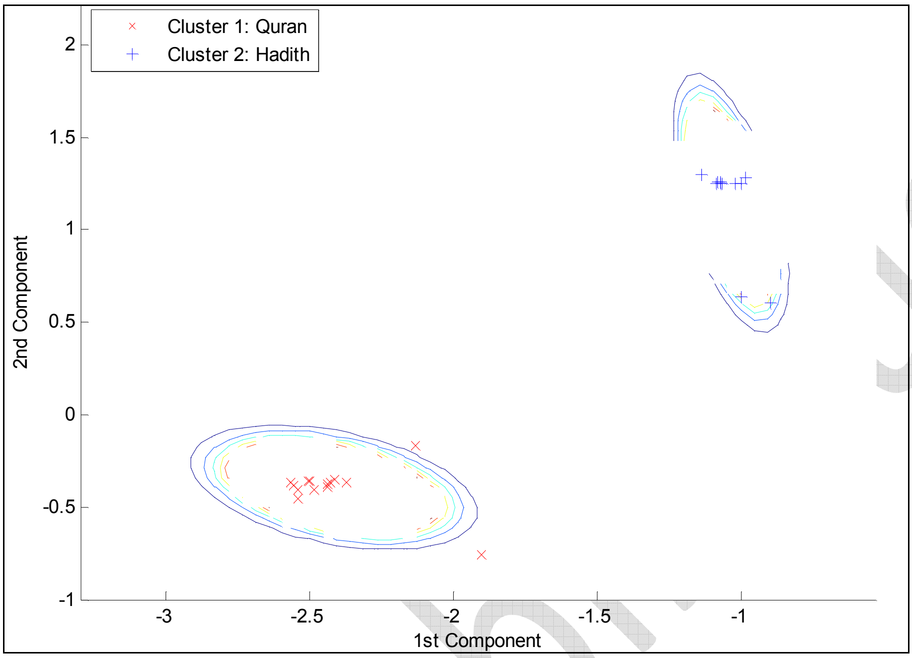

Fig. 2.a: GMM clustering in 2D representation using two components.

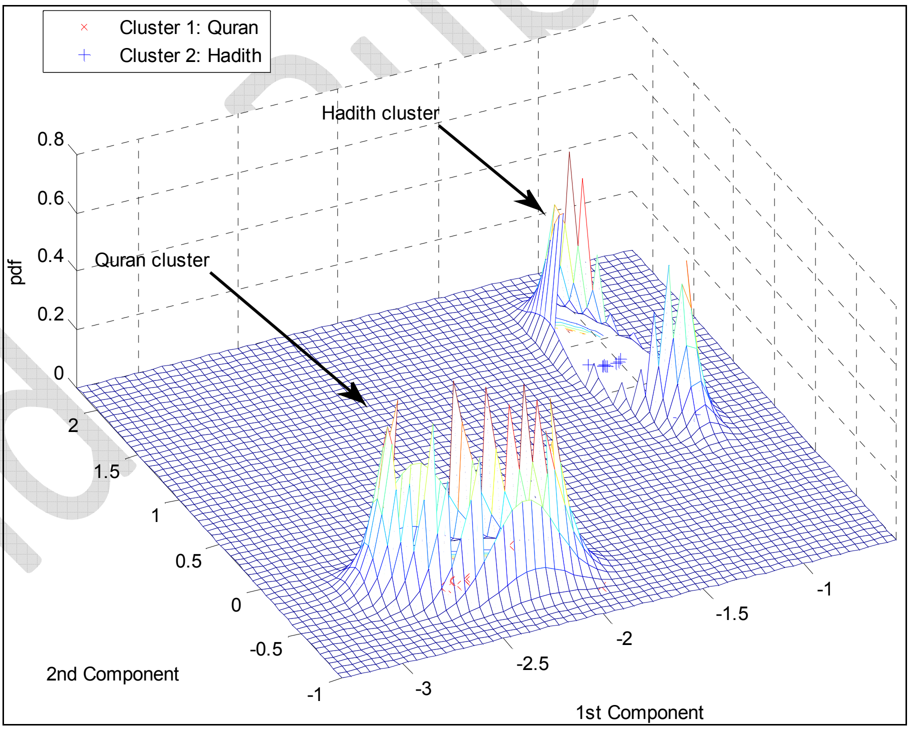

Fig. 2.b: GMM clustering in 3D. The 3rd dimension represents the probability density function.

3.3 Self-Organizing Map based clustering

3.3.1 Definition

A Self-organizing Map is a data visualization technique developed by Teuvo Kohonen in the early 1980's [Kohonen 1990] [Tambouratzis 2003]. The SOMs map multidimensional data onto lower dimensional subspaces, where geometric relationships between points indicate their similarity. SOMs generate subspaces with an unsupervised learning neural network trained with a competitive learning algorithm. The SOM learning tries to make the different parts of the network respond similarly to certain input patterns. This is partly motivated by how visual, auditory or other sensory information is handled in separate parts of the cerebral cortex in the human brain.

The weights of the neurons are initialized either to small random values or sampled evenly from the subspace spanned by the two largest principal component eigenvectors. The network must be fed by a large number of example vectors that represent, as close as possible, the kinds of vectors expected during mapping. The examples are usually administered several times as iterations. The training utilizes competitive learning. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed. The neuron whose weight vector is most similar to the input is called the best matching unit. The weights of the best matching unit and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance from the best matching unit. The update formula for a neuron v with weight vector $W_v(s)$ is

$$W_v(s+1) = W_v(s) + \Theta(u,v,s)\, \alpha(s)\big(D(t) - W_v(s)\big) \tag{3}$$

where "s" is the step index, "t" an index into the training sample, "u" is the index of the BMU for D(t), α(s) is a monotonically decreasing learning coefficient and D(t) is the input vector; Θ(u, v, s) is the neighbourhood function that gives the distance between the neuron u and the neuron v in step s. Depending on the implementations, t can scan the training data set systematically (*t = 0, 1, 2...T-1, then repeat, T being the training sample's size*), be randomly drawn from the data set, or implement some other sampling method such as jackknifing.

The main advantage of using a Self-Organizing Map is that the data is easily interpreted and understood. The reduction of dimensionality and grid clustering makes it easy to observe similarities in the data. The major disadvantage of the SOM is that it requires necessary and sufficient data in order to develop meaningful clusters. Moreover, lack of data will usually add randomness to the groupings.

3.3.2 Application of the SOMs clustering on the two ancient books

A Self-Organizing Map has been performed on the two ancient books (*see figures 3.a and 3.b*). According to those figures, we can easily notice that there are mainly 2 clusters. Hence, in the 3D figure below (figure 3.a), representing the Distance matrix (*inter-distances*), we can see that there are 2 distinct dark regions (*representing low inter-distances*). Those 2 dark regions involve the presence of 2 distinct clusters, since every black area represents a cluster in this case.

In the 2D figure below (*figure 3.b*), a Self-Organizing Map (SOM) using 3 PCA components has been performed. Here, the U-matrix is shown on the left, and an empty grid named 'Labels' is shown on the right. In the left figure (*U-matrix*), we can note 2 main clusters in white (*light colours represent clusters*). The black (*or dark*) cells represent boundaries between clusters, unlike in the previous figure. Hence, one big cluster is visible at the right bottom and another big one at the left top. In the middle figure, the different cells have been labelled (*with regards to the book origin*) by using 2 colours (*red for the Quran and green for the Hadith*), showing what text segments should belong to each cluster. Furthermore the distribution of the data set has been added to the map by using the corresponding hit-histograms. In fact an important tool in data analysis using SOM is called hit-histogram. It is formed by taking a data set, finding the BMU (*Best Matching Unit*) of each data sample from the map, and increasing a counter in a map unit each time it is the BMU (*Best Matching Unit*). The hit-histogram shows the distribution of the data set on the map. Here, the hit-histogram for the whole data set is calculated and visualized on the U-matrix.

Once again, we notice that the Quran samples in red are well grouped together and separated from the Hadith samples in green, by a sharp horizontal black (*dark*) line. Consequently, we can see that the SOM clustering leads to the same conclusion as previously, which is: The two books should have two different authors (*or at least two different writing styles*).

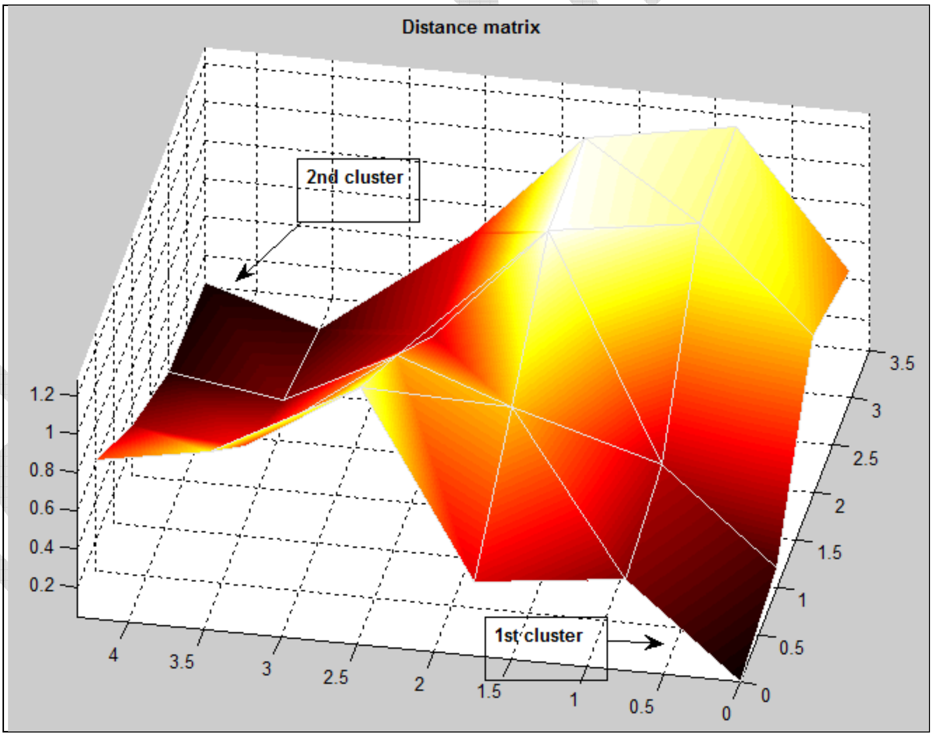

Fig. 3.a: 3D representation of the Distance matrix (*inter-distances*). We can see 2 distinct dark regions (i.e. *low inter-distances*). The 2 dark regions involve the presence of 2 distinct clusters.

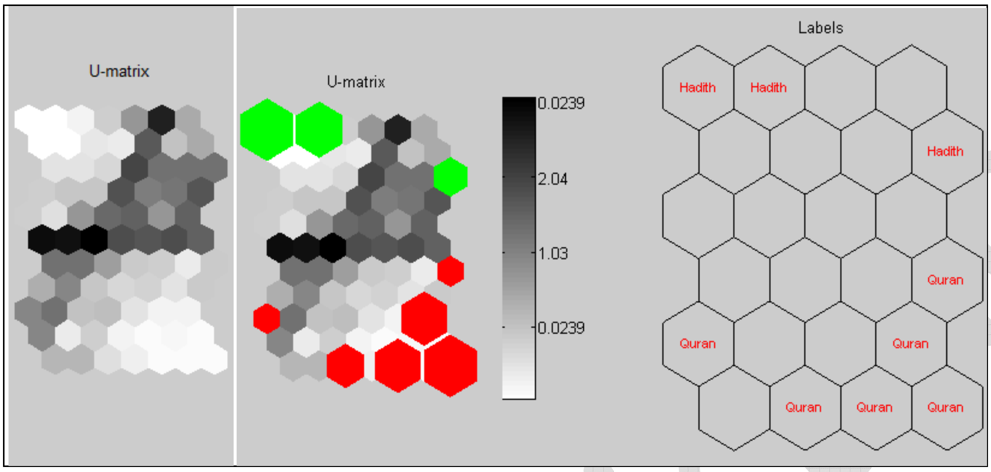

Fig. 3.b: 2D Self-Organizing Map (*SOM*) using 3 PCA components. In the left figure, we can see 2 main clusters in white. In this representation, the light colours represent clusters. Hence, one cluster is visible at the right bottom and another one at the left top. In the middle figure, the different cells have been labelled by using 2 colours (*green for the Hadith and red for the Quran*), showing what are the cells that should belong to each cluster. In the right, we have only the labels of the different SOM cells.

4 Discussion

In this investigation, we proposed a new set of linguistic features that are original and not used previously by other scientific communities. Furthermore, we proposed a new graphical/visual way to analyse the authorship authenticity of a document (*i.e. Visual Analytics*) by using three approaches: PCA, GMM and SOM clustering techniques. The visual analytical approach revealed much more information since it does not only show the distinction between the author styles, but also sheds light on how consistent was that distinction. That consistency could be visually estimated thanks to the 3D or 2D separation distance between text samples. So, the visualization diagrams of the clustering techniques are interesting since they allow one to see how near or far a particular text sample is from an author's style.

That is, seven new stylometric features have been proposed, and three visual analytics based clustering approaches have been employed:

- The first approach (*i.e. PCA*) is not a clustering method, but it is only used for feature reduction and low dimensional mapping. The resulting 3D representation, in figure 1, shows the position of the text samples thanks to their 3 first PCA coordinates. We can observe that the Quran texts are located in the left, whereas the Hadith ones are located in the right and that the two text groups are quite separated. We also notice that Hadith samples are more condensed than Quran ones. The PCA representation suggests that the two books have two different author styles.

- The second approach, namely GMM, is a clustering technique based on mixture models. In our case it is used with only 2 components (*2D representation*). As we can see in figures 2.a and 2.b, two main GMM-based clusters have been obtained: Quran cluster at the left bottom (*grouping all Quran segments*) and Hadith cluster at the right top (*grouping all Hadith segments*). Consequently, and thanks to this 2D representation, the two books appear to have two different author styles.
- The third approach (*i.e. SOM*) is a self organizing neural network, which makes a 2D representation of the different possible clusters in an interesting way, since it gives a rich amount of information regarding the clusters and their consistency. The resulting SOM mapping (*figures 3.a and 3.b*) shows a dark horizontal region separating the different SOM cells into two main regions: Top and Bottom. The bottom area contains Quran segments, while the top area contains Hadith segments. Furthermore, one can observe that Hadith is represented by a *big* sub-cluster in the left and a small one in the right (top area), whereas Quran is represented by a big sub-cluster in the right and a small one in the left (*bottom area*). However, in the overall, the two main clusters are well separated one from the other: Quran samples in the bottom and Hadith samples in the top, which implies that there are 2 different author styles: one author style common to all Quran segments and another author style common to all Hadith segments.

We recall that every book has been segmented into several text segments and that there is no prior information on how could be the general clusters configuration.
That is, knowing that there are two sets of texts: $\{Q_1, Q_2, \ldots Q_N\}$ and $\{H_1, H_2, \ldots H_M\}$, which are extracted from 2 different books: Quran and Hadith respectively, it is quite evident to get interesting information from the number of obtained clusters and the text segments contained within each cluster. For instance:

a) If we get only 1 cluster, this means that probably the different texts are written by the same author;

b) However, if we get 2 clusters and all the Q_i texts are grouped into 1 cluster and all the H_j texts are grouped into another distinct cluster, this implies that the two books are probably written by two different authors.

That is, by exploring the results section and by observing the entire clusters and texts disposition in those clusters, we easily see that all the results, we got, correspond to the second case (*case b*). Consequently and statistically speaking, it appears that the two investigated books (*Quran and Hadith*) have 2 different writing styles, which suggests the hypothesis of 2 different authors.
Hence, it appears that it would be interesting to associate the visual analytics to authorship authenticity analysis for getting a better interpretation of the results and a more consistent decision.

References

[Blascheck 2016] T. Blascheck ; M. John ; K. Kurzhals ; S. Koch ; T. Ertl. VA2: A Visual Analytics Approach for Evaluating Visual Analytics Applications. IEEE Transactions on Visualization and Computer Graphics. Volume: 22, Issue: 1, Jan. 31 2016.

[Edge 2017] D. Edge ; N. H. Riche ; J. Larson ; C. White. Beyond Tasks: An Activity Typology for Visual Analytics. Published in: IEEE Transactions on Visualization and Computer Graphics, Date of Publication: 29 August 2017. DOI: 10.1109/TVCG.2017.2745180.

[Ellis 2010] G. Ellis and F. Mansmann, VisMaster, Visual Analytics. Mastering the Information Age. Chapter 2 http://www.vismaster.eu/book/chapter-2-visual-analytics. Editor: (Scientific Coordinator of VisMaster) Daniel Keim Jörn Kohlhammer, Edition of 2010.

[Endert 2014] Endert, A., Hossain, M.S., Ramakrishnan, N. et al. The human is the loop: new directions for visual analytics J Intell Inf Syst (2014) 43: 411. doi:10.1007/s10844-014-0304-9

[Ibrahim 1996] I. A. Ibrahim. A brief illustrated guide to understanding Islam. Library of Congress, Catalog Card Number: 97-67654, Published by Darussalam, Publishers and Distributors, Houston, Texas, USA. Web version: http://www.islam-guide.com/contents-wide.htm, ISBN: 9960-34-011-2.

[Jiexun 2006] Li, J., Zheng, R., and Chen, H. (2006). From fingerprint to writeprint. Communications of the ACM, vol 49, No 4, April 2006, pp. 76-82.

[Kohonen 1990] T. Kohonen. The self-organizing map. Proceedings of the IEEE (Volume:78 , Issue: 9) Invited Paper, pp 1464 – 1480 DOI: 10.1109/5.58325 1990.

[McLachlan 2001] G. J. McLachlan, S. K. Ng, D. Peel. On Clustering by Mixture Models,Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Munich, March 14–16, 2001, pp 141-148]

[McLachlan 2003] G.J. McLachlan, D. Peel, R.W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis, Volume 41, Issues 3–4, 28 January 2003, Pages 379–388

[Mills 2003] D. E. Mills. Authorship Attribution Applied to the Bible. Master thesis, Graduate Faculty of Texas, Tech University, 2003.

[Nasr 2013] S. H. Nasr, Encyclopædia Britannica Online. http://www.britannica.com/eb/article-68890/Quran, Last access in 2013.

[Norusis 2008], M. Norusis. Cluster Analysis, Chapter 16.,pp:361-391. SPSS 17.0 Statistical Procedures Companion, Marija Norusis , 2008. Pearson editor, Published in 2008.

[Ranatunga 2013] S.P.K Ranatunga. Finding Efficient Linguistic Feature Set for Authorship Verification. Journal of Computer Science, 2013. Vol. 1, No. 1 (2013) pp 35-43.

[Sayoud 2015-a] H. Sayoud, Title: Segmental Analysis Based Authorship Discrimination between the Holy Quran and Prophet's Statements. Digital Studies Journal. 2015: 2014-2015, Open Issue. ISSN: 1918-3666.

[Sayoud 2015-b] H. Sayoud, A Visual Analytics based Investigation on the Authorship of the Holy Quran . The 6th International Conference on Information Visualization Theory and Applications (IVAPP'2015), March 11-14, 2015, pp 177-181.

[Shlens 2003] J. Shlens, 2003, A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS - Derivation, Discussion and Singular Value Decomposition. Version 1, 2003. www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf.

[Stamatatos 2001] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," Computers and the Humanities, Vol. 35, No. 2, pp. 193–214, 2001.

[Stamatatos 2009] E. Stamatatos 2009. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, Vol. 60, No. 3, pp. 538-556, 2009, Wiley.

[Tambouratzis 2003] G. Tambouratzis, G. Hairetakis, S., Markantonatou & G. Carayannis (2003). Applying the SOM Model to Text Classification According to Register and Stylistic Content. International Journal of Neural Systems, 13(1), 1-11.

[Tambouratzis 2004-a] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis & D. Tambouratzis (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 1: Diglossia in Stylistic Analysis. Literary and Linguistic Computing, 19(2), 197-220.

[Tambouratzis 2004-b] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis & D. Tambouratzis (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 2: Extending the Feature Vector to Optimize Author Discrimination. Literary and Linguistic Computing, 19(2), 221-242.

[Tambouratzis 2005] G. Tambouratzis (2005). Assessing the Effectiveness of Feature Groups in Author Recognition Tasks with the SOM Model.. IEEE Transactions on Systems, Man & Cybernetics – Part C: Applications and Reviews, 36(2), 249-259.

[Tambouratzis 2007] G. Tambouratzis & M. Vassiliou (2007). Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments. Literary and Linguistic Computing, 22(2), 207-224.

[Van 2013] J.J.van Wijk. Evaluation: A Challenge for Visual Analytics. Computer journal ISSN 0018-9162, vol. 46, no7, pp. 56-60, 2013.

[Wi1 2012] Quran. The free encyclopedia. Wikipedia, Last modified in 2011, http://en.wikipedia.org/wiki/ Quran

[Wi2 2014] Cluster analysis, Wikipedia. Last modified on November 12, 2014, http://en. wikipedia.org/wiki/Cluster_analysis

[Wi3 2014] Visual Analytics, Wikipedia. http://en.wikipedia.org/wiki/Visual_analytics]. last modified on October 09, 2014.