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

Halim Sayoud(✉)

USTHB University, Algiers, Algeria
halim.sayoud@uni.de, halim.sayoud@gmail.com

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 on, 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 separated clusters: namely a Quran cluster and Hadith cluster. The clusters distinction corresponds to a clear authorship difference between the two investigated documents, which implies that the two books (i.e. Quran and Hadith) come from two different Authors.

Keywords: Artificial intelligence · Data mining · Visual analytics
Natural language processing · Authorship attribution · Quran authorship

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 [1]. 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.

M. Mouhoub et al. (Eds.): IEA/AIE 2018, LNAI 10868, pp. 158–164, 2018.
https://doi.org/10.1007/978-3-319-92058-0_15

Authorship Discrimination (AD) [2], 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 [3]. 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 [4], and the Hadith (i.e. statements of the Prophet). Our corpus consists of the two 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. We can quote four main types: Vocabulary based Features, Syntax based Features, Orthographic based features and Characters based features.

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*). Those 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). 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». 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. 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. That is, if somebody has a son called Youssof for instance, then it is possible to call him *Aba*-Youssof, which can be translated into *Father-of*-Youssof. This fact is often noticed in verbal communications, when somebody talks with his companions.

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. As other 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 proposed: the COST parameter [5].

2.5 Word Length Frequency

The fifth feature is the word length frequency, which 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 choose n < 11*).

2.6 Frequency of the Coordination Conjunction «و» (Meaning AND)

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*).

2.7 Frequency of the Conjunction «و» at the Beginning of Sentence

Herein we are still interested in the frequency of the 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: "And now, what should we do?".

3 Visual Analytics Based Clustering Methods

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 (*i.e. cluster*) are more similar to each other than to those in other groups [6]. On the other hand, visual analytics [1, 7], which is a combination of several fields (*i.e. 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. 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 (*i.e. Quran vs Hadith*). 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 the two ancient books: Quran and Hadith. However, since the sizes of the two books are different, we segmented 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 the same size and the medium size is about 2076 words per text.

3.1 Principal Components Analysis

A PCA representation of the data, using the 3 most important eigenvectors, is given in Fig. 1. We can notice that all the Quran documents are grouped together in the right side, while all the Hadith ones are separately grouped in the left side.

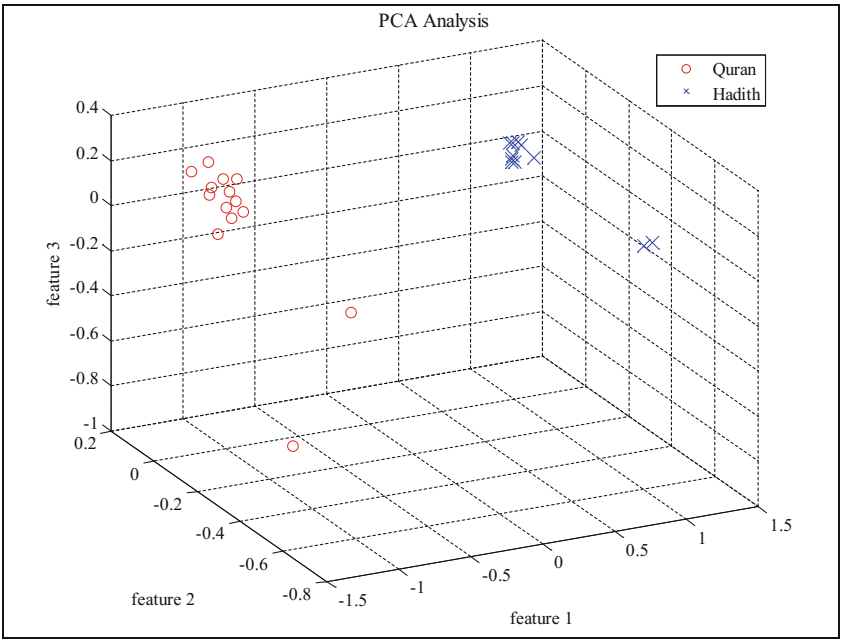

Fig. 1. PCA representation of the Quran (circles) and Hadith (crosses).

3.2 Gaussian Mixture Model Based Clustering

A GMM based clustering is performed after PCA reduction into the 2 most important components. 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. The Gaussian mixtures are represented by different 3D gaussians surrounding the two clusters (Fig. 2). This fact confirms, once again, that the writing styles of the 2 books are probably different.

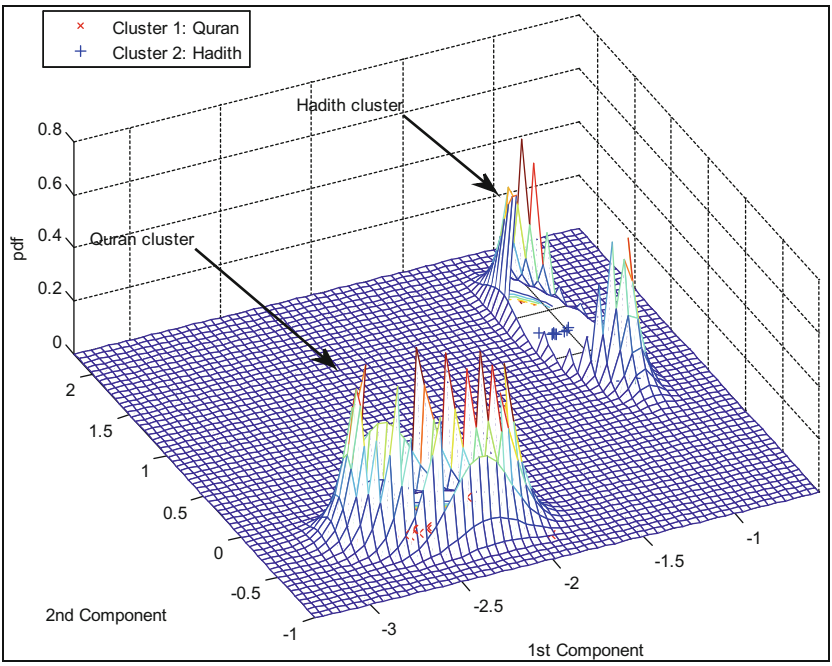

Fig. 2. GMM clustering in 3D. The 3rd dimension represents the probability density function.

3.3 Self-Organizing Map Based Clustering

In Fig. 3, a Self-Organizing Map (SOM) using 3 PCA components is performed. The U-matrix is shown on the left, and a grid named Labels is shown on the right.

In the left 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*). 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 representing a boundary between the two classes. Consequently, we can see that the SOM clustering leads to the same previous conclusion: the two books should have two different authors.

4 Discussion

In this investigation, we have proposed a new set of linguistic features that are original and not used previously. Furthermore, we have proposed a new graphical way to analyse the authorship authenticity of a document by using three approaches: PCA, GMM and SOM techniques. The different results led to the following conclusions:

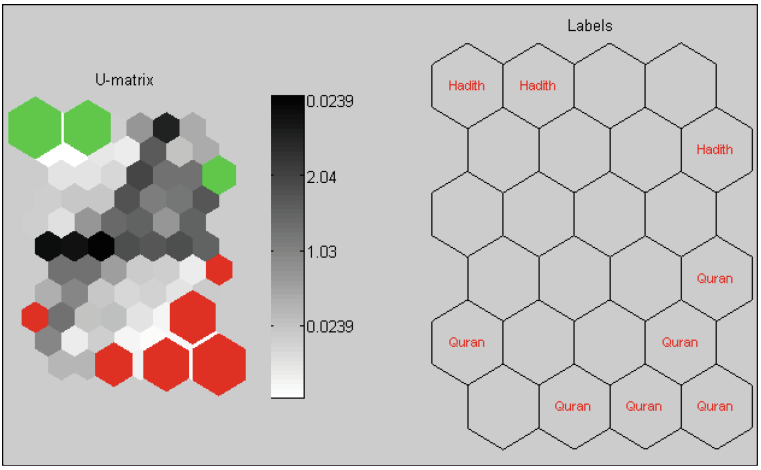

Fig. 3. 2D Self-Organizing Map (*SOM*). We can see 2 main clusters: one cluster is visible at the right bottom and another one at the left top. The different cells have been labelled by using 2 colours (*green for the Hadith and red for the Quran*). The dark lines represents boundaries. (Color figure online)

- Visual Analytics is interesting and promising in the field of authorship attribution.
- Although the first approach (*i.e. PCA*) is not a clustering method, the resulting 3D representation suggests that the two books have two different author styles.
- The second approach, namely GMM, is a clustering technique based on gaussian mixture models. According to the 3D representation, the two books appear to have two different author styles, too.
- The third approach (*i.e. SOM*) is a self organizing neural network, which makes a 2D representation of the different possible clusters. The resulting mapping shows that there are also two different author styles: one for the Quran and one for the Hadith.

Consequently, it appears that the two investigated books (*Quran and Hadith*) have 2 different writing styles, which suggests the hypothesis of 2 different authors.

References

1. Blascheck, T., John, M., Kurzhals, K., Koch, S., Ertl, T.: VA2: a visual analytics approach for evaluating visual analytics applications. IEEE Trans. Vis. Comput. Graph. **22**(1), 61–70 (2016)
2. Sayoud, H.: Segmental analysis based authorship discrimination between the Holy Quran and Prophet's statements. Digital Stud. J. 2014–2015 (2015)
3. Sayoud, H.: A visual analytics based investigation on the authorship of the Holy Quran. In: International Conference on Information Visualization Theory and Applications (IVAPP'2015), 11–14 March 2015, pp. 177–181 (2015)

4. Ibrahim, I.A.: A brief illustrated guide to understanding Islam. Library of Congress, Darussalam Publishers, Houston. www.islam-guide.com/contents-wide.htm
5. Sayoud, H.: Author discrimination between the Holy Quran and Prophet's statements. Literary Linguist. Comput. **27**(4), 427–444 (2012)
6. Norusis, M.: Cluster analysis. In: SPSS 17.0 Statistical Procedures Companion, Marija Norusis, pp. 361–391. Pearson editor (2008). Chap. 16
7. Ellis, G., Mansmann, F.: VisMaster, Visual Analytics. In: Mastering the Information Age. Scientific Coordinator of VisMaster. Daniel Keim Jörn Kohlhammer (2010). Chap. 2