A Visual Analytics based Investigation on the Authorship of the Holy Quran

Halim Sayoud
USTHB University, Algiers, Algeria

Keywords: Visual Analytics, Authorship Attribution, Pattern Recognition, Natural Language Processing, Religious Books.

Abstract: In this paper, we present a visual analytics based investigation for the task of authorship attribution of the holy Quran with regards to the Hadith Author (*the Prophet*). This can be seen as an authorship discrimination task between the two religious books: Quran vs Hadith. The first book represents the Divine book written by Allah (*God*) as claimed by the Prophet Muhammad, whereas the second one represents a collection of certified Prophet's statements.
Two visual analytics clustering methods are employed, namely: a Hierarchical Clustering and Fuzzy C-mean Clustering. On the other hand, seven types of NLP features are combined and normalized by PCA reduction before the classification process. The visual analytics results have revealed interesting results in 2D and 3D disposition. In summary, they show two main clusters in both experiments: Quran cluster and Hadith cluster; and the disposition of the resulting clusters corresponds to a clear authorship distinction between the two religious books.

1 INTRODUCTION

In the Islamic religion, the Quran is defined as the Divine book that was written by Allah and transmitted to the Prophet Muhammad by the Anger Gabriel (Wi1, 2011) (Nasr, 2013) (Ibrahim, 2014). Also, as claimed and confirmed by the Prophet, the holy Quran was written by God (*Allah*) and only transmitted to him (*the Prophet*). Furthermore, in the Quran, a clear verse says: « O Messenger (Muhammad)! transmit (the Message) which has been sent down to you from your Lord. And if you do not, then you have not conveyed his Message. Allah will protect you from people. Allah do not guide the people who disbelieve » [5:67].

So, basically there is no dispute on the origin of the Quran between the Islamic scholars, since the two concerned authors: Allah and his Prophet agree and confirm that it is really written by Allah and not by the Prophet.

However, some doubts on the origin of the Divine book assume that it could be written by the Prophet (Al-Shreef, 2009).

So, now, a raising question would be: "Was the Quran written by the Prophet?"

To respond to that question, it is important to handle the problem with a great delicacy and a maximum of scientific rigor.

The Quran is taken in its entirety in electronic unicode form. It should be conform with the official Saudi holy Quran, which is agreed by the Saudi ministry of religious affairs and which represents the universal book recognized by the Islamic community.

Similarly, for the Christian religion, there exist several disputes about the origin of some texts of the Bible. Such disputes are very difficult to solve due to the delicacy of the problem, the religious sensitivity and because the texts were written a long time ago. One of the purposes of stylometry is authorship attribution, which is the determination of the author of a particular piece of text for which there is some dispute about its writer, as reported by Mills (Mills, 2003). Hence, 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 disputed the authorship of the Pauline Pastoral Epistle 1 Timothy. As a result, other German speaking theologians, namely, Baur and Holtzmann, initiated similar studies of New Testament books (Mills, 2003).

Sayoud H..
A Visual Analytics based Investigation on the Authorship of the Holy Quran.
DOI: 10.5220/0005355601770181
In *Proceedings of the 6th International Conference on Information Visualization Theory and Applications* (IVAPP-2015), pages 177-181
ISBN: 978-989-758-088-8

As mentioned previously, 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 making them differentiable, we tried to make some experiments of authorship discrimination between the Quran and some Prophet's statements in order to see whether the Quran was written by the Prophet Muhammad or not. For this purpose, 7 types of features are extracted and 2 different clustering methods based on Visual Analytics are employed.

2 STYLOMETRIC FEATURES

Several linguistic features have been proposed in the field of authorship attribution. 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. Those seven features are collected from the two religious books and normalized by the maximum so that the different numerical values will range approximately between 0 and 1.

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 (*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 (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 main task of authorship discrimination (ie. Quran vs Hadith).

As for the clustering methods, in the present survey, we have used two different methods separately and tried to observe and comment the resulting clusters thoroughly. The employed methods are: Hierarchical Clustering and Fuzzy C-mean Clustering.

3.1 Dataset Description

The two books have been segmented into 25 several text segments (14 for the Quran and 11 for the Hadith). Furthermore, there is no intersection between them and there is no prior information on how could be the general configuration of the clusters (resulting clustering).

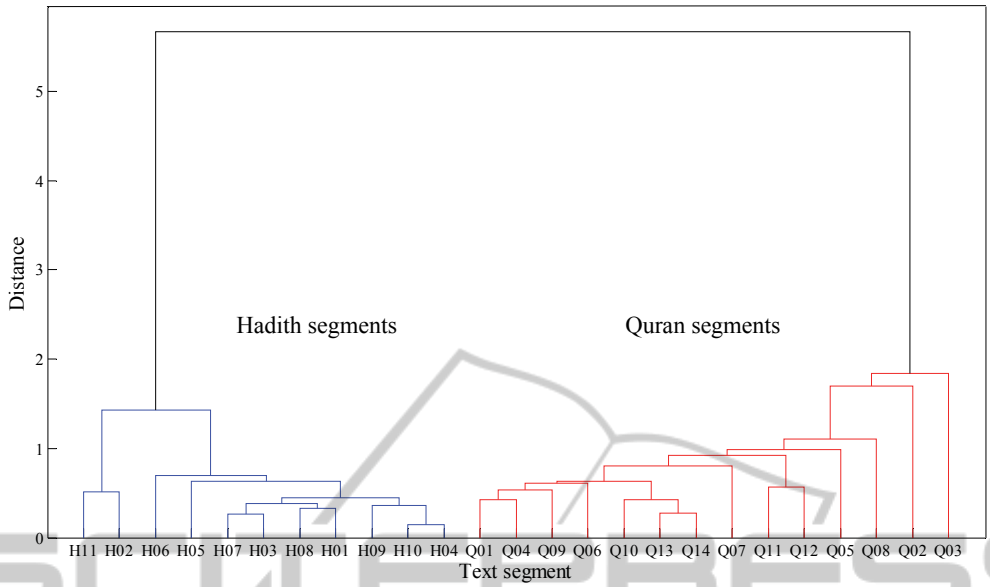

Figure 1: Results of the Hierarchical Clustering.

The global text dataset is then decomposed into two sets of texts: {Q1, Q2, …Q14} for the Quran and {H1, H2, …H11} for the Hadith.

3.2 Hierarchical Clustering

Our first method is based on a hierarchical clustering. The hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters (Greenacre, 2014). In general, there are two types:

-Agglomerative clustering: This is a "bottom up" approach, where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

-Divisive clustering: This is a "top down" approach, where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In our case, we used the first clustering type with a Manhattan distance measure, which is defined below. If we assume that X and Y re two vectors, then Manhattan distance (between those 2 vectors) is given by the following equation:

$$d(X,Y) = \sum_i |x_i - y_i| \quad (1)$$

The resulting linkage of the different documents is called “Dendrogram”. It represents the different possible clusters in a graphical way. By observing the dendrogram, it will be possible to estimate the actual number of clusters and the corresponding documents for each cluster, since all similar documents should be linked together with a consistent linkage.

The Result of the Hierarchical clustering, also called dendrogram, is given in the following figure.

3.3 Fuzzy C-means Clustering

Our second method is based on a Fuzzy C-mean clustering. Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic (Suganya, 2012). So, every point has a degree of belonging to clusters, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be *in the cluster* to a lesser degree than points in the center of cluster. That is, any point x has a set of coefficients giving the degree of being in the kth cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$C_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m} \quad (2)$$

In 3D or 2D dimensions, Fuzzy C-mean can provide a graphical representation of the different samples and the corresponding clusters to which they should belong. This representation allows separating the different samples with regards to their similarities automatically and in a visual manner.

The Result of the Fuzzy C-mean clustering, in

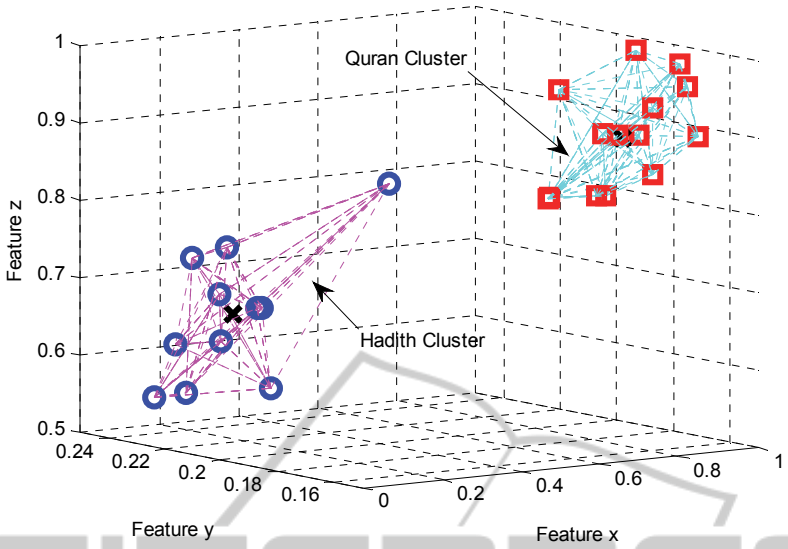

Figure 2: Results of the Fuzzy C-mean Clustering.

3D representation, is given in the following figure.

4 RESULTS SUMMARIZATION

In this investigation, two visual analytics based clustering approaches have been employed to make a visual authorship clustering of 25 religious text segments.

- In the first approach (Hierarchical clustering), the resulting dendrogram shows two separated sharp clusters (the Quran cluster in the right and Hadith cluster in the left). We can see that there is no intersection between the different clusters and that the final linkage is extremely weak since the corresponding distance is very large. This result shows that there are probably two authors: Quran Author and Hadith Author.
- In the second approach (Fuzzy C-mean clustering), which is an automatic clustering technique, the resulting 3D representation shows two main clusters: a Quran cluster located at the top right area and Hadith cluster located at the bottom left area of the 3D representation. Although the Quran cluster is more condensed, the two sets of text segments have been automatically organized into 2 sharp clusters (with different symbol markers), showing that there are two main authors: Quran Author and Hadith Author and that the two authors are different.

5 DISCUSSION

Some experiments of authorship discrimination have been conducted on two religious books: the holy Quran and Hadith, in a visual analytics way. And as described in the beginning of this manuscript, several original features are used to make a stylometric comparison between the two books. On the other hand, the task of comparison is ensured by two clustering approaches based on visual analytics techniques: namely, a Hierarchical Clustering and Fuzzy C-mean Clustering.

Furthermore, every clustering method is performed alone and the resulting clusters are commented regardless of the other classifiers results. We recall that every book has been segmented into 25 several text segments (14 for the Quran and 11 for the Hadith) and that there is no prior information on how could be the general configuration of the clusters (resulting clustering).

That is, knowing that there are two sets of texts: {Q1, Q2, …Q14} and {H1, H2, …H11}, 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:

i. If we get only 1 cluster, this means that probably the different texts are written by the same author (one author);
ii. Also, if we get several clusters, but some Quran texts are grouped with some Hadith ones

(in a same cluster), this means that some Quran texts were probably written by the Hadith author;

iii. However, if 2 clusters appear in the clustering display and all the Quran texts are grouped in one cluster and all the Hadith texts are grouped in another distinct cluster, this will implies that the two books (Quran and Hadith) are written by 2 different authors or at least with 2 different styles.

That is, by exploring the results section and by observing all the clusters and texts disposition in those clusters (and since the topics and genres are quite similar for the two books), we easily see that all the results, we got, correspond to the third case (case iii). In other words, all the clustering methods led to 2 distinct clusters (one cluster containing the Quran texts and another one containing the Hadith texts) in a visual way.

Consequently and statistically speaking, it appears that the two investigated books (Quran and Hadith) belong to 2 different authors (or at least 2 different writing styles).

REFERENCES

Al-Shreef A. Is the Holy Quran Muhammad's invention? http://www.quran-m.com/firas/en1/index.php?option=com_content&view=article&id=294:is-the-holy-quran-muhammads-invention-&catid=51:prophetical&Itemid=105. Last access in Nov. 2012.

Ellis G. and Mansmann F., 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.

Greenacre M., Hierarchical cluster analysis, online documentation, Chapter 7, pp : 7.1-7.11. Last visit on November 23, 2014. http://www.econ.upf.edu/~michael/stanford/maeb7.pdf.

Ibrahim I. A.. 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. Last access in 2014.

Juola P.. JGAAP: Authorship Attribution. Foundations and TrendsR in Information Retrieval Vol. 1, No. 3 (2006) 233–334, Now Publisher.

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.

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

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

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

Peng F., Shuurmans D., Keselj V., and Wang S., "Language independent authorship attribution using character level language models", In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 267-274, 2003.

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

Stamatatos E. 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.

Suganya R. and Shanthi R., Fuzzy C- Means Algorithm-A Review, International Journal of Scientific and Research Publications, Volume 2, Issue 11, November 2012, www.ijsrp.org.

Wi1 2011. 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, from Wikipedia website. http://en.wikipedia.org/wiki/Visual_analytics. last modified on October 09, 2014.