ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 1, No.19 , January 2016
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Text Anomalies Detection Using Histograms of Words

Abdulwahed Almarimi[1] and Gabriela Andrejková[2]

[1,2]Institute of Computer Science, Faculty of Science, P. J. Šafárik University in Košice
04001 Košice, Slovakia
abdoalmarimi@gmail.com
gabriela.andrejkova@upjs.sk

Abstract

Authors of written texts mainly can be characterized by some collection of attributes obtained from texts. Texts of the same author are very similar from the style point of view. We can consider that attributes of a full text are very similar to attributes of parts in the same text. In the same thoughts can be compared different parts of the same text. In the paper, we describe an algorithm based on histograms of a mapped text to interval $\langle 0,1 \rangle$. In the mapping, it is kipped the word order as in the text. Histograms are analyzed from a cluster point of view. If a cluster dispersion is not large, the text is probably written by the same author. If the cluster dispersion is large, the text will be split in two or more parts and the same analysis will be done for the text parts. The experiments were done on English and Arabic texts. For combined English texts our algorithm covers that texts were not written by one author. We have got the similar results for combined Arabic texts. Our algorithm can be used to basic text analysis if the text was written by one author.

Keywords: *Authorship attribution, stylometry*, anomaly detection, histogram

1. Introduction

In the text processing, there are solved many problems connected to an authorship of texts, for example external plagiarism, internal plagiarism, authorship verification, document verification, authorship attribution. The problems are followed in a PAN competition [http://pan.webis.de/], benchmarked texts are on web page [1], [2] and [3]. The Authorship Attribution (AA) problem is formulated as a problem to identify the author of the given text from the group of potential candidate authors. The solutions of AA problem are based on basic features extracted from the text (statistics information, stylistics information, syntax information and semantics information). Some interesting approaches to solving of the problem can be found in [4], [5], [6],[7],[8] and [9].
Some different situation is in the Text Anomaly Detection (TAD) problem. The problem is formulated as an identification of text parts they have unseen behavior in a comparison to full text or to the other parts of the text. The result on anomalies could answer to the question "if the given text was written by one author or it has some parts probably written by the other authors". In the paper we developed an algorithm to cover some anomalies in the texts using histograms of mapped texts to time sequences and modified by a kernel smoothing function. The motivation to the algorithm we found in [6].

The paper is written in the following structure: The second section describes some background and statistics of some analyzed Arabic and English texts. In the third section, it is described our developed method, the algorithm TAD_Histo. The fourth section contains results for analyzed texts. In the conclusion, we formulate summary of results and the plan of the following research.

2. Basic Background and Text Statistics

In the text anomaly detections we used English texts from benchmark [1] and Arabic texts from [2], [3].

2.1 English Texts

We illustrate some statistical analysis of 5 English analyzed texts in Table 1. We show the number of words, the number of letters and the number of different words, the number of words by the length 3 and 4 with percentage. More statistics we described in the paper [10].

Table 1: Statistics of 5 English Texts,

The number	E1	E2	E3	E4	E5
of words	176598	132020	125487	106359	93085
of letters	874761	607783	498696	471566	417899
of diff. words	22954	19268	15853	15321	12929
words by the length 3	**38733** 21.93%	**27599** 20.90%	**23780** 18.95%	**26497** 24.91%	**19875** 21.35%
words by the length 4	26321 14.90%	20909 15.83%	18429 14.68%	19224 18.07%	18520 19.89%

In the Fig. 1 we show graphs of frequencies for five Arabic and English texts. In the left panel there are Arabic texts and in the right panel English texts. In the

Arabic texts the highest percentage of occurrences is for words of the length 3 and 4. In the English texts the highest percentage of occurrences is for words of the length 3. In English texts the majority of words has the length 1-15, bigger than Arabic texts where the majority words has the length 1-10.

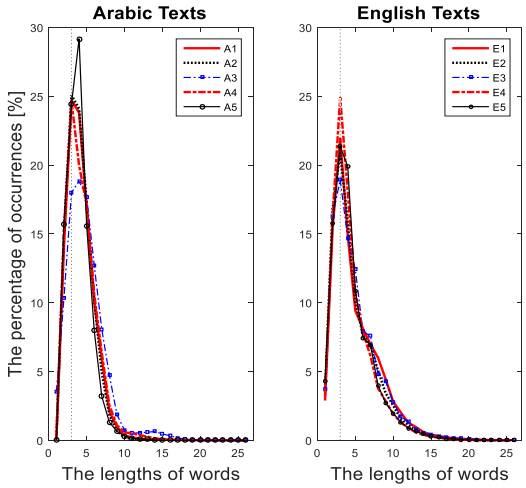


Fig. 1. The percentage of occurrences according to the lengths of words in 5 Arabic and 5 English texts. We prepared the figure in [11].

2.2 Arabic Texts

In the Table 2, we illustrate the number of words, letters and different words, we describe the number of words by the length 3 and 4 with percentage of five Arabic texts. More statistics we described in the paper [10].

Table 2: Statistics of 5 Arabic Texts,

The number	A1	A2	A3	A4	A5
of words	94197	48358	51938	31656	39340
of letters	395065	198019	247448	135573	152905
of diff. words	14110	9061	25755	10098	7036
Words by the length 3	**23287** 24.72%	**12130** 25.08%	9353 18.00%	**7795** 24.62%	9619 24.45%
Words by the length 4	22426 23.80%	11653 24.09%	**9779** 18.82%	6324 19.97%	**11476** 29.17%

2.3 Basic Background

We will use the following symbols:

- Γ - a finite alphabet of letters; $|\Gamma|$ is the number of letters in Γ ; in our texts: Γ_A is Arabic alphabet and Γ_E is English alphabet.
- V - a finite vocabulary of words in the alphabet Γ presented in the alphabetic order; $|V|$ - the numbers of words in the vocabulary V ;
- D - text; a finite sequence of words, D = $\langle w_1,...,w_n\rangle; w_i \in V; N -$ the number of words in the texts.
- $D = \langle d_1 d_2 ... d_{|D|}\rangle; |D|$ - the number of symbols in the text D;

3. Histograms Method

The analysis using n-gram profiles method [11] do not keep the sequence of the words, it follows occurrences of the n-grams (sequences of letters) in texts. If the vocabulary V is alphabetically ordered, then it is possible to do its mapping to integer numbers 1…|V|. The texts should be considered as integer valued time series, the sequences of the numbers of words in the vocabulary V. According to [5] and [12] the sequences of words can be followed in time using weighted bag of words (lowbow). Lowbows can follow a track of changes in histograms connected to words through all text. Histograms will be done in the interval $\langle 0,1\rangle$ and it is necessary to map texts into the interval $\langle 0,1\rangle$. A length normalized text x_D is a function $x_D : \langle 0,1\rangle \times V \to \langle 0,1\rangle$ such that

$$\sum_{j \in V} x_D(t, j) = 1, \forall t \in \langle 0,1\rangle.$$

If f_D^j is the frequency of $j \in V$ in the text D then $x_D(t, j) = f_D^j / N$ of word *j* in the mapping position *t*. The mapping to the interval is important because of the different lengths of texts.

The main idea behind the locally weighted bag of words framework [12] is to use a local smoothing kernel to smooth the original word sequence temporally. The first version of our modified algorithm was published in [5]. We have developed a new modification in the last 3 steps and it can be formulated in the following steps.

The algorithm TAD_Histo:

Step 1:

To map a text *D* to the interval $\langle 0,1\rangle$. Let $t = (t_1, t_2, ..., t_N) = (1/N, 2/N, ...N/N)$ be the vector of values from $\langle 0,1\rangle$, $\sum_{j=1}^{N} t_j = (N+1)/2$. Each t_j should be associated to the word in the position *j* in the text.

Mapping of the text is

$$MD(t) = \langle md_{t1}, md_{t2}, ...md_{tN}\rangle,$$

where $md_{t_i} = f_D^j / N$, *i* is a word in the text *D* and *j* is index of word I in the vocabulary *V* .

Step 2:

Let $K^{s}_{\mu,\sigma}(x):\langle 0,1\rangle \to R$ be some kernel smoothing function with location parameter $\mu, \mu \in \langle 0,1\rangle$ and a scale parameter σ. We take k positions of parameter $\mu, (\mu_1, \mu_2, ...\mu_k)$, such that $\sum_{j=1}^{k} k^{s}_{\mu,\sigma}(t_j) = 1.$

It is possible to use Gaussian Probability Density Function (PDF) (7) restricted to the interval $\langle 0,1\rangle$ and renormalized

$$N(x;\mu;\sigma) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad (1)$$

We will use a modification of the function N, the function (2)

$$K^{s}_{\mu,\sigma}(x) = \begin{cases} \frac{N(x,\mu,\sigma)}{\phi(1,\mu,\sigma)-\phi(0,\mu,\sigma)}, & if\ x \in \langle 0,1\rangle, \\ 0, & otherwise. \end{cases} \quad (2)$$

where $\phi(x)$ is a Cumulative Distribution Function (CDF)

$$\phi(x,\mu,\sigma) = (1 + er\ f(\frac{x-\mu}{\sigma\sqrt{(2)}})), \quad (3)$$

where $er\ f(x) = \frac{1}{\sqrt{(pi)}} \int_{-x}^{x} \exp(-t^2) dt.$

Compute vectors $K^{s}_{\mu,\sigma}(t)$ for each positions $\mu_j, j = 1,...,k$ and chosen σ.

Step 3:

Compute local modified vectors LH^{i}_{D} for each position $\mu_j, j = 1,...,k$ as follows:

$$LH^{j}_{D}(t) = MD(t) \times K^{s}_{\mu_j,\sigma}(t) \quad (4)$$

LH^{i}_{D} present the sequences of vectors for a computation of histograms usable for some possible analysis of the texts.

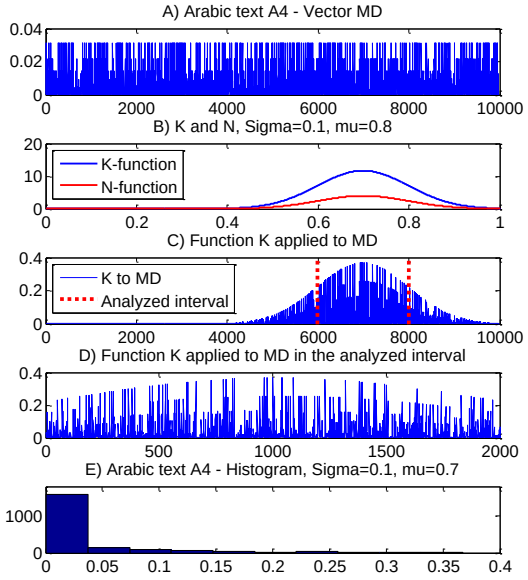
Step 4:

a) Reduce LH^{i}_{D} vectors to analyzed intervals $\langle ibeg, iend\rangle$. In the intervals the values of the function K are higher than some constant C_K. The value of C_K is developed in experiments.
Let the reduced LH^{i}_{D} be $LH^{i}_{D}r$.

b) Compute histograms to $LH^{i}_{D}r$ in *q* equidistant intervals. We will get *k* histogram points in *q*-dimensional space.

Step 5:

Cluster analysis will be applied to cover if all histogram points belong to the same cluster. If all histogram points are in one compact cluster then the analyzed text is probably written by one author if the points belong to more clusters then some anomalies were found in the text and analyzes will continue by splitting the text into two or more parts.
Compute the center C of the cluster and analyze distances histogram points from the center.



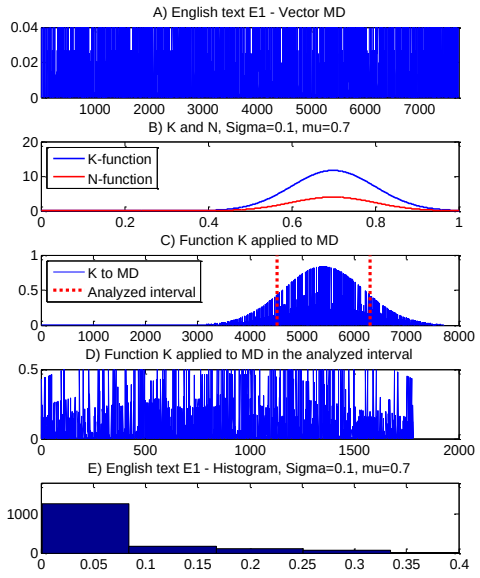


Fig. 2. The illustration of the algorithm on Arabic text A4 and English text E1. In the panels A), the vectors MD are constructed for text words. The prepared smooth functions $K^{s}_{\mu,\sigma}$ and N, where sigma=0.1 and mu=0.7, are shown in the panels B). The panels C) show the application of function K to vectors MD. In the panels D) there are shown values of the analyzed intervals and the down panels E) represent the histograms of the result in the panels C).

The steps 1-3, 4a) of the algorithm are shown in Fig. 2 using Arabic text A4 and English text E1.
The step 4b) is illustrated in Fig. 3 using Arabic text A4. In Fig. 3, there are shown some applications of Kμ function to MD vector for different values μ in the first and the third columns. Histograms are plotted in the second and the fourth columns.

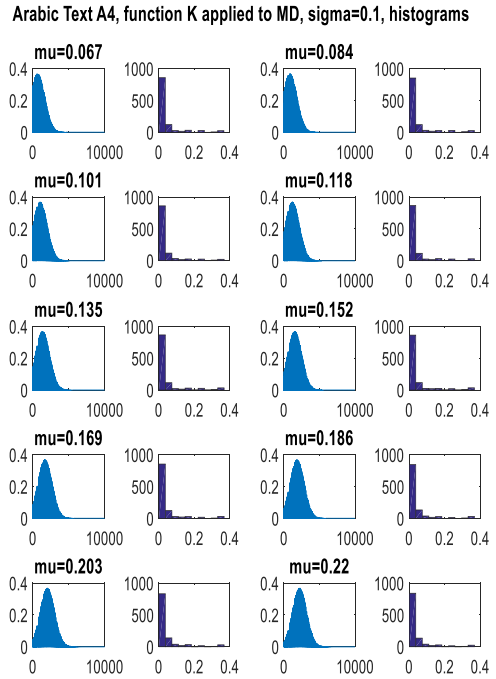


Fig. 3. The panels in the first and third column illustrate application of the function K with different values μ to MD of text A4. The second and fourth column show histograms of the analyzed histograms.

The step 5 of the algorithm is illustrated in Fig. 4 using Arabic text A4 and English text E1. To a visualization of the histogram points, *q* – dimensional space was reduced to 3 dimensional using dimensions 2, 3 and 4. The first interval of $LH_D^i r$ histogram contains all values closed to 0, it means we did not use them in the visualization but in the evaluation complete histogram points were used.

4. Evaluation

The analysis of Arabic text A4 and English text E1 is shown in Figs. 2-4. According to the presented visualization of results it is visible that the points are in one cluster.

In our experiments we used the following values of constants: C_K=10 for Arabic texts and C_K=8, 9 for English texts, q=10, σ = 0.1, μ_1 = 0.05, μ_{i+1} = μ_i+0.17, i=1, 2,… 49. The number of prepared histograms for each text was 50.
The results of 6 Arabic and 6 English texts are shown in the Tables 3 and 4. Five English texts were chosen from [9] and five Arabic texts from [7], [8]. In the Tables, there are coordinates of the centers, the distance and the index of the histogram point with the maximal distance d_{max} from the center and the number of points in the bigger distance than 50% from d_{max}.

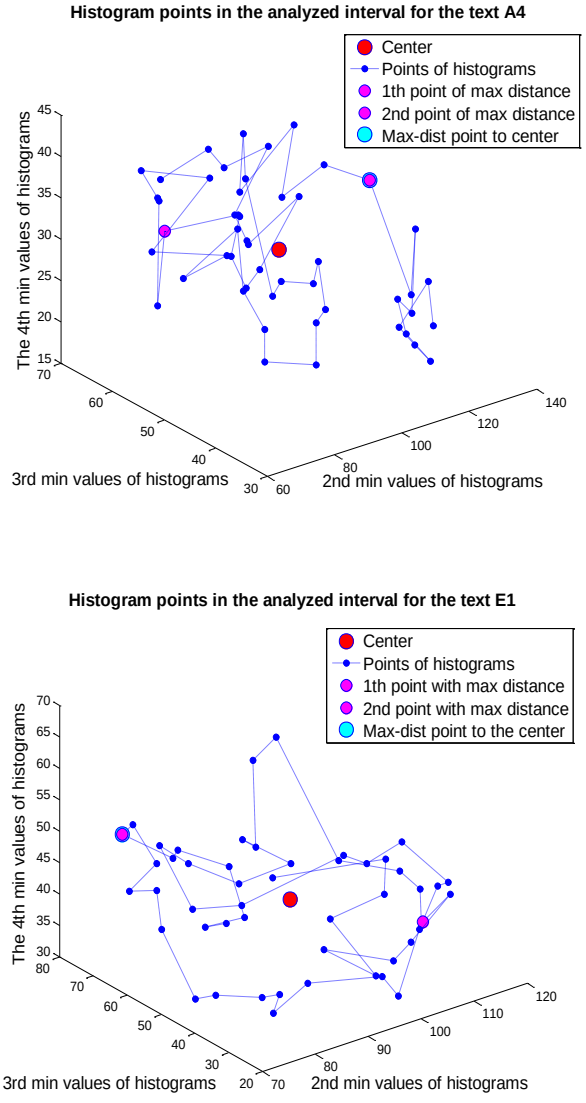


Fig. 4. The first panel illustrates histogram points of Arabic text A4, the second panel shows histogram points of English text E1. Except histogram points, we show the centers of the clusters, the points with the maximal distance from the center and two pints with the maximal distance.

Our first idea is to compare the numbers of histogram point in the distance less than 50% of d_{max} – d50l to the number of histogram points in the bigger distance than d_{max} – d50b.

Let p50 be

$$p50 = \frac{d50l}{d50b} \qquad (5)$$

If p50>1.00 then the cluster is quite compact else it is necessary to do some new analysis. In the text, there are some anomalies.

Table 3: The results of 5 English original and 1 combined texts.

Texts	Coordinates of the histogram center	d_{max} - Maximal distance between the center and histogram point	50% of distance / p50
E1	[536.2200, 70.3400, 34.8600, 33.0800, 17.4800, 30.2800, 16.3800, 30.5600, 0, 0]	94.5163 H-Point: 1	47.2581 36/14 = 2.5714
E2	[9010.2, 1218.6, 972.5, 283.9, 134.0, 173.9, 206.6, 584.1, 601.0, 0]	498.4786 H-Point: 1	249.2393 46/4= 11.5000
E3	[8773.7, 767.7, 980.5, 416.1, 226.6, 110.9, 564.1, 285.9, 84.3, 0]	428.9243 H-Point: 50	214.4621 39/11= 3.5454
E4	[6962.7, 1061.3, 722.7, 819.8, 302.5, 558.6, 181.6, 291.8, 387.9, 0]	383.1302 H-Point: 32	191.5651 32/18= 1.2777
E5	[6434.2, 954.7, 381.2, 716.9, 294.2, 293.9, 264.2, 143.9, 400.9, 0]	214.7325 H-Point: 1	107.3663 35/15= 2.3333
E1-E5	[1188.7, 163.6, 97.3, 93.4, 44.8,1.5, 90.6, 94.1, 0, 0]	113.7178 H-Point: 26	56.8589 24/26= 0.9230

Table 4: The results of 5 Arabic original and 1 combined texts.

Texts	Coordinates of the histogram center	d_{max} - Maximal distance between the center and histogram point	50% of distance / p50
A1	[7674.2, 1111.5, 565.9, 571.3, 164.4, 183.7, 203.8, 0, 0, 0]	508.2270 H-Point: 49	254.1135 30/20= 1.5
A2	[3864.5, 670.4, 291.8, 160.6, 268.6, 0, 49.3, 87.2]	280.4267 H-Point: 1	140.2133 32/18= 1.7777
A3	[5089.7, 286.9, 97.6, 96.7, 187.8, 38.1, 0, 0, 0, 0]	105.4364 H-Point: 29	52.7182 33/17= 1.9411
A4	[862.0, 97.62, 52.22, 29.9, 16.48, 8.06, 14.88, 1.94, 19.86, 0]	48.9918 H-Point: 11	24.4959 27/23= 1.1739
A5	[2574.0, 229.0, 229.5, 114.8, 114.6, 143.2, 170.9, 96.4, 87.7, 0]	201.8314 H-Point: 1	100.9157 39/11= 3.5454
A1-A3	[2394.5, 230.6, 192.2, 69.9, 1.5, 79.9, 42.8, 0, 0, 0]	341.5122 H-Point: 50	170.7561 24/26= 0.9230

Tested texts not written by one author were prepared as a combination of 2 texts. One part was taken from the first text and the second part was taken from the second text. It was clear that for us that the combined text was not written by one author. The Arabic combined text A1-A3 and the English combined text E1-E5 are analyzed in Fig. 4. It is possible to see different structure of vector MD in both parts and using function K to construct histograms does important some parts according to μ.

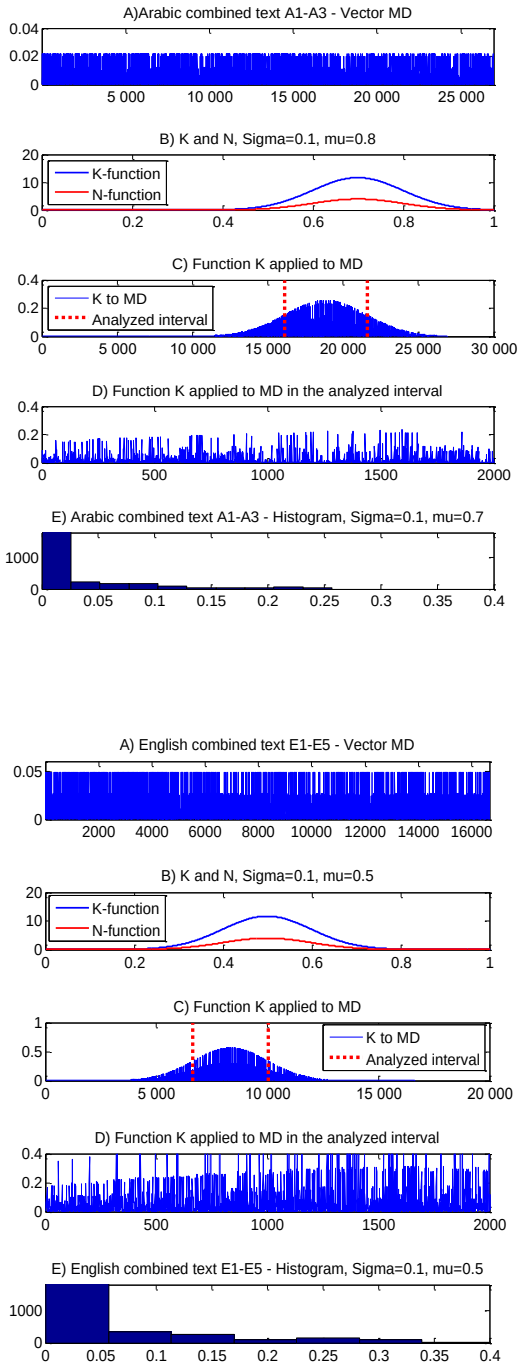


Fig. 5. The illustration of the algorithm on Arabic combined text A1-A3 and English combined text E1-E5. In the panels A), the vectors MD are constructed for text words. The prepared smooth functions $K^{s}_{\mu,\sigma}$ and N, where sigma=0.1 and mu=0.7, are shown in the panels B). The panels C) show the application of function K to vectors MD. In the panels D) there are shown values of the analyzed intervals and the down panels E) represent the histograms of the result in the panels C).

The analysis of histogram points is plotted in Fig. 5. It is visible that the histogram points are arrange in two clusters. The value p50 for Arabic combined text is 0.9230 and the English combined text is 0.9230. It means our method covers in both texts some anomalies.

In the experiments, we found that the constant C_K is different for Arabic and English texts.

5. Conclusion

In the paper, we developed the modified algorithm to cover anomalies in the paper. It is based on histograms computed on mapped texts but it is the order of word in the text is not disturb. We illustrate results for 6 Arabic and 6 English texts. Our method finds anomalies in artificial combined texts. Our next plan is to do some statistics on benchmarked texts.

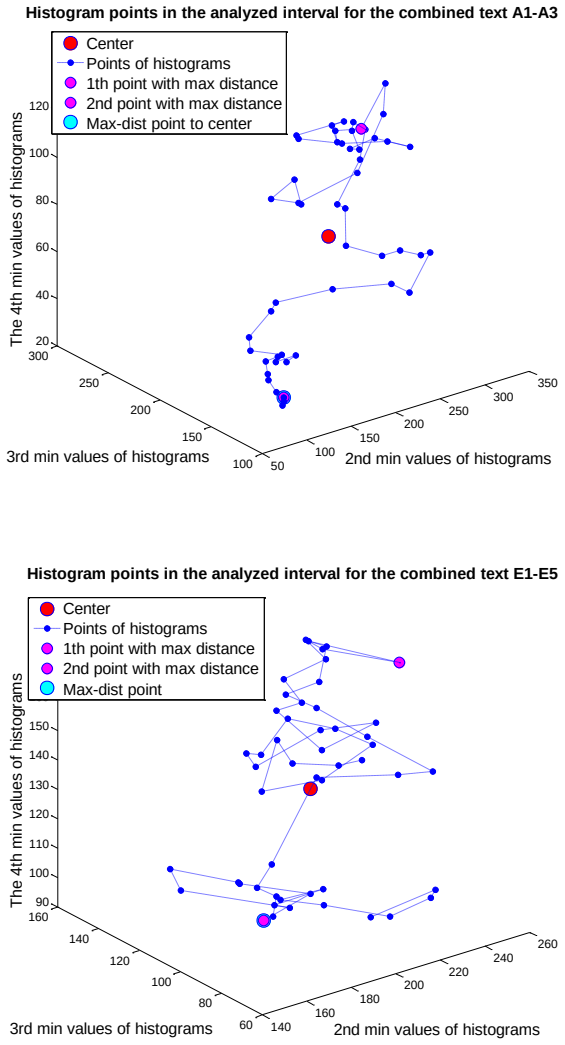


Fig. 6. The top panel illustrates histogram points of Arabic combined text A1-A3, the down panel shows histogram points of English combined text E1-E5. Except histogram points, we show the centers of the clusters and the points with the maximal distance from the center.

Acknowledgements

The research is supported by the Slovak Scientific Grant Agency VEGA, Grant No. 1/0142/15.

References

[1] pan-plagiarism-corpus-2011.part1.rar, ttp://www.uniweimar.de/en/media/chairs/webis/corpora/pan-pc-11/ .

[2] King Saud University Corpus of Classical Arabic. http://ksucorpus.ksu.edu.sa .

[3] I. Bensalem, P. Rosso, S. Chikhi: A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection. CLEF 2013, LNCS 8138, pp. 53-58, 2013.

[4] A. Neme, J.R.G. Pulido, A. Muñoz, S. Hernández, T. Dey: Stylistics analysis and authorship attribution algorithms based on self-organizing maps**,** Neurocomputing, 147 5 January 2015, pp. 147–159.

[5] E. Stamatatos: Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools, 15.5, pp. 823-838.

[6] H. J. Escalante, T. Solorio, M. Montes-y-Gomez: Local Histograms of Character N-grams for Authorship Attribution. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 288-298, Portland, Oregon, June 19-24, 2011. c 2011 Association for Computational Linguistics.

[7] G. Oberreuter, G. LHuillier, S. A. R´ıos, and J. D. Velasquez: Approaches for Intrinsic and External Plagiarism Detection. Notebook for PAN at CLEF 2011 .

[8] F. I. Haj Hassan, M. A. Chaurasia: N-Gram Based Text Author Verification. IPCSI vol. 36 (2012), IACSIT press, Singapore.

[9] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60(3) (2010), pp. 538-556.

[10] A. Almarimi, G. Andrejková: Discrepancies Detection in Arabic and English Documents, ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 5, No.17 , September 2015. ISSN : 2322-5157, pp. 69-75.

[11] A. Almarimi, G. Andrejková: Document Verification using N-grams and Histograms of Words. IEEE 13th International Scientific Conference on Informatics, November 18-20 Poprad Slovakia, pp. 21-26.

[12] G. Lebanon, Y. Mao, J. Dillon: The locally weighted bag of words framework for document representation. Journal of Machine Learning Research 8 (2007), pp. 2405-2441.

First Author: Master of Science from P. J. Šafárik University in Košice, Institute of Computer Science, Faculty of Science, now PhD. Student.

Second Author: Associate Professor, Deputy director for academic affairs in of Computer Science, Faculty of Science, P. J. Šafárik University in Košice. She is interested in artificial neural networks and in stringology.