SPECTRAL ANALYSIS APPLICATIONS TO THE LITERARY TEXTS STUDY

Oleg Granichin, Natalia Kizhaeva
Faculty of Mathematics and Mechanics
Saint Petersburg State University
Saint Petersburg, Russia
o.granichin@spbu.ru, natalia.kizhaeva@gmail.com

Zeev Volkovich
Department of Software Engineering
ORT Braude College of Engineering
Karmiel, Israel
vlvolkov@braude.ac.il

Abstract

This paper presents an approach for dynamic modeling of writing process. A text is divided into sequence of sub-texts that are described through occurrences of character N-grams. The Mean Dependence similarity measures the association between a present sub-text and a number of preceding ones and transforms a text into a time series, which is supposed to be weak stationary if the text is created using single writing style. A periodogram of this signal estimates its Power Spectral Density providing a spectral attribute of the style. Numerical experiments demonstrate high ability of the proposed method in authorship identification and the revealing of writing style evolution.

Key words

Writing style, Authorship Attribution, Spectral Analysis

1 Introduction

The phenomena of Big Data led in turn to the use of the Internet information field in virtually all areas of human activity. The increasing amount of digital text data results in active delelopment of tools for data collection and processing, the infrastucture and services for use of this data in scientific reasearch, politics and economy. One of the most promising areas of modern use of Big Data is the rapid development of technologies for revealing "hidden" information in large amounts of unstructured data that exist in digital or text form. This does not mean that the information is deliberately hidden - just the threshold of "information noise" in such conglomerates of different formats is usually large and it does not allow us to quickly identify the existing semantic links. Cluster analysis of social processes and sociocultural factors of politics in different countries allows evaluation and interpretation of events in the last decade and detection of tension in society that leads to mass riots. And furthermore nowadays the appropriate data analysis of social media and telecommunication can be used to predict terrorist threats [2] At the same time the research of the texts themselves, especially fictional or historical ones, can uncover new facts about the authors, their lives, the time period and the influence of sociocultural events on their work.

Nowadays textual data is the source of Big Data. Powerful Natural Language Processing, Machine Learning and Data Mining techniques allow conducting extensive research in the language field. Such instruments for text analysis are used separately or as a part of larger complex analytical systems. Their application in philology, literary criticism, in the construction of behavioral models often give unexpected results, since they allow us to reveal fairly subtle connections and regularities not only in technical and technological fields, but also in a saturated emotional sphere - and at the same time, on the basis of study of style, the manner of using linguistic means and etc., to see how the world perception of the authors of the text changes, how their visual means develop and the dictionary is enriched on a large time interval. Methods based on frequency lists are known for more than a century and mostly used for the analysis of static structure of a text. Such a structure can provide information for attribution in cases of unknown or doubtful authorship of the text, e.g. apocrypha, historical documents.

Recently, spectral analysis of weakly structured data, particularly literary texts, gained much attention. Methods of spectral estimation find application in various fields of natural sciences and make it possible to recognize patterns that are difficult to find by observation or measurement methods. Social and political events, for example, have analogues in physical world, such as the accumulation of stresses in the earth's crust, the appearance, propagation and growth of seismic waves, resonance phenomena leading crustal faults and volcanic activity.

It is worth noting that unlike dynamic physical process where significant parameters change over time, complete text as a set of elements (characters, digits, punctuation marks) constructing semantic combinations (words, paragraphs, pages) remains unchanged. If we change this set of elements it will result in a completely different text. This makes it impossible to apply tools that are used to analyze physical process directly. The mentioned semantic combinations can construct stable internal sructure of the text that can be identified by frequency methods and characterize the author's writing style. In literature the stylistic features remain fairly stable and are often re-used by the author. This sustainability underlies in the authorship attribution algorithms. Nevertheless methods of time series analysis can be applied to texts. In particular, in the study of texts in which it is possible to find hidden "temporary" (periodic) regularities. Therefore the spectral analysis of time series can reveal hidden information in literary texts and give it a new interpretation.

The following section provides basic theory of spectral analysis. The main algorithm is designed in Section 3. Section 4 describes numerical experiments. We conclude our paper with a discussion in Section 5.

2 Spectral Analysis

In this section we present some basic concepts of spectral analysis (see, for example [11]). The spectral analysis of time series is one of the widespread technique in statistical signal processing. Whenever a signal wavers over a stable state, it recognizes frequency fluctuations that dominate the signal by means of decomposition in terms of a linear combination of the basic trigonometric functions with diverse frequencies and amplitudes. The Power Spectral Density (PSD) expresses how power of such a stationary process is dispersed within these features.

Let us take a continuous weakly stationary process $x(t)$ and consider a Fourier integral

$$\mathbb{G}_T(\omega)=\frac{1}{2T}\int_{-T}^{T} x(t)\exp\left(-i\omega t\right)\, dt.$$

PSD is defined as

$$\mathbb{P}(\omega) = \lim_{T\to\infty} \left|\mathbb{G}_T(\omega)\right|^2 .$$

Averaging over multiple realizations, we get the general description of PSD

$$\mathbb{P}(\omega) = \lim_{T\to\infty} \mathbf{E}\left(\left|\mathbb{G}_T(\omega)\right|^2\right).$$

Recall, that a stochastic process $x(t)$ is weakly stationary if

1. Its mean function $\mathbf{E}(x) = m_x(t) = m_x(t+\tau)$ for all $\tau \in \mathbb{R}$, that is the mean function $m_x(t)$ is constant.
2. The autocovariance function

$$\begin{aligned}\mathbf{E}[(x(t_1)-m_x(t_1))(x(t_2)-m_x(t_2))] &= C_x(t_1,t_2)\\ &= C(t_1-t_2,0)\end{aligned}$$

depends only on difference between t_1 and t_2. This implies that the autocorrelation depends only on $\tau = t_1 - t_2$, i.e. $\boldsymbol{\gamma}(t_1,t_2) = \boldsymbol{\gamma}(\tau)$.

In this case PSD and $\boldsymbol{\gamma}(\tau)$ are a Fourier transform pair:

$$\mathbb{P}(\omega) = \int_{-\infty}^{\infty} \boldsymbol{\gamma}(\tau)\exp\left(-i\omega\tau\right) d\tau.$$

A periodogram provides an estimator for PSD. Assume that a zero mean signal $x(t)$ is sampled at N points to produce values

$$\mathbf{x} = \{x_n,\ n = 0,1,...,N-1\}$$

with the sampling interval Δ. For simplicity, we suppose that N is an even number. Discrete Fourier transform of $\mathbf{x}$ is expressed as

$$X_k(\mathbf{x}) = \sum_{n=0}^{N-1} x_n w_n \exp\left(\frac{2\pi i n k}{N}\right),\ k = 0,...,N-1, \tag{1}$$

where $i = \sqrt{-1}$, and w_n is a window function that changes more gradually from zero to a maximum and then back to zero as n ranges from 0 to $N-1$. The periodogram is defined at $\left(\frac{N}{2}+1\right)$ frequencies

$$f_k = \frac{k}{N\Delta} = 2f_c\frac{k}{N},\ k = 0,...,\frac{N}{2},$$

where f_c is the Nyquist frequency, by the following way:

1. $P_{\mathbf{x}}(0) = P_x(f_0) = \frac{1}{W^2}\left|X_0(\mathbf{x})\right|^2.$
2. $P_{\mathbf{x}}(k) = P_x(f_k) = \frac{1}{W^2}\left(\left|X_k(\mathbf{x})\right|^2 + \left|X_{N-k}(\mathbf{x})\right|^2\right),\quad k = 1,...,\left(\frac{N}{2}-1\right).$
3. $P_{\mathbf{x}}(\frac{N}{2}) = P_{\mathbf{x}}(f_{\frac{N}{2}}) = \frac{1}{W^2}\left|X_{\frac{N}{2}}(\mathbf{x})\right|^2.$

for

$$W = N\sum_{n=0}^{N-1} w_n^2.$$

The standard estimator corresponding to $w_n = 1$, $n = 0, ..., N-1$ is asymptotically unbiased, but the variance does not go to 0 as N grows. To remedy this situation windowing is applied [10]. We use the following version of the Parzen window:

$$w_n = \begin{cases} 1 - 6\left(\frac{|m|}{N/2}\right)^2 + 6\left(\frac{|m|}{N/2}\right)^3, & 0 \le |m| \le \frac{(N-1)}{4} \\ 2\left(1 - \frac{|m|}{N/2}\right)^3, & \text{otherwise} \end{cases},$$

where $m = n - (N-1)/2$ for $n = 0, 1, ..., N-1$.

3 Spectral Attributing

Let us consider a collection $\mathbb{D}$ of documents with a similarity function S defined on $\mathbb{D} \times \mathbb{D}$ as follows:

1. $0 \le S(\mathcal{D}_1, \mathcal{D}_2) \le 1$ for all $\mathcal{D}_1, \mathcal{D}_2 \in \mathbb{D}$.
2. $S(\mathcal{D}, \mathcal{D}) = 1$ for all $\mathcal{D} \in \mathbb{D}$.

We do not suggest that $S(\mathcal{D}_1, \mathcal{D}_2) = 1$ implies equality of $\mathcal{D}_1$ and $\mathcal{D}_2$. In the framework of our model, we consider each document $\mathcal{D} \in \mathbb{D}$ as a series of m sequential documents:

$$\mathcal{D} = \{\widehat{\mathcal{D}}_1, ..., \widehat{\mathcal{D}}_m\}. \tag{2}$$

Thus, $\mathcal{D}$ is the concatenation of $\widehat{\mathcal{D}}_1, ..., \widehat{\mathcal{D}}_m$. The Mean Dependence characterizing the mean relationship between a chunk $\widehat{\mathcal{D}}_i$, $i = T+1, ...m$ and the set of its T "precursors" is defined similarly to [18] as

$$ZV_{T,S}(\widehat{\mathcal{D}}_i) = \frac{1}{T} \sum_{\widehat{\mathcal{D}} \in \Delta_i} S(\widehat{\mathcal{D}}_i, \widehat{\mathcal{D}}), \tag{3}$$

where $\Delta_i = \left\{\widehat{\mathcal{D}}_{i-j},\ j = 1, ..., T\right\}$ the set of its T "precursors" of $\widehat{\mathcal{D}}_i$ An example of $ZV_{T,S}$ is presented in Fig. 1. Here, the values of $ZV_{T,S}$ are computed for the sequential compound of the seven books in the "Harry Potter" series written by J. K. Rowling with the six novels in "Rama" series by Arthur C. Clarke. The biggest pit in the graph corresponds to the border between the collections, and smaller ones designate borders between single books. However, the general behavior of the constructed time series inside the series appears to be quite stationary demonstrating oscillation around a central level.

This confirms our conception that each text $D \in \mathbb{D}$ regarded as an outcome of "a random number generator" reflecting authors personal characteristics exposes the same writing style if and only if the sequence:

$$Y_i(\mathcal{D}) = ZV_{T,S}(\mathcal{D}),\ i = T+1, ...m \tag{4}$$

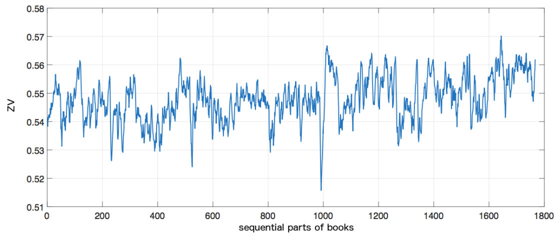

Figure 1. Graph of $ZV - value$.

is a weakly stationary one. Therefore we consider spectral attributes of a writing style as the relevant features of an appropriate periodogram. For example, the following algorithm 1 evaluates the number of the writing styles in a collection $\mathbb{D}$.

Algorithm 1 Main Algorithm

Input:

$\mathbb{D}-$ collection to be considered;

$\mathcal{D} = \{\widehat{\mathcal{D}}_1, ..., \widehat{\mathcal{D}}_m\}-$ representation (2) of each document in the collection;

T – delay parameter in (3);

S – similarity measure in (3);

$k^\star$ – maximum number of styles in the collection;

Cl – clustering algorithm;

CLV – cluster validation index;

Procedure

1: Transform each document $\mathcal{D}_j \in \mathbb{D}$ into a time series $\mathbb{Y}(\mathcal{D}_j) = \{Y_i(\mathcal{D})\}$ according to (4).

2: Calculate a periodogram $P_{\mathbb{Y}(\mathcal{D}_j)}$ of each time series $\mathbb{Y}(\mathcal{D}_j)$.

3: **for** $k = 2$ *to* k^* **do**

4: $\quad c = Cl(\{P_{\mathbb{Y}(\mathcal{D}_j)}\}, k)$;

5: $\quad ind_k = CLV(c)$;

6: **end for**

7: The number of styles is chosen as the optimal number of clusters according to the index value ind_k $\{k = 2, .., k^*\}$.

Application of a partitioning algorithm, like the most representative K-means [8] and Partition Around Medoids methods [7], in the proposed procedure together with a cluster validation approach makes sense if the number of documents in $\mathbb{D}$ is sufficiently large. Otherwise a hierarchical algorithm can be applied in order to recognize the clusters.

4 Numerical experiments

A similarity function is an important part of the proposed approach. In the current paper we form it based on the N-grams model. N-grams methods are widely used in the text mining (see, for example, [14]). We create an 3-gram vocabulary containing all 3-grams presenting in $\mathbb{D}$ after omitting the stop words - a set of very common, high frequent words. Subsequently, each chunk $\widehat{\mathcal{D}}$ is represented as a histogram $h\left(\widehat{\mathcal{D}}\right)$ of these vocabulary, and the following similarity is calculated

$$S(\widehat{\mathcal{D}_1}, \widehat{\mathcal{D}_2}) = \rho(h\left(\widehat{\mathcal{D}_1}\right), h\left(\widehat{\mathcal{D}_2}\right)),$$

where ρ is the Spearman's ρ (see e.g., [3]), which treats the distributions $h\left(\widehat{\mathcal{D}_1}\right)$ and $h\left(\widehat{\mathcal{D}_2}\right)$ as a kind of ordinal data where frequencies are interpreted as the rank positions. This method has been successfully applied to visual word histogram relationship evolution (see, for example [5]), and for clustering genomes within the compositional spectra approach [1].

The standard Euclidean distance as DM and the delay parameter T equals to 10, and $m = 256$ are used.

4.1 Comparison of book collections

First, we consider three known book collections in order to study the evolution of their inner style and we use the single-linkage hierarchial clustering [4] in all experiments.

4.1.1 The Foundation Series The series was written in approximately thirty years period by American author Isaac Asimov and includes the books: "Prelude to Foundation" (1988), "Forward the Foundation" (1993), "Foundation" (1951), "Foundation and Empire" (1952), "Second Foundation"(1953), "Foundation's Edge" (1982) and "Foundation and Earth" (1986) denoted as $F1, \ldots, F7$ according to the chronological order. It is natural to suggest that the underlying writing style of the cycle has being changed within the time-consuming period of its creation.

Let us consider two periodograms in Fig. 2 constructed for the first published book "Foundation" (red) and the last published book "Forward the Foundation" (blue).

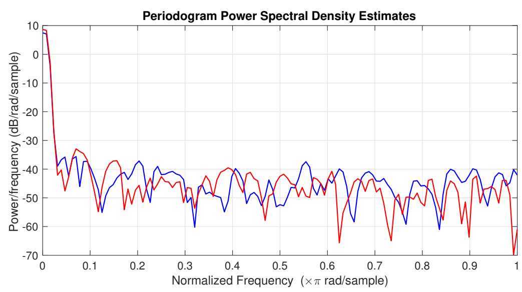

Figure 2. Periodograms constructed for the first (red) and the last (blue) published books.

Generally speaking, the curves are similar, but the power of the red one appears to be more concentrated in the low rates area. It might be assumed that the style became "wrigglier" over time. Alteration of the style is also detected in the results of the hierarchical clustering (see, dendrogram in Fig. 3).

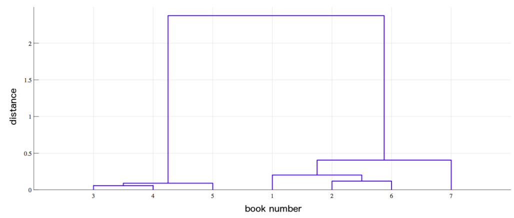

Figure 3. Hierarchal clustering of the Foundation series.

Two clusters distinguishable by own inner style are clearly identified here. The first one contains three novels, "Foundation", "Foundation and Empire", "Second Foundation" are published in 1951, 1952 and 1953 correspondingly. Other four books published in 1988, 1993, 1982, and 1986 are located in the second group. Apparently, a considerable gap between the writing periods (about 30 years) causes a significant change in the series style.

4.1.2 The Rama series The "Rama" Series is a set of six science fiction novels (in this paper denoted as $(R1 - R6)$) initiated in 1973 by the acclaimed novel "Rendezvous with Rama" by Arthur Clarke. The following table presents the pairwise distances between periodograms in the "Rama" series.

Table 1. Distances between periodograms in the Rama series.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$
$R1$	0.00	2.28	2.11	2.65	1.92	1.72
$R2$	2.28	0.00	0.18	0.37	0.35	0.56
$R3$	2.11	0.18	0.00	0.54	0.19	0.39
$R4$	2.65	0.37	0.54	0.00	0.73	0.93
$R5$	1.92	0.35	0.19	0.73	0.00	0.20
$R6$	1.72	0.56	0.39	0.93	0.20	0.00

The maximum value (2.65) is the distance between $R1$ and $R4$. The corresponding periodograms are exhibited in Fig. 4 and are marked in red and blue respectively.

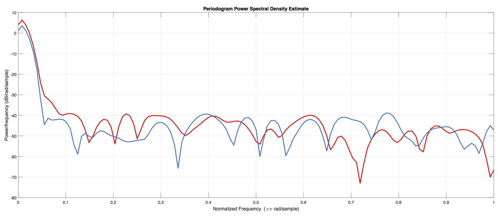

Figure 4. Periodograms constructed for $R1$ and $R4$.

By analogy with the previously considered "Foundation" series, a formerly published book demonstrates smoother style.

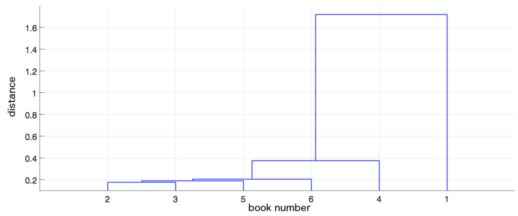

Figure 5. Hierarchal clustering of the Rama series.

A hierarchical clustering of the cycle given in Fig. 5 clearly reveals the collection structure. At the outset, only the first book,"Rendezvous with Rama" was written solely by Arthur C. Clarke. He paired up with Gentry Lee for the following three novels "Rama II", "The Garden of Rama", and "Rama Revealed". There is a common point of view that Gentry Lee did the real composing, while Arthur C. Clarke mainly edited, and the focus and style of the mutual novels ($R2 - R4$) are very dissimilar from those of the first one.

4.1.3 The Epistles Collection In this section, we evaluate seven written in the common Greek language texts from the New Testament (collection of Epistles) considering the figure and teachings of Christ: "Corinthians 1", "Corinthians 2", "Romans", "Philippians", "Thessalonians 1", "Galatians", and "Hebrew" (denoted as $B1 - B7$). Let us consider the distance matrix between the periodograms in the collection.
A graphical representation of this matrix is given in Fig. 6.

So, we can anticipate a two-cluster structure, which is confirmed by the following dendrogram.

Table 2. Distances between periodograms in collection of Epistles.

	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$	$B7$
$B1$	0.00	0.14	0.29	3.48	2.78	2.23	0.31
$B2$	0.14	0.00	0.42	3.34	2.64	2.09	0.44
$B3$	0.29	0.42	0.00	3.76	3.07	2.52	0.06
$B4$	3.48	3.34	3.76	0.00	0.71	1.25	3.78
$B5$	2.78	2.64	3.07	0.71	0.00	0.55	3.08
$B7$	2.23	2.09	2.52	1.25	0.55	0.00	2.53
$B7$	0.31	0.44	0.06	3.78	3.08	2.53	0.00

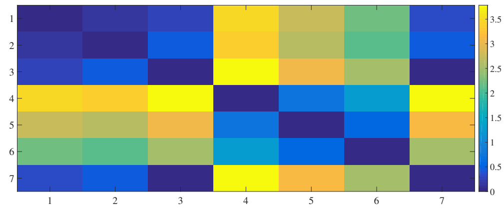

Figure 6. Graphical representation of the distance matrix.

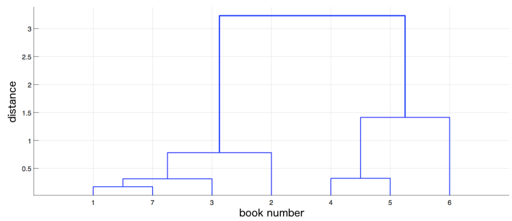

Figure 7. Hierarchal clustering of collection of Epistles.

Two clusters are attained here:

1. "Corinthians 1", "Corinthians 2", "Romans", "Hebrew"
2. "Philippians", "Thessalonians 1", "Galatians"

The Undisputed Epistles: "Romans", "Corinthians 1", "Corinthians 2", "Thessalonians 1", "Galatians", "Philippians", and "Philemon" are respected to be a genuine work of Paul by most scholars [9]. Our process splits the group into two. Note, that the same result has been obtained in [15] using a completely different method. However, the last part ("Hebrew") was isolated in this paper into a separate cluster.

4.2 Authorship determination of single books

In this section several experiments are provided to evaluate the method's ability to discover authorship of single books. We take several books written by Arthur C. Clarke and Isaac Asimov and compare them with the "Foundation" and "Rama" series using Algorithm

1. The clustering algorithm PAM, offered in [7], is applied here with the cluster validation Silhouette method [13]. The silhouette value measures how points in a cluster are closer to each other in contrast to points falled into other clusters. Points with large positive silhouette values around $+1$ are well clustered; those with negative silhouette values are located in a wrong cluster. The average of the silhouette values computed for all points characterizes the partition quality. A partition with the highest silhouette value is desired since such a partition means it provides the best compact clusters separated as well as possible. The reference collection is the concatenation of the studied earlier $F1, .., F7$ and $R1, .., R6$ books with the maximum tested number of clusters $k^* = 5$.

4.2.1 Books by Arthur C. Clarke

At the first step three following books by Arthur C. Clarke are considered:

1. "Odyssey two"
2. "A Space Odyssey"
3. "A Fall of Moondust"

In all experiments the number of clusters is chosen equal to two, and the following clusters are composed:

1. $G_1 = \{AC, F3, F4, F5, R1, R2, R3, R4, R5, R6\}$.
2. $G_2 = \{F1, F2, F6, F7\}$.

AC is one of the tested books. It is possible to see that our method assigns all tested texts to the correct style, and the partition provided for the foundation series coincidences with one obtained in Section 4.1.1.

4.2.2 Books by Isaac Asimov

In these experiments the following books by Isaac Asimov are compared with reference collection

1. "Nemesis"
2. "The End of Eternity"
3. "The Gods Themselves"

For the first book the optimal number of clusters is recognized as three and following clusters are produced:

$$G_1 = \{F3, F4, F5, R1, R2, R3, R4, R5, R6\}.$$
$$G_2 = \{F1, F2, F6, F7\}.$$
$$G_3 = \{Nemesis\}.$$

The silhouette values are presented in Fig. 8.

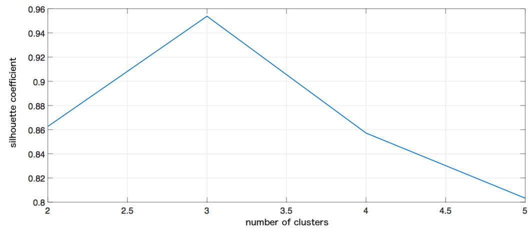

Figure 8. Graph of the silhouette measure values calculated for testing of the Nemesis authorship.

It should be mentioned that a three cluster structure is essential here. However, within clustering into two clusters, G_2 and G_3 merge into one group. Similar partitions arise in two last cases, where the optimal numbers of clusters are evaluated to be two. Thus, all three books are recognized as ones written by Isaac Asimov, but the style of "Nemesis" is sufficiently different from the style of the "Foundation" series.

5 Conclusion

In this paper a new method describing spectral characteristic of writing style is proposed. A text is converted into a time series, which is assumed to be weakly stationary if the text is composed using the same style. The Power Spectral Density of this process estimated by means of a periodogram expresses a spectral pattern of the style. The provided numerical experiments demonstrate a high ability of the proposed method to trace evolution of the style trough time and to resolve author verification tasks.

6 Acknowledgement

This work was supported by RFBR (projects 16-07-00890 and 17-51-53053).

References

Bolshoy A., Volkovich Z., Kirzhner V., Barzily Z. (2010) *Genome Clustering: From Linguistic Models to Classification of Genetic Texts*. Springer Science & Business Media.

Ferrara, E., et al. (2016) Predicting online extremism, content adopters, and interaction reciprocity. *International Conference on Social Informatics*. Springer International Publishing.

Gibbons, J.D., Kendall, M.(1990) Rank correlation methods. In *Edward Arnold*

Hastie, T., Tibshirani, R. and J. Friedman, J. (2009) Hierarchical clustering. In *The elements of statistical learning*. 2.

Ionescu, R. T., Popescu, M. (2015), PQ kernel: A rank correlation kernel for visual word histograms, In *Pattern Recognition Letters*, vol. 55, pp. 51–57.

Janet, E. (1968). *The Composing Processes of Twelfth Graders*. Urbana: NCTE.

Kaufman, L., Rousseeuw, P. (2009) *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

MacQueen, J. (1967) Some Methods for classification and Analysis of Multivariate Observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*. pp. 281–297.

Perkins, P. (1988) *Reading the new testament: An introduction*, Paulist Press.

Prabhu, K.M.M. (2013) *Window functions and their applications in signal processing*. CRC Press.

Priestley, M.B. (1981) Spectral analysis and time series". Academic press.

Reid, J. (1993). *Teaching ESL Writing*. New Jersey: Prentice Hall Regents

Rousseeuw, P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In *Journal of computational and applied mathematics*. 20. pp. 53–65.

Suen, Ch. (1979) N-gram statistics for natural language understanding and text processing. In *IEEE transactions on pattern analysis and machine intelligence*. 2. pp. 164–172.

Shalymov, D., Granichin O., Klebanov L. and Volkovich Z. (2016) Literary writing style recognition via a minimal spanning tree-based approach. In *Expert Systems with Applications* 61. pp. 145–153.

Volkovich, Z. (2016) A Time Series Model of the Writing Process, In *Proceedings of Machine Learning and Data Mining in Pattern Recognition*. New York, NY, USA. July 16-21.

Volkovich, Z., Avros, R. (2016) Text Classification Using a Novel Time Series Based Methodology. In *20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES 2016*. York, United Kingdom, Sept. 5-7.

Volkovich, Z., Granichin O., Redkin O. and Bernikova, O. (2016) Modeling and visualization of media in Arabic. In *Journal of Informetrics*. 10 (2). pp. 439–453.