Pattern Recognition 77 (2018) 45–64

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Patterning of writing style evolution by means of dynamic similarity

Konstantin Amelin[a], Oleg Granichin[a,b], Natalia Kizhaeva[a], Zeev Volkovich[c,*]

[a] *Faculty of Mathematics and Mechanics and Research Laboratory for Analysis and Modeling of Social Processes, Saint Petersburg State University, Saint Petersburg, Russia*
[b] *Institute of Problems in Mechanical Engineering RAS, Saint Petersburg, Russia*
[c] *Department of Software Engineering, ORT Braude College of Engineering, Karmiel, Israel*

ARTICLE INFO

Article history:
Received 7 June 2017
Revised 18 October 2017
Accepted 9 December 2017
Available online 11 December 2017

Keywords:
Patterning
Writing style
Text mining
Dynamics

ABSTRACT

This paper suggests a new methodology for patterning writing style evolution using dynamic similarity. We divide a text into sequential, disjoint portions (chunks) of the same size and exploit the Mean Dependence measure, aspiring to model the writing process via association between the current text chunk and its predecessors. To expose the evolution of a style, a new two-step clustering procedure is applied. In the first phase, a distance based on the Mean Dependence between each pair of chunks is evaluated. All document chunks in a pair are embedded in a high dimensional space using a Kuratowski-type embedding procedure and clustered by means of the introduced distance. In the next phase, the rows of the binary cluster classification documents matrix are clustered via the hierarchical single linkage clustering algorithm. By this way, a visualization of the inner stylistic structure of a texts' collection, the resulting classification tree, is provided by the appropriate dendrogram. The approach applied to studying writing style evolution in the "Foundation Universe" by Isaac Asimov, the "Rama" series by Arthur C. Clarke, the "Forsyte Saga" of John Galsworthy, "The Lord of the Rings" by John Ronald Reuel Tolkien and a collection of books prescribed to Romain Gary demonstrates that the suggested methodology is capable of identifying style development over time. Additional numerical experiments with author determination and author verification tasks exhibit the high ability of the method to provide accurate solutions.

1. Introduction

The rapidly growing number of digital sources in the virtual space prompts the development of intelligent systems for handling of these data. Vast practical problems arise in such areas as plagiarism detection, identification of threat authorship, and computer forensics. The analysis of authorship and writing style transformations is one of the emerging tools suitable for numerous applications in these fields.

Writing style conveys a writer's outline of attendance and represents an individual embodiment of the general writing process composed from many uncertain and attaching phases, which are commonly recognized as Pre-writing, Drafting and Writing, Sharing and Responding, Revising and Editing, and Publishing (see, for example, [1]). The writing style may vary over time even among the documents created by the same author, and these changes can be caused by modifications in the creative intention, influences of colleagues, changes in the social state, and so on. This would naturally lead to a dynamic patterning of the writing style and its inherent evolution. However, most of the existing methods (see a partial review in Section 2) do not take this fact into account and only study the results of the writing process depicted by the considered texts.

A characteristic of the writing process dynamics has been introduced [2] as a part of the modeling and visualization problem for media in Arabic. This method has adequately pointed out the changes in social state, which were reflected in variations of the newspaper style. Modifications of the mentioned approach were proposed in [3–5].

Using this methodology each document is divided into sequential, disjoint portions (chunks) of the same size, and whole document or its chunk is represented as a distribution of suitably chosen N-grams (usually, 3-grams). The association of the current text with its several predecessors is evaluated by employing the Mean Dependence technique presented here, which averages text similarity or dissimilarity with a precursor's set. This overall approach provides a time series representation of a consecutive document collection, and conclusions concerning the style behavior are made with respect to the corresponding characteristics of the consequent time series. From the model standpoint these features actually appear to be the attributes of the writing style. For example, the oscillation of this measure around a certain level indicates the style

[*] Corresponding author.
E-mail addresses: konstantinamelin@gmail.com (K. Amelin), o.granichin@spbu.ru (O. Granichin), natalia.kizhaeva@gmail.com (N. Kizhaeva), vlvolkov@braude.ac.il (Z. Volkovich).

https://doi.org/10.1016/j.patcog.2017.12.011
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

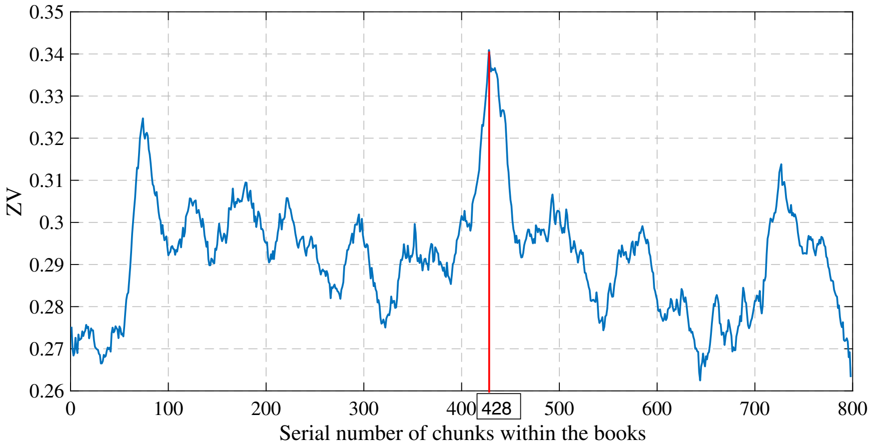

Fig. 1. Example of $ZV_{T,\mathrm{Dis},L}$ graph.

consistency, and its significant deviation points identify alterations in the style. However, the styles of non-adjacent segments may coincide, so an additional pairwise comparison procedure has to be used in order to distinguish the styles. We follow this generic outline in current research.

This paper is devoted to the task of pattern recognition, namely to the application of described general methodology to dynamic patterning of the writing style evolution. Note that this problem is different from the known author verification problem where a set of documents created by a single author is provided, and the purpose is to check if examined text was composed by same author. This task is usually resolved by construction of the author profile and comparing the examined documents to this reference standard. In our case the situation is different. As noted, the writing style of the same author can evolve over time. Hence, the desired decision tool has to be sufficiently specific to recognize changes in style affected by its own evolution, while remaining adequately general, like the mentioned profile, in order to disregard variations associated only with changes in the genre, topic, etc.

We treat this problem in the following way. As was mentioned above, a text (document) or a collection of documents under consideration is divided into a series of sequential sub-texts (chunks), and vector representation of the document chunks is built upon the content-free words that commonly “glue” together the terms in the text body. Joint occurrences of the content-free word can provide valuable stylistic evidence of authorship [6,7], and quantify the influence degree of different historical periods for a given author [8].

Further, in order to expose the style evolution, a two-step clustering procedure is applied. In the first phase, by using the Mean Dependence technique the distance between each pair of the chunks is computed, which is calculated for each chunk with respect to its own precursors and to the precursors of another chunk. This measure is fed into a clustering procedure in order to verify whether a pair of documents was written in the same style (style verification). Afterwards, the chunks are embedded into a high dimensional space using a Kuratowski-type embedding procedure, and the result is clustered by means of the introduced distance. The provided embedding allows one to improve the clustering accuracy, similarly to the famous Kernel trick. Finally, a single text is assigned to a cluster that is consistent with the majority voting of its own chunks, and a binary decision (whether style is the same or not) is made.

In the next phase, the rows of obtained binary classification matrix are once again clustered by the hierarchical single linkage clustering algorithm based on the Hamming distance, which in this case coincides with the classical Euclidean distance. The resulting classification tree displayed by the hierarchical single linkage clustering dendrogram presents a visualization of the stylistic structure of a set. The idea behind this operation is to allocate the documents in accordance with their connections to the rest of text collection. We apply the developed method to the analysis and exhibition of writing style evolution in the fiction book series and demonstrate that proposed methodology is trustworthy and capable of properly identifying style changes over time. We also discuss a feasibility of applying the methods connected to the sequential data clustering for our task.

The last group of experiments with the author identification procedure demonstrates the ability of our method to successfully recognize the author, relying on a relatively limited amount of text. In this case, the text fragments are similarly grouped into the number of clusters, that corresponds to the number of discovered styles. It should be noted that insufficiently separated combinations of the source documents may appear. In attempt to exclude such collections from the classification process, we use the adjusted Rand index to estimate the correspondence between splitting of each source document across the obtained partition and the underlying assignment. Only the combinations demonstrating high agreement expressed by a sufficiently large value of the adjusted Rand index are involved in the analysis of the examined text. As a result, short text portions drawn at random from the books written by the authors of the previously considered series, yet not belonging to these series, were assigned to the correct author.

This paper is organized as follows. Section 2 contains the review of related works. Section 3 describes the presented methodology. Section 4 includes the results of numerical experiments. The last section is devoted to the conclusions and discussion of the future research directions.

2. Related works

The field of authorship attribution aims to determine the author of a certain unidentified document in question by analyzing a provided collection of documents created by a number of known candidates. This field was derived from analysis of comprehensive text reading involving documents of anonymous or questionable authorship. There is a long history of research in this area and the most prominent surveys of various methods are given in [6] and [9].

The measure of deviation used for quantitative evaluation of the text dissimilarity proves to be the key part of any quantitative authorship attribution algorithm. Burrows's Delta [10] is one of the

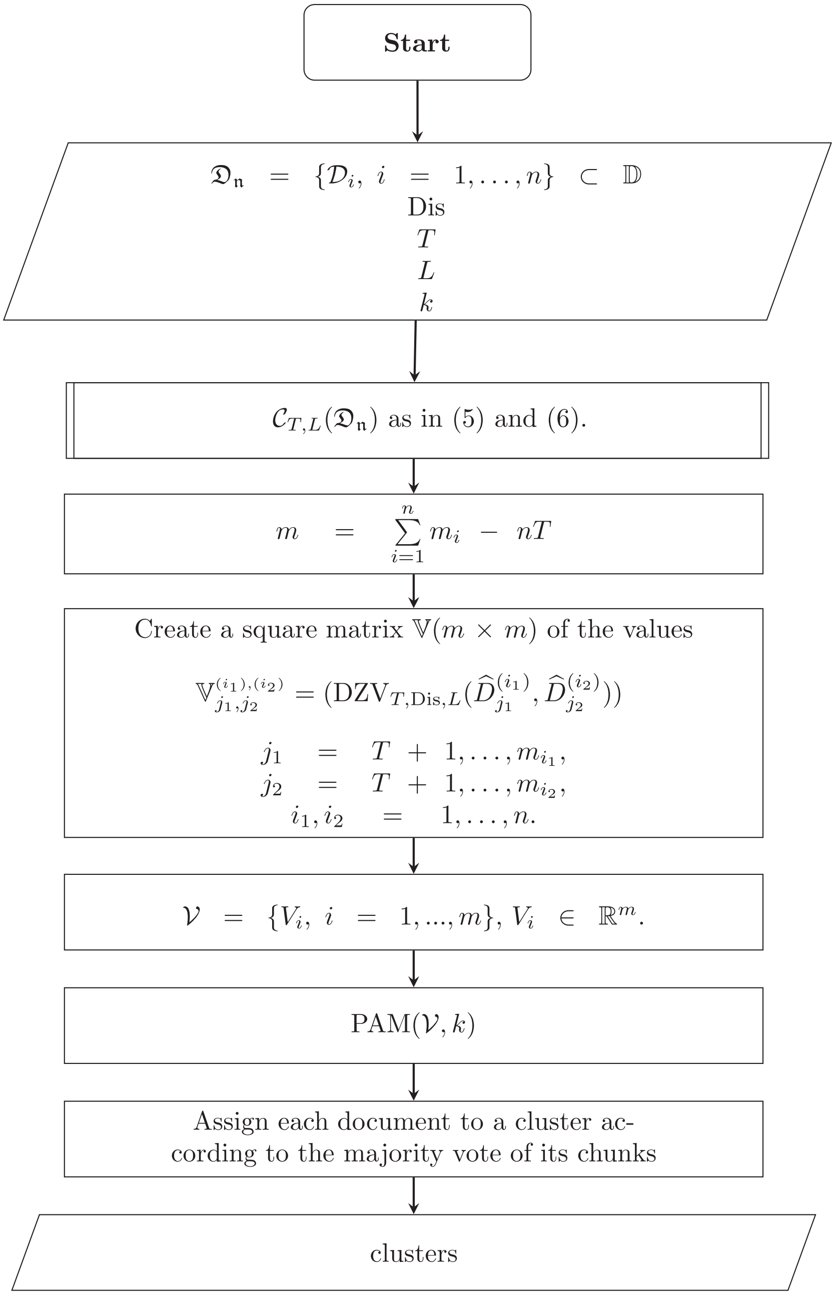
Start
$\mathfrak{D}_\mathfrak{n} = \{\mathcal{D}_i,\ i = 1,\dots,n\} \subset \mathbb{D}$
Dis
T
L
k
$\mathcal{C}_{T,L}(\mathfrak{D}_\mathfrak{n})$ as in (5) and (6).
$m = \sum_{i=1}^{n} m_i - nT$
Create a square matrix $\mathbb{V}(m \times m)$ of the values
$\mathbb{V}^{(i_1),(i_2)}_{j_1,j_2} = (\mathrm{DZV}_{T,\mathrm{Dis},L}(\widehat{D}^{(i_1)}_{j_1}, \widehat{D}^{(i_2)}_{j_2}))$
$j_1 = T + 1,\dots,m_{i_1},$
$j_2 = T + 1,\dots,m_{i_2},$
$i_1, i_2 = 1,\dots,n.$
$\mathcal{V} = \{V_i,\ i = 1,\dots,m\},\ V_i \in \mathbb{R}^m.$
$\mathrm{PAM}(\mathcal{V}, k)$
Assign each document to a cluster according to the majority vote of its chunks
clusters

Fig. 2. Flowchart of Algorithm 1.

most recognized measures of stylistic difference. Since its first appearance in 2002, various modifications have been proposed and tested [11–13]. The Normalized Compression Distance parameter is also successfully applied to the text clustering and is used for the evaluation of computational costs of authorship attribution tasks (see, e.g. [14,15]).

Numerous algorithms utilize the word-based features and they can be classified into three categories. The algorithms of the first type are considering documents as a group of functional words (for example, content-free words), neglecting the content words, since they tend to be strongly associated with the document topics [16]. The second type of methods employs the conventional bag-of-words approach only considering the content words to be the document features [17]. Algorithms of the bag-of-words type are based on the assumption that style mainly depends on the occurrence probability distribution of words, phrases, or any other relevant structures [18]. They are applicable when there is an explicit connection between authors and topics. In this regard one can mention a known method of locally discriminative topic modeling [19].

The last type of methods considers word *N*-gram features presented by sequences containing *N* words or characters [20]. Character *N*-grams located at the character level appear to be a significant feature for stylistic analysis. This representation is tolerant to grammatical errors, has low computational cost and is appropriate for various languages as it allows one to avoid complex preprocessing (e.g. tokenization in case of oriental languages). The proper choice of the *N*-gram length *N* is the key aspect of this approach. A larger *N* values allow one to take the contextual data and subject of the text into account, while also leading to the dimensionality enlargement. A smaller *N* values increase the sensitivity to the sub-word information, but lacks the ability to evaluate a wider

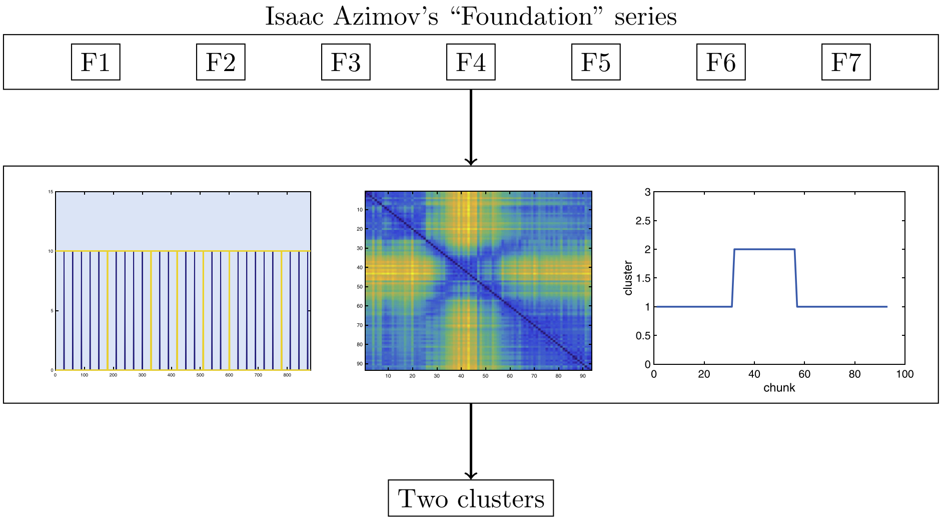

Fig. 3. Illustration of Algorithm 1.

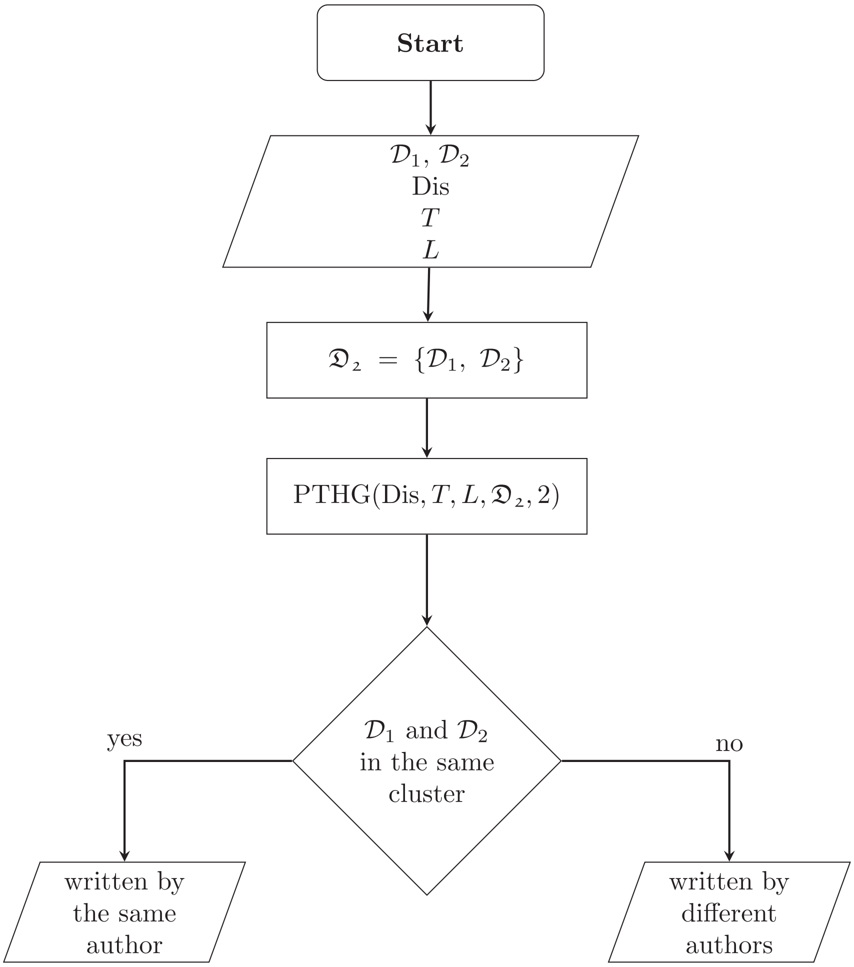
Start
$\mathcal{D}_1$, $\mathcal{D}_2$
Dis
T
L
$\mathfrak{D}_2 = \{\mathcal{D}_1, \mathcal{D}_2\}$
PTHG(Dis, T, L, $\mathfrak{D}_2$, 2)
$\mathcal{D}_1$ and $\mathcal{D}_2$ in the same cluster
yes
no
written by the same author
written by different authors

Fig. 4. Flowchart of Algorithm 2.

context. In order to expose syntactical information, which is inherently suitable for style determination, syntactic *N*-grams have been introduced in [21–23].

Hybrid methods combine several types of features (see, for example [24]), thus exploiting both stylistic and topic features simultaneously. As it was noted in [25], there is no universal feature that is able to clearly separate different author styles. Thus, it is necessary to analyze an incredibly wide set of features involving many approaches [6] in order to get the appropriate result. In the framework of author verification problem, the writing style becomes the most important definitive feature of the considered text [9,26]. The problem of authorship verification only considers a single candidate author [27]. However, since any certain author identification problem can be reduced into a sequence of authorship verification

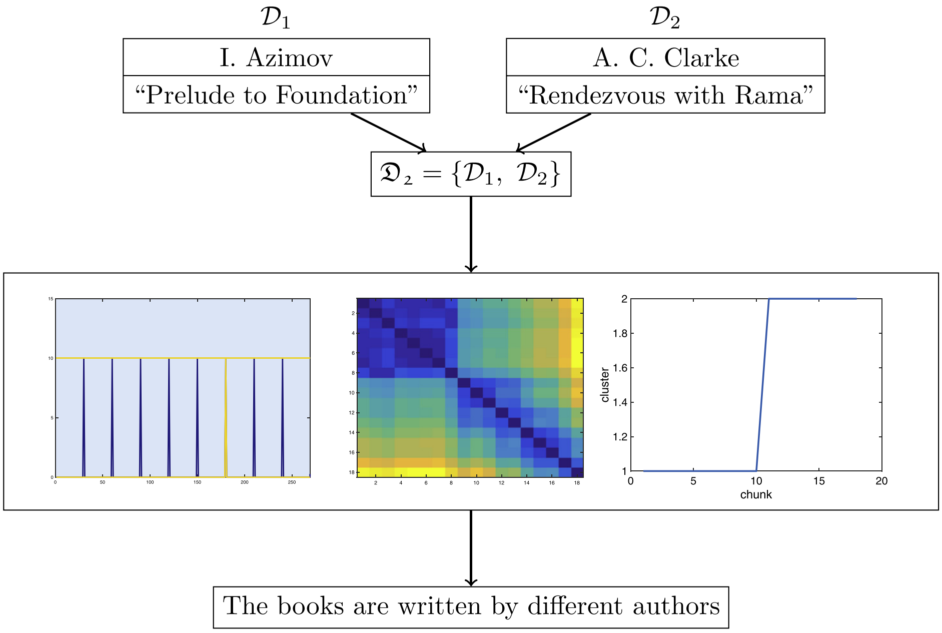

Fig. 5. Illustration of Algorithm 2.

problems, the latter should be regarded as the fundamental one [28,29].

There are two principal methodologies of studying the author verification problem: intrinsic and extrinsic. The intrinsic approach operates only with the provided texts (one of acknowledged authorship and one being examined) and leads to a one-class classification problem [30–33]. Such problems also appear in the plagiarism detection area (see, e.g. [5,34–37]). Extrinsic methods transform the verification task into a binary-classification problem. In this respect, the notable Impostors Method [38] have to be mentioned. The decision in regard to the authorship of the examined document is made by determining whether the document with known authorship is more similar to the one under examination in comparison to the documents from the impostors set. While the method proves to be effective in general, its applicability has several limitations. For instance, it might be problematic to separate the same-author and different-author pairs when the documents under investigation belong to the different genres or topics [28].

As for the occurrence of content-free words in the documents, the large-scale stylometric analysis of literature was proposed in [8]. A new methodology for the literary style recognition was proposed in [39] and [40]. In this approach, text is considered as an output of a random number generator corresponding to the given author. From this standpoint, two styles would be distinguished by means of a multivariate two-sample test procedure applied in appropriate manner.

3. Methodology

3.1. Mean dependency

Let us consider $\mathbb{D}$ as a collection of finite-length texts (documents) on a given alphabet, and take a distance function (semi-metric) Dis: $\mathbb{D} \times \mathbb{D} \longrightarrow [0, +\infty)$ on $\mathbb{D}$ such that for all $\mathcal{D}_1, \mathcal{D}_2 \in \mathbb{D}$ there holds:

- $\mathrm{Dis}(\mathcal{D}_1, \mathcal{D}_2) \geq 0$ (non-negativity).
- $\mathrm{Dis}(\mathcal{D}_1, \mathcal{D}_2) = \mathrm{Dis}(\mathcal{D}_2, \mathcal{D}_1)$ (symmetry).
- $\mathrm{Dis}(\mathcal{D}, \mathcal{D}) = 0$ (reflexivity).

In general case it is not implied that if $\mathrm{Dis}(\mathcal{D}_1, \mathcal{D}_2) = 0$ then $\mathcal{D}_1 = \mathcal{D}_2$.

We divide a document $\mathcal{D} \in \mathbb{D}$ into a series of sequential disjoint parts, named chunks, of the same size L:

$$\mathcal{D} = \{\widehat{D}_1, \ldots, \widehat{D}_m\}. \tag{1}$$

Thus, the document $\mathcal{D}$ is the concatenation of m parts $\widehat{D}_1, \ldots, \widehat{D}_m$. Let us introduce the Mean Dependence, characterizing the mean distance between a chunk $\widehat{D}_i$, $i = T+1, \ldots, m$ and the set $\Delta_{i,T} = \left\{\widehat{D}_{i-j}, j = 1, \ldots, T\right\}$ of its T "precursors":

$$\mathrm{ZV}_{T,\mathrm{Dis},L}(\widehat{D}_i, \Delta_{i,T}) = \frac{1}{T} \sum\nolimits_{\widehat{D} \in \Delta_{i,T}} \mathrm{Dis}(\widehat{D}_i, \widehat{D}). \tag{2}$$

Under the proposed model the text body is considered as an output of a "random number generator", which reflects the author's writing style. Hence, the sequence $\mathrm{ZV}_{T,\mathrm{Dis},L}$ oscillating around a certain level, and the association of a given text chunk with its "precursors" remains on the approximately same level if a document is written in a single writing style. This approach is illustrated in Fig. 1 giving an example of $\mathrm{ZV}_{T,\mathrm{Dis},L}$ graph.

This graph is built for values of $\mathrm{ZV}_{T,\mathrm{Dis},L}$ calculated for the concatenation of three first books of the "Rama" series by Arthur C. Clarke (see, e.g. [41]) and the first three books of the "Foundation Universe" by Isaac Asimov (see, e.g. [42]). The x-axis represents the sequential number of a chunk in the concatenation, and the y-axis represents the values of $\mathrm{ZV}_{T,\mathrm{Dis},L}$. The largest peak corresponds to the border between the book series, and smaller ones actually designate borders between the different books within the series. Large values of $\mathrm{ZV}_{T,\mathrm{Dis},L}$ appear as the result of evaluation of $\mathrm{ZV}_{T,\mathrm{Dis},L}$ for the opening pieces of novels or cycles against a set of precursors, which are partially composed from chunks belonging to the previous book or series.

3.2. Partitioning texts into homogeneous groups

An important component of the proposed approach is a distance measure, designed to perform the separation of text parts into homogeneous groups in accordance with their writing style. The diversity of chunks may be evaluated via the Mean Dependency in the following way:

$$\mathrm{DZV}_{T,\mathrm{Dis},L}(\widehat{D}_i, \widehat{D}_j) = \left|A_{i,i} + A_{j,j} - A_{i,j} - A_{j,i}\right|, i, j > T, \tag{3}$$

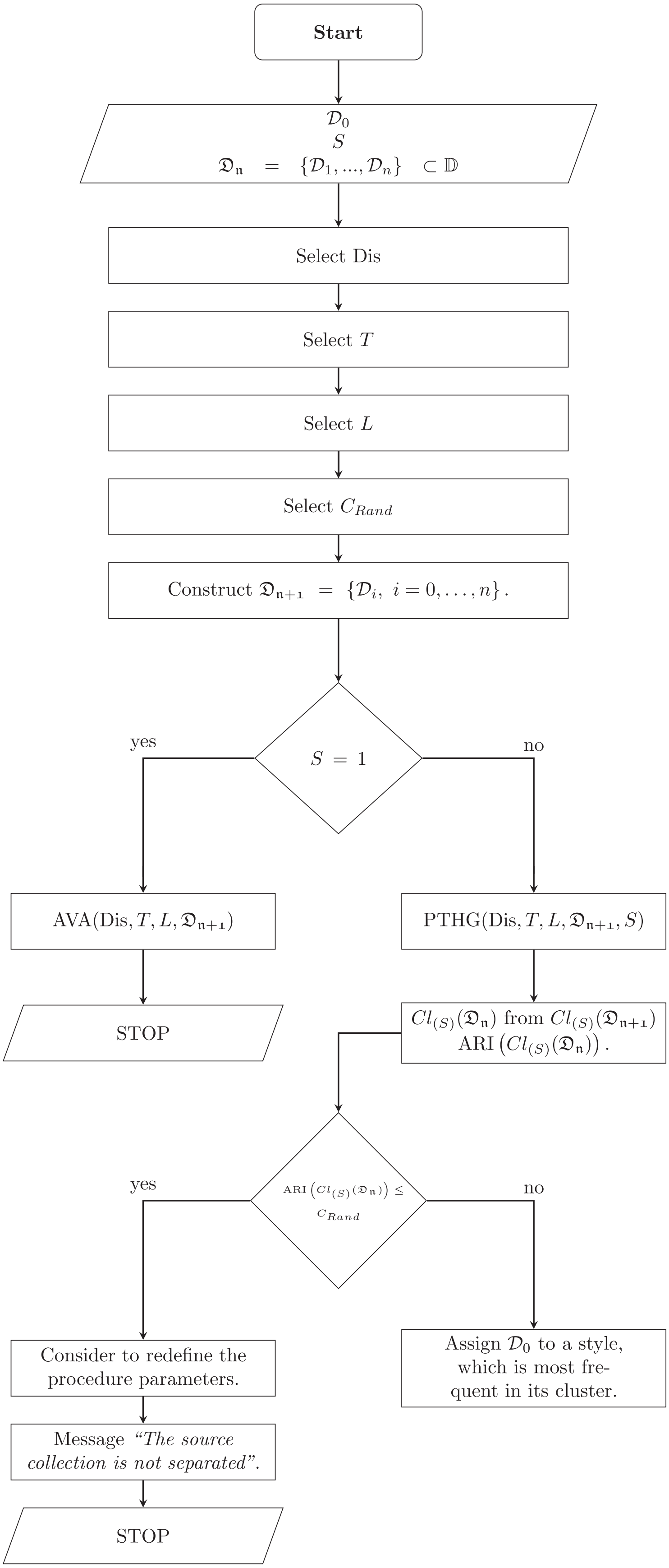
Start
$\mathcal{D}_0$
S
$\mathfrak{D}_{\mathfrak{n}} = \{\mathcal{D}_1, ..., \mathcal{D}_n\} \subset \mathbb{D}$
Select Dis
Select T
Select L
Select C_{Rand}
Construct $\mathfrak{D}_{\mathfrak{n+1}} = \{\mathcal{D}_i,\ i = 0, \ldots, n\}$.
$S = 1$
yes
no
$\mathrm{AVA}(\mathrm{Dis}, T, L, \mathfrak{D}_{\mathfrak{n+1}})$
$\mathrm{PTHG}(\mathrm{Dis}, T, L, \mathfrak{D}_{\mathfrak{n+1}}, S)$
STOP
$Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}})$ from $Cl_{(S)}(\mathfrak{D}_{\mathfrak{n+1}})$
$\mathrm{ARI}\left(Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}})\right)$.
$\mathrm{ARI}\left(Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}})\right) \leq C_{Rand}$
yes
no
Consider to redefine the procedure parameters.
Assign $\mathcal{D}_0$ to a style, which is most frequent in its cluster.
Message *"The source collection is not separated"*.
STOP

Fig. 6. Flowchart of Alg. 3.

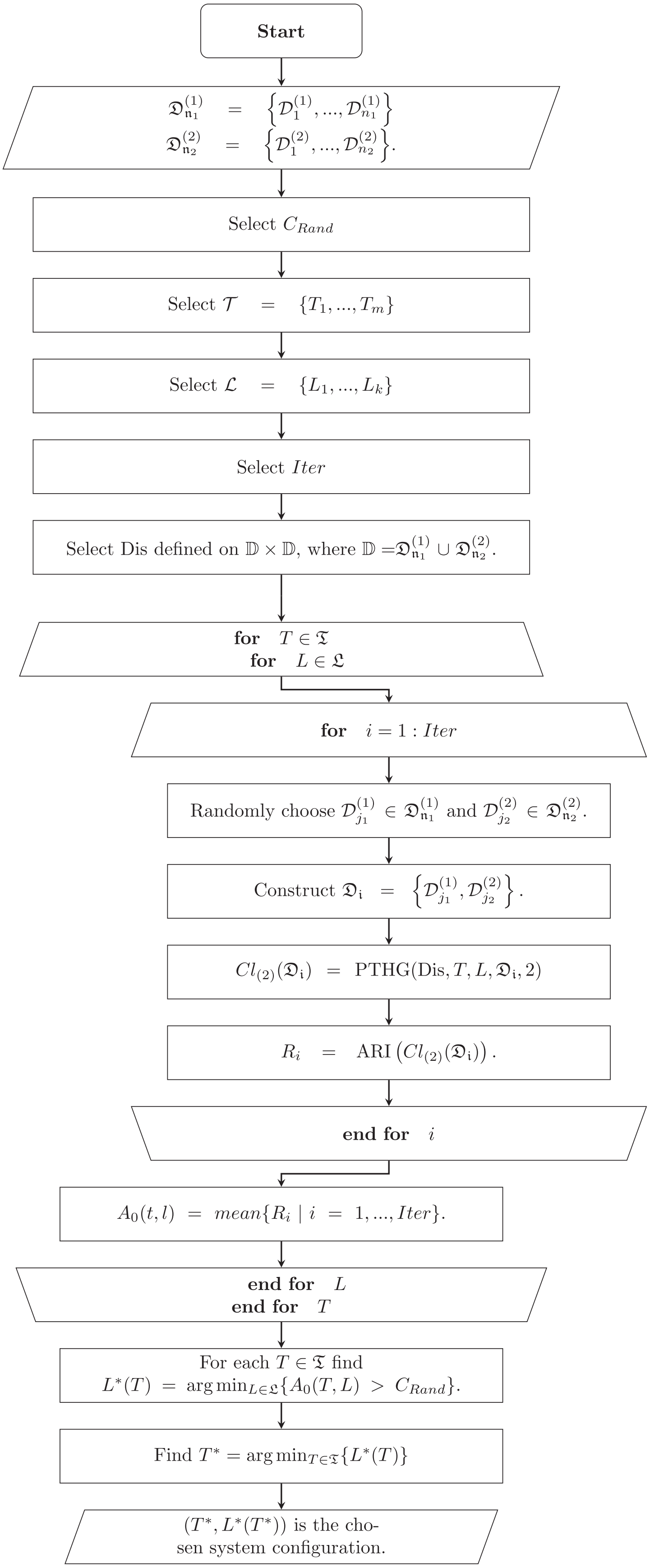
Start
$\mathfrak{D}^{(1)}_{\mathfrak{n}_1} = \left\{\mathcal{D}^{(1)}_1, ..., \mathcal{D}^{(1)}_{n_1}\right\}$
$\mathfrak{D}^{(2)}_{\mathfrak{n}_2} = \left\{\mathcal{D}^{(2)}_1, ..., \mathcal{D}^{(2)}_{n_2}\right\}.$
Select C_{Rand}
Select $\mathcal{T} = \{T_1, ..., T_m\}$
Select $\mathcal{L} = \{L_1, ..., L_k\}$
Select $Iter$
Select Dis defined on $\mathbb{D} \times \mathbb{D}$, where $\mathbb{D} = \mathfrak{D}^{(1)}_{\mathfrak{n}_1} \cup \mathfrak{D}^{(2)}_{\mathfrak{n}_2}$.
for $T \in \mathfrak{T}$
for $L \in \mathfrak{L}$
for $i = 1 : Iter$
Randomly choose $\mathcal{D}^{(1)}_{j_1} \in \mathfrak{D}^{(1)}_{\mathfrak{n}_1}$ and $\mathcal{D}^{(2)}_{j_2} \in \mathfrak{D}^{(2)}_{\mathfrak{n}_2}$.
Construct $\mathfrak{D}_{\mathfrak{i}} = \left\{\mathcal{D}^{(1)}_{j_1}, \mathcal{D}^{(2)}_{j_2}\right\}.$
$Cl_{(2)}(\mathfrak{D}_{\mathfrak{i}}) = \text{PTHG}(\text{Dis}, T, L, \mathfrak{D}_{\mathfrak{i}}, 2)$
$R_i = \text{ARI}\left(Cl_{(2)}(\mathfrak{D}_{\mathfrak{i}})\right).$
end for i
$A_0(t, l) = mean\{R_i \mid i = 1, ..., Iter\}.$
end for L
end for T
For each $T \in \mathfrak{T}$ find
$L^*(T) = \arg\min_{L \in \mathfrak{L}}\{A_0(T, L) > C_{Rand}\}.$
Find $T^* = \arg\min_{T \in \mathfrak{T}}\{L^*(T)\}$
$(T^*, L^*(T^*))$ is the cho-
sen system configuration.

Fig. 7. Flowchart of Alg. 4.

where

$$A_{a,b} = \mathrm{ZV}_{T,\mathrm{Dis},L}(\widehat{D}_a, \Delta_{b,T}),\ a = i, j;\ b = i, j. \tag{4}$$

It is easy to see that it is a semi-metric, which meets the requirements of the three semi-metric axioms stated above (same as Dis). Apparently, we estimate the likeness of two chunks or, more precisely, the likeness of the chunk styles, by comparing each one of them to their corresponding precursors. It is reasonable to assume that if the chunks are similarly associated with sets of precursors, then they should belong to the similar writing style. The grouping of chunks in accordance with their own style naturally assumes applying the clustering procedure. However, the fact that $\mathrm{DZV}_{T,\mathrm{Dis},L}$ is actually not a metric, might lead to the ambiguity of the clustering process. In order to overcome this, we use a Kuratowski-type embedding procedure [43].

Let us consider a set of n documents:

$$\mathfrak{D}_{\mathrm{n}} = \{\mathcal{D}_i, i = 1, \ldots, n\ \} \subset \mathbb{D},$$

which are going to be sorted into k groups. Let us divide the documents into chunks selecting such value for the chunk size L that the number of chunks in each document is larger than T:

$$\mathcal{D}_i = \{\widehat{D}_1^{(i)}, \ldots, \widehat{D}_{m_i}^{(i)}\},\ \ m_i > T, i = 1, \ldots, n, \tag{5}$$

and take the total set of chunks:

$$\mathcal{C}_{T,L}(\mathfrak{D}_{\mathrm{n}}) = \left\{\widehat{D}_j^{(i)}, j = T+1, \ldots, m_i, i = 1, \ldots, n\right\}. \tag{6}$$

The embedding into the Euclidean space $\mathbb{R}^m$ equipped by the standard Euclidean distance $\|\cdot\|_m$ is made as follows:

$$\pi : (\mathcal{C}_{T,L}(\mathfrak{D}_{\mathrm{n}}), \mathrm{DZV}_{T,\mathrm{Dis},L}) \to (\mathbb{R}^m, \|\cdot\|),$$

where for $j = T+1, \ldots, m_i,\ i = 1, \ldots, n$

$$\pi\left(\widehat{\mathcal{D}}_j^{(i)}\right) = (\mathrm{DZV}_{T,\mathrm{Dis},L}(\widehat{D}_j^{(i)}, \widehat{D}_p^{(q)})),\ p = T+1, \ldots, m_q, q = 1, \ldots, n,$$

which induces a new metric on $\mathcal{C}_{T,L}(\mathfrak{D}_{\mathrm{n}})$ via the standard Euclidean distance $\|\cdot\|_m$, $m = m_1 + \ldots + m_n - nT$ in $\mathbb{R}^m$. In this embedding procedure each chunk is represented as a vector with coordinates corresponding to its $\mathrm{DZV}_{T,\mathrm{Dis},L}$ distances to all $\mathcal{C}_{T,L}(\mathfrak{D}_{\mathrm{n}})$ members.

For example, if there are only three chunks in $\mathcal{C}_{T,L}(\mathfrak{D}_{\mathrm{n}})$ such that

- $\mathrm{DZV}_{T,\mathrm{Dis},L}\left(\widehat{D}_1^{(1)}, \widehat{D}_1^{(2)}\right) = 0.5,$
- $\mathrm{DZV}_{T,\mathrm{Dis},L}\left(\widehat{D}_1^{(1)}, \widehat{D}_1^{(3)}\right) = 1,$

then, in this case $\pi(\widehat{D}_1^{(1)}) = (0, 0.5, 1)$. On the one hand, this procedure maps the data into a high-dimensional vector space, but on the other hand, in this space, similarly to the famous Kernel trick (see, for example, [44]), a linear classifier (for instance, the classical K-means algorithm) can be applied.

In our study we employ the Partitioning Around Medoids (PAM) algorithm [45], which is considered to be more robust in comparison to the K-means approach. The proposed procedure is presented as follows (see, Alg. 1).

A flowchart of the algorithm is the following:

Fig. 3 illustrates from left to right the intermediate results of Algorithm 1 obtained as an example for the seven books of the "Foundation" series by Isaac Azimov considered in detail in Section 4.2.2. The left picture exemplifies division of the books, where the blue bars symbolize the borders between sequential chunks in a book, while the yellow ones indicate the borders between consequent books. The next picture represents the matrix $\mathbb{V}$ in the Matlab fashion such that greater distances correspond to brighter colors. The right picture is a graph demonstrating the final cluster assignment of the chunks into two clusters discussed in detail in Section 4.2.2.

Algorithm 1 PTHG (Partitioning Texts into Homogeneous Groups).

```
Input:
    Dis - distance function defined on $\mathbb{D} \times \mathbb{D}$.
    $T$ - delay parameter.
    $L$ - chunk size
    $\mathfrak{D}_{\mathrm{n}} = \{\mathcal{D}_i,\ i = 1, \ldots, n\} \subset \mathbb{D}$ - documents collection.
    $k$ - number of groups.
Procedure:
    Construct $\mathcal{C}_{T,L}(\mathfrak{D}_{\mathrm{n}})$ according to (5) and (6).
    Create a square matrix $\mathbb{V}$ of $m \times m$ size
        $m = \sum_{i=1}^{n} m_i - nT$
    of the values
        $\mathbb{V}_{j_1,j_2}^{(i_1),(i_2)} = (\mathrm{DZV}_{T,\mathrm{Dis},L}(\widehat{D}_{j_1}^{(i_1)}, \widehat{D}_{j_2}^{(i_2)}))$
    for $j_1 = T+1, \ldots, m_{i_1},\ \ j_2 = T+1, \ldots, m_{i_2},\ i_1, i_2 = 1, \ldots, n.$
    Treat the matrix rows as a set of vectors $\mathcal{V} = \{V_i, i = 1, ..., m\}$ in
    the Euclidian space $\mathbb{R}^m$.
    Cluster the set $\mathcal{V}$ into $k$ clusters using the PAM algorithm.
    Assign each document to a cluster according to the majority
    vote ofits chunks.
```

We apply this clustering procedure to the Author Verification task in the following manner. Let us suppose that we have two texts $\mathcal{D}_1$ and $\mathcal{D}_2$, which are going to be tested for being written by the same author. We split the documents and group the chunks in two clusters using Algorithm 1. The conclusion is made according to the cluster assignment.

Fig. 5 illustrates intermediate results of Algorithm 2 obtained as an example for first book of the "Foundation" series by Isaac Azimov and for first book in the "Rama" series by Arthur C. Clarke studied in Section 4.2.3. The meaning of the pictures is the same that in Fig. 3.

Algorithm 2 Author Verification Algorithm (AVA).

```
Input:
    $\mathcal{D}_1,\ \mathcal{D}_2 \in \mathbb{D}$ - two texts to compare.
    Dis - distance function defined on $\mathbb{D} \times \mathbb{D}$.
    $T$ - value of the delay parameter $T$.
    $L$ - chunk size.
Procedure:
    Construct a collection $\mathfrak{D}_2 = \{\mathcal{D}_1,\ \mathcal{D}_2\}$.
    Call PTHG$(\mathrm{Dis}, T, L, \mathfrak{D}_2, 2)$ and assign the documents to two
    clusters.
    If $\mathcal{D}_1$ and $\mathcal{D}_2$ are allocated in the same cluster then they are
    assumed to be written by the same author, otherwise they are
    not.
```

In turn, a general algorithm summarizing the proposed methodology is presented. This clustering algorithm (Algorithm 3) is designed for the author identification procedure in case the document in question $\mathcal{D}_0$ should be verified regarding its alleged authorship in a collection:

$$\mathfrak{D}_{\mathrm{n}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\} \subset \mathbb{D}.$$

The main idea is the same: the document under investigation and the training collection are divided into chunks, and clustered. The examined document is assigned to a cluster (i.e. style) with the highest winning rate of corresponding chunks.

The number of different authors is usually known and assumed to coincide with the number of different styles (S) in the collection. Note that the presence of S different styles in the source collection

Algorithm 3 Author Identification Algorithm (AIA).

```
Input:
    $\mathcal{D}_0$ - the document in question.
    $S$ - number of different styles in the collection.
    $\mathfrak{D}_{\mathfrak{n}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\} \subset \mathbb{D}$ - documents assigned to $S$ author styles.
Procedure:
Select Dis - distance function defined on $\mathbb{D} \times \mathbb{D}$.
Select $T$ - value of the delay parameter $T$.
Select $L$ - chunk size.
Select $C_{Rand}$ - threshold for significance of the Adjusted Rand Index.
Construct a new collection $\mathfrak{D}_{\mathfrak{n}+1} = \{\mathcal{D}_i,\ i = 0, \ldots, n\}$.
if $S = 1$ then
    Call AVA$(\mathrm{Dis}, T, L, \mathfrak{D}_{\mathfrak{n}+1})$ to compare the styles of $\mathcal{D}_0$ and $\mathfrak{D}_{\mathfrak{n}}$.
    STOP
else
    Call PTHG$(\mathrm{Dis}, T, L, \mathfrak{D}_{\mathfrak{n}+1}, S)$ and obtain a partition $Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}+1})$.
    Construct $Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}})$ from $Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}+1})$ and calculate $\mathrm{ARI}\big(Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}})\big)$.
    if $\mathrm{ARI}\big(Cl_{(S)}(\mathfrak{D}_{\mathfrak{n}})\big) \le C_{Rand}$ then
        Consider to redefine the procedure parameters.
        Message "The source collection is not separated".
        STOP
    else
        Assign $\mathcal{D}_0$ to a style, which is the most frequent in its cluster.
    end if
end if
```

has to be assured, since several authors may write documents in collaboration or the chunk size L may be selected in an inappropriate manner such that the styles cannot be distinguished using the chosen configuration of parameters. In order to evaluate the separation of styles for a clustering solution we use the adjusted Rand index [46].

Originally, the Rand index [47] appeared in classification problems where grouping outcomes are compared to a "Ground Truth" categorization. The value range of the Rand index lies between zero and one. Zero value specifies the complete disagreement of two data partitions on any pair of items. If both partitions are exactly the same, then the Rand index is equal to one. The main inconvenience of the Rand index is that its expected value for two random partitions is not constant.

The Adjusted Rand Index (ARI) is based on the generalized hyper-geometric distribution, such that partitions are collected randomly with a fixed number of elements in each cluster. The expected value of this index is zero for independent partitions, and its maximal value is equal to unity for identical ones.

Consider two following distributions of a document collection:

1. Partition constructed according to the predefined document styles

   $$\mathfrak{D}_{\mathfrak{n}} = \bigcup_{m=1}^{S} S_m(\mathfrak{D}_{\mathfrak{n}}),$$

   where $S_m(\mathfrak{D}_{\mathfrak{n}})$ consists of the documents with style S_m, $m = 1, \ldots, S$.
2. The separation $Cl_{(S)}$ of the documents into S clusters obtained by means of Algorithm 1:

   $$\mathfrak{D}_{\mathfrak{n}} = \bigcup_{i=1}^{S} Cl_{(S),i}(\mathfrak{D}_{\mathfrak{n}}).$$

We create a contingency table composed from the all quantities

$$n_{si} = \left| S_m(\mathfrak{D}_{\mathfrak{n}}) \cap Cl_{(S),i}(\mathfrak{D}_{\mathfrak{n}}) \right|,\ s,\ i = 1, \ldots, S,$$

and introduce:

$$n_{s\cdot} = \sum_{i=1}^{S} n_{si},\ \ n_{\cdot i} = \sum_{s=1}^{S} n_{si}.$$

The Adjusted Rand Index (ARI) is

$$\mathrm{ARI}\big(Cl_{(S)}\big) = \frac{\sum_{s,i}^{S}\binom{n_{si}}{2} - \sum_{s=1}^{S}\binom{n_{s\cdot}}{2}\sum_{i=1}^{S}\binom{n_{\cdot i}}{2}/\binom{n}{2}}{\frac{1}{2}\left(\sum_{s=1}^{S}\binom{n_{s\cdot}}{2} + \sum_{i=1}^{S}\binom{n_{\cdot i}}{2}\right) - \sum_{s=1}^{S}\binom{n_{s\cdot}}{2}\sum_{i=1}^{S}\binom{n_{\cdot i}}{2}/\binom{n}{2}}.$$

We consider a clustering $Cl_{(S)}$ to be conventional if $\mathrm{ARI}(Cl_{(S)}) > C_{Rand}$, where C_{Rand} is a given threshold.

Note, that there is a new parameter C_{Rand} involved in the algorithm. Its purpose is to estimate the ability of clustering procedure to separate the training set. If the value of ARI calculated for the source collection does not exceed a given threshold C_{Rand} then we cannot assume the training collection to be reliable for the current configuration of parameters.

3.3. Distance construction

Within the proposed approach the choice of the distance function is essential for suitable distinguishing of the different writing styles. Formally, measures such as the Levenstein distance (or the edit distance) [48] can be used. In the text mining domain it is more suitable to convert texts into the probability distributions and to measure the distance between them, subsequently. In our context, we introduce a transformation F, which maps all the documents belonging to $\mathbb{D}$ into the set $\mathcal{P}_M$ of all probability distributions on $[0, 1, 2, \ldots, M]$:

$$\boldsymbol{P} = \{p_i, i = 0, 1, 2, \ldots, M\}, \quad p_i \ge 0, \quad \sum_{i=0}^{M} p_i = 1,$$

and consider

$$\mathrm{Dis}(\mathcal{D}_1, \mathcal{D}_2) = \mathrm{dis}(\mathrm{F}(\mathcal{D}_1), \mathrm{F}(\mathcal{D}_2)),$$

where M is a natural number, and dis is a distance function (a simple probability metric) defined on $\mathcal{P}_M \times \mathcal{P}_M$.

The theory of probability metrics is presented in [49] and [50]. A comprehensive survey of distance/similarity measures between probability densities can be found in [51]. As usual, a transformation F is constructed by means of the Vector Space Model. Each document is described by the table of term frequencies in contrast to the vocabulary representation, containing all the words (or "terms") in all of the documents contained in the corpus. Thus, the model disregards grammar and particular order of terms but retains the collection of terms. The tables are interpreted as vectors in a linear space with dimensionality equal to the size of vocabulary.

In the Bag of Words model a document is represented as the distribution of words, where stop-words are usually removed in order to reduce the spatial dimensions. The Keywords Model is a derivative from the latter. In this case the bag contains only particular selected words, instead of including every term from the text corpus. As for the N-grams model, the vocabulary includes every N-gram in the corpus, with an N-gram being a connecting sequence of N characters from a text occurring in a slide window of length N. The N-gram based approaches are widely applied in the area of text retrieval tasks.

3.4. Feature selection

Feature selection is a process of picking a subset of distinctive features, appropriate to the particular problem under investigation. In the studied problem, the non-informative terms appear in minor fractions of chunks with relatively low frequencies. Therefore, the separation algorithms are not sensitive to the presence of such terms within a given chunk, since their occurrence rates are low for all chunks involved. Naturally, the number of such terms may increase as the chunk size L becomes smaller. We evaluate the merit of a given term based on their average occurrence in the whole corpus:

$$S(w_i) = \text{average}\{f(w_i, \mathcal{D}),\ \mathcal{D} \in \mathbb{D}\},$$

where $f(w_i, \mathcal{D})$ is the frequency of the term w_i in a document $\mathcal{D} \in \mathbb{D}$. During the next step only terms belonging to the set

$$IW(T) = \{S(w_i) > Tr\}, \tag{7}$$

are involved in construction of the Vector Space Model. Here, Tr is a predefined threshold. Evidently, the most crucial parameters in the proposed methodology are the delay T and the size of chunks L. The problem of appropriate feature combination selection is ill-posed, since various parameter configurations could lead to the identical behavior of the system. It is clear that larger values of T and L should hypothetically lead to more stable results. However, on the other hand, the number of text chunks may decrease to such degree that $ZV_{T,\text{Dis},L}$ will no longer reflect the style dynamics, and the majority vote classifier will become unreliable. In this regard, it is necessary to keep a balance between the parameter values and the number of chunks when choosing the parameter configuration.

In the spirit of [52], we propose to seek the parameter values, which provide an appropriate separation of document sets belonging to inherently different styles. This idea is implemented in the following algorithm (Algorithm 4):

Let us take two collections written in different styles

$$\mathfrak{D}^{(1)}_{\mathfrak{n}_1} = \left\{\mathcal{D}^{(1)}_1, \ldots, \mathcal{D}^{(1)}_{n_1}\right\},\ \mathfrak{D}^{(2)}_{\mathfrak{n}_2} = \left\{\mathcal{D}^{(2)}_1, \ldots, \mathcal{D}^{(2)}_{n_2}\right\}$$

with two sets of possible parameters values

$$\mathcal{T} = \{T_1, \ldots, T_m\},\ \mathcal{L} = \{L_1, \ldots, L_k\}$$

These groups can be preferred resting upon our previously collected knowledge or the general perception. We repeat several times (parameter *Iter* in the algorithm) the same procedure.

For each combination of $T \in \mathfrak{T}$ and $L \in \mathfrak{L}$ two documents $\mathcal{D}^{(1)}_{j_1} \in \mathfrak{D}^{(1)}_{\mathfrak{n}_1}$ and $\mathcal{D}^{(2)}_{j_2} \in \mathfrak{D}^{(2)}_{\mathfrak{n}_2}$ are chosen at random, divided into chunks and clustered using the Algorithm 1, with purpose of determining the ARI between initial and obtained partitions. After the completion of iterations when the average value $A_0(T,\ L)$ of ARI is found, the following value is attained for each $T \in \mathfrak{T}$

$$L^*(T) = \arg\min_{L \in \mathcal{L}} \{A_0(T, L) > C_{Rand}\}$$

and

$$T^* = \arg\min_{T \in \mathcal{T}} \{L^*(T)\}.$$

Here, C_{Rand} is predefined threshold. The pair $(T^*,\ L^*(T^*))$ is the chosen system configuration.

Algorithm 4 Parameters Selection

```
Input:
    Two document collections assigned to two different styles$\mathfrak{D}^{(1)}_{\mathfrak{n}_1} = \left\{\mathcal{D}^{(1)}_1, ..., \mathcal{D}^{(1)}_{n_1}\right\}$ and $\mathfrak{D}^{(2)}_{\mathfrak{n}_2} = \left\{\mathcal{D}^{(2)}_1, ..., \mathcal{D}^{(2)}_{n_2}\right\}$.
Procedure:
    Select $C_{Rand}$ - threshold for significance of the adjusted Rand index.
    Select $\mathcal{T} = \{T_1, ..., T_m\}$ - a set of the tested values of $T$.
    Select $\mathcal{L} = \{L_1, ..., L_k\}$ - a set of the tested values of $L$.
    Select $Iter$ - number of iterations.
    Select Dis - distance function defined on $\mathbb{D} \times \mathbb{D}$, where $\mathbb{D} = \mathfrak{D}^{(1)}_{\mathfrak{n}_1} \cup \mathfrak{D}^{(2)}_{\mathfrak{n}_2}$.
    for $T \in \mathcal{T}$ do
        for $L \in \mathcal{L}$ do
            for $i = 1 : Iter$ do
    Randomly choose $\mathcal{D}^{(1)}_{j_1} \in \mathfrak{D}^{(1)}_{\mathfrak{n}_1}$ and $\mathcal{D}^{(2)}_{j_2} \in \mathfrak{D}^{(2)}_{\mathfrak{n}_2}$.
    Construct $\mathfrak{D}_\mathfrak{i} = \left\{\mathcal{D}^{(1)}_{j_1}, \mathcal{D}^{(2)}_{j_2}\right\}$.
    Call PTHG$(\text{Dis}, T, L, \mathfrak{D}_\mathfrak{i}, 2)$ and obtain a partition $Cl_{(2)}(\mathfrak{D}_\mathfrak{i})$.
    Calculate $R_i = \text{ARI}\big(Cl_{(2)}(\mathfrak{D}_\mathfrak{i})\big)$.
            end for
    Calculate $A_0(t, l) = \text{mean}(R_i | i = 1, ..., Iter)$.
        end for
    end for
    For each $T \in \mathcal{T}$ find
        $L^*(T) = \arg\min_{L \in \mathcal{L}} \{A_0(T, L) > C_{Rand}\}$.
    Find
        $T^* = \arg\min_{T \in \mathcal{T}} \{L^*(T)\}$
    The pair $(T^*, L^*(T^*))$ is the chosen system configuration.
```

4. Numerical experiments

4.1. Experiments setup

4.1.1. Vector space model

All calculations are performed in the Matlab environment. Similarly to [8], in this paper we employ the content-free word approach as the basis for the Vector Space Model. Content-free words can be considered as a kind of stylistic “glue” of the language, because they do not convey semantic meaning on their own, however they establish the link between the terms that do. As it was mentioned earlier, joint occurrences of the content-free words can provide valuable stylistic evidence for authorship verification [6,7]. This approach was successfully used in the analysis of quantitative patterns of stylistic influence [8].

Frequency of the content-free word occurrences calculated for a large number of authors and texts over a long period can reflect temporal trends in styles. Moreover, frequency vectors of the content-free words provide topic-independent literal characteristics of a text and are correspondingly distributed among the authors working in adjacent periods. Though this research was performed resting upon a large number of books from the Project Gutenberg Digital Library corpus, we apply content-free word approach in our study since it appears to be suitable for considering the style evolution. A list of 307 content-free words used in our experiment is presented in this article and contains prepositions, articles, conjunctions, auxiliary verbs, some common nouns and pronouns.

4.1.2. Distance

Resting upon the similarity of their shapes, we would like to characterize similarity between the distributions $\boldsymbol{P} = \{p_i, i = 0, 1, \ldots, M\} \in \mathcal{P}_M$ and $\boldsymbol{Q} = \{q_i, i = 0, 1, \ldots, M\} \in \mathcal{P}_M$ obtained using the Vector Space Model, in assumption that the distribution forms naturally characterize the manner of term incorporation in the document style. In this paper we compare two following distances:

- *The Spearman's Correlation Distance* (see, e.g. [53,54]) is defined as

  $$S(\boldsymbol{P}, \boldsymbol{Q}) = 1 - corr(R(\boldsymbol{P}), R(\boldsymbol{Q})) = 1 - \rho(\boldsymbol{P}, \boldsymbol{Q}),$$

  where ρ is the Spearman's ρ (see, e.g. [55]):

  $$\rho(\boldsymbol{P}, \boldsymbol{Q}) = 1 - \frac{6\sum_{i=0}^{M}(R(p_i) - R(q_i))^2}{(M+1)\left((M+1)^2 - 1\right)}.$$

  If the value of Spearman correlation is high ($\rho \approx 1$ and $S \approx 0$) then the variables demonstrate a comparable ranking, while the low value ($\rho \approx -1$ and $S \approx 2$) means that the ranks are opposed. A function R maps each distribution $\boldsymbol{P} = \{p_i, i = 0, 1, \ldots, M\} \in \mathcal{P}_M$ to $(1, \ldots, M+1)$ such that $R(p_i)$ is the rank (position) of p_i in the ranked array $\boldsymbol{P}$. If several probabilities appear to have tied values, then their ranks are computed as the average one. Note that this definition is slightly different from those given in ([56], p. 211 and p. 309), where the Spearman ρ distance is the Euclidean metric on permutations (rankings).
- *The Canberra Type Distance* [57]:

  $$C(\boldsymbol{P}, \boldsymbol{Q}) = \sum_{i=0}^{M}\left(\frac{2(p_i - q_i)}{p_i + q_i}\right)^2.$$

  This measure, which is closely associated with the Canberra distance dissimilarity, is very popular. It was successfully used for classification based on the Common N-Grams [30,31], in plagiarism detection area [34,36] and so on.

4.1.3. Clustering

Clustering is an unsupervised tool that suggests enhanced interpretations of the underlying data structure via partitioning it into homogeneous groups. Competitive discussions of the modern clustering techniques can be found in [58] and [59]. Roughly speaking, clustering methods fall into two categories of hierarchical and partitioning ones.

The clustering procedure separates the data points into given number of disjoint groups, usually via minimization of certain objective function. The Mean Squared Error (MSE) is by far the most popular objective function employed for partition clustering. This function evaluates the mean-squared distance of data points from the nearest centroid (cluster center). Euclidean sum-of-squares clustering, appearing ones the squared Euclidean distance used, is an *NP*-hard problem [60].

The famous K-means algorithm provides a suboptimal solution of this problem. In its most popular version, in order to decrease the value of the objective function this algorithm generates clusters directly in attempt to learn groups by their random initialization combined with iterative moving points between subsets. From the probabilistic point of view, the algorithm tries to detect dense areas in the data within the Gaussian Mixture Model (see, e.g. [58]).

The key benefit of K-means approach is that local minima for any initial centroid set can always be reached. The main weakness of K-means is that the obtained solution essentially depends on the conditions of process initialization, thus it is not able to solve global clustering problems. Numerous methods were proposed to initialize the K-means process (see, e.g. [58,61,62]). However, no method has been yet recognized as a superior one.

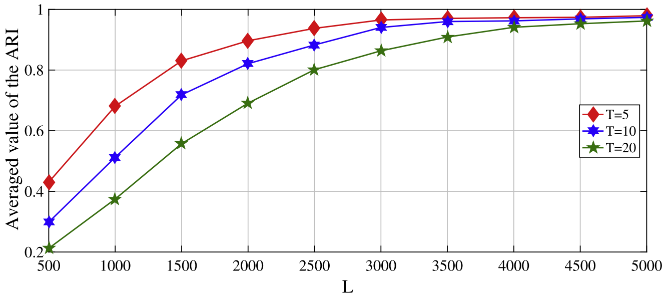

Fig. 8. The averaged values of adjusted Rand index obtained for the Spearman's Correlation Distance.

According to [58], iterative optimization is the most common technique used for seeking the optimal partitions. Generally speaking, the strategy is to relocate points from a group to an alternative group incrementally, trying to improve the value of the objective function. This idea was actually implemented in various iterative clustering procedures (see, e.g. [63]). Although the method can only ensure a local solution, it helps to avoid the so-called artificially stable clusters and to activate more suitable configurations.

Another prominent weakness of the method is connected to the fact that the arithmetic mean value obtained as a cluster representative (centroid) is not robust in respect to the outlying points. The K-medoids clustering approach related to the K-means method was considered as an attempt to highlight this problem.

K-medoids methodology aims to minimize the sum of divergences between all corresponding cluster items and the data item that was appointed as the cluster center. As a result, a whole cluster is represented by a single belonging point. Such a selection of medoids (cluster centers) is affected by the major fraction of elements within a cluster and, consequently, is more robust in comparison to the K-means. In particular, it is less sensitive to the outliers (see, e.g. [59]).

For our study we employ the most common implementation of K-medoids clustering, namely the Partitioning Around Medoids (PAM) algorithm [45]. Despite the fact, that PAM has the quadratic complexity with respect to the number of items. The algorithm is initiated with a starting set of medoids and then attempts to swap them with the non-medoids points, aiming to reduce the value of the object function. The algorithm stops when no possible change is left in the items assignments.

4.1.4. Parameter selection

In order to determine the appropriate parameter values, the Algorithm 4 is applied. In the future experiments the following book series were taken as two a priori different text collections:

- The “Foundation Universe” by Isaac Asimov (see, e.g. [42]).
- The “Rama” series by Arthur C. Clarke (see, e.g. [41]).

These collections contain 7 and 6 books, correspondingly. The Vector Space Model was constructed as described in Section 4.1.1 using a threshold $Tr = 0.5$ in (7). Each book from the first collection is compared with each book from another collection (42 comparisons). This is slightly different from the Algorithm 4, where randomly selected documents were related. We test three values of the delay parameter $\mathcal{T} = \{5, 10, 20\}$ with ten sequential values of the chunk sizes $\mathcal{L} = \{500, 1000, \ldots, 5000\}$. Fig. 8 presents three graphs of the adjusted Rand index calculated for chosen values of T using the Spearman's Correlation Distance.

Very close results are obtained for the Canberra Type Distance (see Fig. 9).

Assuming $C_{Rand} = 0.9$ in the Algorithm 4, we get $L^*(T) = 2500$, 2000, 1500. So, $T^* = 20$ and $L^*(T^*) = 2000$.

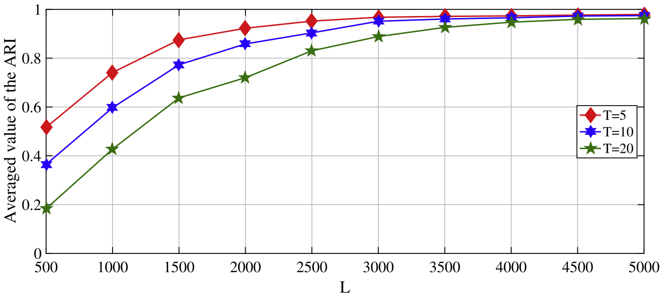

Fig. 9. The averaged values of adjusted Rand index obtained for the Canberra Type Distance.

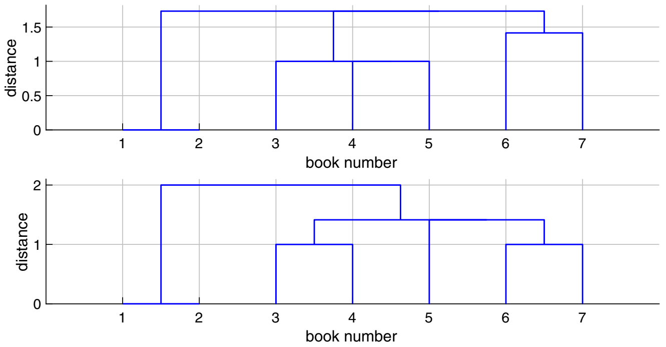

Fig. 10. Dendrograms of the “Foundation” series hierarchy.

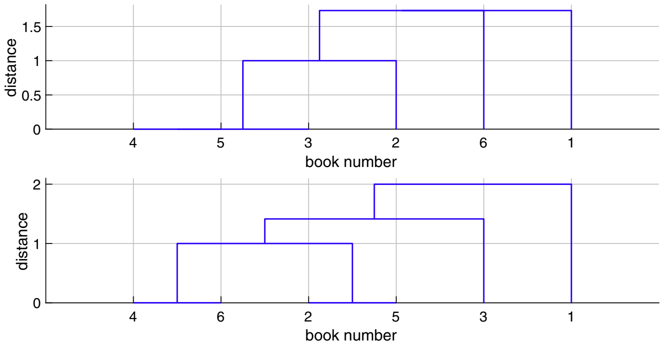

Fig. 11. Dendrograms of the “Rama” series hierarchy.

Thus, we use the following parameter values in our experiments:

- $T = 20$,
- $L = 2000$,
- $Tr = 0.5$.

4.2. Evolution of the style in book series

A book series is a set of several tomes, organized together in an arranged collection on account of certain common features. Series are formed to share a common scenery, story arc or a group of characters by means of referencing some preceding events. Thus, the books from a given series are usually published sequentially, in accordance with their internal chronology. Often, the principal characters (the series skeleton), develop across the series, although it does not influence on the central plot. Some authors do not write their books in the chronological order, publishing each book independently of internal chronology of the plot. Therefore, the writing style of a series may evolve, following the changes of the authors' attitude or alterations of genre. In this section, we apply the proposed methodology to expose the evolution of writing style and to divide a series into style-consistent periods.

We analyze the following four book series:

- “Foundation Universe” by Isaac Asimov (see, e.g. [42]).
- “Rama” series by Arthur C. Clarke (see, e.g. [41]).
- “Forsyte Saga” by John Galsworthy (see, e.g. [64]).
- “The Lord of the Rings” by John Ronald Reuel Tolkien (see, e.g. [65]).

Additionally, a set of twelve available in the internet books prescribed to a famous novelist Romain Gary is studied.

4.2.1. Two-step clustering and results visualization

The two-step cluster analysis is a scalable clustering methodology constructed to manage very large data sets [66]. The common approach consists of two main steps. Initially, a partition algorithm like the *K*-means is applied, in order to form the so-called small “pre-clusters”. The number of clusters can be beforehand determined or evaluated using a cluster validation technique. The obtained clusters are expected to be sufficiently consistent but not too small, since they are treated at the next step as separate observations. Afterwards, a procedure of hierarchical agglomerative clustering consecutively combines the “pre-clusters” into the homogeneous groups. An agglomerative hierarchical procedure starts from the singleton clusters and aggregates them into groups until a stopping criterion is met. No item is moved from a constructed cluster to another one. In this paper we propose different procedure designed in the spirit of the two-step cluster methodology.

At the first stage, books from a series are compared to each other by the Algorithm 2 (see, Section 3.2). This procedure assigns documents to styles based on a partition clustering technique accompanied by the major voting. The results are presented via a binary square matrix, where ’1’-s indicate the corresponding pair of books found to have different styles. However, at this point we would like to classify the books' similarity by means of the overall relationship between the styles. Namely, we intend to form books into groups resting upon their similarity or dissimilarity with all books in the series. To this effect the row clusters of the obtained binary classification matrix is created, using the single linkage agglomerative hierarchical algorithm based on the Hamming distance. In our case of binary vectors it coincides with the standard Euclidean distance. The process is performed until all items are collected into one single cluster. This hierarchical clustering procedure yields a nested structure of the styles.

The obtained dendrogram, named resulting classification tree in our method, reveals a visualization of the writing style evolution. Further, we present such trees via dendrogram plots, where the *y*-axis represents the distances between the conjoint items.

4.2.2. The “Foundation Universe” by Isaac Asimov

The “Foundation Universe” is the legendary collection of science fiction books by Isaac Asimov, which had been published during the period between 1950 and 1993. The consequent books had the arbitrary order in respect to the internal chronology of a series. The plot of the original series, which was centered around the mathematician Hari Seldon and the development of his plan, was later merged with other Asimov cycles. The “Author's Note” to the “Prelude to Foundation” proposes the timetable of the original “Foundation” series, and it is also said there that “they were not written in the order in which (perhaps) they should be read”. The book “Forward the Foundation” is not mentioned in this list, as it have not yet been published. However, based on its contents this novel is usually put into the second position on the internal timescale of the series. We arrange the original “Foundation” series in the following order:

- “Prelude to Foundation” (denoted as $F1$) (1988),

Table 1
Comparison of the “Foundation” series using the S distance.

	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$
$F1$	0	0	1	1	1	1	1
$F2$	0	0	1	1	1	1	1
$F3$	1	1	0	1	0	1	1
$F4$	1	1	1	0	0	1	1
$F5$	1	1	0	0	0	1	1
$F6$	1	1	1	1	1	0	1
$F7$	1	1	1	1	1	1	0

Table 2
Comparison of the “Foundation” series using the C distance.

	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$
$F1$	0	0	1	1	1	1	1
$F2$	0	0	1	1	1	1	1
$F3$	1	1	0	0	0	1	1
$F4$	1	1	0	0	1	1	1
$F5$	1	1	0	1	0	1	0
$F6$	1	1	1	1	1	0	0
$F7$	1	1	1	1	0	0	0

- “Forward the Foundation” (denoted as $F2$) (1993),
- “Foundation” (denoted as $F3$) (1951),
- “Foundation and Empire” (denoted as $F4$) (1952),
- “Second Foundation” (denoted as $F5$) (1953),
- “Foundation’s Edge” (denoted as $F6$) (1982),
- “Foundation and Earth” (denoted as $F7$) (1986).

Tables 1 and 2 represent the results of style comparison obtained using the S and C distances, correspondingly.

Table 1 highlights the following clusters: {$F1$, $F2$}, {$F3$, $F4$, $F5$}, {$F6$} and {$F7$}. The first two rows and first two columns (the first cluster) in Table 1 contain only ‘0’s. The block corresponding to the second cluster is composed from seven ‘0’s and only two ‘1’s. The sixth and seventh columns contain only ‘1’s except for the diagonal elements. The classification tree given in Fig. 10 (top panel) reaffirms this partition result.

As one can see from Table 2 and Fig. 10 (bottom panel), the partition provided by the C distance is slightly different. Here, the second cluster contains {$F3$} and {$F4$}, and “ Second Foundation” is moved to {$F5$, $F6$, $F7$}. This may be connected to the fact that Asimov tried to finish the series with “Second Foundation”, however admirers persuaded him to write the sequel.

The “Foundation” initially consisted of eight small sections, which had been published between May 1942 and January 1950. The first tome of the series entitled “Foundation” and issued in 1951 consists of the main four stories and single ancillary story, which takes place after the main ones. The rest of the pairwise combined stories formed the “Foundation and Empire” (1952) and the “Second Foundation” (1953) tomes. This collection known as the “Foundation Trilogy” exactly coincides with the second cluster in the obtained partition. These three books form the same cluster in both partitions.

The fourth tome entitled “Foundation’s Edge” was written after a 30-year pause in 1982 and was accompanied with “Foundation and Earth” later in 1986. In this volume, Asimov tries to bring together all three novels “Robot”, “Empire” and “Foundation” into a unified “Universe” and to offer the “Galaxia” notion as an integrated collective mind. This pair of books comprises the third cluster in the second partition and two separate groups in the first one. This fact can be related to the difference in intent of these books. Afterwards, Asimov wrote two prequels that comprise the first cluster. Thus, both partitions obtained via our method perfectly suit the evolution of writing style.

Table 3
Comparison of the “Rama” series using the S distance.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$
$R1$	0	1	1	1	1	1
$R2$	1	0	0	0	0	0
$R3$	1	0	0	0	0	1
$R4$	1	0	0	0	0	1
$R5$	1	0	0	0	0	1
$R6$	1	0	1	1	1	0

Table 4
Comparison of the “Rama” series using the C distance.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$
$R1$	0	1	1	1	1	1
$R2$	1	0	0	0	0	0
$R3$	1	0	0	1	0	1
$R4$	1	0	1	0	0	0
$R5$	1	0	0	0	0	0
$R6$	1	0	1	0	0	0

4.2.3. The “Rama” series by Arthur C. Clarke

This book series includes six novels:

- “Rendezvous with Rama” (denoted as $R1$) (1972),
- “Rama II” (denoted as $R2$) (1989),
- “The Garden of Rama” (denoted as $R3$) (1991),
- “Rama Revealed” (denoted as $R4$) (1993),
- “Bright Messengers” (denoted as $R5$) (1995),
- “Double Full Moon Night” (denoted as $R6$) (1999).

“Rendezvous with Rama” is the first novel written personally by Arthur C. Clarke and published in 1972. Arthur C. Clarke paired up with Gentry Lee for ($R2 - R4$) books. According to [67], these books were actually written by Gentry Lee, while Arthur C. Clarke was mainly providing the editing recommendations. The next two novels $R5$ and $R6$ were written by Gentry Lee alone. Tables 3 and 4 represent the results of the style comparison obtained using the S and C distances, correspondingly.

The following Fig. 11 displays dendrograms of the series hierarchy for two distances (S-top panel and C-bottom panel), correspondingly.

Reviewing the obtained results we note that the source novel “Rendezvous with Rama” ($R1$) is completely different from other books of the series, as to be expected. In both tables the first row and first column are composed from ‘1’s except for the first element. This initial novel was awarded on several occasions, but the following books did not receive the same critical acclaim. Table 3 together with Fig. 11 (top panel) shows a three cluster structure {$R1$}, {$R2 - R5$} and {$R6$}. This result is in good agreement with the fact that the books ($R2 - R5$) were published at constant rate of one book per two years, but the last book ($R6$) was published after a four years pause.

According to the allocation of ‘0’s and ‘1’s the classification based on the C distance yields three clusters {$R1$}, {$R2 - R3$} and {$R4 - R6$}. The corresponding classification tree offers the following split (see, Fig. 11 (bottom panel)): {$R1$}, {$R3$}, {$R2$, $R5$} and {$R4$, $R6$}. These two partitions are poorly matched and do not agree with the series creation process.

4.2.4. The “Forsyte Saga” by John Galsworthy.

The famous “Forsyte Saga” by John Galsworthy was announced under that title for the first time in 1922. It includes three novels and two interludes written during the period between 1906 and 1921. Galsworthy created a sequel to the series, named “A Modern Comedy”, between 1924 and 1928. An additional sequel trilogy, “End of the Chapter”, which is actually a spin-off from the pre-

Table 5
Comparison of the “Forsyte Saga” series using the *S* and *C* distances.

	*For*1	*For*2	*For*3	*For*4	*For*5	*For*6	*For*7
*For*1	0	1	0	1	1	1	1
*For*2	1	0	1	1	1	1	0
*For*3	0	1	0	0	1	1	1
*For*4	1	1	0	0	1	0	1
*For*5	1	1	1	1	0	0	0
*For*6	1	1	1	0	0	0	0
*For*7	1	0	1	1	0	0	0

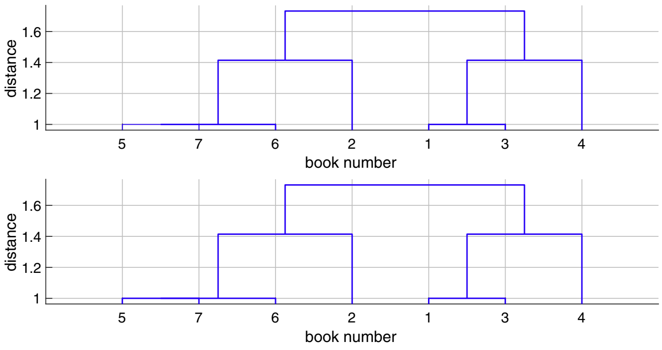

Fig. 12. Dendrograms of the “Forsyte Saga” series hierarchy.

viously written stories, was issued in 1931–1933. We analyze the following titles:

1. **The “Forsyte Saga”**
   - “The Man of Property” (novel denoted as *For*1) (1906),
   - “Indian Summer of a Forsyte” (interlude denoted as *For*2) (1918),
   - “In Chancery” (novel denoted as *For*3) (1920),
   - “To Let” (novel denoted as *For*4) (1921).
2. **“End of the Chapter”**
   - “Maid In Waiting” (novel denoted as *For*5) (1931),
   - “Flowering Wilderness” (novel denoted as *For*6) (1932),
   - “Over the River (One more River)” (novel denoted as *For*7) (1933).

We do not take a very short interlude “Awakening” published in 1920 into account. Both of the considered distance functions *S* and *C* provide the same classification results presented in the Table 5.

First of all, from Table 5 one can see that the writing styles of the second sub-series are similar to each other, yet they are different from styles of other books in the collection. Thus, these books have formed a cluster of their own. The styles of the book pairs {*For*1, *For*3} and {*For*3, *For*4} can be successfully distinguished. These three manuscripts comprise the next cluster. The remaining single interlude naturally falls into its own self-containing cluster. The obtained hierarchy of the series is properly validated and given in Fig. 12. The style evolution of the cluster is {*For*1, *For*3, *For*4} is clearly outlined. At the first stage {*For*1, *For*3} is constructed, and afterwards {*For*4} is appended to the cluster. All works are accurately divided in accordance with time period of their creation. Fig. 2, 4, 6 and 7

4.2.5. *“The Lord of the Rings” by John Ronald Reuel Tolkien*

“The Lord of the Rings” is an epic high fantasy story created by John Ronald Reuel Tolkien as a sequel to his previous fantasy novel “The Hobbit” published in 1937. We analyze the following five titles:

- “The Hobbit” (denoted as *T*1) (1937),
- “The Fellowship of The Ring” (denoted as *T*2) (1954),

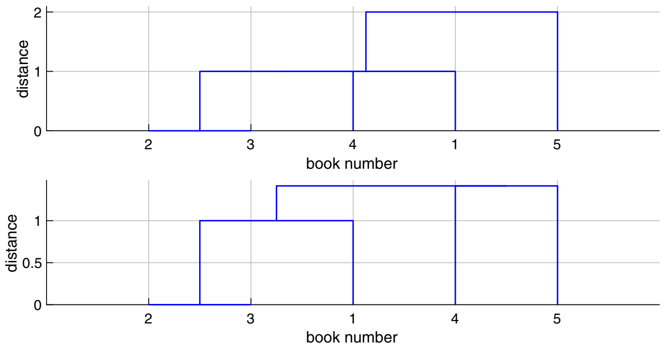

Fig. 13. Dendrograms of the “Lord of the Rings” series hierarchy.

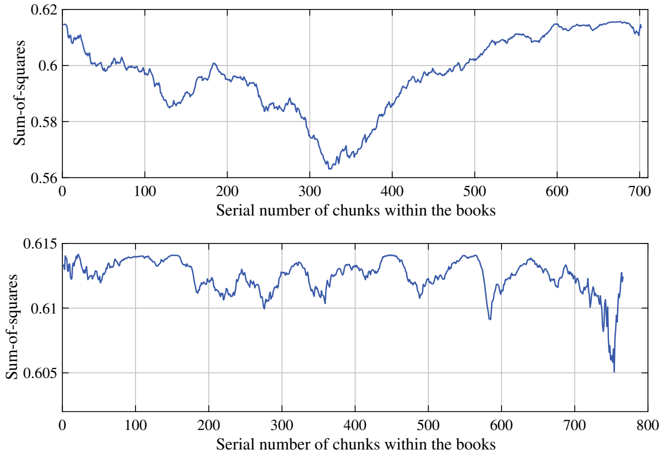

Fig. 14. Examples of the Euclidean sum-of-squares error graphs.

Table 6
Comparison of the “Lord of the Rings” series using the *S* distance.

	*T*1	*T*2	*T*3	*T*4	*T*5
*T*1	0	0	0	1	1
*T*2	0	0	0	0	1
*T*3	0	0	0	0	1
*T*4	1	0	0	0	1
*T*5	1	1	1	1	0

Table 7
Comparison of the “Lord of the Rings” series using the *C* distance.

	*T*1	*T*2	*T*3	*T*4	*T*5
T1	0	0	0	1	1
T2	0	0	0	0	1
T3	0	0	0	0	1
T4	1	0	0	0	0
T5	1	1	1	0	0

- “The Two Towers” (denoted as *T*3) (1954),
- “The Return of The King” (denoted as *T*4) (1955),
- “The Silmarillion” (denoted as *T*5) (1977).

Tables 6, 7 and Fig. 13 demonstrate the obtained results.

From Table 6 one can see that books *T*2, *T*3 and *T*4 (the core part of the series) constitute a purely homogeneous cluster (just ’0’-s in the corresponding block of the matrix). Such result is to be expected, since the novels were created by splitting single unpublished text into three parts. Predictably, the novel *T*1 (“The Hobbit”) is closely connected to this cluster. Nevertheless, the style of this book actually differs from the style of *T*4. Finally, the last book *T*5 is positioned quite far, separately from all the others novels. It may be explained by the fact that this novel, named “The Silmarillion", was compiled and issued later by Tolkien's son, Christopher Tolkien, in 1977, with help of G. G. Kay. He had to write several parts himself in order to fix the discrepancies in the plot. The main distinction in the classification provided by the *C* distance is similarity of the last book of the trilogy *T*4 and "The Silmarillion" novel. However, the general merit of the series is preserved.

4.2.6. Romain Gary novels

Romain Gary (Roman Kacew) is a well-known Jewish-French author published, as many critics believe, under the pseudonyms of Émile Ajar, Shatan Bogat, Rene Deville and Fosco Sinibaldi, directed two movies, fought in the air force, and represented France as a consul. It is conventionally considered that he is the only person to have won the Prix Goncourt under his own name ("Les Racines du ciel" – 1956) and under the pseudonym Émile Ajar ("La Vie devant soi" – 1975). although some critics are not sure that the second book was written by him. We analyzed the following available in the internet novels denoted as $RG1, \ldots, RG12$.

As Romain Gary:

- "Éducation européenne" (translated as "Forest of Anger", reprinted as "Nothing Important Ever Dies" and "A European Education") (1945) [68]
- "Le Grand Vestiare" (translated as "The Company of Men") (1949) [69]
- "Les Racines du ciel" (translated as "The Roots of Heaven") (1956) [70]
- "La Promesse de l'aube" (translated as "Promise at Dawn") (1961) [71]
- "Chien blanc" (self-translation of the novel "White Dog") (1970) [72]
- "Charge d'âme" (self-translation of the novel "The Gasp") (1977) [73]
- "Les Clowns lyriques" (self-translation of the novel "The Colours of the Day") (1979) [74]

As Émile Ajar:

- "Gros-Câlin" (not translated in English, the title means "Big Cuddle") (1974) [75]
- "La Vie devant soi" (translated as "Madame Rosa" and later re-released as "The Life Before Us") (1975) [76]
- "Pseudo" (1976) [77]
- "L'Angoisse du roi Salomon" (translated as "King Salomon") (1979) [78]

As Shatan Bogat:

- "Les Têtes de Stéphanie" (translated as "Direct Flight to Allah") (1974) [79]

All considered texts are written in French, and a problem appearing here is as such that there is not any acceptable list of the content-free words in French. Instead of this collection, we use a set of stop words. These words are typically cleaned out within text mining approaches because they are basically a collection of the extremely common used words in any language to be seen of minor value in documents classification (see, e.g. [18] , chap. 15). From this point of view, stop words play role similar, but not identical, to the role of the content-free words gluing informative terms in a text. No single generic list of stop words exists. We operate in our experiments with a list presented in [80].

Note that the books under study are not a series related to a common plot. Probably therefore the proposed Two-step Clustering does lead to consequential results, since provided pairwise comparisons indicate differences in the style between almost all books. However, Algorithm 1 makes it possible to describe the inner structure of the considered collection. The same cluster structure is revealed for the both considered distances:

- $\{RG1 - RG7, RG12\}$.
- $\{RG8 - RG11\}$.

Table 8
Distribution of the averaged distance.

Novel	DIST
1	0.54
2	0.37
3	0.43
4	0.46
5	0.41
6	0.47
7	0.39
8	0.47
9	0.65
10	0.42
11	0.55
12	0.50

As can be seen, the second cluster contains only novels "written by Émile Ajar". So, the procedure hints to different in style between the novels written under this pseudonym and the rest of the considered books' collection. It is curious, in this connection, to comprehend how two awarded with the Prix Goncourt books ($RG3$ and $RG9$) lie inside the collection. To this end let construct from the books' divisions: $D_i = \{\widehat{D}_1^{(i)}, \ldots, \widehat{D}_{m_i}^{(i)}\},\ i = 1, \ldots, 12$ a new metric

$$\mathrm{DIST}_{i_1,i_2} = \underset{j_1,j_2}{\mathrm{average}}\left(\mathbb{V}_{j_1,j_2}^{(i_1),(,i_2)}\right),\ \ i_1, i_2 = 1, \ldots, 12,$$

which delivers the average DZV distance between the chunks of each two novels.

Table 8 exhibits the total metric-distance of each one of the books to other books of the collection found using the Spearman's Correlation Distance. The results obtained for the Canberra Type Distance are very similar.

The first awarded book ($RG3$) properly lies within the collection heard. In the next step, we apply an approach detects the outlier values using Thopson's Tau method [81], which finds that $RG9$ (the second laureate of the Prix Goncourt) is the truest outlier. Therefore, the awarded books are completely different in their own style. The first one corresponds absolutely to the general style of the author, and the second one is written in a fully different style. The carried out formal analysis does not allow surely deducing anything about the authorship of $RG9$. The conclusion may be founded on additional research including not formal stylistic study.

4.3. Clustering of sequential data as an alternative approach

A key ingredient of the proposed method is a time series representation of a text evolution. Analogous text description appears in the plagiarism detection tasks [35]. Here, a text is also divided in chunks that are imaged as distributions of suitably chosen *N*-grams. These "*N*-grams profiles" are compared with one obtained for whole document aiming to detect essential fluctuations in the style. The principal supposition is as such as that, there is a leading text's author, who mainly wrote the document. The approach demonstrated high ability to discover the style variations in relatively small text's portions. However, it is hardly expected to trace effectively the style evolution, since the method is inherently constituted to find deviations from the underline template, which can temporary change.

Another method based on a time series representation is proposed in [82] for a new computational approach for tracking and detecting statistically significant linguistic shifts in the meaning and usage of words. A time series constructed to reflect word usage exposes linguistic modifications by allocation of change points

Table 9
Comparison of the “Foundation” series using the *S* distance and a sequential clustering.

	*F*1	*F*2	*F*3	*F*4	*F*5	*F*6	*F*7
*F*1	0	1	0	1	1	1	0
*F*2	1	0	1	1	1	1	1
*F*3	0	1	0	0	1	1	1
*F*4	1	1	0	0	1	1	1
*F*5	1	1	1	1	0	0	0
*F*6	1	1	1	1	0	0	0
*F*7	0	1	1	1	0	0	0

Table 10
Comparison of the “Rama” series using the *S* distance and a sequential clustering.

	*R*1	*R*2	*R*3	*R*4	*R*5	*R*6
*R*1	0	1	1	1	1	1
*R*2	1	1	0	1	0	1
*R*3	1	0	0	0	0	1
*R*4	1	1	0	0	0	0
*R*5	1	0	0	0	0	1
*R*6	1	1	1	0	1	0

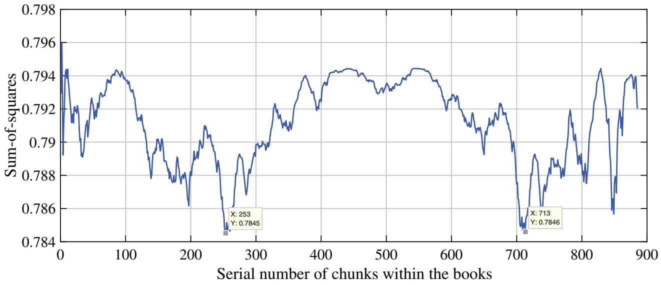

Fig. 15. Graph of the Euclidean sum-of-squares in comparison *R*2 with itself.

of the series. This method can be apparently used for tracing of a writing style evolution by applying an ensemble technique summarizing the behaviors of separate words that definitely leads to a more complicated computational model.

Note that a partition of time series of any Sequential Data is essentially recognised by its change points. The Sequential Data methodology that takes advantage of a time series measurements is one of the most intensively studied subjects in the area of pattern recognition (see, e.g. [83], [84] Chapter 13, [85]). Dealing with clustering of such data, we expect that the desired clusters will contain the connected item segments. The classical clustering algorithms typically are ill-suited to provide a partition since they do not take the inherent sequential structure into account. A lot of different algorithms have been proposed to handle this problem. A thorough review can be found in [86].

A distinguished Warped *K*-Means method proposed in this article solves the problem via the iterative optimization of the Euclidean sum-of-squares (see, Section 4.1.3), while adding a strict sequential constraint in the classification step.

The Mean Dependence ZV method suggests a time series representation of a text. It appears very natural to apply the sequential clustering methodology aiming to split a document into intervals of homogeneous writing style. We discuss such an approach in this section.

In such a manner ZV is calculated for a concatenation of two texts and afterwards the texts are divided into two clusters using a sequential clustering. If the majority of the texts' volume belongs to a single cluster then the styles are accepted as identical, otherwise they are recognized as different. Due to the arrangement of cluster attachment, which actually appears to be a connected segment, the optimization of the Euclidean sum-of-squares can be explicitly undertaken by means of straight exhaustive search across all possible segment borders. We use this approach in this study instead of the Warped *K*-Means method.

4.3.1. *The “Foundation Universe” by Isaac Asimov*

As the first example of the methodology being discussed, we consider a style classification of the “Foundation Universe” by Isaac Asimov. The result of the pairwise book comparisons from the series is given in Table 9.

Fig. 14 demonstrates two typical examples of the Euclidean sum-of-squares error graphs appearing during the comparison procedure.

The graph presented on the top panel corresponds to comparison of *F*1 and *F*2. The global minimum point lies very close to the border between the volumes and thus the writing styles of this pair are recognized as different. The second graph obtained via comparison of *F*1 and *F*7 has the global minimum point near the end of the united document. The majority of the texts' chunks are assigned to the cluster located before the discussed point and therefore the styles are accepted as identical.

In accordance with the square blocks in the table filled only by ‘0’s, Table 9 suggests the following clusters : {*F*1}, {*F*2}, {*F*3, *F*4} and {*F*5, *F*6, *F*7}. The second cluster {*F*3, *F*4} can be considered as consistent, since both books composing it were written and published within a sufficiently small time interval. The group {*F*5, *F*6, *F*7} appears to be artificial. Books *F*5 and *F*6 published around thirty years after the first one are significantly different from *F*5 in terms of their style and plot. Moreover, comparison of all series within the framework of two-step clustering procedure described earlier (the Hamming distance between the rows) reveals that according to this table *F*6 is more similar to *F*5 than to *F*7. This is quite an unexpected result, given the years of these books publication are 1982, 1953 and 1986 correspondingly.

4.3.2. *The “Rama” series*

The second example that we consider is the “Rama” series by Arthur C. Clarke. The clustering procedure outcome is presented in Table 10.

First of all, as it was expected, the style of the book *R*1 is completely different from the styles in the rest of collection. There is only a single non self-contained cluster {*R*4, *R*5, *R*6}. Its style can be barely interpreted, because, as it was discussed in Section 4.2.3, *R*5 and *R*6 were written by Gentry Lee alone, so the last book (*R*6) was published after a gap of four years. For *R*4 Arthur C. Clarke was offering only the general editing suggestions.

Interesting phenomena arise when *R*2 is compared with itself and the styles are recognized as different. An appropriate graph of the sum-of-squares is showed in Fig. 15.

One can clearly identify two adjacent optimal points: the first one is located at position 253 with the sum-of-squares value of 0.7845; the second one appears at position 713 with the sum-of-squares value of 0.7846. Logic suggests that second point is more appropriate since in this case the styles of two texts are not distinguishable. On the other hand, from the formal point of view the first location has to be chosen. Presence of such optimal points can hypothetically indicate the fact that this novel was written by two different authors. Note, that the Algorithm 2 assigns about 75% of the text volume to a single cluster.

Summarizing the above, one can conclude that being directly applied to the studied task the methodology of sequential data clustering leads to less appropriate results. Hopefully, an attempt to incorporate a dynamic distance like DZV into the method can improve the performance. This approach seems to be very promising but it needs more detailed consideration, which cannot be pro-

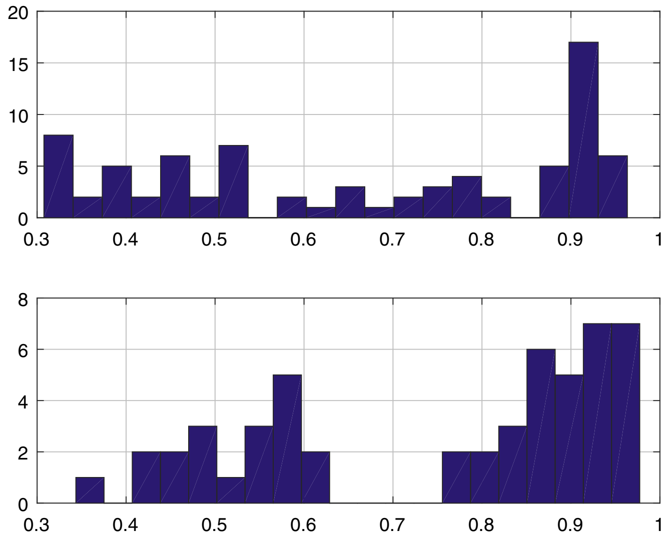

Fig. 16. Histograms of *ARI*.

Table 11
Comparison of single books to the series.

	R	F	For	R	F	For
AC	**23**	2	5	**23**	2	5
NEM	0	**30**	0	0	**28**	2
WM	0	1	**29**	0	2	**28**
RSH	1	11	18	5	14	11

vided within the context of this article. We are going to study this problem in our future research.

4.4. Experiments with the author identification procedure

In this section we present experiments with the Algorithm 3 described in Section 3. The set consisting of three first books from the collection studied in the previous subsection is used as the training source. The material comprising the documents under investigation is drawn from the following books, which do not belong to any of studied collections:

- “2010: Odyssey two” by Arthur C. Clarke (denoted as *AC*), published in 1982 as the sequel to the 1968 novel “2001: A Space Odyssey”,
- “Nemesis” by Isaac Asimov (denoted as *NEM*), published in 1989. As it was declared by the author in the Author’s Note: “This book is not part of the Foundation Series, the Robot Series, or the Empire Series. It stands independently”,
- “The White Monkey” by John Galsworthy (denoted as *WM*), published in 1924 as the first novel in John Galsworthy’s second “Forsyte trilogy”,
- “Immortality, Inc.” by Robert Sheckley (denoted as *RSH*), published in 1959.

The author identification procedure is implemented in the following mode. At first, one of the books from this list is selected as the text source for the examination. During every iteration, a single book from each of the series is randomly chosen, and the investigated document is drawn as a random sequential sub-text of the source, with length of $(T+40)L$. Then the Algorithm 3 is applied. An iteration is considered successful if the value of ARI calculated for the source collection is greater than a threshold value $C_{Rand} = 0.8$. The process stops when thirty successful iterations are collected. The results are presented in the following Table 11.

First three columns correspond to the *S* distance, and the least three columns relate to the *C* distance. As one can see all standalone novels written by the corresponding collection authors are properly assigned to the correct collections. On the other hand, the book *RSH* written by an author not belonging to any of the source collections has no clear affiliation. The histograms of the ARI calculated for *AC* before accumulating thirty successful iterations are given in Fig. 16 The *S* and *C* distances match up to the top and bottom panels correspondingly. The second distribution is shifted towards unity, and the number of the trials (51) is smaller compared to the fist case. This tendency applies to all experiments, i. e. the process converges faster if the *C* distance is used. The last novel in the list (*RSH*) is significantly separated from others because it is not written by any of the collection authors. The assigned values for *RSH* have a more even distribution over the sources. To quantify the degree of scattering, we calculate its *p*-value computed as the probability of the maximal assignment value to be greater of 0.5:

$$p = \Phi\left(\frac{f_0 - 0.5}{0.5}\sqrt{30}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution. This *p*-value is used in the Hypothesis Testing procedure that verifies if the sample proportion is greater than 0.5 for a sufficiently large sample size (see, e.g. [87]). The corresponding values of 0.8633 and 0.3575 are less than the default significance level 0.95. Hence, the book cannot be definitely allocated to any collection.

5. Conclusion

This paper presents a novel, simple and efficient methodology intended to model the evolution of the author’s writing style. Using the Mean Dependence values, we process the documents under consideration to represent them as a time series. Therefore, when a document is produced with the single writing style this sequence should oscillate around a certain constant level. The presence of significant variation points in this sequence indicates a possible alteration of the writing style. A new distance function constructed using this feature and incorporated in a clustering procedure allows one to categorize writing styles into a number of homogeneous groups. Application of this procedure to the comparison of books from a single series demonstrates its fair ability to trace changes in the style of a series, also in good agreement with the literary criticism. Resting upon this classification, we propose a new tree-type inner representation of the book series. Preferably, the Spearman based distance should be employed at this stage. The experimental trials of the constructed author identification procedure exhibit its high reliability for the tasks of identifying the appropriate author from a given set, especially while using the Canberra type distance.

In the future, we plan to investigate a procedure intended to accommodate the model and its parameter configuration to the corpus structure in an effort to classify relatively short documents and to take language differences into account. Another prominent research direction is a study of possible applications of the sequential clustering methodology.

Acknowledgement

This work was supported by Russian Science Foundation (project 16-19- 00057). The authors would like to thank the anonymous reviewers for serious and constructive suggestions and comments that helped us to significantly improve the manuscript.

References

[1] J. Alred, C. Brusaw, W. Oliu, Handbook of Technical Writing, Ninth Edition, St. Martin's Press, 2008.
[2] Z. Volkovich, O. Granichin, O. Redkin, O. Bernikova, Modeling and visualization of media in arabic, J. Inform. 10 (2) (2016) 439–453.
[3] Z. Volkovich, A time series model of the writing process, in: Machine Learning and Data Mining in Pattern Recognition, Springer, 2016, pp. 128–142.
[4] Z. Volkovich, R. Avros, Text classification using a novel time series based methodology, in: Proceedings of the 20th International Conference Knowledge-Based and Intelligent Information & Engineering Systems: KES-2016, Procedia Computer Science, 2016, pp. 53–62.
[5] D. Lemberg, A. Soffer, Z. Volkovich, New approach for plagiarism detection, Int. J. Appl. Math. 29 (3) (2016) 365–371.
[6] P. Juola, Authorship attribution, Found. Trends Inf. Retr. 1 (3) (2006) 33–334.
[7] J. Binongo, Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution, Chance 6 (2) (2003) 9–17.
[8] J.M. Hughes, N.J. Foti, D.C. Krakauer, D.N. Rockmore, Quantitative patterns of stylistic influence in the evolution of literature, in: Proceedings of the National Academy of Sciences, 109, 2012, pp. 7682–7686.
[9] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 538–556.
[10] J.F. Burrows, Delta: a measure of stylistic difference and a guide to likely authorship, Lit. Linguist. Comput. 17 (2002) 267–287.
[11] S. Argamon, Interpreting burrows's delta: geometric and probabilistic foundations, Lit. Linguist. Comput. 23 (2) (2008) 131–147.
[12] D. Hoover, Testing burrows's delta, Lit. Linguist. Comput. 19 (4) (2004) 453–475.
[13] S. Stein, S. Argamon, A mathematical explanation of Burrows's Delta, in: Proceedings of the Digital Humanities Conference, 2006, pp. 207–209.
[14] W. Oliveira, E. Justino, L. Oliveira, Comparing compression models for authorship attribution, Forensic Sci. Int. 228 (1) (2013) 100–104.
[15] D. Cerra, M. Datcu, P. Reinartz, Authorship analysis based on data compression, Pattern Recogn. Lett. 42 (Supplement C) (2014) 79–84.
[16] Y. Zhao, J. Zobel, Effective and scalable authorship attribution using function words, in: Proceedings of Asia Information Retrieval Symposium, 2000, pp. 174–189.
[17] J. Diederich, J. Kindermann, E. Leopold, G. Paas, Authorship attribution with support vector machines, Appl. Intell. 19 (1) (2003) 109–123.
[18] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 2003.
[19] H. Wu, J. Bu, C. Chen, J. Zhu, L. Zhang, H. Liu, C. Wang, D. Cai, Locally discriminative topic modeling, Pattern Recogn. 45 (1) (2012) 617–625.
[20] F. Peng, D. Schuurmans, V. Keselj, S. Wang, Augmenting naive bayes classifiers with statistical languages model, Inf. Retr. Boston 7 (2004) 317–345.
[21] G. Sidorov, Non-continuous syntactic n-grams, Polibits 48 (1) (2013) 67–75.
[22] G. Sidorov, Should syntactic n-grams contain names of syntactic relations, Int. J. Comput. Linguist. Appl. 5 (1) (2014) 139–158.
[23] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernandez, Syntactic n-grams as machine learning features for natural language processing, Expert Syst. Appl. 41 (3) (2014) 853–860.
[24] R.M. Coyotl-Morales, L. Villasenor-Pineda, M. Montes-y Gomez, P. Rosso, Authorship attribution using word sequences, in: Proceedings of Iberoamerican Congress on Pattern Recognition, 2006, pp. 844–853.
[25] J. Rudman, The state of authorship attribution studies: some problems and solutions, Comput. Hum. 31 (1998) 351–365.
[26] M. Kestemont, K. Luyckx, W. Daelemans, T. Crombez, Cross-genre authorship verification using unmasking, Engl. Stud. 93 (3) (2012) 340–356.
[27] K. Luyckx, W. Daelemans, Authorship attribution and verification with many authors and limited data, in: Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 513–520.
[28] M. Koppel, Y. Winter, Determining if two documents are written by the same author, J. Am. Soc. Inf. Sci. Technol. 65 (1) (2014) 178–187.
[29] E. Stamatatos, W. Daelemans, B. Verhoeven, P. Juola, A. Lopez, M. Potthast, B. Stein, Overview of the author identification task at pan 2015, in: Proceedings of CLEF (Working Notes), 2015.
[30] V. Keselj, F. Peng, N. Cercone, C. Thomas, N-gram-based author profiles for authorship attribution, in: Proceedings of the Conference Pacific Association for Computational Linguistics, 2003, pp. 255–264.
[31] M. Jankowska, V. Keselj, E.E. Milios, Proximity based one-class classification with common *n*-gram dissimilarity for authorship verification task, in: Proceedings of CLEF 2013 Evaluation Labs and Workshop, 2013.
[32] J. Frery, C. Largeron, M. Juganaru-Mathieu, UJM at CLEF in author verification based on optimized classification trees, in: Proceedings of CLEF 2014, 2014.
[33] O. Halvani, M. Steinebach, An efficient intrinsic authorship verification scheme based on ensemble learning, in: Proceedings of the 9th International Conference on Availability, Reliability and Security, 2014, pp. 571–578.
[34] M. Kestemont, K. Luyckx, W. Daelemans, Intrinsic plagiarism detection using character trigram distance scores, in: Proceedings of PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2012 Conference, 2011, p. 8.
[35] G. Oberreuter, J. Velàsquez, Text mining applied to plagiarism detection: the use of words for detecting deviations in the writing style, Expert Syst. Appl. 40 (9) (2013) 3756–3763.
[36] E. Stamatatos, Intrinsic plagiarism detection using character *n*-gram profiles, in: Proceedings of SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, 2009, pp. 38–46.
[37] H. Zhang, T. Chow, A coarse-to-fine framework to efficiently thwart plagiarism, Pattern Recogn. 44 (2) (2011) 471–487.
[38] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol. 60 (1) (2009) 9–26.
[39] D. Shalymov, O. Granichin, L. Klebanov, Z. Volkovich, Literary writing style recognition via a minimal spanning tree-based approach, Expert Syst. Appl. 61 (2016) 145–153.
[40] O. Granichin, N. Kizhaeva, D. Shalymov, Z. Volkovich, Writing style determination using the KNN text model, in: Proceedings of 2015 IEEE International Symposium on Intelligent Control (ISIC), 2015, pp. 900–905.
[41] A.C. Clarke, G. Lee, The Complete Rama Omnibus, Gollancz, 2011.
[42] I. Asimov, The Complete Isaac Asimov's Foundation Series Books 1–7, Mass Market Paperback, 2016.
[43] C. Kuratowski, Quelques problèmes concernant les espaces métriques non-séparables, Fundam. Math. 25 (1) (1935) 534–545.
[44] T. Hofmann, B. Schölkopf, A. Smola, Kernel methods in machine learning, Ann. Stat. 36 (3) (2008) 1171–1220.
[45] L. Kaufman, P.J. Rousseeuw, Finding Groups in data: An Introduction to Cluster Analysis, John Wiley, 1990.
[46] L. Hubert, P. Arabie, Comparing partitions, J. Class. 2 (1) (1985) 193–218.
[47] W. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.
[48] V. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Sov. Phys.-Dokl. 10 (1966) 707–710.
[49] V.M. Zolotarev, Modern Theory of Summation of Random Variables, Walter de Gruyter, 1997.
[50] S. Rachev, Probability Metrics and the Stability of Stochastic Models, 269, John Wiley & Son Ltd, 1991.
[51] S.H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, Int. J. Math. Models Methods. Appl. Sci. 1 (4) (2007) 300–307.
[52] C.S. Cai, J. Yang, S.W. Shulin, A Clustering Based Feature Selection Method Using Feature Information Distance for Text Data, Springer International Publishing, 2016.
[53] R.T. Ionescu, M. Popescu, PQ Kernel: a rank correlation kernel for visual word histograms, Pattern Recogn. Lett. 55 (2015) 51–57.
[54] A. Bolshoy, Z. Volkovich, V. Kirzhner, Z. Barzily, Genome Clustering: From Linguistic Models to Classification of Genetic texts, Springer Science & Business Media, 2010.
[55] M.G. Kendall, J.D. Gibbons, Rank Correlation Methods, Edward Arnold, 1990.
[56] M. Deza, E. Deza, Encyclopedia of Distances, Springer, 2009.
[57] G.N. Lance, W.T. Williams, Computer programs for hierarchical polythetic classification (ææsimilarity analysesg), Comput. J. 9 (1) (1966) 60–64.
[58] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification (2nd ed), Wiley, 2001.
[59] P. Berkhin, A survey of clustering data mining techniques, in: Grouping Multidimensional Data - Recent Advances in Clustering, Springer, 2006, pp. 25–71.
[60] A.D. Deshpande, A. Hansen, P.P. Preyas, NP-Hardness of euclidean sum-of-squares clustering, Mach. Learn. 75 (2) (2009) 245–248.
[61] F. Cao, J. Liang, G. Jiang, An initialization method for the *k*-means algorithm using neighborhood model, Comput. Math. Appl. 58 (3) (2009) 474–483.
[62] Z. Volkovich, J. Kogan, C.K. Nicholas, Sampling methods for building initial partitions, in: Grouping Multidimensional Data - Recent Advances in Clustering, Springer, 2006, pp. 161–185.
[63] I. Dhillon, Y. Guan, J. Kogan, Iterative clustering of high dimensional text data augmented by local search, 2002, pp. 131–138.
[64] J. Galsworthy, The Forsyte Saga, Oxford University Press; 1 edition, 2008.
[65] J.R.R. Tolkien, The Hobbit and the Lord of the Rings, Houghton Mifflin Harcourt, 2012.
[66] T. Chiu, D. Fang, J. Chen, Y. Wang, C. Jeris, A robust and scalable clustering algorithm for mixed type attributes in large database environment, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2001, pp. 263–268.
[67] G. Zebrowski, Arthur C. Clarke looks back on the lifetime of influences that led him to become a science-fiction Grand Master, Sci-Fi Wkly 10 (2008) 10–15.
[68] R. Gary, Éducation européenne, Gallimard, 1972.
[69] R. Gary, Le grand vestiare, Gallimard, 1985.
[70] R. Gary, Les racines du ciel, Gallimard, 1972.
[71] R. Gary, La promesse de l'aube, Gallimard, 1973.
[72] R. Gary, Chien blanc, Gallimard, 1972.
[73] R. Gary, Charge d'âme, Gallimard, 1997.
[74] R. Gary, Les clowns lyriques, Gallimard, 1989.
[75] E. Ajar, Gros-Câlin, Gallimard, 1977.
[76] E. Ajar, La vie Devant Soi, Gallimard, 2017.
[77] E. Ajar, Pseudo, Gallimard, 2004.
[78] E. Ajar, L'Angoisse du Roi Salomon, Gallimard, 1987.
[79] R. Gary, Les Têtes de Stéphanie, Gallimard, 2013.
[80] French stopwords, (https://www.ranks.nl/stopwords/french).
[81] R. Thompson, A note on restricted maximum likelihood estimation with an alternative outlier model, J. R. Stat. Soc. Ser. B: Methodol. 47 (1985) 53–55.
[82] V. Kulkarni, R. Al-Rfou, B. Perozzi, S. Skiena, Statistically significant detection of linguistic change, in: Proceedings of the 24th International Conference on World Wide Web, in: WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2015, pp. 625–635.

[83] Y. Xiong, D.-Y. Yeung, Time series clustering with ARMA mixtures, Pattern Recogn. 37 (8) (2004) 1675–1689.

[84] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[85] J. Calvo-Zaragoza, J. On, An efficient approach for interactive sequential pattern recognition, Pattern Recogn. 64 (Supplement C) (2017) 295–304.

[86] L.A. Leiva, E. Vidal, Warped *k*-means: an algorithm to cluster sequentially-distributed data, Inf. Sci. (Ny) 237 (2013) 196–210.

[87] K.H. Zou, J.R. Fielding, S.G. Silverman, C.M. Tempany, Hypothesis testing i: proportions, Radiology 226 (3) (2003) 609–613.

Dr. Konstantin Amelin is working as the PostDoc at the Software Engineering Department of Saint Petersburg State University, St. Petersburg, Russia. He was born in 1986 in St. Petersburg, Russia. Dr. Amelin received his Ph.D. degree in Software Engineering in 2012. His current research interests include Multi-Agent Technology, Embedded Systems and Data Mining. He published 3 books and about 10 papers in referred international journals.

Dr. Oleg Granichin is working as the Professor (Full) at the Software Engineering Department of Saint Petersburg State University, St. Petersburg, Russia. He was born in 1961 in St. Petersburg, Russia. Dr. Granichin received his Ph.D. degree in Mathematical Cybernetics in 1985 and Doctor Degree in System Analysis, Control and Data Processing in 2001. His current research interests include Dynamical Systems, Data Mining and Randomized Algorithms. He published 7 books and more than 60 papers in referred international journals.

Ms. Natalia Kizhaeva is Ph.D. student at the Software Engineering Department of Saint Petersburg State University, St. Petersburg, Russia. She was born in 1991 Stavropol, Russia. Her current research interests include Natural Language Processing, Data Mining and Clustering Algorithms.

Dr. Zeev Volkovich is working as the Professor (Full), the Head of the Software Engineering Department of ORT Braude College, Karmiel, Israel. He is also Affiliate Full Professor in Institute of Applied, Mathematics of Middle East Technical University, Ankara, Turkey and Affiliate Adjunct Full Professor in Department of Mathematics and Statistics of University of Maryland (UMBC), Baltimore, USA. Dr. Volkovich received his Ph.D. degree in Probability Theory in 1982. His current research interests include Data Mining, Pattern Recognition and Clustering Algorithms. He published 4 books and about 100 papers in referred international journals.