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/

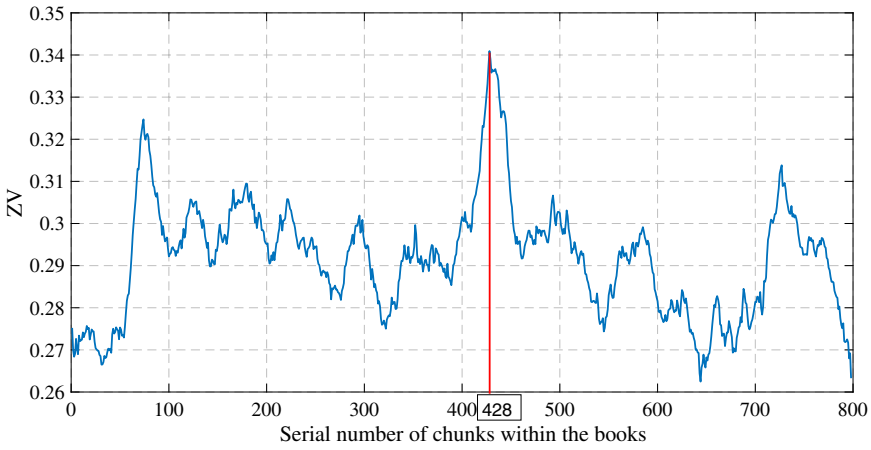

Fig. 1. Example of $ZV_{T,\mathrm{Dis},L}$ graph.

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