Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 53 – 62

Procedia
Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Text Classification Using a Novel Time Series Based Methodology

Zeev Volkovich[a] and Renata Avros[a]*

[a]*Ort Braude College of Engineering, Karmiel 21982, Israel*

Abstract

This paper discusses a novel time series methodology for writing process modeling, taking into account the dependency between sequentially written text parts. A series of consecutive sub-documents of a given document are represented via histograms of the appropriately chosen terms. To characterize the document overall style and its fluctuations, a new feature named the Mean Dependence is introduced. This similarity measure quantifies the association between a current sub-document and numerous earlier composed ones. So, such a collection of sub-documents is represented as a time series of the Mean Dependence development. The series change points naturally link to the style changes. Two possible approaches constructed within the general methodology are discussed. The first one intended to study media sources, is constructed to detect change points of media associated with social life transformations. Consequently, the homogeneous periods are detected using a new distance based on the Mean Dependence. The proposed methodology is applied to analysis of editorial texts published in the Egyptian "Al-Ahraam" and succeeds to indicate several important events connected to the "Arab Spring". The second approach, based on the strictly stationary model of time series, is applied to authorship verification. Numerical experiments demonstrate high ability of the proposed methods to recognize an authorship and to expose writing style evolution.

Keywords: Text Classification; Time Series Text Model; Authorship Verification

* Corresponding author. Tel.: +972-4-990-1994; fax +972-4-990-1852.
E-mail address: vlvolkov@braude.ac.il

doi:10.1016/j.procs.2016.08.095

1. Introduction

With the huge number of on-line documents abounding on the Internet, text classification has come to be one of the crucial methods for treatment and systematizing text data. Such tasks arise in the authorship recognition, automatic media content analysis, plagiarism detection and other areas. The main weakness of the methods used in these fields is that the association between texts is frequently evaluated without any regard to their developing process. Such as, for newspapers the similarity assessment of the issues is not accompanied by any connection to the early published ones. One of the common viewpoints of the human writing process views this process as composition of four key elements: planning, drafting, editing, and writing the final draft. Thus, it is natural to presume that dependency between sequential written text parts is remained at the almost uniform level if the text is composed by the same author, or in case of an official newspaper the social situation is relatively stable. This dynamical modeling of the writing process is the main advantage of the proposed method.

This paper discusses a novel time series methodology to the writing process modeling taking into account the dependency between sequentially written text parts. A series of consecutive sub-documents of a given document are represented via histograms of the appropriately chosen terms. To characterize the document overall style and its fluctuations, a new feature named the Mean Dependence is introduced. This similarity measure quantifies association between a current sub-document and numerous earlier composed ones. So, such a collection sub-documents is represented as a time series of the Mean Dependence development. The series change points naturally link to points of the style changes. Two possible approaches constructed within the general methodology are discussed. The first one intended to study of the media sources, is constructed to detected change points of media associated with the social life transformations. Consequently, the homogeneous periods are detected using a new distance based on the Mean Dependence. The proposed methodology is applied to analysis of editorial texts published in the Egyptian "Al-Ahraam" newspaper and successes to indicate several important events connected to the "Arab Spring". The second approach based on the strictly stationary model of time series is applied to the authorship verification. Numerical experiments demonstrate high ability of the proposed methods to recognize an authorship and to expose of a writing style evolution.

The rest of the paper is organized in the following way. A short review of related works is presented in Section 2. The proposed general methodology is stated in Section 3. Section 3 demonstrates an application of the suggested methodology to analysis of editorial texts published in the Egyptian "Al-Ahraam" newspaper for the periods containing some important events in the political life of the appropriate society, particularly, the "Arab Spring". A modification of the methodology based on a strictly stationary model of the Mean Dependency development is given in Section 4 together with an approach to the author verification problem and the appropriate numerical experiments.

2. Related Works

The field of authorship attribution originates from stylometry, analyzing texts for evidence of authenticity, authorial identity, and other questions. The task has a long history and an overview of different methods in chronological order is given in surveys [1,2]. One of the essential parts in quantitative authorship attribution algorithms is the distance measure quantifying the similarity between the texts. Burrow's Delta[3] is the most recognized measure of stylistic difference. Since its first appearance in 2002, various modifications have been proposed. The performance tests for Delta and its variants are provided in [4]. The compression-based measures, like application independent Normalized Compression Distance, are successfully applied to the text clustering. The comparison of numerous compression models for authorship attribution is performed in [5]. Despite their universality these measures are computationally expensive and therefore difficult to use in practice. To reduce the computational complexity in [6] a new model is developed.

Character *N*-grams are proved to be strong features for stylistic analysis [7]. This representation is tolerant to grammatical errors, computationally cheap and applicable to various languages as it allows avoiding hard preprocessing (e.g. tokenization for oriental languages). A significant point in this approach is the choice of *N*. A larger *n* is able to capture contextual information and topic of the text, but it leads to dimensionality growth. A smaller *N* can capture subword information but fails to consider context. To reflect syntactical information, which is naturally useful for style determination, syntactic *N*-grams are presented in [8].

The writing style of a document is the essential evidence once causing the problem of author verification, responding the problem whether given documents have been written by the same author [9]. As usual, this task is limited here by