Neurocomputing 136 (2014) 201–212

Contents lists available at ScienceDirect

Neurocomputing

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

Text style analysis using trace ratio criterion patch alignment embedding

Peng Tang, Mingbo Zhao*, Tommy W.S. Chow

Department of Electronic Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 25 June 2013
Received in revised form
2 January 2014
Accepted 6 January 2014
Communicated by X. Gao
Available online 29 January 2014

Keywords:
Text style analysis
Trace ratio criterion patch alignment embedding
Style markers
Text clustering

ABSTRACT

An effective algorithm for extracting cues of text styles is proposed in this paper. When processing document collections, the documents are first converted to a high dimensional data set with the assistant of a group of style markers. We also employ the Trace Ratio Criterion Patch Alignment Embedding (TR-PAE) to obtain lower dimensional representation in a textual space. The TR-PAE has some advantages that the inter-class separability and intra-class compactness are well characterized by the special designed intrinsic graph and penalty graph, which are based on discriminative patch alignment strategy. Another advantage is that the proposed method is based on trace ratio criterion, which directly represents the average between-class distance and average within-class distance in the low-dimensional space. To evaluate our proposed algorithm, three corpuses are designed and collected using existing popular corpuses and real-life data covering diverse topics and genres. Extensive simulations are conducted to illustrate the feasibility and effectiveness of our implementation. Our simulations demonstrate that the proposed method is able to extract the deeply hidden information of styles of given documents, and efficiently conduct reliable text analysis results on text styles can be provided.

1. Introduction

The tasks on text analyzing, such as text classification and document categorization, have gained a prominent status in the information systems field, due to the availability of documents created by the World Wide Web and the increasing demand to retrieval them by flexible means [1]. In these text analyzing tasks, papers are usually measured and classified according to their contents, or topics, and genres, or types. A third type of text analysis can exist. For example, different writers can write in different style when they describe the same thing, and experienced English readers usually have no difficulty in telling whether an essay was written by native English speakers or the non-native ones by looking at the authors' wordings and writing styles which differ from its genre or topic. Here, the cues to differentiate the native and English as a Secondary Language (ESL) speakers are writing text styles. Therefore, the style cues of documents can be an effective measure in automatic text processing tasks. In this research, we will analyze the text styles using style markers and machine learning approaches.

Document styles are obviously affected by topics and genres of documents. Consequently, approaches that handle topic-based or genre-based features in documents are of great help to our research on text analysis. The Term Frequency (TF), as well as Term Frequency-Inverse Document Frequency (TF-IDF) is applied to most document models like Vector Space Models or Probabilistic Models [2,3]. It can be noticed that simple textual features, like word (or term) frequency and length of sentences, are the most widely used. For instance, word length and sentence length features have been used to test the genre classes and authorship [4–6]. Tweedie has pointed out that the richness of vocabulary highly depends on text length and is very unstable [7]. Many works have been established with the assistance of POS tagging. For example, POS tagging features are used to detect text genre [8–12]. *N*-gram mixed with POS tagging are used to investigate the influence of syntax structure [12]. Feldman et al. extracted text genre features with POS histograms and machine learning technologies [13]. Biber [14] defined "style markers", regarded as a formal definition of style of texts, as a set of measurable patterns. Kessler identified four generic cues on the bases of style markers [8]. It is also believed that writers can be a determining factor for writing habits. There are also research work focusing on identifying the authorship of given documents. Style markers are utilized to dealing with unrestricted text for an authorship-based classification, and a 50% or above accuracy has been reported when a 10-author corpus are processed [15]. In [15], multiple regression and discriminant analysis are employed to analyse genres of documents. Similar approaches are also applied in web document classification [16–18]. In these analyses, the style

* Corresponding author. Tel.: +852 34427756; fax: +852 27887791.
E-mail addresses: ptang@ee.cityu.edu.hk (P. Tang), mzhao4@cityu.edu.hk (M. Zhao), eetchow@cityu.edu.hk (T.W.S. Chow).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2014.01.012

markers of text, together with HTML tags and entities are considered as textual features. By using regression and discriminate analysis, differences among given documents can be revealed. Using the occurrence frequency of the most widely used words from a training corpus as style markers has also been studied [19,20,15]. Textual features and self-organizing maps are used for text classification [1,21]. Proximity-based information between words to extract extra features of documents is also widely used for retrieving information. Petkova and Croft propose a document representation model based on the proximity between occurrences of entities and terms [22]. Lv and Zhai propagate the word count using the so-called Positional Language Model to obtain a virtual propagated word count and applied to other language models [23]. Different from the above method, our proposed method models a given text as a lexicon of weighted word pairs. In this paper, the weight of word pair, calculated by using proximity-based kernels in many applications, refers to the closeness between the two terms of the word pairs. Syntactic features performs better than simple textual statistics such as word frequency and length of sentences in genre classification [20,14]. It is reported, however, that the syntactic dependent features are computationally expensive and time-consuming [8]. To balance the computational performance and the effectiveness of analysis result, we use POS bigrams and trigrams, which can encode useful syntactic information [24].

Most above-mentioned approaches on document analysis contain two parts. First, they generate a matrix by using a set of style markers. Second, use regression, discriminate analysis, classifiers, or other machine learning methods to evaluate the results. Hence, if we want to improve the text analysis results, two approaches can be applied: (1) by exploring more effective features and (2) by utilizing more advanced machine learning methods that can better make use of hidden information in the original data. In this paper, we mainly focus the latter approach. Specifically, to better represent the features of text or documents, we first collected several corpuses using real-life textual data and existing popular corpuses for text analysis in different scenarios. In this way, document analysis is transformed into a feature extraction problem with a high-dimensional and non-Gaussian data set. We then develop an effective approach to handle such data set for document analysis.

However, dealing with high-dimensional data has always been a major problem for pattern recognition. Hence, dimensionality reduction techniques can be used to reduce the complexity of the original data and embed high-dimensional data into low-dimensional data, while keeping most of the desired intrinsic information [25,26]. Over the past decades, many dimensionality reduction methods have been proposed [27,28]. PCA pursues the direction of maximum variance for optimal reconstruction [29,30]. For linear supervised methods, LDA and its variants find the optimal solution that maximizes the distance between the means of the classes while minimizing the variance within each class [31]. Due to the utilization of label information, LDA can achieve better classification results than those obtained by PCA if sufficient labeled samples are provided.

To find the intrinsic manifold structure of the data, nonlinear dimensionality reduction methods such as ISOMAP [25], Locally Linear Embedding (LLE) [26], Laplacian Eigenmap (LE) [32] were developed. These methods preserve the local structures and look for a direct non-linearly embedding the data in a global coordinate. For example, ISOMAP aims to preserve global geodesic distances of all pairs of measurements; LLE uses linear coefficients, which reconstruct a given measurement by its neighbors, to represent the local geometry; LE is able to preserve the proximity relationships by using an undirected weighted graph to indicate neighbor relations of pair-wise measurements. But it is worth noting that all the above methods suffer from the out-of-sample problem [33]. To deal with the problem, He et al. [33] developed the Locality Preserving Projections (LPP) in which a linear projection matrix is used for mapping new-coming samples and extend LE to its linear version.

The aforementioned methods are developed based on the specific knowledge of field experts for their own purposes. Recently, Yan et al. [27] demonstrate that several dimension reduction methods (e.g. PCA, LDA, ISOMAP, LLE and LE) can be unified in a graph-embedding framework, in which the statistical and geometrical properties of the data are encoded as graph relationships. Zhang et al. [28] further reformulated several dimension reduction methods into a unified patch alignment framework (PAF), which consists of two parts: local patch construction and whole alignment, and showed that the above methods are different in the local patch construction stage and share an almost identical whole alignment stage. In addition, this general framework, which is also originally used by local tangent space alignment (LTSA) [34], has been widely used in different fields such as correspondence construction [35], image retrieval [36] and distance metric learning [37] by constructing different patches corresponding to different applications.

In general, most of the above methods are unsupervised and they do not use label information. However, label information is of great importance when handling classification problem. In addition, though LDA can achieve promising performance as a supervised method, it is developed based on the assumption that the samples in each class follow a Gaussian distribution. In many applications such as text classification problems, samples in a data set, however, may follow a non-Gaussian distribution that cannot satisfy the above assumption. Without this assumption, the separation of different classes may not be well characterized which results in degrading the classification performance [31]. To solve this problem, some supervised dimensionality reduction methods have adopted the idea from the mentioned unsupervised manifold methods for better preserving the discriminative information. These methods usually start from the local structure of data and preserve the geometric information provided by data points and the label information. Typical methods include Supervised Locality Preserving Projection (SLPP) [38], Discriminative Locality Alignment (DLA) [28], Stable Orthogonal Local Discriminant Embedding (SOLDE) [39], Sparse Neighbor Selection and Sparse-Representation-based Enhancement (SNS-SRE) [40], Unsupervised Transfer Learning based Target Detection (UTLD) [41], etc.

To better unveil the hidden information in the given high dimensional data created by a group of specified style markers, a fast trace ratio criterion patch alignment embedding (TR-PAE) method is introduced. Our proposed method has the advantages that the inter-class separability and intra-class compactness are well characterized by the special designed intrinsic graph and penalty graph, which are based on discriminative patch alignment method. This strategy is essential for extracting the deeply hidden information about text styles in document collections. Another advantage is that the proposed method is based on trace ratio criterion, which directly represents the average between-class distance and average within-class distance in the low-dimensional space. This advantage is helpful to directly obtain intuitive text analysis results. In this paper, we have performed extensive study using style marker collections, our collected corpuses and the proposed TR-PAE. The simulation results show that using our method, we can distinguish various styles of documents of different genres. Meanwhile, the styles of British and American writing English, as well as English from Asian areas, can be separated. Moreover, our proposed algorithm can separate news items collected from the same media and of the same genre, but composed in different decades.

The rest of this paper is organized as follows. The corpuses and textual features, i.e. the style marker collections, collected and used in our study are addressed in Section 2. In Section 3, we briefly overview the conventional linear discriminant analysis. Our proposed fast trace ratio criterion patch alignment embedding