Text style analysis using trace ratio criterion patch alignment embedding

ARTICLE *in* NEUROCOMPUTING · JULY 2014

Impact Factor: 2.08 · DOI: 10.1016/j.neucom.2014.01.012

Mingbo Zhao

City University of Hong Kong

Tommy W S Chow

City University of Hong Kong

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 method is subsequently elaborated. In Section 4, data sets used in visualization and classifications, experimental configurations and corresponding experimental results are detailed. We conclude our algorithm in Section 5.

2. Corpuses and features used

To our knowledge, there are few corpuses aiming for analyzing styles of text or documents. Most existing corpuses are for classification by the contents, topics, types, or genres of the documents. To conduct our text style analysis, we collected several corpuses using real-life textual data and existing popular corpuses for text analysis in different scenarios. The components of Corpus 1, Corpus 2 and Corpus 3 are detailed in Tables 1–3, respectively. The Corpus 1 consists of a group of materials covering various contents and genres, aiming to examine the performance of our proposed algorithm for analyzing the styles among different topics and genres. News and reportage items covering different English media, i.e. the British and American English media as the native English media, and the English media in the CJK (Chinese Japanese and Korean) area as the non-native English media, form Corpus 2. This corpus is designed for testing our algorithm when dealing with documents of the same genre but in composed in different regions. The Corpus 3 includes news and reports from the same English media and covering the same topics. We collect this corpus mainly for validating the feasibility of using our algorithm under such harsh conditions.

We choose some popular style markers of documents to construct a high dimensional data set [18,15,42]. Here, each style marker is considered as a textual feature concerning of token-level, lexical-level, structural-level. The 200 style markers used in our study are detailed in Table 4, which means we construct a 200-dimensional data set by using them. Token-level features include the classic term frequencies for the words, numbers, punctuations, and special symbols. Lexical-level features consist of both function words and content words, POS tagged tokens and some useful word using statistics that indicate the high and low frequency words in a corpus. Date, time, telephone, e-mail, abbreviation terms are also extracted by using regular expression. Structural-level features are some combination of POS taggers, word collocation information among words and some other textual statistics. In our study, we obtain the listed style markers with the assistant of python scripts and the python package NLTK.

Compared to features used in [18,15,42], we removed features related to topics and genres so that the bias caused by topics and genres can be minimized. Moreover, features concerning of frequent/function words are enhanced. Such methods also aim to eliminate the bias caused by genres and topics of different documents.

3. Feature extraction based on fast trace ratio criterion linear discriminant analysis

3.1. Related work

3.1.1. Review of trace ratio linear discriminant analysis

LDA uses the within-class scatter matrix S_w to evaluate the compactness within each class and between-class scatter matrix S_b to evaluate the separability of different classes. The goal of LDA is to find a linear transformation matrix $W \in R^{D\times d}$, for which the trace of between-class scatter matrix is maximized, while the trace of within-class scatter matrix is minimized. Let $X=\{x_1,x_2,\dots x_l\} \in R^{D\times l}$ be the training set, each x_i belongs to a class $c_i=\{1,2,\dots c\}$. Let l_i be the number of data points in the ith class and l be the number of data points in all classes. Then, the between-class scatter matrix S_b, within-class scatter matrix S_w, and total-class scatter matrix S_t are defined as follows:

$$S_t=\sum_{i=1}^{c}\sum_{x\in c_i}(x-\mu)(x-\mu)^T$$

Table 3
The Components of Corpus 3.

Component contents	Count
New York Times News and Reports of 1980s	2000
New York Times News and Reports of 1990s	2000
New York Times News and Reports of 2000s	2000

Table 1
The Components of Corpus 1.

Component contents	Count
Novels (Pride and Prejudice, Gone with the Wind, Wuthering Heights, The Call of the Wild)	4
English diaries by Chinese Learners from a Chinese website for Chinese users	200
Newsgroup topics from *20 Newsgroups*	400
Abstracts of undergraduate students Final year project (FYP) reports of an Asian University	400
Native English news and reports available on CNN.com and Guardian.co.uk	500
CJK (Chinese, Japanese and Korean) English news and reports available from Chinese, Hong Kong and Japanese Media	500
Popular science readings (The Greatest Show on Earth, A Briefer History of Time, Genome)	3

Table 2
The Components of Corpus 2.

Component contents	Count
New York Times News and Reports	2000
The Wall Street Journal News and Reports	2000
CNN News and Reports	2000
Guardian News and Reports	2000
The Times News and Reports	2000
Telegraph News and Reports	2000
CJK English news and reports available from Chinese, Hong Kong and Japanese Media	2000

Table 4
The selected style markers in our study.

Feature used	Description
Token-level	
A1	Number of words
A2	Number of distinct words
A3	Number of punctuation
A4	Average sentence length
A5	Average number of words per sentence
A6	Average word length
A7	Number of candidate sentences/number of characters
A8	Number of digit
A9	Number of distinct stemmed word/number of distinct original word
A10	Average frequency rank of sentences
Lexical-level	
B1–B100	Frequency of CONTENT words/total frequency of content words; for 100 most frequently used content words
C1–C20	most frequently used function words/total frequency of function words; for 20 most frequently used function words
D1–D20	Frequency of PUNCTUATION/total frequency of punctuation for 20 most frequently used punctuation marks
E1	Number of usual words/total number of words (frequency of usual word 1000 in the training corpus)
E2	Number of unusual words/total number of words (frequency of unusual words = 1 in the training corpus)
E3	Unique number of words/total number of words (Vocabulary richness)
E4	Number of POS words/total number of words for 9 POS: noun, pronoun, adjective, verb, adverb, interjection, modifier, postposition, verbal-ending
F1–F5	Number of chunks for 5 expressions: date, time, telephone, e-mail, abbreviation
Structural-level	
G1	Overall intimacy of word pairs
H1–H20	Number of phrase/total number of phrases in a document for 17 phrases: NP, VP, AJP, AUXP, AVP, CONJP, SENT, IMPR, etc.
I1–I20	Average number of words per phrase for 17 phrases: NP, VP, AJP, AUXP, AVP, CONJP, SENT, IMPR, etc.

$$S_w = \sum_{i=1}^{c} \sum_{x \in c_i} (x-\mu_i)(x-\mu_i)^T$$

$$S_b = \sum_{i=1}^{c} l_i(\mu_i-\mu)(\mu_i-\mu)^T \tag{1}$$

where $\mu_i = 1/l_i \sum_{x_i \in c_i} x_i$ is the mean of the data points in the ith class, and $\mu = 1/l_i \sum_{x_j \in c_i} x_j$ is the mean of the data points in all classes. The original formulation of LDA, called Fisher LDA [31], can only deal with binary classification. Two optimization criteria can be used to extend Fisher LDA to solve the multi-class classification problem. The first one is in the ratio trace form (we refer it as LDA):

$$W^* = \arg\max_{W} \mathrm{Tr}\{(W^T S_w W)^{-1} W^T S_b W\} \tag{2}$$

and the second one is in the trace ratio form (we refer it as TR-LDA):

$$W^* = \arg\max_{W^T W = I} \mathrm{Tr}(W^T S_b W)/\mathrm{Tr}(W^T S_w W) \tag{3}$$

The optimal solution of LDA can be formed by the top eigenvectors of $S_w^{-1} S_b$. On the other hand, the optimization problem of TR-LDA in Eq. (3) has no close-form solution and has to calculate it by an Iterative Trace Ratio method (ITR) [43]. Specifically, if W_t denotes the solution at the tth iteration, then at the $(t+1)$th solution, W_{t+1} can be formed by the top eigenvectors of $S_b - \lambda_t S_w$, where $\lambda_t = \mathrm{Tr}(W_t^T S_b W_t)/\mathrm{Tr}(W_t^T S_w W_t)$. This procedure can be proved to converge to the globally optimal solution given any initialization W_0 [43].

The two objective functions have advantages and disadvantages. The ratio trace form is computationally more efficient than the trace ratio from. On the other hand, the physical meaning of the trace ratio form is clearer than that of the ratio trace form because the numerator and denominator of the objective function in the trace ratio form represent the average between-class distance and average within-class distance in the low-dimensional space, respectively. A more detailed comparison between the two objective function can be seen in [44,45].

Recently, several works are proposed to accelerate to convergence of ITR algorithm [46–48]. We in the previous work [49] have proposed a more efficient algorithm, called improved ITR algorithm (iITR), to solve this problem. The iITR algorithm has transformed the trace ratio problem into a linear fractional programming (LFP) problem, which can be generally solved by Dinkelbachs algorithm. Specifically, let $b = \{b_1, b_2, \ldots, b_D\} \in [0,1]$ be the selected vector, i.e. if the ith eigenvector is selected, then $b_i = 1$; otherwise $b_i = 0$. $f = \{f_1, f_2, \ldots, f_D\} \in R^{1\times D}$, $g = \{g_1, g_2, \ldots, g_D\} \in R^{1\times D}$ with each element satisfying $f_i = w_i^T S_b w_i$ and $g_i = w_i^T S_w w_i$, the basic steps of the iITR algorithm for solving trace ratio problem are listed in Table 5.

3.1.2. *Review of patch alignment for dimensionality reduction*

The patch alignment framework, which is originally used by local tangent space alignment (LTSA) [34], consists of two parts [28]: local patch construction and whole alignment. In the local patch construction stage, different methods have different optimization criteria. Each patch is constructed by one sample and its related samples according to both the characteristics of the data set and the objective of the methods. In the whole alignment stage, all part optimizations are integrated together to form a consistent global coordinate for all independent patches.

Specifically, given a certain sample x_j from $X = [x_1, \ldots, x_N] \in R^{D\times N}$, its local patch is formed by combining x_j and its K nearest neighbor samples as $X_j = [x_j, x_{j_1}, \ldots, x_{j_K}] \in R^{D\times(K+1)}$. For each X_j, its corresponding low-dimensional representation is denoted as $Y_j = [y_j, y_{j_1}, \ldots, y_{j_K}] \in R^{d\times(K+1)}$. The optimization for the local patch is defined by

$$\arg\min_{F_j} \mathrm{Tr}(F_j L_j F_j^T). \tag{4}$$

where $L_j \in R^{(K+1)\times(K+1)}$ encodes the objective function that describes the local geometry for the ith patch. Here L_j varies with different DR methods and the detail information about how to choose L_j for different methods can be seen in [28].

Second, in the whole alignment stage, all achieved local patches will be aligned together to form a global optimization. For each patch X_j, there is a low-dimensional representation Y_j. All Y_j can be unified together as a whole one by assuming that the coordinate for the ith patch $Y_j = [y_j, y_{j_1}, \ldots, y_{j_K}]$ is selected from the

Table 5
iITR algorithm for solving the trace ratio problem.

```
1. Initialize $\lambda_0 = 0$.
2. Compute the eigen-decomposition of $S_b - \lambda_t S_w$ as $(S_b - \lambda_t S_w) w_i = \tau_i w_i$, where $w_i (i = 1, 2, \ldots D)$ is the eigenvector of $S_b - \lambda_t S_w$.
3. Calculate $f_i = w_i^T S_b w_i$ and $g_i = w_i^T S_w w_i$ for $i \in \{1, 2, \ldots, D\}$ and initialize $\gamma_0 = \lambda_t$ and $b^0 = [b_1^0, b_2^0, \ldots b_D^0]$ be a zero vector, iteratively solving the following sub-problem until convergence:
   • Sort $f_i - \gamma_k g_i$ and set $b_i^k = 1$ corresponding to the $d$ largest value of $f_i - \gamma_k g_i$, $b_i^k = 0$ otherwise.
   • Update $\gamma_{i+1} = b^k f^T / b^k g^T$.
   • If $b^k = b^{k-1}$, output $b^* = b^k$ and $\gamma^* = b^* f^T / b^* g^T$.
4. Form $W_t$ by choosing the $d$ eigenvectors of $w_i$ with $b_i^* = 1$ and Update $\lambda_{t+1} = \gamma^*$.
5. Iterate the steps (2–4) until $|\lambda_{t+1} - \lambda_t| < \varepsilon$. Output $W^*$.
```

global coordinate $Y = [y_1, \ldots, y_N]$ such that

$$Y_j = Y S_j. \tag{5}$$

where $S_j \in R^{N \times (K+1)}$ is the selection matrix defined as $(S_j)_{mn} = 1$ if $x_m \in X_j$ and x_m is the nth element in the patch of X_j; $(S_j)_{mn} = 0$, otherwise. Then Eq. (4) can be rewritten as

$$\arg\min_{Y_j} \operatorname{Tr}(Y S_j L S_j^T Y^T). \tag{6}$$

By summing over all part optimization in Eq. (6), we can obtain the whole alignment as

$$\begin{aligned} \arg\min_{Y} \sum_{j=1}^{N} \operatorname{Tr}(Y S_j L S_j^T Y^T) &= \arg\min_{Y} \operatorname{Tr}\left(Y \sum_{j=1}^{N} (S_j L S_j^T) Y^T \right) \\ &= \arg\min_{Y} \operatorname{Tr}(Y L Y^T), \end{aligned} \tag{7}$$

where $L = \sum_{j=1}^{N} S_j L_j S_j^T \in R^{N \times N}$ is the alignment matrix [28]. For the linearization, we assume that there is a linear projection matrix W satisfying $Y = W^T X$ and add a constraint $W^T W = I$ to make the projection matrix W unique, then Eq. (7) becomes

$$\arg\min_{W^T W = I} \operatorname{Tr}(W^T X L X^T W). \tag{8}$$

The problem in Eq. (8) can be solved by performing the Eigenvalue Decomposition of XLX^T and the optimal solution of Eq. (8) is formed by the eigenvectors associated with d smallest eigenvalues.

The patch alignment framework has shown that many DR methods are different in the local patch construction stage and share an almost identical whole alignment stage, which also motivates us to construct special local patch for the application of text style analysis.

3.2. The proposed algorithm

The above algorithm in Table 5 is developed with the assumption that the samples in each class follow a Gaussian distribution. However, in many text classification problems, samples in a data set may follow a non-Gaussian distribution that cannot satisfy the above assumption [27,28,50,51]. Without this assumption, the separation of different classes may not be well characterized by the scatter matrices, causing the classification results to be degraded [31]. To solve this problem, some researchers have developed new scatter matrices of S_w and S_b to characterize the intra-class compactness and inter-class separability [27,28]. In this paper, inspired by the work in [28], S_b and S_w are constructed using the specially designed intrinsic graph $W^s \in R^{l \times l}$, and penalty graph, $W^p \in R^{l \times l}$ (see Fig. 1 for details). Specifically, for a given sample x_i, we denote its k_1 nearest neighbor samples with the same class label as $X_j^s = \{x_{j_1}^s, x_{j_1}^s, \ldots, x_{j_{k_1}}^s\} \in R^{D \times k_1}$, $(x_{j_1}^s = x_j)$ and its k_2 nearest neighbor samples with different class labels as $X_j^p = \{x_{j_1}^p, x_{j_1}^p, \ldots, x_{j_{k_2}}^p\} \in R^{D \times k_2}$. The intrinsic graph W^s and penalty graph W^p can be defined by $w_{ji}^s = 1/k_1$, if $x_i \in N^s(x_j)$; $w_{ji}^s = 0$, otherwise, $w_{ii}^p = 1/k_2$, if $x_i \in N^p(x_j)$; $w_{ji}^p = 0$, otherwise. We also denote the low-dimensional representation of x_i, X_j^s and X_j^p as $f_j \in R^d$, $F_j^s = \{f_{j_1}^s, f_{j_1}^s, \ldots, f_{j_{k_1}}^s\} \in R^{d \times k_1}$ and $F_j^p = \{f_{j_1}^p, f_{j_1}^p, \ldots, f_{j_{k_2}}^p\} \in R^{d \times k_2}$. In such low-dimensional space, we hope that the distances between f_j and its k_1 nearest neighbor samples with the same class label are as small as possible, while the distances between f_j and its k_2 nearest neighbor samples with different class labels are as large as possible. Hence it is reasonable to maximize the following objective function, which is based on a trace ratio criterion:

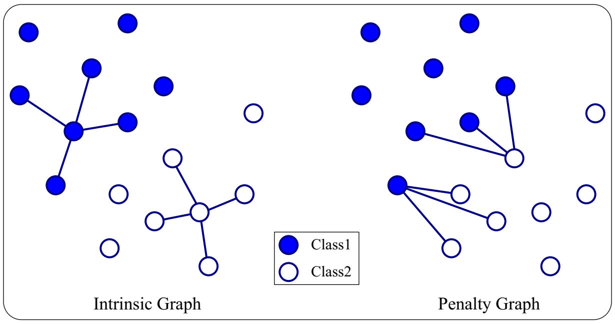

Fig. 1. Intrinsic graph and penalty graph: in the intrinsic graph, each sample is connected to its k_1 nearest neighbors within the same class, while in the penalty graph each sample is connected to its k_2 nearest neighbors in different classes.

$$\max \frac{\sum_{j=1}^{l} \sum_{m=1}^{k_2} \| f_j - f_{j_m}^p \|^2 w_{jm}^p}{\sum_{j=1}^{l} \sum_{m=1}^{k_1} \| f_j - f_{j_m}^s \|^2 w_{jm}^s} = \max \frac{\sum_{j=1}^{l} (1/k_2) \sum_{m=1}^{k_2} \| f_j - f_{j_m}^p \|^2}{\sum_{j=1}^{l} (1/k_1) \sum_{m=1}^{k_1} \| f_j - f_{j_m}^s \|^2} \tag{9}$$

Here, if we further denote $F_j^p = F S_j^p$, $F_j^s = F S_j^s$, where $F \in R^{d \times l}$ is the low-dimensional representation of the data matrix X , $S_j^p \in R^{l \times k_2}$, $S_j^s \in R^{l \times k_1}$ are the patch selection matrix of x_j satisfying $(S_j^p)_{mn} = 1$, if $x_m \in X_j^p$ and x_m is the n th element in the patch of X_j^p; $\left(S_j^p\right)_{mn} = 0$, otherwise, $x_m \in X_j^s$, if $x_m \in X_j^s$ and x_m is the n th element in the patch of X_j^s; $(S_j^s)_{mn} = 0$, otherwise. Then, we can rewrite Eq. (9) in a matrix form as

$$\begin{aligned} &\max \frac{\sum_{j=1}^{l} \operatorname{Tr}\left(\frac{1}{k_2} F S_j^p \begin{bmatrix} -e_{k_2} \\ I_{k_2} \end{bmatrix} [-e_{k_2}^T \quad I_{k_2}] S_j^p T F^T \right)}{\sum_{j=1}^{l} \operatorname{Tr}\left(\frac{1}{k_1} F S_j^s \begin{bmatrix} -e_{k_1} \\ I_{k_1} \end{bmatrix} [-e_{k_1}^T \quad I_{k_1}] S_j^s T F^T \right)} \\ &= \max \frac{\operatorname{Tr}\left(F \sum_{j=1}^{l} \left(\frac{1}{k_2} S_j^p \begin{bmatrix} -e_{k_2} \\ I_{k_2} \end{bmatrix} [-e_{k_2}^T \quad I_{k_2}] S_j^p T \right) F^T \right)}{\operatorname{Tr}\left(F \sum_{j=1}^{l} \left(\frac{1}{k_1} S_j^s \begin{bmatrix} -e_{k_1} \\ I_{k_1} \end{bmatrix} [-e_{k_1}^T \quad I_{k_1}] S_j^s T \right) F^T \right)} \\ &= \max \frac{\operatorname{Tr}\left(F \sum_{j=1}^{l} S_j^p L_j^p S_j^p T F^T \right)}{\operatorname{Tr}\left(F \sum_{j=1}^{l} S_j^s L_j^s S_j^s T F^T \right)} = \max \frac{\operatorname{Tr}(F L^p F^T)}{\operatorname{Tr}(F L^s F^T)} \end{aligned} \tag{10}$$

where $e_s \in R^{1\times s}$ is a unit vector with size s, $I_s \in R^{s\times s}$ is an identity matrix and L_j^p, L_j^s, L^p, L^s satisfy:

$$L_j^p = \frac{1}{k_2}\begin{bmatrix} -e_{k_2} \\ I_{k_2} \end{bmatrix}[-e_{k_2}^T \quad I_{k_2}] \quad L^p = \sum_{j=1}^{l} S_j^p L_j^p S_j^p T$$
$$L_j^s = \frac{1}{k_1}\begin{bmatrix} -e_{k_1} \\ I_{k_1} \end{bmatrix}[-e_{k_1}^T \quad I_{k_1}] \quad L^s = \sum_{j=1}^{l} S_j^s L_j^s S_j^s T \tag{11}$$

In order to calculate the projection matrix, we constrain that the low-dimensional representation F lies in the linear subspace formed by the data matrix X, i.e. $F = W^T X$, then by replacing F into Eq. (10), we can reformulate the objective function of Eq. (10) as

$$\max \frac{\mathrm{Tr}(W^T X L^p X^T W)}{\mathrm{Tr}(W^T X L^s X^T W)} = \max \frac{\mathrm{Tr}(W^T \overline{S_b} W)}{\mathrm{Tr}(W^T \overline{S_w} W)} \tag{12}$$

where

$$\overline{S_b} = XL^pX^T, \overline{S_w} = XL^sX^T. \tag{13}$$

The above strategy for constructing the scatter matrices is able to avoid the drawback that samples in each class must follow a Gaussian distribution (such as in LDA and TR-LDA), as the constructed scatter matrices are to preserve the local discriminative structure rather than the global discriminative structure, which can better handle the data sampled from a non-Gaussian distribution. In this paper, we calculate the scatter matrices in the proposed TR-LDA method using the above strategy. Hence, by simply performing notation substitutions in Table 5, i.e. $S_b \to \overline{S_b}$ and $S_w \to \overline{S_w}$, we present our algorithm as in Table 6.

4. Experimental results and analysis

4.1. High-dimensional data set preparation

The corpuses are required to be transformed into high dimensional data sets for discriminate analysis. First, documents in each corpus are cut into slices. Each slice contains n sentences. In this study, the value of n is set to 100 [42]. When there are less than n sentences in a document, this document will be merged to other documents then to be cut. Therefore, a full-text novel can generate many slices. Such a procedure ensures enough text style information contained in slices. The bias caused by different length of text can thus be avoided. After the documents are cut into slices, each slice are transformed into a 200 dimensional vector by utilizing the style markers described in Section 4.1. Consequently, the data are visualized in a text space using different machine learning approaches.

4.2. Experimental configurations

In this section, the proposed TR-PAE algorithm is examined by visualizing the corpuses in low dimensional textual spaces. Here, six approaches: PCA [30], LDA, SLPP [38], TR-LDA, DLA [28], and our proposed TR-PAE are compared in visualization performance. In addition, to quantitatively evaluate the results, the following evaluation metric is introduced and applied.

Table 6
The proposed algorithm.

```
Input: Data matrix $X$, parameter $k_1$ and $k_2$
Output: Optimal projection matrix $W^*$
Algorithms:
(1) Form the scatter matrix $\overline{S_b}$ and $\overline{S_w}$.
(2) Calculate the projection matrix $W$ using Table 1 with $\overline{S_b}$ and $\overline{S_w}$.
(3) Output the optimal projection matrix $W^*$.
```

Clustering performance is evaluated by comparing the obtained cluster label of each data item with that provided by the data corpus. The accuracy rates and normalized mutual information (MI) metric [52–55] are used to evaluate the clustering results. Given a point x_i, let r_i and f_i be the obtained cluster label and the provided class label. The clustering accuracy (AC) is defined as

$$AC = \frac{\sum_{i=1}^{N} \delta(f_i,\ \mathrm{Map}(r_i))}{N} \tag{14}$$

where N is the total amount of data, $\delta(a,b)$ is the delta function which $\delta(a,b)=1$ if $a=b$ and $\delta(a,b)=0$ otherwise, and $\mathrm{Map}(r_i)$ is the permutation mapping function, mapping each r_i to the equivalent label in the data set. Let C denote the set of clusters obtained from the ground truth and C' denote the set of clusters obtained from our method. Their MI metric $MI(C,C')$ is defined by

$$MI(C,C') = \sum_{c_i \in C, c_j' \in C'} \Pr(c_i, c_j') \log \frac{\Pr(c_i, c_j')}{\Pr(c_i)\Pr(c_j')} \tag{15}$$

where $\Pr(x)$ is the probability that a point randomly selected from the data set belongs to the cluster X, $\Pr(x,y)$ is the joint probability of x and y. To simplify comparisons among different clusters, we employ normalized MI defined by

$$MI(\overline{C}, C') = \frac{MI(C,C')}{\max(H(C), H(C'))}. \tag{16}$$

Here, $H(X)$ is the entropy of X. It is easy to be observed that $MI(\overline{C}, C')$ ranges from 0 to 1, i.e. $MI(\overline{C}, C') = 0$ if two sets are totally independent and 1 if the two sets are identical.

4.3. Experimental results

4.3.1. Visualizing documents of different genres

It is a fact that styles can vary little between two different genres, while completely different styles can exist in the same genre. In some cases, it is difficult to recognize which style a given document should belong to. Therefore, the styles should vary continuously, not discretely. From this angle, we can expect that the visualization results should contain many overlapped elements. Nevertheless, in most cases, the text styles change with genres of documents.

We firstly show the 2-D embedding of the data set generated from Corpus 1 using style markers in Fig. 2. In our simulations, we apply all points to construct both the supervised and unsupervised parts. Observing the results, we find the following.

(1) From the visualization results we can intuitively observe that TR-PAE, TR-LDA and DLA outperform SLPP, LDA and PCA. PCA, the unsupervised algorithm, cannot separate each component at all.

(2) Our proposed TR-PAE (Fig. 2(a)), SLPP (Fig. 2(c)) and LDA (Fig. 2(e)) can correctly scatter CJK English Media and CJK FYP report together, because the two components are composed in formal English by Asian and expected in similar styles. Also, the three methods provide a clear separation between formal (reports, academic writings) and informal writing styles (newsgroup topics, novels, etc.).

(3) Compared to TR-LDA and DLA, our TR-PAE as shown in Fig. 2 (a) can distribute Native and CJK English Media in different areas with small overlapped parts, while TR-LDA and DLA mass Native and CJK English Media together. We here should notice that the two components Native English Media and CJK English Media, though they have differences, cannot be separated completely, i.e. they overlap partially, because nowadays considerable items of non-native English media are of professional quality, the difference between native and non-native English media is getting smaller. From this point of view, we can say that the TR-PAE algorithm is able to reveal the intrinsic structure of the original corpus more correctly than TR-LDA and DLA.

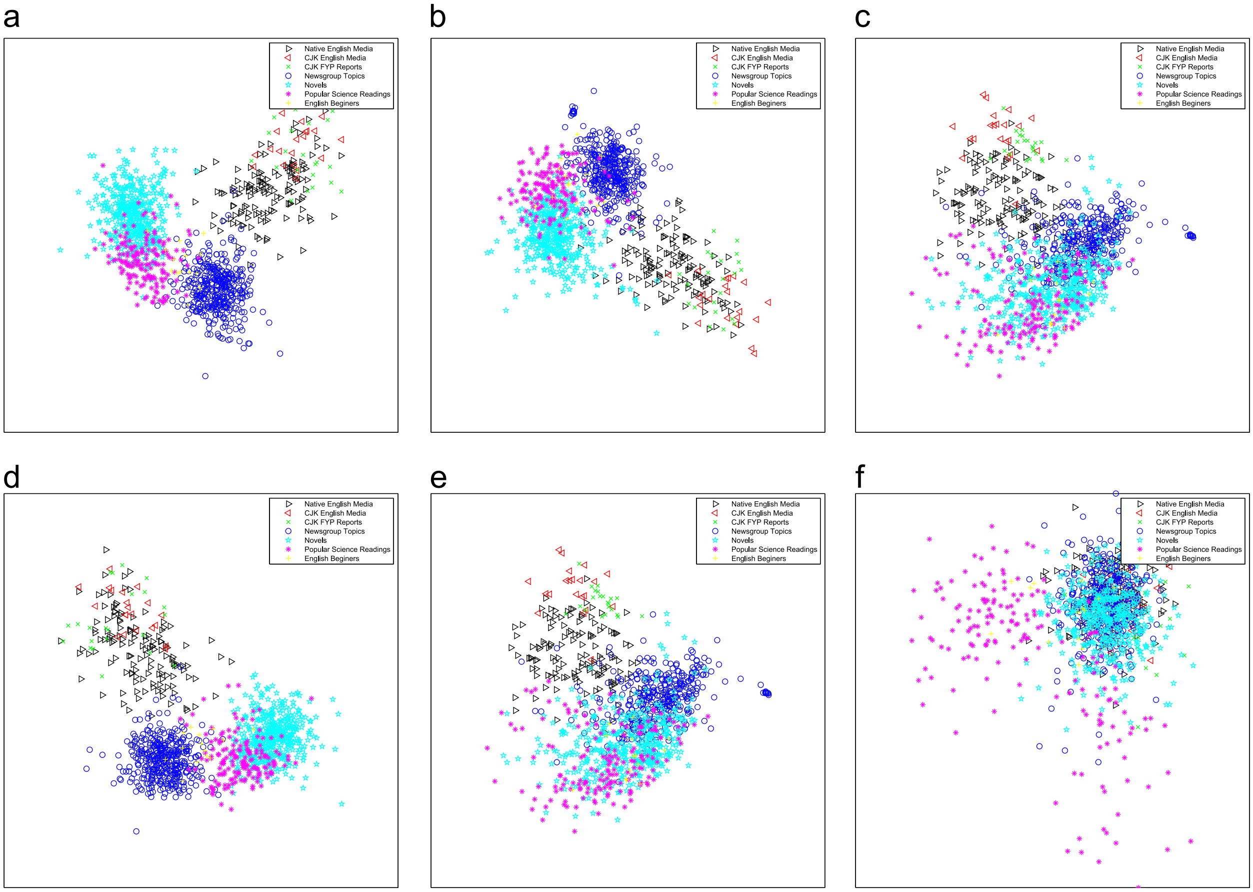

Fig. 2. Visualization of Corpus 1. (a) TR-PAE, (b) TR-LDA, (c) SLPP, (d) DLA, (e) LDA and (f) PCA.

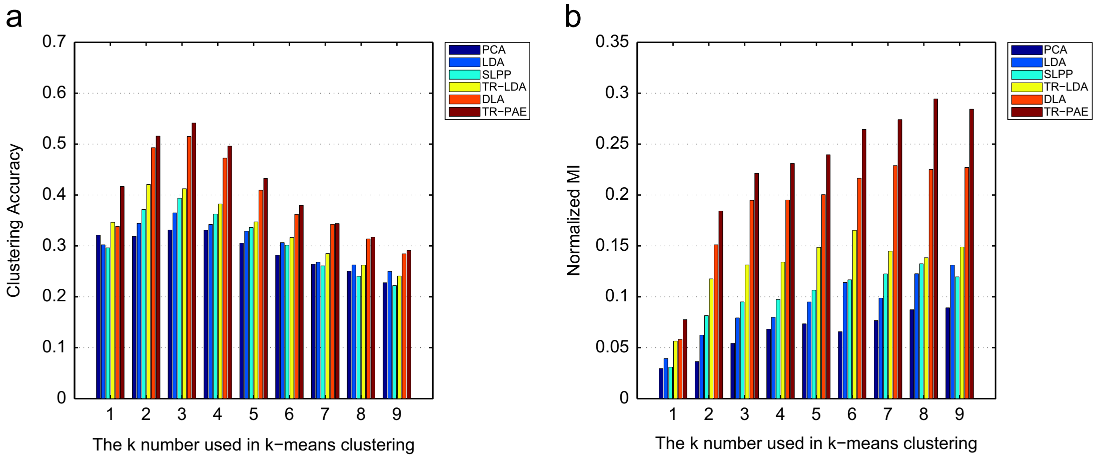

Fig. 3. Clustering evaluation result of each method on Corpus 1. (a) Clustering accuracy and (b) normalized mutual information.

We also employ the clustering evaluation measure to have quantitative comparison among the clustering performances of each approach. In our simulations, the clustering measurement process induced by the *k*-means algorithm is conducted as follows. First, the original data are embedded into a low-dimensional output space; then, *k*-means is performed. For each *k* value used in the k-means algorithm, the *k*-means clustering is applied 200 times with different initial parameters. The averaged clustering accuracy and normalized MI over different *k* values are recorded in Fig. 3. We have the observation that the clustering performance is greatly improved with our proposed TR-PAE compared with the other methods.

4.3.2. *Visualizing documents of the same genre*

It is usually difficult to separate different styles probably existing in the same genre, because the styles in documents of the same genre are generally very similar. Therefore, analyzing text styles in the same genre is challenging. In this part, two simulations on styles varying with different regions and different time are performed to validate our proposed algorithm.

4.3.2.1. Visualizing American, British and Asian English. Five components of Corpus 2, i.e. two popular American English media, two popular British English media, and Asian English media, are selected to perform the simulation. Here, we only choose the news and reportage items on the same topics, e.g. international news, policy and sports, from popular English media to avoid the style drifts caused by different sources and topics. We display the 2-D representation of test data set in Fig. 4 from which the following observations can be found:

(1) In contrast to the visualization results of documents of different genres, the components of Corpus 2 are largely overlapped with each other, which hints that the styles of each components are more similar than the components in Corpus 1.

(2) Different components are virtually overlapped by using TR-LDA (Fig. 4(b)), SLPP (Fig. 4(c)), LDA (Fig. 4(e)) and PCA (Fig. 4(f)), which indicate that these methods are unable to separate the five components.

(3) Our proposed TR-PAE as shown in Fig. 4(a) and DLA (Fig. 4(b)) perform better than the other four methods. In the visualization results of TR-PAE and DLA, the Asian English is separated clearly, while the American and British media are largely overlapped. This indicates that, compared to British and American English, Asian English has significant difference from British and American English, i.e. native English.

(4) It is interesting that in the visualization result of TR-PAE which is displayed in Fig. 4(a), the two American English media are overlapped. Meanwhile, the two British English media are

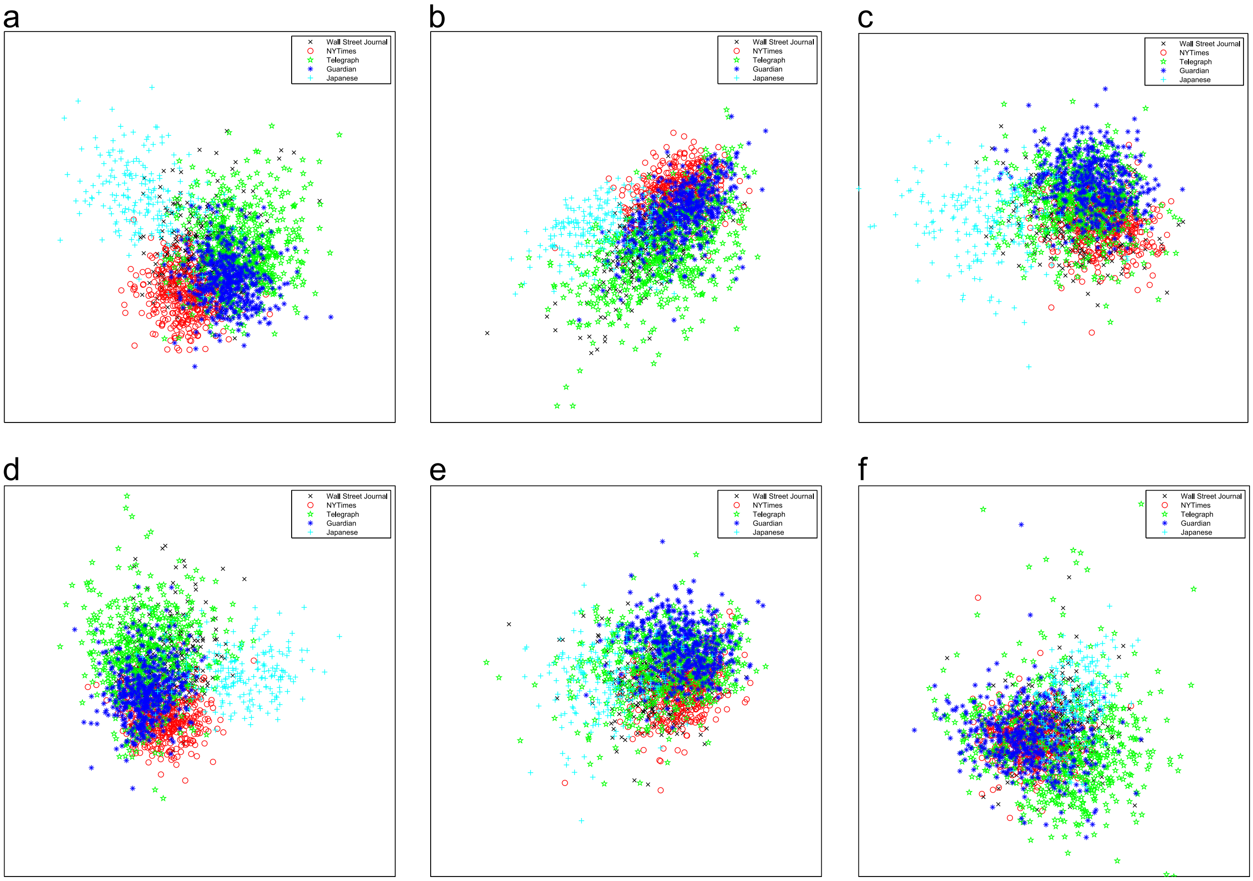

Fig. 4. Visualization of American, British and Asian English. (a) TR-PAE, (b) TR-LDA, (c) SLPP, (d) DLA, (e) LDA and (f) PCA.

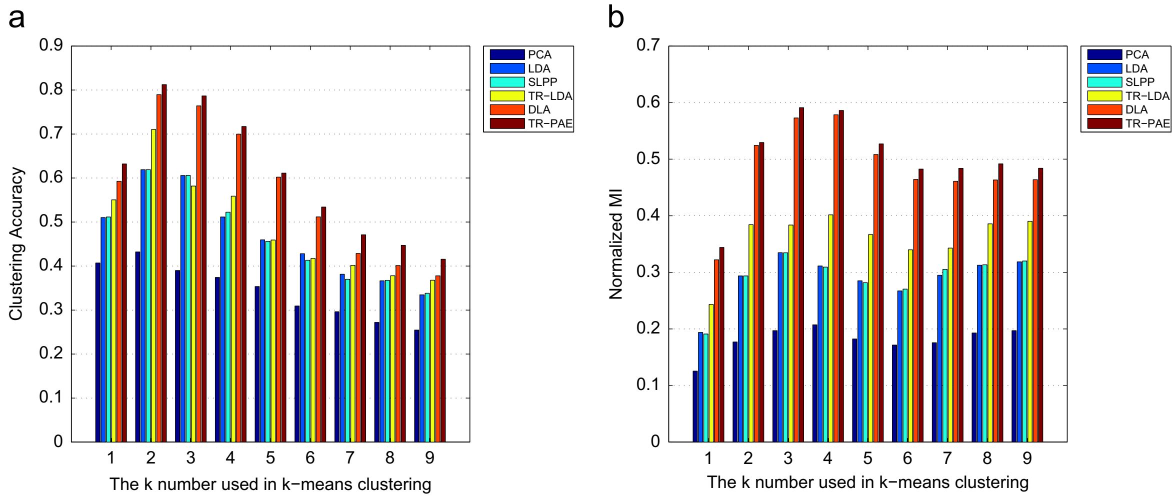

Fig. 5. Clustering evaluation result of each method on American, British and Asian English. (a) Clustering accuracy and (b) normalized mutual information.

overlapped. In contrast to TR-PAE, the American and British media are largely overlapped when using other approaches.

The numerical analysis result of different methods using AC and NMI illustrated in Fig. 5 also shows that our proposed TR-PAE outperforms the other approaches.

4.3.2.2. Visualizing news items of different decades. We have studied the validation of the performance of our proposed method in analyzing the text styles of different genres, and of the same genre but in different regions. In this part, another experiment is designed to explore the feasibility of applying the proposed TR-PAE to analyze the style changes of documents of different decades. We utilize three components of Corpus 3, i.e. New York Times News and Reports of 1980s, 1990s and 2000s, to conduct our simulation. The visualization results are shown in Fig. 6. From Fig. 6, we can observe the following list:

(1) We can intuitively observe that only the proposed TR-PAE as shown in Fig. 6(a) can separately visualize the three different components from each other. The components are totally overlapped when other five methods are employed.

(2) Note that the styles of the three components vary continually. Therefore, the three components cannot be told apart

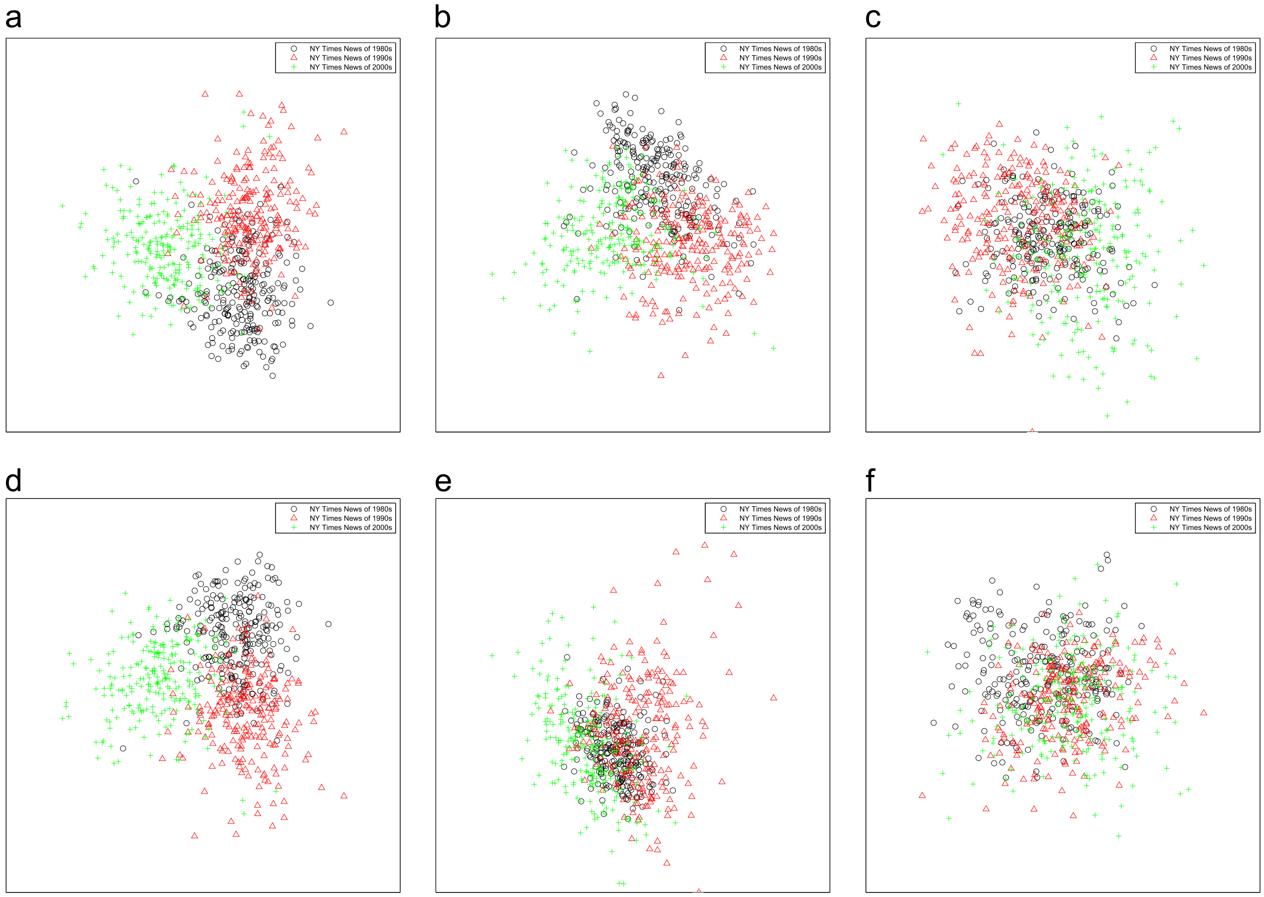

Fig. 6. Visualization of news of NY Times. (a) TR-PAE, (b) TR-LDA, (c) SLPP, (d) DLA, (e) LDA and (f) PCA.

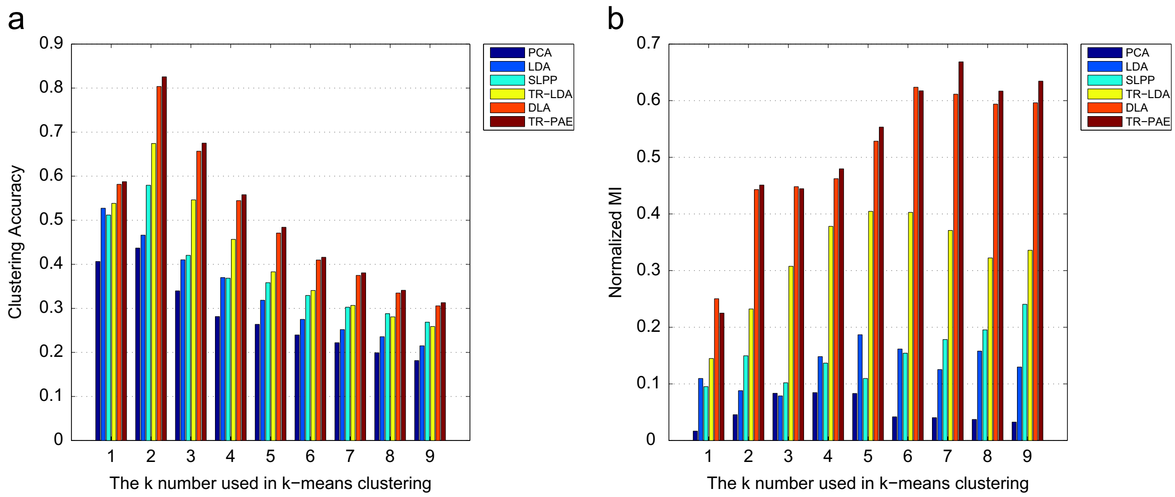

Fig. 7. Clustering evaluation result of each method on news and reports of different decades. (a) Clustering accuracy and (b) normalized mutual information.

from each other. It is expected that there should be an overlap between two components. We can see that all the six methods cannot completely separate the components. Our TR-PAE (Fig. 6 (a)), however, have the least overlapped area.

The evaluation using AC and NMI displayed in Fig. 7 once again shows that the proposed TR-PAE have better performance when analyzing the text styles.

We list the clustering results in Table 7. From Table 7, we can observe that for all test corpuses, TR-PAE and DLA work better than the other methods by delivering higher AC and NMI values. Though TR-PAE and DLA are comparative, in most cases as is shown in Figs. 3, 5 and 7, and Table 7, our proposed TR-PAE outputs higher AC and NMI than DLA.

Table 7
Performance comparisons on three corpuses.

Simulation Setting						
Method	Corpus 1 (7 classes)		Corpus 2 (5 classes)		Corpus 3 (3 classes)	
	AC	NMI	AC	NMI	AC	NMI
PCA	0.2641	0.0766	0.3536	0.1823	0.3398	0.0830
LDA	0.2681	0.0987	0.4596	0.2851	0.4102	0.1867
SLPP	0.2607	0.1225	0.4561	0.2817	0.4203	0.1096
TR-LDA	0.2850	0.1449	0.4592	0.3667	0.5462	0.4046
DLA	0.3423	0.2289	0.602	0.5083	0.6568	0.5286
TR-PAE	0.3438	0.2741	0.6111	0.5271	0.6750	0.5533

4.3.3. *Numerical comparison of dimensionality reduction based on classification*

The simulation settings are as follows: we randomly select 100 samples per class ($l_i = 100$) from each data set as training set (about 50% samples in each data set chosen as training set) and the remaining samples as test set. For SLPP, we use the supervised version of LPP in [38] (LPP2), where the similarity matrix S between any pair-wise samples is defined as (we use the same strategy in [38]) $S_{ij} = x_i^T x_j/(\|x_i\| \cdot \|x_j\|)$, if x_i and x_j belong to the same class; $S_{ij} = 0$, otherwise. For DLA and TR-PAE, there are two parameters k_1 and k_2 involved, which represent the number of nearest neighbor samples with the same class label and that with different class labels, respectively. We use five-fold-cross validation to determine them and the candidate sets are $\{4, 8, \ldots, l_i\}$ for k_1 and $\{4, 8, \ldots, l-l_i\}$ for k_2, where l_i and l are the total number of samples in the ith class and whole training set. All methods used labeled set in the output reduced space to train a nearest neighborhood classifier for evaluating the accuracies of test set. The average accuracies over 20 random splits under different dimensionality for each data set are shown in Fig. 8, which is used to select the best subspace dimensionality. Then, Table 8 reports the final classification accuracies with the best subspace dimensionality.

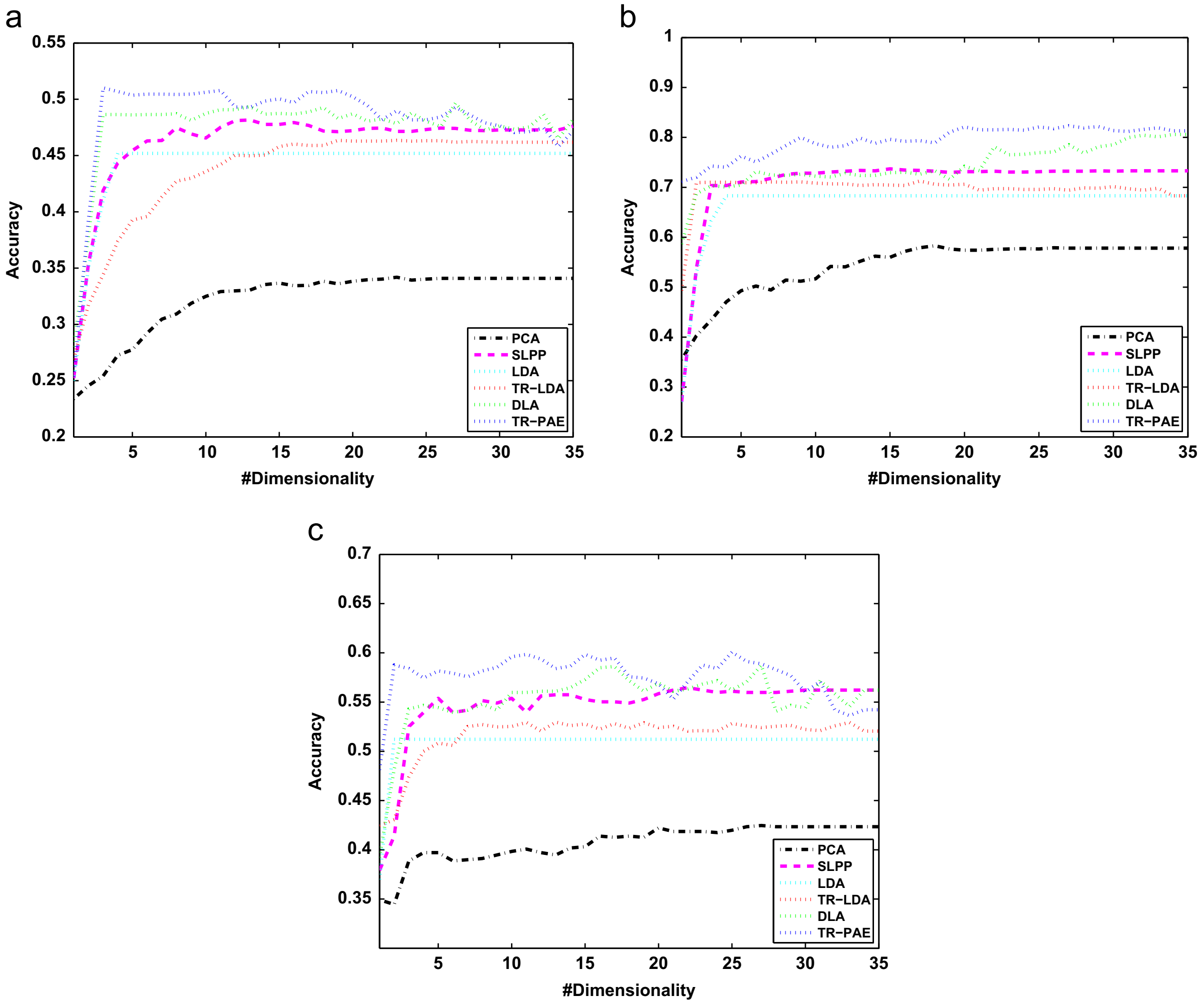

Fig. 8. Average accuracies under different dimensionality. (a) Corpus 1, (b) Corpus 2 and (c) Corpus 3.

Table 8
Average accuracy on the test set (the numbers in dim show the best subspace dimensionality for PCA, LDA, TR-LDA, SLPP, DLA and TR-PAE).

Methods	Corpus 1			Corpus 2			Corpus 3		
	Mean	std	dim	mean	std	dim	mean	std	dim
PCA	0.3419	0.02512	23	0.5829	0.04596	18	0.4246	0.03054	27
LDA	0.4519	0.01899	4	0.683	0.03987	4	0.5122	0.02843	2
TR-LDA	0.4635	0.01553	19	0.7125	0.03167	17	0.5297	0.02468	13
SLPP	0.4816	0.01652	18	0.737	0.03531	15	0.5646	0.02196	22
DLA	0.4966	0.01793	27	0.806	0.02982	34	0.5873	0.02269	27
TR-PAE	0.5101	0.01638	3	0.8232	0.0263	27	0.6008	0.02108	25

In summary, we can obtain the following observations: (1) as an unsupervised method, PCA cannot perform well as it cannot grasp the discriminative information embedded in the training set. All the supervised methods are superior to the unsupervised method such as PCA by about 20–30% in all the data sets due to the utilizing the label information. This also indicates that the labeled samples are important in handling classification problem. (2) SLPP, DLA and the proposed TR-PME are superior to other supervised methods such as LDA and TR-LDA, e.g. TR-PME outperform LDA by about 10% in the classification results of News items of different decades. The superiority can also achieve 5–8% in other classification results. This enhancement is believable due the reason that the three methods can characterize the local discriminative structure of different data sets, which is better than the global structure such as LDA and TR-LDA; (3) the classification accuracies of all methods vary as the reduced dimensionality increase, another important observation is that the proposed TR-PME outperforms other methods.

5. Conclusions

In this paper, we have discussed the text style analysis based on our proposed TR-PAE. By incorporating improved ITR algorithm (iITR), the proposed TR-PAE is capable to obtain reliable cues of text styles in lower dimensional spaces. It 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. This strategy helps to extract the deeply hidden information about text styles in documents. 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. The convenience enables us to obtain intuitive and meaningful observations about the text styles in a visualized way. Our strategy for constructing the scatter matrices in TR-PAE can avoid the drawback that samples in each class must follow a Gaussian distribution. Thus, for a non-Gaussian textual data set, it is better to represent the separation of different classes than the intra-class covariance used in other supervised methods like LDA.

The validity of text style analysis using TR-PAE has been examined by corpuses collected from real online data sets and existing widely used corpuses. First the original corpus is converted by employing style markers to matrix that the machine learning approaches can handle. Consequently, different methods are utilized to obtain the representation in lower dimensional space. From all investigated cases, the text style analysis based on TR-PAE algorithm is capable of producing comprehensive visualization results of multiple components and outperforms the other approaches. Both the visualization results and quantitative evaluation results indicate that the proposed TR-PAE outperforms the conventional approaches. The simulation results show that using our method, various styles of documents of different genres are able to be distinguished. 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 simulations demonstrate that the cues that reveal the variation of text styles caused by changing regions and time cannot be discovered by using conventional methods, while by using our proposed approach, the text style information in such data sets is able to be extracted.

References

[1] R.T. Freeman, H. Yin, Tree view self-organisation of web content, Neurocomputing 63 (0) (2005) 415–446, http://dx.doi.org/10.1016/j.neucom.2004.07.005.

[2] K. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1) (1972) 11–21.

[3] G. Chowdhury, Introduction to Modern Information Retrieval, Facet publishing, London, United Kingdom, 2010.

[4] C. Brinegar, Mark twain and the quintus curtius snodgrass letters: a statistical test of authorship, J. Am. Stat. Assoc. (1963) 85–96.

[5] A. Morton, The authorship of greek prose, J. R. Stat. Soc. Ser. A (Gen.) 128 (2) (1965) 169–233.

[6] B. Brainerd, Weighing Evidence in Language and Literature: A Statistical Approach, vol. 19, University of Toronto Press, Toronto, Ontario, Canada, 1974.

[7] F. Tweedie, R. Baayen, How variable may a constant be? Measures of lexical richness in perspective, Comput. Humanit. 32 (5) (1998) 323–352.

[8] B. Kessler, G. Numberg, H. Schütze, Automatic detection of text genre, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA, 1997, pp. 32–38, http://dx.doi.org/10.3115/976909.979622.

[9] Y. Lee, S. Myaeng, Text genre classification with genre-revealing and subject-revealing features, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information rRetrieval, ACM, 2002, pp. 145–150.

[10] A. Finn, N. Kushmerick, B. Smyth, Genre classification and domain transfer for information filtering, Adv. Inf. Retr. (2002) 349–352.

[11] A. Finn, N. Kushmerick, Learning to classify documents according to genre, J. Am. Soc. Inf. Sci. Technol. 57 (11) (2006) 1506–1518.

[12] R. Clement, D. Sharp, Ngram and Bayesian classification of documents for topic and authorship, Lit. Linguist. Comput. 18 (4) (2003) 423–447.

[13] S. Feldman, M. Marin, M. Ostendorf, M. Gupta, Part-of-speech histograms for genre classification of text, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), IEEE, 2009, pp. 4781–4784.

[14] D. Biber, Dimensions of Register Variation: A Cross-Linguistic Comparison, Cambridge University Press, Cambridge, United Kingdom, 1995.

[15] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic text categorization in terms of genre and author, Comput. Linguist. 26 (4) (2000) 471–495.

[16] S.M. zu Eissen, B. Stein, Genre classification of web pages, in: Advances in Artificial Intelligence (AI 2004), Springer, 2004, pp. 256–269.

[17] E.S. Boese, A.E. Howe, Effects of web document evolution on genre classification, in: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05), ACM, New York, NY, USA, 2005, pp. 632–639, http://dx.doi.org/10.1145/1099554.1099715.

[18] C.S. Lim, K.J. Lee, G.C. Kim, Multiple sets of features for automatic genre classification of web documents, Inf. Process. Manag. 41 (5) (2005) 1263–1276, http://dx.doi.org/10.1016/j.ipm.2004.06.004.

[19] J. Burrows, Word-patterns and story-shapes: the statistical analysis of narrative style, Liter. Linguist. Comput. 2 (2) (1987) 61–70.

[20] H. van Halteren, F. Tweedie, H. Baayen, Outside the cave of shadows: using syntactic annotation to enhance authorship attribution, Comput. Humanit. 28 (2) (1996) 87–106.

[21] D. Merkl, Text classification with self-organizing maps: some lessons learned, Neurocomputing 21 (1) (1998) 61–77.

[22] D. Petkova, W. Croft, Proximity-based document representation for named entity retrieval, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, ACM, 2007, pp. 731–740.

[23] Y. Lv, C. Zhai, Positional language models for information retrieval, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2009, pp. 299–306.

[24] S. Argamon, M. Koppel, G. Avneri, Routing documents according to style, in: Proceedings of the First International Workshop on Innovative Information Systems, 1998.

[25] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[26] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[27] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[28] T. Zhang, D. Tao, X. Li, J. Yang, Patch alignment for dimensionality reduction, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1299–1313.
[29] J. Duchene, S. Leclercq, An optimal transformation for discriminant and principal component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 978–983.
[30] M. Turk, A. Pentland, Face recognition using eigenfaces, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'91), 1991, pp. 586–591, http://dx.doi.org/10.1109/CVPR.1991.139758.
[31] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic press, Waltham, Massachusetts, 1990.
[32] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[33] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[34] Z.-y. Zhang, H.-y. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, J. Shanghai Univ. (English Ed.) 8 (4) (2004) 406–424.
[35] J. Yu, D. Liu, D. Tao, H.S. Seah, Complex object correspondence construction in two-dimensional animation, IEEE Trans. Image Process. 20 (11) (2011) 3257–3269.
[36] J. Yu, D. Liu, D. Tao, H. Seah, On combining multiple features for cartoon character retrieval and clip synthesis, IEEE Trans. Syst. Man, Cybern. Part B: Cybern. 42 (5) (2012) 1413–1427.
[37] J. Yu, M. Wang, D. Tao, Semisupervised multiview distance metric learning for cartoon synthesis, IEEE Trans. Image Process. 21 (11) (2012) 4636–4648.
[38] D. Cai, X. He, J. Han, Using Graph Model for Face Analysis, Department of Computer Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL, 2005.
[39] Q. Gao, J. Ma, H. Zhang, X. Gao, Y. Liu, Stable orthogonal local discriminant embedding for linear dimensionality reduction, IEEE Trans. Image Process. 22 (7) (2013) 2521–2531, http://dx.doi.org/10.1109/TIP.2013.2249077.
[40] X. Gao, N. Wang, D. Tao, X. Li, Face sketch-photo synthesis and retrieval using sparse representation, IEEE Trans. Circuits Syst. Video Technol. 22 (8) (2012) 1213–1226, http://dx.doi.org/10.1109/TCSVT.2012.2198090.
[41] B. Du, L. Zhang, D. Tao, D. Zhang, Unsupervised transfer learning for target detection from hyperspectral images, Neurocomputing 120 (0) (2013) 72–82, http://dx.doi.org/10.1016/j.neucom.2012.08.056, URL: http://www.sciencedirect.com/science/article/pii/S092523121300297X.
[42] P. Tang, T.W. Chow, Recognition of word collocation habits using frequency rank ratio and inter-term intimacy, Expert Syst. Appl. 40 (11) (2013) 4301–-4314, http://dx.doi.org/10.1016/j.eswa.2013.01.003, URL: http://www.sciencedirect.com/science/article/pii/S0957417413000067.
[43] H. Wang, S. Yan, D. Xu, X. Tang, T. Huang, Trace ratio vs. ratio trace for dimensionality reduction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), IEEE, 2007, pp. 1–8.
[44] Y. Jia, F. Nie, C. Zhang, Trace ratio problem revisited, IEEE Trans. Neural Netw. 20 (4) (2009) 729–735.
[45] M. Zhao, Z. Zhang, T.W. Chow, Trace ratio criterion based generalized discriminative learning for semi-supervised dimensionality reduction, Pattern Recognit. 45 (4) (2012) 1482–1499.
[46] L. Zhou, L. Wang, C. Shen, Feature selection with redundancy-constrained class separability, IEEE Trans. Neural Netw. 21 (5) (2010) 853–858.
[47] T. Matsui, Y. Saruwatari, M. Shigeno, An Analysis of Dinkelbach's Algorithm for 0–1 Fractional Programming Problems.
[48] F. Nie, S. Xiang, Y. Jia, C. Zhang, S. Yan, Trace ratio criterion for feature selection, in: Proceedings of the 23rd National Conference on Artificial Intelligence, vol. 2, 2008, pp. 671–676.
[49] M. Zhao, R.H. Chan, P. Tang, T.W. Chow, S.W. Wong, Trace ratio linear discriminant analysis for medical diagnosis: a case study of dementia, IEEE Signal Process. Lett. 20 (5) (2013) 431–434.
[50] M.E. Newman, Power laws, pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351.
[51] S.J. Pan, J.T. Kwok, Q. Yang, Transfer learning via dimensionality reduction., in: AAAI, vol. 8, 2008, pp. 677–682.
[52] D. Cai, X. He, J. Han, Document clustering using locality preserving indexing, IEEE Trans. Knowl. Data Eng. 17 (12) (2005) 1624–1637.
[53] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Adv. Neural Inf. Process. Syst. 18 (2006) 507.
[54] Z. Zhang, T. Chow, M. Zhao, M-isomap: orthogonal constrained marginal isomap for nonlinear dimensionality reduction, IEEE Trans. Cybern. 43 (1) (2013) 180–191.
[55] Z. Zhang, T.W. Chow, M. Zhao, Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization, IEEE Trans. Knowl. Data Eng. 25 (5) (2013) 1148–1161, http://dx.doi.org/10.1109/TKDE.2012.47.

Peng Tang is currently pursuing the Ph.D. degree at the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. He received his B.Eng. degree from the Department of Information Science and Technology, Shandong University, China, in 2007, and Master degree from the Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China, in 2010. His current interests include document retrieval, computational intelligence, pattern recognition, and their applications.

Mingbo Zhao received the Ph.D. degree in computer engineering from the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR in January 2013. He also received his B.Sc. and M.Sc. degrees from Shanxi University, Shanxi, PR China in 2005 and 2008, respectively. He is currently a Senior Research Assistant at the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR. His current interests include data mining, machine learning, pattern recognition, and their applications.

Tommy W.S. Chow received the B.Sc. (First Hons.) and Ph.D. degrees from the University of Sunderland, Sunderland, UK. He joined the City University of Hong Kong, Hong Kong, as a Lecturer in 1988. He is currently a Professor in the Electronic Engineering Department. His research interests are in the area of Machine learning including Supervised and unsupervised learning, Data mining, Pattern recognition and fault diagnostic. He worked for NEI Reyrolle Technology at Hebburn, England developing digital simulator for transient net- work analyser. He then worked on a research project involving high current density current collection system for superconducting direct current machines, in collaboration with the Ministry of Defense (Navy) at Bath, England and the International Research and Development at Newcastle upon Tyne. He has authored or coauthored of over 120 technical papers in international journals, 5 book chapters, and over 60 technical papers in international conference proceedings.