December 2016. Vol.5 / pp.87-102

A Quantitative Study of Chinese Writing Style, Based on *New Youth*

Ching-Syang Jack YUE
Li-Hsing HO
Yan-Yan PAN*

Abstract

The term “Big Data” has become very popular in recent years, and this concept has also had a significant influence on the study of languages. By utilizing vast archives of digital text and newly developed analysis methods, linguists now have more options to do research. In particular, we can cross-check several sources at the same time, and in greater detail than was previously possible. The digital humanities are a rapidly growing research field all around the world. But even with the help of computers, humans are still central to such studies, and we rely on expert opinion to determine the key elements for data analysis. Perhaps this is the main reason why text mining has only developed relatively slowly, and one of the main difficulties encountered in conducting text mining is the data input. For computing purposes, texts are a form of unstructured data which first needs to be quantified, and since, so far, there has been no Standard Operating Procedure (SOP) for the quantification process, experts play an important role in selecting relevant information (*i.e.* variables) for data analysis. In this study, our goal is to establish an SOP for text mining, and then to use it to study the historical change in Chinese writing style, from classical Chinese to modern Chinese, which occurred about a century ago. The study material is Volumes 1-7 of *New Youth*

* Ching-Syang Jack Yue (csyue@nccu.edu.tw) is Professor of Statistics at National Chengchi University. Li-Hsing Ho (lillianlhho@gmail.com) is Assistant Professor of Chinese Literature at National Tsing Hua University. Yan-Yan Pan (christinafun1128@163.com) is Master of Statistics at National Chengchi University.

(新青年), by the end of which, modern vernacular Chinese had almost completely replaced classical Chinese. We adapt the idea of unsupervised learning from statistical learning theory to define and identify important variables. In particular, we use the approach of Exploratory Data Analysis (EDA) to evaluate potential variables as candidates for differentiating between language styles. Also, following a previous study of ours, numerous variable and data reduction methods are needed. Thus, we use principle component analysis to reduce the number of variables, and then apply classification methods, such as logistic regression, to judge whether the style of an article is closer to classical or to modern Chinese. Also, to avoid over-parameterization (*i.e.* using more variables than are necessary), we use cross-validation to select the most feasible model. This cross-validation separates the data into a training set (or in-sample) and a testing set (or out-sample); the training set is used to construct the model and this model is then applied to the testing set to calculate the model's accuracy. Our study shows a gradual change in the writing style of *New Youth* articles from classical Chinese to modern Chinese. Thus, only 1% of the articles in Volume 1 are classified as modern Chinese compared with 98% of those in Volume 7, and about 60% of the articles in an intermediate volume, number 4, are classified as modern Chinese. Our model has a prediction accuracy for the articles in Volume 4 of about 84%, as determined by cross-checking with expert Chinese linguists. The results of our quantitative numerical analysis are clearly promising, suggesting that this approach should be continued for the study of modern Chinese writing.

Keywords

classical and modern Chinese, *New Youth*, species diversity, structured and unstructured data, logistic regression

Introduction

Big data has become a very popular talking point in recent years, and experts in many fields are taking note of its influence. The quantitative analysis of languages and text (*i.e.* text mining) has also become popular, but this technique has only developed relatively slowly, mainly because the analysis of text, as compared to other kinds of data, requires additional effort, most notably a knowledge of the languages involved. For computational purposes, there are basically two types of data: structured and unstructured. Structured data are those with a high degree of organization, such as data stored in Excel spreadsheets, and unstructured data are those without such organization. Most recorded information is unstructured data (about 80～90%), but the majority of data analyses are still applied to structured data.

In order to conduct a quantitative analysis of unstructured data, we first need to give them a structure, but there is no single "correct" approach through which such structure should be supplied, and this issue is a serious hindrance to the computational analysis of unstructured data. In this study, we propose a quantitative approach to providing Chinese texts with a structure, and use the variables defined in this approach to analyse Chinese writing style. Specifically, we are interested in comparing classical and modern Chinese; the transition between these two styles is most clearly shown by the influential *New Youth* magazine, and the May Fourth Movement of 1919 was a key event in dividing classical from modern Chinese writing style.

New Youth (a.k.a. *La Jeunesse*) was published in 11 volumes, and our previous study has demonstrated that modern vernacular Chinese had almost completely replaced classical Chinese as the main written language by the end of Volume 7 (published in 1920). Thus, we can use Volumes 1-7 to study the process of transition which gave rise to the modern Chinese writing style. We can draw a useful analogy between biological species and key phrases: following a change of habitat, old species are replaced by new ones, and this resembles the observed change of writing style. The concept of ecological habitat can therefore be used to explore the way in which writing style changes over time, through a quantitative analysis which compares the differences in writing style between classical Chinese and modern Chinese.

Specifically, we compare two kinds of approach, supervised learning and unsupervised learning. The first analyses variables assigned by the expert opinion of humanities scholars, suitably adapted for data analysis. The second is essentially data-driven, applying the concepts and methods used in lexical analysis. In addition, we adapt some ideas from ecology, analysing the changes in writing style according to the principles of species diversity and evolution. We can treat words as analogous to species, and each individual writing style as a distinct ecological system.

Feature selection plays a central role in distinguishing writing style, and we therefore adapt the idea of unsupervised learning from statistical learning theory to define and identify the important variables. In particular, we use the notion of Exploratory Data Analysis (EDA), proposed by the renowned statistician J. W. Tukey in 1977, to evaluate potential variables as candidates for differentiating the language styles of articles in *New Youth*.[1] Also, following a previous study of writing style in Chinese, numerous variable and data reduction methods are needed. Thus, we use principle component analysis to reduce the number of variables, and then apply classification methods, such as logistic regression and classification trees, to judge whether the style of an article is closer to classical or to modern Chinese. Also, to avoid over-parameterization (*i.e.,* using more variables than are necessary), we use cross-validation to select the most feasible model. This cross-validation separates the data into a training set (or in-sample) and a testing set (or out-sample); the training set is used to construct the model and this model is then applied to the testing set to calculate the model's accuracy.

Methods and Data

The data considered in this study are articles from *New Youth,* an important Chinese magazine of the 1910s and 1920s, useful for studying the spread of modern Chinese in response to the May Fourth Movement of 1919. The first seven volumes are used in this study,[2] since from Volume 8 *New Youth* openly supported communism, having been influenced by the 1917 Russian Revolution, and from Volume 10 it became the official journal of the Chinese Communist Party. Our goal was to study the language change of the early volumes of *New Youth,* particularly the change of writing style from classical to modern Chinese. Specifically, we speculate that Volumes 1 and 2 of *New Youth* were not influenced by the May Fourth Movement in 1919 and can thus be considered as typical of the prevailing classical Chinese writing style, whereas the writing style of Volume 7 is close to the modern Chinese.

[1] See Tukey (1977).

[2] See Ho, *et al.* (2014).

We therefore apply the values of 0 and 1 (signifying modern and classical Chinese) to the articles from Volume 7 and from Volume 1 respectively, to distinguish their relative degree of influence by the May Fourth Movement. Thus, we can use logistic regression, with the target value ranging between 0 and 1, to determine whether articles are influenced by the May Fourth Movement. Based on the selected variables, we construct a regression model to classify articles and apply the model to evaluate the articles for their closeness to classical or modern Chinese.

The logistic regression uses independent variables ($x_i, i = 1, 2, \dots, k$) to assign the target variable (y) a value between 0 and 1,

$$y = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}} \tag{a}$$

or

$$\log(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \tag{b}$$

if we define $E(y) = p$.[3] The key to applying logistic regression is usually to select appropriate variables and their functional forms, and this is especially difficult in the case of text mining and other soft (unstructured) data. To address this issue, we used EDA to select the variables, as described in the following section.

When logistic regression is used for binary classification, with the value 0 and 1 indicating the respective groups, we usually assign observations with fitted values smaller than 0.5, the threshold value, to the "0" group and those larger than 0.5 to the "1" group. The accuracy of logistic regression is measured by $\frac{a+d}{a+b+c+d}$, where the values *a*, *b*, c, and *d* are defined in Table 1. The observations associate to *a* and *d* are correctly classified, while those associate to *b* and *c* are not. There is also an empirical consideration: if the size of the "0" and "1" groups differs greatly, then the threshold value can be set to some other value than 0.5.

[3] See Agresti (1990).

〈Table 1〉 Evaluating Fitting Accuracy

		Fit	
		0	**1**
True	0	*a*	*c*
	1	*c*	*d*

When there are many independent variables, we usually apply variable (or data) reduction techniques, such as Principal Component Analysis (PCA), for variable selection. PCA can be used to determine the minimum number of independent variables required for regression analysis. Of course, if our goal is the accuracy of logistic regression, we might not need to consider PCA. Note that PCA is a prominent method of multivariate analysis[4] which can also be used to interpret the regression result, but after applying PCA the original independent variables are transformed, so that the interpretation of the regression model may also change.

Since the logistic regression model for the data from Volumes 1 and 7 is constructed first, and then this model is applied to data from Volume 4, we need to evaluate the stability of regression model. This is achieved by cross-validating to make sure that the model is not unduly data-sensitive. The data is divided into k equal parts and k-1 parts (the training set) are used to construct the model, which is then applied to the remaining part (the testing set). The model is deemed to be stable if the model accuracies (Table 1) of the training and testing sets are about the same. This k-fold evaluation can also be repeated several times, by randomly separating the data, which optimizes the scope of the available data.

Exploratory Data Analysis (EDA)

For computational purposes, two types of big data can be identified: “hard” or structured data, and “soft” or unstructured data. Structured data possesses a high degree of organization, examples include library catalogues (with the publication dates, publication places, authors, *etc.,* of books) and population census records (with the birth dates, addresses, family details, *etc.,* of

[4] See Johnson and Dean (2007).

people). This type of data can usually be quantified without difficulty and are easily captured, stored, and analysed.

Unstructured data, by contrast, lacks any obvious structure, so that quantifying them usually requires specific knowledge of the application domain. Thus, the process of converting such data into numeric values is inherently subjective, which results in a wide variation in the values obtained, depending on who does the conversion and on which problem they are addressing. Most textual data are unstructured and all computational analysis thereof, *i.e.* text mining, needs a lot of extra work. We need to create a relational structure for such textual data before plugging them into logistic regression. So far, there has been no standard operating procedure for structuring textual data, but we believe that EDA is a promising approach for this purpose, and this section explains how we applied it.

The function of EDA is to discover the data properties, such as computing the sample average and sample variance, through basic statistical analysis. For textual data, the basics of EDA include the number of words, the number of different words (or phrases) and their distribution. Table 2 shows the numbers of words and different words for all 11 volumes of *New Youth*. In general, the later volumes have more words but not more phrases. This property can also be well described by species diversity indices, such as the Simpson index and the Shannon entropy, which are defined as $\theta_s = \sum_i p^2$ and $\theta_E =- \sum_i p_i \ln(p_i)$, respectively, where *p* is the proportion of phrase *i*. Note that the larger the Shannon entropy is, the larger the species (or phrase) diversity. Conversely, a smaller value of the Simpson index indicates more diversity. In other words, the statistics in Table 2 imply less diversity in later volumes.

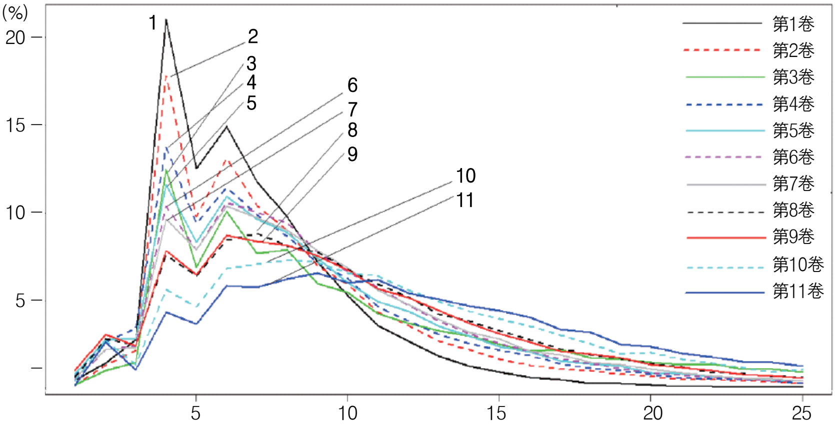
In addition to the counts of words and phrases, we can also use other attributes of Chinese texts for style classification. Thus, the early volumes of *New Youth* are stylistically classical, while the later volumes are stylistically modern, and among the main differences between these two writing styles are the sentence length and the common function words used. For determining sentence length, the punctuation marks “, . ; ! ?” (comma, period, semicolon, exclamation mark, and question mark) are used to separate sentences. In Volume 1 the proportion of sentences containing 4, 5, or 6 words totals about 50%, and in Volume 7 the proportion of sentences containing 4, 5, 6, 7, or 8 words totals about 50%, clearly

〈Table 2〉 Words Count of *New Youth*

Volume	No. of Words	No. of Phrases	Simpson Index	Shannon Entropy
1	248,833	4,379	0.004568	6.654036
2	291,848	4,344	0.004500	6.649539
3	290,038	4,227	0.004954	6.541824
4	305,020	4,298	0.004172	6.539378
5	343,519	4,125	0.004672	6.461579
6	389,407	3,848	0.005749	6.348547
7	586,942	3,850	0.006053	6.328604
8	461,731	3,753	0.006035	6.320355
9	437,748	3,745	0.005574	6.322103
10	342,778	2,980	0.005700	6.177278
11	489,223	3,093	0.005712	6.212699

indicating how the later volumes have longer sentences and larger variances: figure 1 shows frequency of particular sentence lengths for all the volumes. The average number of words (and the corresponding standard deviation) is 7.07 (1.07) for Volume 1, and 9.27 (1.99) for Volume 7. This clear increase in words used in the later volumes matches our expectation that classical Chinese expression is simpler and more concise.

〈Figure 1〉 The Frequency of Number of Words in a Sentence



Some other variables indicative of species richness are also considered in this study, in addition to those introduced above, including the number of new phrases per 1,000 words, the 10 most common single words, and the 10 most common two-word phrases. New phrases can be considered a proxy representing lexical depth, similar in function to the Type-Token Ratio (TTR). Since more words are usually associated with a smaller TTR, we measure this quantity per 1,000 words, as a form of normalization.

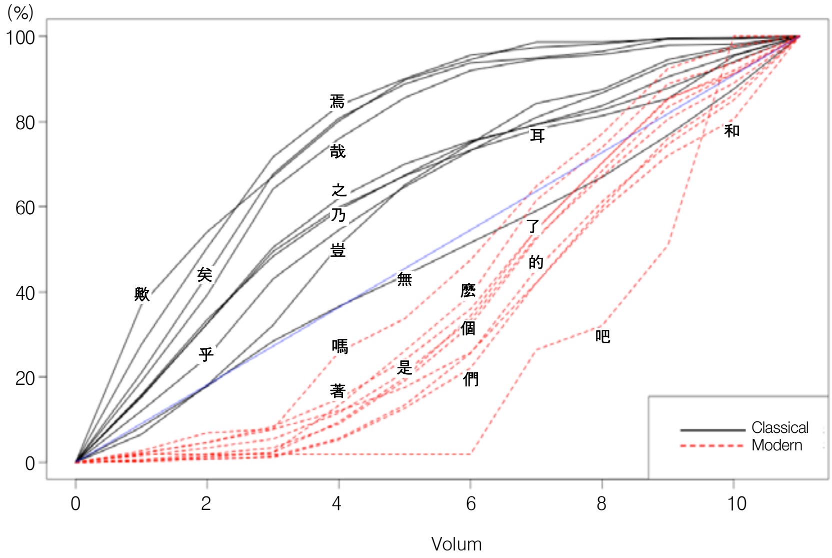
〈Table 3〉 10 Common Function Words

	Classical Chinese	**Modern Chinese**
Words	矣乎焉歟哉耳豈之乃無	的是們個了和麼著嗎吧
Volume 1 proportion	3.6%	0.7%
Volume 7 proportion	0.5%	8.8%
Volumes 1-7 proportion	2.4%	7.3%

The variables mentioned above derive from unsupervised (data-driven) learning, and we should also consider variables derived from supervised learning, *i.e.* suggestions made by experts. Function words are one of the most popular choices for the study of writing style. Here we adapt the idea of Ho, *et al.* (2014), choosing 10 common function words typical of classical Chinese, and 10 which are typical of modern Chinese. Again, we use the statistics for Volumes 1 and 7 as a demonstration, and as expected, there are more classical function words in Volume 1 and more modern function words in Volume 7 (Table 3): it seems that the pattern of function words usage is very different.

We can see from this figure how the usage of these 20 function words varies with time, the trends being calculated in a similar way to the Gini index, by accumulating the percentages of word usage. Thus, lines above the blue line indicate words used more in the early volumes, while lines below the blue line show that the words are used more in the later volumes. Except for the word "無," all classical Chinese function words appear above the blue line and all modern Chinese function words are below the blue line: it seems these 20 function words are good candidates to differentiate between the two Chinese writing styles under analysis.

〈Figure 2〉 Frequencies of 20 Common Function Words



〈Table 4〉 Variables in the Logistic Regression

Types	**Variables**
Words & Phrases	Total numbers of words and phrases
	New phrases per 1,000 words
	The cumulative number of phrases
	Simpson Index
	Shannon Entropy
Sentence	Average sentence length
	Variance of sentence length
Function Words	10 classical and 10 modern Chinese
Common Words	10 most common single words and 10 most common two-word phrases

All the variables considered are summarized in Table 4: and they have been used to differentiate the writing styles of the articles in *New Youth*. In the next section, these variables are applied to construct a logistic regression model.

Data Analysis of *New Youth*

Using the textual data from *New Youth* to construct the logistic regression model, we first focus on verifying whether it can provide an accurate and stable assessment between modern and classical Chinese, labeled "0" and "1" respectively, and then we check if we can distinguish modern Chinese articles from classical Chinese ones. Also, we also evaluate if the constructed model is stable and reliable by cross-validation.

Table 5 shows the classification results of articles from Volumes 1 and 7, which have 162 and 132 articles respectively. Of the total 294 articles, 289 are classified correctly, *i.e.* 98.3%. In addition to the issue of classification accuracy, we also need to check the stability of the regression model via cross-validation. The regression model is first built from on training data and then applied to the testing data. The fitting accuracies of training data and testing data are recorded separately, and these two numbers should be close if the model is stable. For each simulation run, we randomly separate the training data (90%) and the testing data 10%. Table 6 shows the averages and their standard errors of the fitting accuracy for the training and testing data from 100 such simulation runs. Apparently, the regression model is fairly stable since it has very similar fitting accuracy (and small standard errors) for training and testing data. Note that the larger standard error of the testing data reflects its smaller size.

〈Table 5〉 Classification Results of Volumes 1 and 7

		Fit	
		Volume 1 (Classical)	**Volume 7 (Modern)**
True	Volume 1 (Classical)	160	2
	Volume 7 (Modern)	3	129

〈Table 6〉 Cross-Validation Results of Volumes 1 and 7

	Fitting Accuracy	
	Average	**Standard Error**
Training	96.10%	0.07%
Testing	95.95%	0.31%

Since there are quite a lot of variables (see Table 4), we could also employ PCA to reduce the number of variables in the regression model, without sacrificing the accuracy of classification. The number of principal components (or variables) is found to be 3 or 4, much smaller than the list in Table 4. However, the variables selected in this technique are linear combinations of the original variables and it is usually difficult to give a useful interpretation for these linear combinations. Thus, if our goal is to provide an interpretation of the classification with respect to the original variables, then we suggest not using PCA for variable reduction.

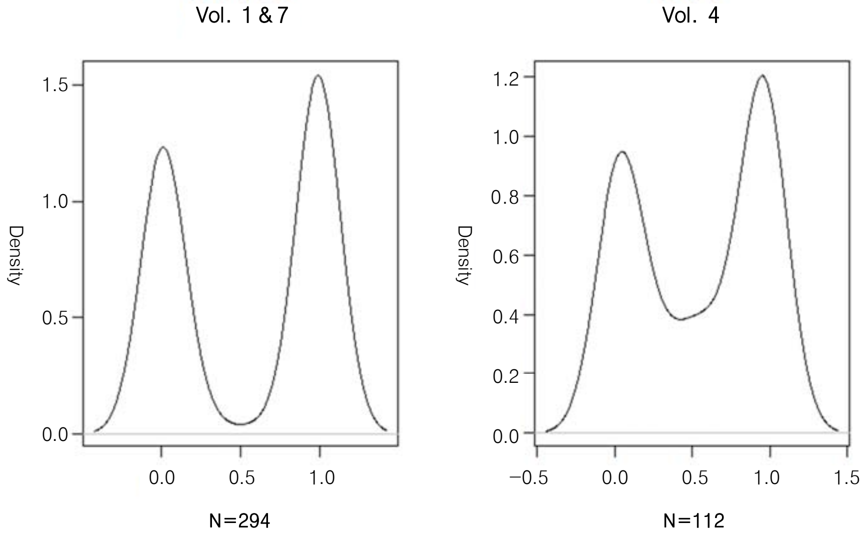
〈Table 7〉 Classifications of Articles in Volume 4

		Fit	
		Classical	**Modern**
True	Classical	34	0
	Modern	13	32

In the next step, we apply the constructed regression model to the articles in Volume 4, comparing these results with expert opinions about the writing style. Of the 79 articles, 66 are correctly classified (83.54% accuracy), as shown in Table 7. All of the articles judged by human experts to be classical Chinese are correctly classified by our regression model, while the fitted accuracy of articles judged to be modern Chinese is 72.73%. Looking at these results in another way, we might say that the articles labelled as modern Chinese by logistic regression are truly modern Chinese. Although these numbers are interesting, they probably imply that the variables chosen have some systematic deficiency, and are therefore insufficient to cleanly differentiate classical and modern Chinese.

Figure 3 shows the detailed results of the classification process, after kernel smoothing. Fitted values closer to 0 and 1 (modern and classical Chinese) indicate stronger confidence in the logistic model, and values closer to 0.5 indicate ambiguous decisions. The fitted results from Volumes 1 and 7 are all close to 0 or 1, but a moderate proportion of those from Volume 4 are around 0.5, which accords with the results of classification accuracy (Tables 5, 6, and 7).

〈Figure 3〉 Classification Results for Volumes 1 & 7, and 4



Conclusion and Discussion

2015 marked the 100th anniversary of the first publication of *New Youth*. By employing the methods of the digital humanities to examine the stylistic language changes of the earlier volumes, we hope to make some new contributions, both to the study of *New Youth* and to the analysis of the writing styles used in modern Chinese. Our fitting results are stable and their variances are fairly reliable, as determined by cross-validation. This suggests that the writing styles of Volumes 1 and 7 are very different, and establishes that using a logistic regression model with the variables selected can easily differentiate between the articles from Volumes 1 and 7: numerical analysis is thus confirmed as a feasible approach.

In addition, we learn that the writing styles of articles (as identified by logistic regression) in *New Youth* gradually changes from classical Chinese to modern Chinese, from Volume 1 to Volume 7. Thus, when we apply the constructed model to Volumes 2, 68% of articles are classified as classical Chinese, and when we apply it to Volume 6, 88% of articles are classified as modern Chinese. It seems that the results of the quantitative analysis agree with the trend, as perceived by human experts, that the writing style of *New Youth* changed significantly during this period. The results of our study encourage further studies of Chinese writing style via the methods of big data and the digital humanities: this numerical approach offers the promise of a new methodology for conducting research in social sciences and the humanities.

As to the variables used for classification analysis, this study only referenced variables concerned with the numbers of words and phrases and the length of sentences. Of course, it would be better to also include information related to the article content, but this would require more work. We suggest that the EDA-type procedure used in this study to select content-related variables is a useful approach to avoid researchers' subjective bias when choosing which words to use. Thus, we can first summarize the most common words or phrases, and then decide which of these best represent the meaning of an article and should be used for classification. Care is needed in this process, however, since the use of different words does not necessarily indicate a different meaning or style. For example, synonyms and antonyms may be used to express the very same idea, according to the author's mood.

It is generally estimated that about 80-90% of the data recorded in any organization are unstructured. Unstructured data, according to Wikipedia, is information lacking a predefined data model or which is unorganized. Converting textual data into numerical data (*i.e.* giving unstructured data a structure) is the first and probably the most important step required to enable computational analysis. Once the texts are transformed into numerical data, we can apply well-developed data-mining techniques to the textual data. But despite the efforts of past studies to provide a standard operating procedure for structuring unstructured data, no consensus has been reached on how best to convert textual data, and in practice there is considerable variation in data transformation methodology.

Text mining often requires multidisciplinary knowledge, and this is one of the most important reasons why textual data transformation has proved difficult. In fact, for all kinds of big data studies, teamwork is essential, and especially for textual data. Such research often exceeds the abilities of a single person or of experts from a single discipline. For example, the identification of key words and phrases needs domain knowledge in the appropriate application field, as well as experience in quantitative analysis and data structures. In the future then, we anticipate that interdisciplinary collaboration, as exemplified by the researchers contributing to this paper, will become much more frequent.

Bibliography

Agresti, Alan. 1990. *Categorical Data Analysis*. New York: Wiley.

Hastie, Trever, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer.

Ho, Li-hsing, Ching-syang Yue and Wen-huei Cheng. 2014. "From Classical Chinese to Modern Chinese: A Study of Function Words from *Xin Qing Nian*" (從文言到白話: 《新青年》雜誌語言變化統計研究). *Journal of the History of Ideas in East Asia* (東亞觀念史集刊) 7.

Johnson, Richard A. and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. New York: Pearson.

Karlgren, Bernhard. 1952. "New Excursions in Chinese Grammar." *Bulletin of the Museum of Far Eastern Antiquities* 24.

Mosteller, Frederick and David Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Boston: Addison-Wesley.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27.

Thisted, Ronald and Bradley Efron. 1986. "Did Shakespeare Write a Newly-Discovered Poem?" *Biometrika* 74: 3.

Tukey, John W. 1977. *Exploratory Data Analysis*. Boston: Addison-Wesley.