Automated Chinese Essay Scoring From Topic Perspective Using Regularized Latent Semantic Indexing

Shudong Hao, Yanyan Xu
School of Information Science and Technology
Beijing Forestry University
Beijing, China
shudongh@acm.org
xuyanyan@bjfu.edu.cn

Hengli Peng
Institute of Educational Measurement
Beijing Language and Culture University
Beijing, China
penghl6402@aliyun.com

Kaile Su
College of Mathematics Physics and Information Engineering
Zhejiang Normal University
Zhejiang, China
Institute for Integrated and Intelligent Systems
Griffith University
Brisbane, Australia
k.su@griffith.edu.au

Dengfeng Ke*
Institute of Automaton
Chinese Academy of Sciences
Beijing, China
dengfeng.ke@ia.ac.cn

***Abstract*—Finding out an effective way to score Chinese written essays automatically remains challenging for researchers. Several methods have been proposed and developed but limited in the character and word usage levels. As one of the scoring standards, however, content or topic perspective is also an important and necessary indicator to assess an essay. Therefore, in this paper, we propose a novel perspective – topic, and a new method integrating topic modeling strategy called Regularized Latent Semantic Indexing to recognize the latent topics and Support Vector Machines to train the scoring model. Experimental results show that automated Chinese essay scoring from topic perspective is effective which can improve the rating agreement to 89%.**

Keywords—document understanding; automated Chinese essay scoring; topic modeling application; classification application

I. INTRODUCTION

Automated essay scoring is becoming increasingly popular in the field of educational evaluation technology. Comparing to the conventional human rater method, computer-assistant scoring has advantages over three aspects: less human labor, less subjective factors and higher agreement rates [1]. Therefore, since 1970s when Page firstly broke the dawn of the field of automated English scoring, this technique has captured the interests of both the educators and researchers. After that, several successful scoring systems, such as Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater, IntelliMetric and Bayesian Essay Test Scoring System (BETSY), have been applied into industry widely and successfully [2]. Because of the characteristic of Chinese which leads to the difficulties in processing texts, the research on automated scoring for Chinese essays started later than that for English and is developing slowly [3].

To evaluate an essay written in any languages, the raters often focus on such three aspects: language proficiency, structure arrangement and contents [4]. Since we focus on Chinese written essays, we take MHK (the minorities-oriented Chinese level test) for an example [5]. In MHK, language proficiency is the basic requirement that includes using Chinese characters correctly, choosing words or phrases properly, and writing sentences grammatically; structure arrangement requires test takers to arrange passages logically, in order to support the idea effectively; contents focus on the topic the test taker chooses.

From these three assessment perspectives, researchers have studied several techniques for automated essay scoring for Chinese by virtue of previous that for English. For language proficiency, word-level-based method [6] and Latent Semantic Analysis (LSA) [7] etc., have been introduced. Furthermore, incremental LSA [8] has also been proposed to reduce the computational time and memory usage. The lack of scoring from structure arrangement and contents, however, leads to an incomplete assessment.

Considering these current limitations, we propose the following assumptions: *1)* the topic(s) hidden in an essay is related to the score the writer gets; and *2)* there may exist more than one topic in an essay collection. The first assumption is reasonable, because the content depends on the topic(s), and that is one of the scoring standards. The second assumption leads to that we can apply topic modeling strategies which are popular in text mining, information retrieval and natural language processing [9], into automated Chinese essay scoring, making it possible to score essays from topic perspective.

Latent Dirichlet Allocation (LDA), a representative and successful method in topic modeling, was firstly proposed by David M. Blei [10]. Some other improved approaches such as probabilistic Latent Semantic Indexing (pLSI) and Non-negative Matrix Factorization (NMF) are also studied and applied later as well. Based on some assumptions needed to be maintained during computations, these methods are difficult to be scaled to process big datasets. Recently, a new method is proposed by Quan Wang et al. called Regularized Latent Semantic Indexing (RLSI) which can handle with scalability problems [11-12]. Therefore, we use RLSI learning algorithm to recognize the latent topics in the datasets. This method not

* *corresponding author*.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.533

only performs better according to experimental results, but can be applied into big datasets in the future, as the scale of MHK is becoming larger.

The rest of the paper is organized as follows. In section II, we introduce our proposed method, and we present the experimental results and some discussions in section III. At last, in section VI, we conclude this paper and point out the future work.

II. THE PROPOSED METHOD

In this paper, we propose a method of automated Chinese essay scoring from topic perspective, whose flowchart is shown in Fig. 1. The whole processing consists of three parts: dataset preparation and text-preprocessing, topic feature extraction, and scoring model training, which will be described as follows.

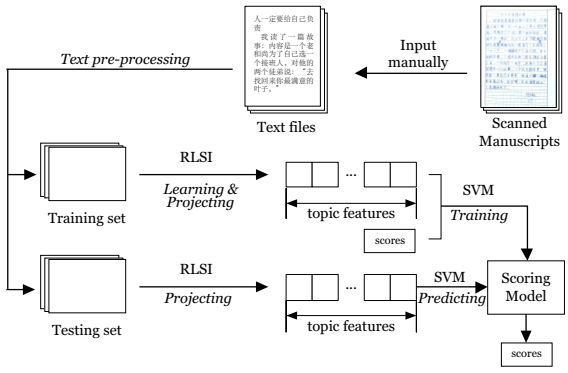


Fig. 1. The flowchart of the proposed method.

A. Dataset preparation and text pre-processing

Raw materials for datasets are scanned manuscripts that are needed to be typed into texts manually.

In text pre-processing, firstly, we use our previous research, Weighted Finite-State Transducer (WFST), to segment the essays to words with relatively high accuracy [13]. WFST is a strategy which segments sentences, and detects and corrects the erroneous characters in Chinese sentences simultaneously. In automated essay scoring, we use WFST for segmentation without corrections for erroneous characters.

Next, with the elimination of stopwords (such as auxiliary verbs) and rare words (occur only one or two times in the whole dataset), we get the term-document matrix **D** where each column is an *M*-dimensional term-based *essay vector* ***d***. *M* is the number of different terms and an essay vector denotes an written essay. The element $d_{i,j}$ is the frequency of the term *i* appeared in the essay ***j***.

For this occurrence-based matrix, we use TF-IDF strategy to weight it in order to achieve a better performance [14]. Generally, TF-IDF weights each element $d_{i,j}$ in the matrix as:

$$d_{i,j} = TF_{i,j} \times IDF_{i,j} \tag{1}$$

where TF is the term frequency and IDF is the inverse document frequency. Therefore, each element $d_{i,j}$ in the matrix **D** will be weighted as:

$$d_{i,j} = \frac{count_{i,j}}{\sum_{m=1}^{M} count_{m,j}} \times log \frac{N}{1 + DF_i} \tag{2}$$

where $count_{i,j}$ denotes the number of term *i* occurred in the essay vector ***j*** and DF_i the number of essay vectors containing the term *i*; *N* is the number of essay vectors in a certain collection, training set or testing set, and it is easy to figure out according to the context.

B. Topic feature extraction

In this subsection, we will introduce how to apply RLSI learning algorithm to automated Chinese essay scoring. Given a term-document matrix **D**, RLSI will generate two matrices, the term-topic matrix **U** and topic-document matrix **V**, that approximate the original matrix **D** and thus minimizes the quadratic loss function:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \left(\|\boldsymbol{D} - \boldsymbol{U}\boldsymbol{V}\|_F^2 + \lambda_1 \|\boldsymbol{U}\|_1 + \lambda_2 \|\boldsymbol{V}\|_F^2 \right) \tag{3}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ regularize **U** and **V**. The dimension of topic is denoted as topic number *K*. Thus, the target can be formalized as an optimization problem where there are two parameters. A common method to solve such problems is to fix one parameter and update the other, and continue this interchange until convergence.

1) Updating **U**

Initializing randomly and fixing the matrix **V**, we start from updating the matrix **U**. In this step, the optimization problem is transformed into:

$$\min_{\boldsymbol{U}} \left(\|\boldsymbol{D} - \boldsymbol{U}\boldsymbol{V}\|_F^2 + \lambda_1 \|\boldsymbol{U}\|_1 \right). \tag{4}$$

In order to compute iteratively or parallel, we rewrite the equation (3) as below:

$$\min_{\{\boldsymbol{u}_m\}} \sum_{m=1}^{M} \left(\|\boldsymbol{d}_m - \boldsymbol{V}^T \boldsymbol{u}_m\|_2^2 + \lambda_1 \|\boldsymbol{u}_m\|_1 \right) \tag{5}$$

where $\boldsymbol{d}_m = [d_{m1}, d_{m2}, \ldots, d_{mN}]^T$ is the essay vector, and $\boldsymbol{u}_m = [u_{m1}, u_{m2}, \ldots, u_{mK}]^T$ is from matrix **U**. Thus, the updating of **U** will be separated into *M* iterations. Using coordinate descent method with soft-threshold operator [15-17], the solution to each element *u* in this problem is:

$$u^*_{mk} = \frac{\left(\left|r_{mk} - \sum_{l \neq k} s_{kl} u_{ml}\right| - \frac{1}{2}\lambda_1\right)_+ \cdot sign(r_{mk} - \sum_{l \neq k} s_{kl} u_{ml})}{s_{kk}} \quad (6)$$

where s is from $\mathbf{S} = \mathbf{V}\mathbf{V}^T$ and r is from $\mathbf{R} = \mathbf{D}\mathbf{V}^T$, and $(\cdot)_+$ is the hinge function [18].

2) Updating V

Similarly, updating $\mathbf{V}$ can be formed as:

$$\begin{aligned} &\min_{\boldsymbol{V}} \|\boldsymbol{D} - \boldsymbol{U}\boldsymbol{V}\|_F^2 + \lambda_2 \|\boldsymbol{V}\|_F^2 \\ &= \min_{\{\boldsymbol{v}_n\}} \sum_{n=1}^{N} \left(\|\boldsymbol{d}_n - \boldsymbol{U}\boldsymbol{v}_n\|_2^2 + \lambda_2 \|\boldsymbol{v}_n\|_2^2 \right) \end{aligned} \quad (7)$$

and separated into N iterations or parallelized computations.

This is a Ridge Regression problem with $\ell 2$-norm regularization, and the solution is:

$$\boldsymbol{v}_n^* = \left(\boldsymbol{U}^T\boldsymbol{U} + \lambda_2 \boldsymbol{I}\right)^{-1} \boldsymbol{U}^T \boldsymbol{d}_n \quad (8)$$

where $\mathbf{I}$ is the identity matrix with corresponding dimension.

3) Projection

After learning, the matrix $\mathbf{U}$ constructs a topic space where any term-based essay vector $\boldsymbol{v}_q$ can be projected and represented as a topic-based essay vector $\boldsymbol{v}_q^*$. Given an essay vector $\boldsymbol{v}_q$, the target is:

$$\min_{\boldsymbol{v}} \left(\|\boldsymbol{v}_q - \boldsymbol{U}\boldsymbol{v}\|_2^2 + \lambda_2 \|\boldsymbol{v}\|_2^2 \right), \quad (9)$$

and the solution is:

$$\begin{aligned} \boldsymbol{v}_q^* &= arg\min_{\boldsymbol{v}} \|\boldsymbol{v}_q - \boldsymbol{U}\boldsymbol{v}\|_2^2 + \lambda_2 \|\boldsymbol{v}\|_2^2 \\ &= \left(\boldsymbol{U}^T\boldsymbol{U} + \lambda_2 \boldsymbol{I}\right)^{-1} \boldsymbol{U}^T \boldsymbol{v}_q . \end{aligned} \quad (10)$$

4) Algorithm

We present the whole RLSI algorithm for automated Chinese essay scoring in Fig. 2. Note that when updating $\mathbf{U}$, we judge the convergence by the difference between two iterations (*deviation*) for each row m:

$$\sqrt{\frac{\sum_{k=1}^{K} \left(u_{mk} - u^*_{mk}\right)^2}{\sum_{k=1}^{K} u_{mk}^2}} = \frac{\|\boldsymbol{u}_m - \boldsymbol{u}_m^*\|_2}{\sqrt{\sum_{k=1}^{K} u_{mk}^2}} \leq 1 \times 10^{-6} \quad (11)$$

where u^*_{mk} is the update value whereas the u_{mk} is the old value. The number of iterations T is an empirical value that can be set by users.

Training:

Input: $\boldsymbol{D} \in \mathbb{R}^{M \times N}$

Output: $\boldsymbol{U} \in \mathbb{R}^{M \times K}$, $\boldsymbol{V} \in \mathbb{R}^{K \times N}$

$\mathbf{V} \leftarrow$ random matrix, $\mathbf{U} \leftarrow \mathbf{0}$;

repeat T times:

  $\mathbf{S} \leftarrow \mathbf{V}\mathbf{V}^T$, $\mathbf{R} \leftarrow \mathbf{D}\mathbf{V}^T$;

  for m = 1 to M:

    $\boldsymbol{u}_m \leftarrow \mathbf{0}$;

    while deviation > 1×10^{-6}

      for k = 1 to K:

        $u_{mk} \leftarrow \frac{\left(\left|r_{mk} - \sum_{l \neq k} s_{kl} u_{ml}\right| - \frac{1}{2}\lambda_1\right)_+ \cdot sign(r_{mk} - \sum_{l \neq k} s_{kl} u_{ml})}{s_{kk}}$

      end for

    end while

  end for

  $\mathbf{V} \leftarrow (\boldsymbol{U}^T\boldsymbol{U} + \lambda_2 \boldsymbol{I})^{-1}\boldsymbol{U}^T\boldsymbol{D}$;

Projection:

Given: term-based essay vector $\boldsymbol{v}_q = [v_{q1}, v_{q2}, ..., v_{qM}]$;

Compute the projection as follows:

$\boldsymbol{v}_q^* = (\boldsymbol{U}^T\boldsymbol{U} + \lambda_2 \boldsymbol{I})^{-1}\boldsymbol{U}^T\boldsymbol{v}_q$.

Fig. 2. The RLSI learning algorithm for automated essay scoring.

C. Scoring model training

After the topic features have been extracted, the scores by human rater (*human scores*) are added to each essay in the training set as classification labels. In our method, the automated scoring is regarded as a multi-class classification problem, and thus, we can use both the human scores and the topic features extracted to train the scoring model.

Since Corinna Cortes and Vladimir Vapnik introduced Support Vector Machines (SVM) [19], it has been a powerful solution to such problems for its accuracy and effectiveness in a wide range, especially for document understanding [20-21]. Therefore, we choose SVM to train the scoring model, and predict the scores of the essays in the testing set. In this paper, we use RBF kernel to map the essay vectors into a higher dimension space:

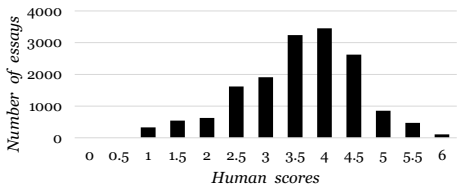
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right) \quad (12)$$

where γ is a parameter that needed to be tuned in our experiment, and $\boldsymbol{x}_i$ is the i-th topic-based essay vector.

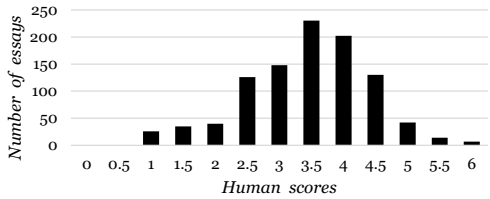
III. EXPERIMENTS

A. Experimental Settings

In the experiments, we use the written essays assigned by a same question and their human scores in MHK tests as the datasets. Omitting the empty essays and those without human scores, we obtain a collection with the size of 16,776. We randomly hold out 1,000 essays as the testing set. Fig. 3 shows the human score distribution of the training set and the testing set.



(a) Training set (total 15,776 essays)



(b) Testing set (total 1,000 essays)

Fig. 3. Human scores distribution of the dataset.

In text pre-processing, we generate a term list from the training set, including 12,974 terms. The RLSI learning program for topic feature extraction is written in C programming language. In scoring model training, we utilize LIBSVM [22] to classify the essays as automated essay scoring. All the programs run on a Linux machine with the CPU 2.9GHz Intel Xeon E5-2690 and 256G memory.

B. Performance Index

Automated essay scoring involves two kinds of performance indices between the human raters and the scoring system: *exact agreement* and *adjacent agreement* [2]. A more general method to measure a scoring system is to use adjacent agreement, which means that minor bias between the predicted score and the human score is acceptable. We call that *rating agreement*.

According to the scoring scale (0 to 6) and the scoring standard of MHK, we set 1 to the acceptable bias, because in reality, when the difference of the scores by two raters is larger than one point, the essay will be sent to the third rater for re-scoring. For different tests and scoring scales, this value can be set by the testers. We propose the rating agreement equation as follows:

$$R.A. = \frac{\sum_{n=1}^{N} 1\{|hs_i - ps_i| \le 1\}}{N} \tag{13}$$

where *hs* and *ps* are the human score and the predicted score respectively, and N is the size of the testing set. The indicator function $1\{\cdot\}$ is used as follows:

$$1\{True\} = 1 \quad and \quad 1\{False\} = 0.$$

All the following experiments are executed in this way: we execute the automated scoring three times and apply the equation (13) to get the rating agreement respectively, using cross-validation to select the parameters C and γ in SVM scoring model to perform best. Then these three rating agreements are averaged as the final results.

C. Parameter Selection

Our target is to construct a scoring model that achieves the best performance, namely, the highest rating agreement. For different datasets, the parameters of models are supposed to be different as well, so they cannot be regarded uniformly. We select the parameters during the RLSI training and the construction of the scoring model respectively.

There are two main parameters in RLSI: the topic number K and the $\ell 1$-norm regularization parameter λ_1. The $\ell 2$-norm regularization parameter λ_2 is trivial since it is imposed on the matrix **V** which is less useful than the matrix **U**, so we set it to 1. Besides, the number of iterations T has little impact on the result, so we set it to 100.

In scoring model training, we use cross-validation to search for the best parameters, C and γ, for each experiment round, according to the SVM optimization equation and equation (12).

1) Topic number

As previously stated, for different applications, the topic number K varies as well. Generally, K is set to 50 to 100 empirically in the field of information retrieval, and the best K is decided by some evaluation formulae. For automated Chinese essay scoring, we determine the best K by rating agreement.

In the experiment, we manipulate the topic number from 25 to 200 at the interval of 25, and Fig. 4 is the experimental results. In Fig. 4, we determine that the rating agreement is best when the topic number is 175. This result validates our assumption that there exist several topics in an essay collection. Especially, the topic number is larger than the empirical value, for there exist more synonyms and ambiguities in Chinese than western languages.

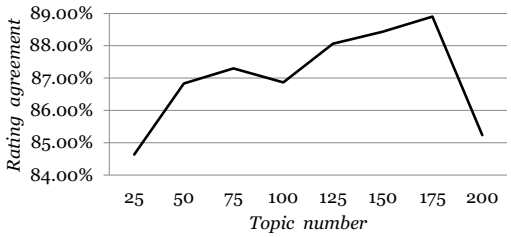


Fig. 4. The selection of topic number.

2) ℓ1-norm regularization parameter λ_1

In RLSI, λ_1 is imposed on the term-topic matrix **U**, which means that it controls the sparsity of the matrix **U**, according to the equation (5) where the hinge function is applied. Additionally, each essay vector is projected to the topic space by the equation (9), meaning that the matrix **U** controls the representative level of the topic space as well.

In our experiments, we set λ_1 from 0.1 to 1.0 at the interval of 0.1. The results are shown in Fig. 5.

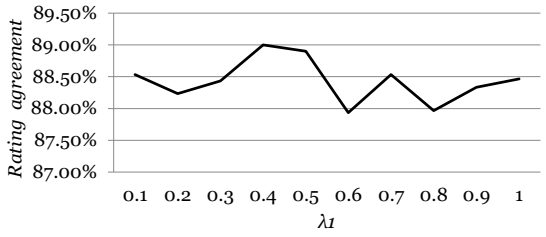


Fig. 5. The selection of λ_1

From Fig. 5, we can see that at the point where λ_1 is 0.4, the rating agreement reaches 89%. This value can be tuned according to different datasets.

D. Results and Discussions

1) Topics discovered by RLSI learning

Firstly, we randomly list five latent topics with their keywords discovered by RLSI learning from 175 topics in the training set, which are shown in Fig. 6.

topic 1	*topic 2*	*topic 3*	*topic 4*	*topic 5*
耐心 patience	成功 success	爸爸 dad	美 beauty	朋友 friends
老师 teacher	了解 realize	知足 happy	机会 chance	优点 merit
儿子 son	勇敢 brave	学会 learn	想法 idea	缺点 weakness
社会 society	知足 happy	生活 life	目标 target	选择 choose
想法 idea	面对 face with	信心 faith	诚实 honest	美 beauty

Fig. 6. Topic words discovered by the RLSI learning.

2) Automated scoring distribution

Secondly, we display the automated scoring distribution in Fig. 7, showing that the SVM-based scoring model recognizes the topic features in the written essay collection and predicts the scores successfully. In this result, we average the predicted scores from three round experiments. The concentration in the score 3.5, however, shows the limitation of conventional SVM, though the result is relatively satisfying.

[1] available at `http://cs.princeton.edu/~blei/lda-c/`

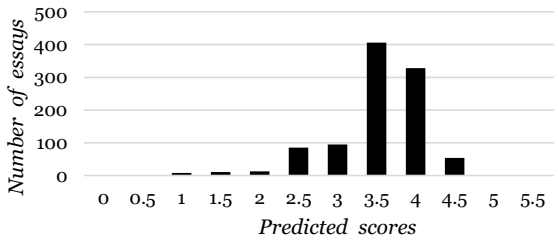


Fig. 7. The automated scoring distribution.

3) Comparison with LDA

Thirdly, in order to make it clearer that RLSI outperforms LDA in the field of automated Chinese essay scoring, we use the LDA tool designed by David M. Blei [1] to extract the topic features.

We initially set α = 1.0, then for each topic number from 25 to 200 at the interval of 25, we use cross-validation to tune the parameters C and γ in the SVM-based scoring model and get the rating agreements. Fig. 8 shows the comparison between LDA and RLSI. From Fig. 8, we can see that RLSI is more suitable than LDA in the field of automated Chinese essay scoring.

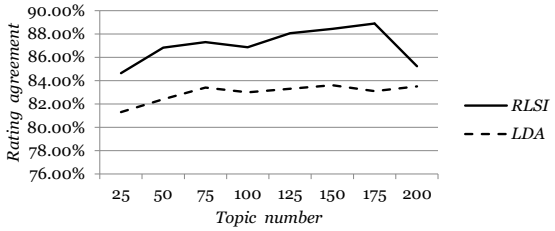


Fig. 8. The comparison between LDA and RLSI.

4) Error Analysis

For the result performed by RLSI, we discuss the possible reasons which lead to erroneous scoring. In the text pre-processing, the written essays are scanned into computer, and someone type the characters into text files manually, as shown in Fig. 1. Then, as previously introduced, we use WFST to segment and eliminate the stopwords. In each of above phases, the errors occur and accumulate, leading to the impreciseness of matrix samples before RLSI learning. Though more advanced technique has been implemented to reduce the errors, they cannot be avoided completely. In addition, Chinese is a language where much more synonyms and ambiguities exist than many western languages do, leading to more noises in feature extraction.

Under the current settings we used in this paper, automated Chinese essay scoring reaches the relatively high

level. With the problems above solved, the rating agreement will get higher, and then scoring from topic perspective using RLSI can be applied effectively.

5) Further discussions

Finally, in [12], the author proposed two RLSI algorithms: batch version and online version. Batch version is used to general tasks, whereas online version can be used to handle missing data and to discover the evolution of the topics. In automated essay scoring, the online version is relatively less useful than the batch version. The algorithm we use in this paper is the batch version without parallel computation, since the dataset we use is not too big. Now that the topic modeling in automated essay scoring is proved applicable according to our experiments, this batch version can be applied in large-scale tests, especially for Chinese language tests such as MHK whose scale is increasing rapidly.

IV. Conclusions and Future Work

In this paper, we implement automated Chinese essay scoring from a topic perspective, using Regularized Latent Semantic Indexing, and achieve the performance of rating agreement of 89%, demonstrating that our application is quite effective.

Considering that this is the first attempt in this field, we propose the future work as follows. First, the huge advantage of RLSI is parallel computation or scalability. As the scale of MHK grows larger continually, RLSI can help to reduce the computational time. Second, topic perspective is proved viable, and then finding out a method that scores the essays from a comprehensive view (combining language proficiency, structure, topic and so forth) will enable the scoring system to predict more genuine scores and achieve at a higher rating agreement.

Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities (No. XS2014023), the Beijing Higher Education Young Elite Teacher Project (No. YETP0768), the National Natural Science Foundation of China (No. 61103152) and the National 973 program in China (No. 2010CB328103). We thank the anonymous reviewers for their insightful comments.

References

[1] Mark D Shermis and Jill C Burstein, "Automated essay scoring: A cross-disciplinary perspective," Psychology Press, 2003.

[2] Semire Dikli, "An Overview of Automated Scoring of Essays," Journal of Technology, Learning, and Assessment, vol. 5, No.1, August 2006.

[3] Yanan Li, "Automated Essay Scoring for Testing Chinese as a Second Language," Beijing Language and Culture University, Beijing, 2006.

[4] Xingyuan Peng, Dengfeng Ke, Zhi Zhao, Zhenbiao Chen, Bo Xu, "Automated Chinese Essay Scoring Based on Word Scores," Journal of Chinese Information Processing, Vol. 26, No. 2, pp. 588-595, March 2012.

[5] Hengli Peng, "The minorities-oriented Chinese level test," China Examinations, pp. 57-59, October 2005.

[6] Dengfeng Ke, Xingyuan Peng, Zhi Zhao, Zhenbiao Chen, and Jinshi Wang, "Word-level-based automated chinese essay scoring method," In Proceedings of National Conference on Man-Machine Speech Communication, pp. 25-29, Xi'an China, 2011.

[7] Yiwei Cao and Chen Yang, "Automated chinese essay scoring with latent semantic analysis," Examinations Research, Vol. 3, No. 1, pp. 63-71, 2007.

[8] Mingqing Zhang, Shudong Hao, Dengfeng Ke, Hengli Peng and Yanyan Xu, "Automated Essay Scoring Using Incremental Latent Semantic Analysis," Journal of Software, accepted.

[9] Ge Xu and Houfeng Wang, "The Development of Topic Models in Natural Language Processing," Chinese Journals of Computers, Vol. 34, No. 8, pp. 1423-1436, August 2011.

[10] David M. Blei, Andrew Y. Ng and Michael I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol. 3, pp. 993-1022, January 2003.

[11] Quan Wang, Jun Xu, Hang Li and Nick Craswell, "Regularized Latent Semantic Indexing," In Proceedings of SIGIR' 11, pp. 685-694, 2011.

[12] Quan Wang, Jun Xu, Hang Li and Nick Craswell, "Regularized latent semantic indexing: A new approach to large-scale topic modeling," ACM Transactions on Information System, Vol. 31, No. 1, Article 5, 44 pages, January 2013.

[13] Shudong Hao, Zongtian Gao, Mingqing Zhang, Yanyan Xu，Hengli Peng, Kaile Su and Dengfeng Ke, "Automated Error Detection and Correction of Chinese Characters in Written Essays Based on Weighted Finite-State Transducer," In Proceedings of 12th International Conference on Document Analysis and Recognition, pp. 763-767, 2013.

[14] Preslav Nakov, Antonia Popova, and Plamen Mateev, "Weight Functions Impact on LSA Performance," In Proceedings of EuroConference RANLP, pp. 187–193, 2001.

[15] Jerome Friedman, Trevor Hastie, Holger Höfling and Robert Tibshirani, "Pathwise Coordinate Optimization," The Annals of Applied Statistics, Vol. 1, No. 2, pp. 302-332, 2007.

[16] Jerome Friedman, Trevor Hastie and Rob Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," Journal of Statistical Software, Vol. 33, No. 1, pp. 1-22, August 2010.

[17] Tongtong Wu and Kenneth Lange, "Coordinate Descent Algorithms for Lasso Penalized Regression," The Annals of Applied Statistics, Vol. 2, No. 1, pp. 224-244, 2008.

[18] P. Pucar and J. Sjöberg, "On the Hinge Finding Algorithm for Hinging Hyperplanes," IEEE Transactions on Information Theory, Vol. 44, No. 3, pp. 1310-1319, May 1998.

[19] Corinna Cortes and Vladimir Vapnik, "Support-Vector Networks," Machine learning, Vol. 20, No. 3, pp. 273-297, 1995.

[20] Min Li, Youngja Park, Rui Ma and He Yuan Huang, "Business Email Classification Using Incremental Subspace Learning," In Proceedings of 21st International Conference on Pattern Recognition, pp. 625 - 628, 2012.

[21] Jayant Kumar, Peng Ye and David Doermann, "Learning Document Structure for Retrieval and Classification," In Proceedings of 21st International Conference on Pattern Recognition, pp. 1558 - 1561, 2012.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.