Text Emotion Classification Research Based on Improved Latent Semantic Analysis Algorithm

Xuren Wang
College of Information Engineering
Capital Normal University
Beijing, China
wangxr@ihep.ac.cn

Qiuhui Zheng
College of Information Engineering
Capital Normal University
Beijing, China
15801413596@163.com

***Abstract*—The emotion classification of text is an important research direction of text mining. Application on emotion text classification, latent semantic analysis algorithm has advantage of small occupied space, applicable to a large scale of text classifications. Compared with the traditional vector space model, latent semantic analysis algorithms reduce the search space for text classification by means of singular value decomposition for term and document matrix. Moreover, latent semantic analysis algorithms solve the problem of words with multiple meanings by analyzing the term at the semantic level. Using an improved latent semantic analysis algorithm to classify the test set by their emotion. The new cluster centroid is the average vector for each emotion category, and access to emotions classification for training dataset by calculating similarity of the average vector and test textual. The experimental results show that the improved latent semantic analysis algorithm have high precision and recall rate as same as the original algorithm, the efficiency of text emotion classification improved 4 percentage points.**

Keywords-Latent Semantic Analysis; Vector Space Model; Text Emotion Classification;

I. INTRODUCTION

Nowadays, vector space model (VSM) is also used to expressing information in text classification. This model can applies to any language theoretically which can split into words. Besides, documents are represented as vectors that the vectors consist of several keywords [1]. More clearly, this model represents unstructured text as vector, so that the text processing can be done by using computer technology and variety of mathematical methods.

The basic idea of traditional vector space model is that text is represented as words elements of vector. Similarity can be generated by calculating cosine value between the two vectors and that also used to text classification [2]. Besides, the traditional vector space model applies to text classification by using the exact word matching between user's inputs and the words in vector space. Moreover, the traditional vector space model suffers from two classical problems of synonymy and polysemy.

LSA(Latent Semantic Analysis), also known as LSI(Latent Semantic Index), is an indexing and retrieval model proposed by Scott Deerwester and Susan T. Dumais et al in 1988 [3]. This algorithm with the traditional vector space model, use vectors to represent the document, and generate vectors and documents similarity by calculating their cosine value. But with the difference of traditional vector space model, latent semantic analysis algorithm map the word and document matrix to latent semantic space, which improve accuracy of the information retrieval and remove some noise in original vector space [4]. Technically, LSA algorithm are based on space vector represent text as same as vector space model, but it reduce the dimension of the term-document matrix by SVD (Singular Value Decomposition), to eliminate the synonyms and polysemy, improve the accuracy of subsequent processing, achieve the purpose of remove noise and information filtering [5].

In this paper, we propose a method that uses an improved LSA algorithm for text emotion classification on ISEAR (International Survey on Emotion Antecedents and Reactions) dataset. The training set according to text sentiment classification random extraction, constructed the term-document matrix and get left singular matrix, singular values vector and right singular matrix after singular value decomposition. Besides, test dataset emotion classification completed by using improved latent semantic analysis algorithm. More specifically, right singular matrix clustered by emotion category and the new cluster centroid is the average vector for each emotion category; completed emotion classification for test set and compares with the standard latent semantic analysis algorithms. The experimental results show that the improved latent semantic analysis algorithms have high precision and recall as same as the original algorithm, the efficiency of text emotion classification improved 4 percentage points.

The rest of this paper is organized as follows. In the next section, we introduce more detail about mainly technologies and evaluation methods about emotion classification of text. Section 3 presents the key ideas of latent semantic analysis and our improved algorithm. Section 4 describes our experiments on ISEAR dataset and presents the result compare with the standard LSA algorithm. The last section summarizes our conclusion and future work.

II. METHOD AND EVALUATION FOR EMOTION CLASSIFICATION OF TEXT

In the text classification system based on vector space mode, text is represented as words elements of vector. Each word is given a weight according to its frequency [6]. Similarity can be generated by calculating cosine value between the two vectors. Besides, the similarity is computed by formula (1).

$$similarity(q,d) = \cos(q,d) = \frac{\sum_{j=1}^{m} W_{qj} W_{dj}}{\sqrt{\sum_{j=1}^{m} W_{qj}^{2}} \sqrt{\sum_{j=1}^{m} W_{dj}^{2}}} \quad (1)$$

where $q=(w_{q1},w_{q2},\ldots,w_{qm})^T$, $d=(w_{d1},w_{d2},\ldots w_{dm})^T$, m is the sum of words for the text set.w_{qj} is the weight of words in w_{dj}, and w_{dj} is the weight of query in d_j.

The main steps of emotion classification of text based on vector space model as follows:

- According to emotion categories of text, test set and training dataset generated by random extracting.
- Training set and test dataset split into words, and represent as words elements of vector.
- Emotion classification of test set completed by calculating similarity between test vector and training vector.

The evaluation measures of text classification are similar to information retrieval, which contains precision and recall.[6]More specifically, precision is the number of correct classification of texts over the actual classification of texts, while recall is the number of correct classification texts which using text classification algorithm over the all of the classification texts. Besides, the precision, recall and F1 are computed by formula (2), formula (3) and formula (3).

$$\text{Precision} = \frac{\text{correct classification of text}}{\text{actual classification of texts}} \quad (2)$$

$$\text{Recall} = \frac{\text{correct classification texts}}{\text{all of the classification texts}} \quad (3)$$

$$\text{F1} = \frac{\text{Precision} * \text{Recall} * 2}{\text{Precision} + \text{Recall}} \quad (4)$$

III. Latent Semantic Analysis Algorithm and Improvement

A. The key technologies of Latent Semantic Analysis Algorithm

LSA algorithm is an information retrieval algebraic model proposed by Scott Deerwester and Susan T. Dumais et al. The paper published has over twenty years, but there are still a large number of references. After twenty years of development and evolution, LSA has been quite comprehensive and mature. LSA algorithm is different from traditional natural language processing and artificial intelligence program; it does not use artificial dictionary, knowledge base and the semantic web. LSA algorithm consider inputs as the original text ,and the algorithm basic idea is map the term-document matrix to latent semantic space by using singular value decomposition [7].

In fact, the LSA algorithm is a dimension reduction method for the VSM. This method is based on singular vector decomposition to decompose term-document matrix into left singular matrix, singular values vector and right singular matrix [8].

TABLE I. Term-Document Matrix

	d1	d2	d3	……	dn
t1	0	0	0	……	1
t2	1	1	0	……	1
……	……	……	……	……	……
tn	1	1	1		1

Table 1 describes more detail about term-document matrix. The process of singular vector decomposition, described as formula (5) and formula (6).

$$A = (a_{ij})_{m \times n} \quad (5)$$

where A is an m*n term-document matrix, a_{ij} is the weight of the i-th word in j-th document. Formally, the singular value decomposition theorem states:

$$A = U \sum V^T \quad (6)$$

where the columns of U are the left singular matrix; Σ has singular values and is diagonal; and V^T has rows that are the right singular matrix.

Calculating the SVD consists of finding the eigenvalues and eigenvectors of AA^T and A^TA [9]. The eigenvectors of A^TA make up the columns of V, the eigenvectors of AA^T make up the columns of U. Also, the singular values in Σ are square roots of eigenvalues from AA^T or A^TA. The singular values are the diagonal entries of the Σ matrix and are arranged in descending order. The singular values are always real numbers. If the matrix A is a real matrix, then U and V are also real [10].

B. Improvement of Latent Semantic Analysis Algorithm

Thorough singular vector decomposition to decompose term-document matrix into left singular matrix, singular values vector and right singular matrix. Each word and document could represent as left and right singular matrix. The term-term matrix and document-document matrix are represented in a same space. Figure 1 is the m*n matrix singular vector decomposition, k is the dimension of the singular value vector.

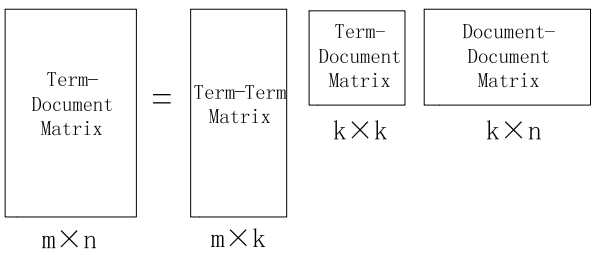

Figure 1. Term-Document Matrix Singular Vector Decomposition

In this paper, we propose a method that uses an improved LSA algorithm for emotion classification of text. More specifically, right singular matrix clustered by emotion categories and the new cluster centroid is the average vector for each emotion classification. For test dataset, it is convenient to get each emotion classification of text by calculating similarity of the average vector and test vector.

The value for average vector is computed by formula (7).

$$\bar{v} = \frac{1}{m}\sum_{i=1}^{m} X^{(i)} \tag{7}$$

where $\bar{v}$ denotes the average vector and the new cluster centroid; m is the number of texts for each emotion category in training set; the training set is represented as $\{X^{(1)}, X^{(2)}, \ldots X^{(m)}\}$.

Emotion classification for test dataset, first it is transformed into vector D_q by formula (8). X_q denotes the columns of test dataset and every column denotes a test textual; T denotes left singular matrix while S denotes right singular matrix after singular value decomposition; S^{-1} denotes the inverse matrix of S.

$$D_q = X_q T S^{-1} \tag{8}$$

For emotion classification of text, only calculate similarity between vector and vector by using improved LSA algorithm. More specifically, the value for similarity is computed by formula (9).

$$Sim(D_q, \bar{v}) = \frac{\sum_{j=1}^{m} W_{D_q j} W_{\bar{v} j}}{\sqrt{\sum_{j=1}^{m} W_{D_q j}^{2}}\sqrt{\sum_{j=1}^{m} W_{\bar{v} j}^{2}}} \tag{9}$$

IV. Experiments and Results

For the experiments, we use ISEAR (International Survey on Emotion Antecedents and Reactions) dataset [11]. This dataset consists of 7666 textual pieces tagged with the most appropriate of seven major emotions (joy, fear, anger, sadness, disgust, shame, and guilt).

A. Data preprocessing

ISEAR dataset provides Microsoft Access mdb file and Clementine SPSS sav file. In our experiments, we use mdb file that provides persistence data through ODBC (Open Database Connectivity). More importantly, the main task of data preprocessing is that remove non ACSII character. After this, ISEAR dataset have valid records for 7581.

TABLE II. Result of Data Preprocessing

Emotion	Number
Anger	1087
Disgust	1082
Fear	1090
Guilt	1076
Joy	1090
Sadness	1083
Shame	1073

B. Generation of Latent Semantic Space

For the experiments, we randomly select 60 for each category of emotions. The training set consists of 420 records with 7 emotions. Besides, we use GNU (GNU's Not Unix) C++ and Scientific Library in our environment of experiment.

C. Emotion Classification and Results

After constructed latent semantic space, we randomly select 100 for each category of emotions that consists of test dataset. Besides, Table 3 is the results of text emotion classification base on improved LSA algorithm for the test dataset.

TABLE III. Test Emotion Classification for Test Dataset Based on Improved LSA Algorithm

Emotion	Precision	Recall	F1
Anger	0.806	0.813	0.405
Disgust	0.744	0.712	0.364
Fear	0.736	0.791	0.381
Guilt	0.738	0.607	0.333
Joy	0.902	0.927	0.457
Sadness	0.916	0.943	0.465
Shame	0.703	0.727	0.357

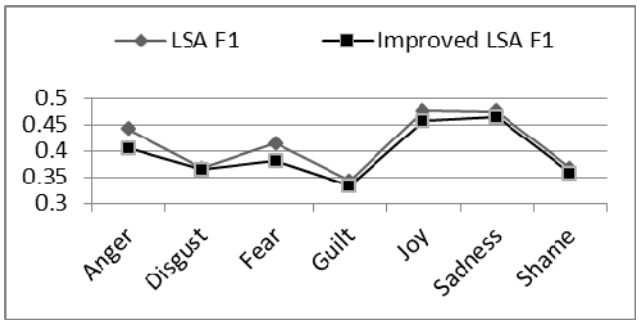

Figure 2. Compaaring the results of the improved LSA algorithm and standard LSA algorithm

Training set has 420 records while test dataset has 700 records. Besides, the term-document has 1515 rows and 420 columns. Moreover, the singular value dimension is 200. More specifically, the standard LSA algorithm consumes 4 hours and 41 minutes while the improved algorithm consumes 4 hours and 29 minutes. Figure 2 shows the comparison between improved LSA algorithm and standard LSA algorithm, which both have similar F1. More importantly, the experimental results show that the improved latent semantic analysis algorithms have high precision and recall rate as same as the original algorithm, the efficiency of text emotion classification improved 4 percentage points.

V. Conclusion and Future Work

We have presented our proposed improved LSA algorithm for emotion classification of text. On one hand, we use LSA algorithm to emotion classification of text and make detail analysis of algorithm principle. On the other hand, the proposed method could apply to emotion classification of text. Evaluation by precision, recall and F1 on the ISEAR dataset shows a promising result. Our

improved LSA algorithm improves 4 percentage points for efficiency of text emotion classification.

This method could be further improved by using a better dimension reduction method such as PLSA, and by taking into account more emotional information. Besides, we will also learn more efficiently from a spare training dataset.

ACKNOWLEDGMENT

This work is supported by Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality, PHR (IHLB), Program Number: PHR201008083.

REFERENCES

[1] *Wu L, Hoi S, C H, Yu, N. Semantics-Preserving Bag-of-Words Models and Applications [J]. IEEE Transactions on Image Processing, 2010,19(7):1098-1920.*

[2] Salton G, Wong A, Yang CS. A Vector Space Model for Automatic Indexing [J］.Communications of the ACM, 1995,18(11) : 613-620.

[3] Dumis S, Fumas G， Landauer T et al. Using Latent Semantic Analysis to Improve Access to Textual Information. Proceedings of Computer Human Interaction, 1988.217-285.

[4] Landauer T K, Foltz P W, Laham D. Introduction to Latent Semantic Analysis. Discourse Processes, 1998,25(2):259-284.

[5] Yungan Xuan, Qinghua Zhu. Research on Tag Semantic Retrieval Based on LSA in Social Tagging System. Ph.D. Thesis， Nanjing University,2011.

[6] Zhigang Chen, Peilian He, Yueheng Sun, Xiaoshen Sun. Research and Implementation of Text Classification Method Based on Vector Space Model. Journal of Computer Applications, 2004.

[7] Xiujie Dong, Lejian Liao. Test Analysis Based on LSA Algorithm, Master Thesis, Beijing Institute of Technology, 2008.

[8] Ziqiang Bao, Zhongyang Xiong. Spectral Clustering Research with LSA on Text Clustering. Master Thesis, Chongqing University, 2010.

[9] Yang Wang, Guisheng Yin. The Intelligent Search Technology Based on Latent Semantic Analysis. Master Thesis, Harbin Engineering University, 2010.

[10] Xiaofang Wu, Shimin Shan. Research on Semantic Analysis of Tag in Social Tagging System， Dalian University of Technology， 2011.

[11] Dung T.Ho, Tru H. Cao. A High-Order Hidden Markov Model for Emotion Detection from Textual Data. Springer,2012.