ATLANTIS PRESS
Advances in Engineering Research, volume 118
2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)

Analysis of large data classification based on knowledge element in micro-blog short text

Xia Wendong[1], Liu yuanfeng[2], Chen deli[1]

1. College of Computer,Jiaying University, Meizhou Guangdong , 514087, China

2. Guangdong Ji Tong Information Development Co., Ltd, Guangzhou Guangdong, 510632, China

Email: 89537987@qq.com

Keywords: behavioral habits; content correction; micro-blog short text; decision; figurative emotion; statistical model

Abstract. A statistical analysis method of figurative emotion of decision on micro-blog short text based on content correction is proposed in the Thesis. Based on emotional analysis and statistics of micro-blog short text, model of figurative method on micro-blog short text based on emotion is created when using decision and corrected by combining with content statistic model. Finally, performance advantages of the proposed algorithm on cosine similarity index are verified by experiments. Classification of the highest similarity pattern of micro-blog short text is realized, and it is corrected by using method of content model. At last, performance advantages of the proposed method on figurative statistics and analysis is verified by experiments.

1. Introduction

In analysis of public opinion, emotional analysis is a difficult task, and its goal is to find a corpus emotion without direct communication, which can also be called opinion mining or emotional extraction [1-2]. Through emotional analysis of micro-blog short text, lot of useful short text [3-4] can be obtained: for example, in e-commerce, the company can promote products through website, blog or social network. Each transaction is made online, and whenever new product short text is published, people will immediately view the short text and leave a comment to express their opinions. Therefore, the role of emotional analysis in network short text mining is becoming more and more important [5].

2 Evaluation of method model based on content

Content-based approach is based on conditional statistics, and use training set to predict score of given micro-blog. Output of the model is scored within the [-5, 5] scoring range.

Micro-blog short text t_k required comments should be firstly subjected to extraction of terms; the form is

$$T_k = \cup\{w_i \mid w_i \in t_k\}_{i=1}^{m_k} \tag{1}$$

In the equation, T_k refers to term set extracted from micro-blog short text t_k; w_i refers to the term belonging to micro-blog short text t_k; m_k refers to number of terms that are extracted from micro-blog short text t_k.

All possible combination can be built by using micro-blog short text t_k. Each combination represents a type of co-occurrence term. Based on co-occurrence term set, the combination can express a specific meaning. All combined terms can be obtained by: (1) considering all possible co-occurrence terms in T_k; (2) calculating distribution of micro-blog scoring in training set according to a given term. It is considered that each combination is a cluster. The number of possible clusters can be calculated by the following equation:

$$C_k = \left\{(\delta_i)_{i=1}^{\gamma_k} \mid \gamma_k = \sum_{j=1}^{m_k}\binom{m_k}{j}\right\} \tag{2}$$

Where, C_k refers to a cluster set; δ refers to a cluster. Each cluster represents a feature vector;

γ_k refers to possible number of combinations created by T_k; definition of m_k is the same as above.

Example 1: For a given set of terms $T_k = \{A,B,C\}$, the set of terms is extracted in micro-blog short text; according to T_k, all clusters owned are:

$$C_k = \{A,B,C\};\{A,B\};\{A,C\};\{B,C\};\{A\};\{B\};\{C\} \quad (3)$$

Each cluster in C_k can be expressed as a feature vector, and its dimension is equal to number of terms in T_k.

$$\vec{\delta} = \{\delta_1,\delta_2,\delta_3,\cdots,\delta_{m_k}\} \quad (4)$$

Example 2: According to cluster set C_k, list of feature vector can be obtained: $(A,B,C) = \{1,1,1\}$; $\{A,B\} = \{1,1,0\}$; $\{A,C\} = \{1,0,1\}$; $\{B,C\} = \{0,1,1\}$; $\{A\} = \{1,0,0\}$; $\{B\} = \{0,1,0\}$; $\{C\} = \{0,0,1\}$.

Each micro-blog short text in Z can be expressed as a vector and clustered into a corresponding cluster in C_k. Micro-blog short text is assigned to a cluster, and the following conditions must be met: (1) distance from a micro-blog short text to a cluster must be the minimum distance with other micro-blog short text. (2) The distance must be less than a given threshold. The distance between micro-blog short text and cluster can be calculated based on the following equation:

$$dis(t_k,\delta) = 1 - \frac{\sum_{i=1}^{m_k}(t_{k_i} \times \delta_i)}{\sqrt{\sum_{i=1}^{m_k}(t_{k_i}^2)} \times \sqrt{\sum_{i=1}^{m_k}(\delta_i^2)}} \quad (5)$$

Where, $dis(t_k,\delta)$ refers to distance from micro-blog short text t_k to cluster δ; definition of m_k is the same as above.

Each cluster has a cluster coefficient that can be calculated based on number of feature items for that cluster. The cluster coefficient is used to display similarity between micro-blog short text in cluster and given micro-blog short text to be analyzed. The higher the similarity among terms is, the greater the value of cluster coefficient obtained is. Therefore, the following equation is defined here to calculate cluster coefficient values:

$$C_\delta = \lambda^{\delta_k} \quad (6)$$

Where, C_δ refers to cluster coefficient value; δ_k refers to number of feature items in a given cluster. In order to obtain best λ value in Equation (6), data set experiment is used to carry out analysis of λ value, as shown in Fig. 1.

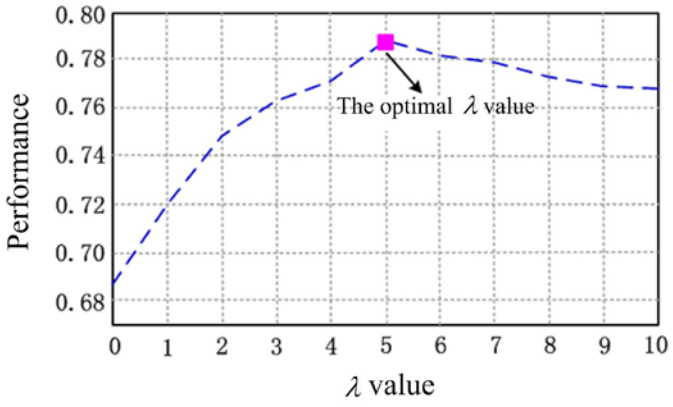

Fig. 1 Content- base Performance Analysis

According to Fig. 1, it can be seen that content-based method module behaves optimally when $\lambda = 5$ and then gradually decreases. This coefficient is reasonable for micro-blog short text, because it has the following characteristics: (1) due to limited length of micro-blog short text, difference between cluster coefficient is not significant. (2) C_δ is a non-linear function that characterizes importance of clusters by considering number of feature items in the cluster.

Example 3: Table 1 shows list of cluster: $\{A,B,C\},\{A,B\},\{A,C\},\{B,C\},\{A\},\{B\},\{C\}$ and its corresponding cluster coefficient values.

Table 1 Cluster and its Cluster Coefficient

Cluster set	Number of feature term	Cluster coefficient
{A、B、C}	3	$C_\delta = 5^3 = 125$
{A、B }	2	$C_\delta = 5^2 = 25$
{A、C}	2	$C_\delta = 5^2 = 25$
{ B、C}	2	$C_\delta = 5^2 = 25$
{A }	1	$C_\delta = 5^1 = 5$
{ B }	1	$C_\delta = 5^1 = 5$
{ C }	1	$C_\delta = 5^1 = 5$

Then, according to scores of micro-blog short text in cluster and cluster coefficients, a histogram is established to represent scoring distribution in training set. Peak of the histogram characterizes optimum of micro-blog short text that may boast emotional score.

Example 4: There are six non-empty clusters. Cluster $\{A,B,C\}$ contains one micro-blog short text and its score $(<t_1,-4.0>)$; cluster $\{A,B\}$ contains two micro-blog short texts and its scores $(<t_2,-3.5>)$, $(<t_3,-3.0>)$; cluster $\{B,C\}$ contains three micro-blog short texts and its scores $(<t_4,-2.5>)$, $(<t_5,-2.0>)$, $(<t_6,0.0>)$; cluster $\{A\}$ contains one micro-blog short text and its score $(<t_7,-1.0>)$; cluster $\{B\}$ contains two micro-blog short texts and its scores $(<t_8,-0.5>)$, $(<t_9,-1.5>)$; cluster $\{C\}$ contains three micro-blog short texts and its scores $(<t_{10},0.5>)$, $(<t_{11},1.0>)$, $(<t_{12},-4.0>)$. Table 2 shows the above data with respective coefficients, and then data in Table 2 is shown in histogram in Fig. 2.

Table 2 Micro-Blog Short Text and its Coefficient

Micro-blog short text	Cluster set	Score	Cluster coefficient
tw_1	{A、B、C}	-4.0	125
tw_2	{A、B }	-3.5	25
tw_3	{A、B }	-3.0	25
tw_4	{ B、C}	-2.5	25
tw_5	{ B、C}	-2.0	25
tw_6	{ B、C}	0.0	25
tw_7	{A}	-1.0	5
tw_8	{B}	-0.5	5
tw_9	{B}	-1.5	5
tw_{10}	{C}	0.5	5
tw_{11}	{C}	1.0	5
tw_{12}	{C}	-4.0	5

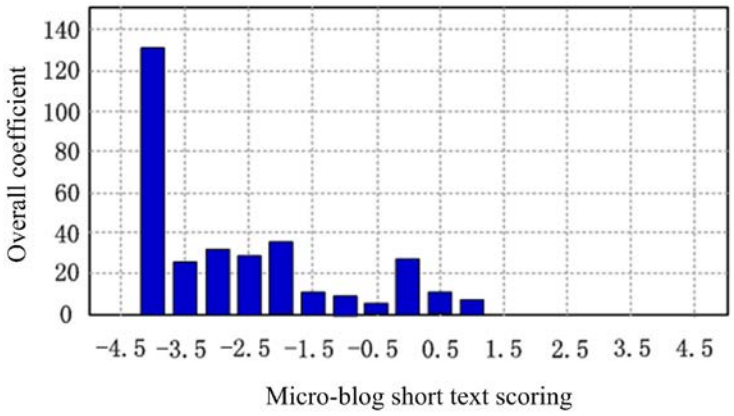

Fig. 2 Histogram of Score Distribution

According to histogram shown in Fig. 2, it can be seen that score of micro-blog short text based on content method is -4.0

3. Evaluation of method model based on emotion

Method model based on emotion mode is to use emotion mode to calculate score of micro-blog short text; this method is used to calculate emotion score of each term and then construct emotion distribution model of term scoring. Decision is used to achieve classification of the highest similar pattern of micro-blog short text.

Each term in training group has a score that expresses emotion of the term, such as positive, negative, or neutral emotion. Score of the micro-blog short text is calculated by using the following two attributes: (1) probability of occurrence of other terms in micro-blog short text; (2) scores of all micro-blog short text including the term. The final score of the term is within the interval of $[-5 \sim 5]$, and according to each micro-blog short text, the following equation can be used to calculate score of the term. In initial stage, score of the term can be obtained by calculating probability $P(S|w)$ that is probability of the highest score of the term :

$$S_w^i = \frac{S_w^{i-1} \times P(S_t \mid w)}{\sum_{j=1}^{n} (S_{w_j}^{i-1} \times P(S_t \mid w_j))} \times S_t \tag{7}$$

Where, S_w^i refers to score of term in Step i; S_t refers to score of micro-blog short text; $P(S_t \mid w)$ refers to probability of score belonging to score for given micro-blog short text; n refers to number of terms in micro-blog short text. The above Step is repeatedly executed until scores of step i and step i-1 are less than the given threshold.

It is assumed that all micro-blog short texts show similar importance, score of term is average value of all scores, namely:

$$S_w = \frac{\sum_{i=1}^{n} S_{wi}}{n} \tag{8}$$

Where, S_w refers to belonged score; interval for taking value is [-5,5]. n refers to number of micro-blog short text that contains the term.

4. Comparison and analysis of experiment

Data of the above micro-blog short texts obtained is divided. Because there are privacy problems, some micro-blog short texts can not be downloaded; a total of 4927 micro-blog short texts are divided into two parts. Data set 1 includes 927 micro-blog short texts, and data set 2 includes 4000 micro-blog short texts. Data set 1, including only figurative micro-blog short text, is used to assess figurative language recognition ability. Data set 2 includes figurative and nonfigurative of two kinds of micro-blog short texts. Figurative emotion analysis of content-based micro-blog short text, figurative emotion analysis of decision and the algorithm are selected for comparison algorithm; comparison results are shown in Fig. 3.

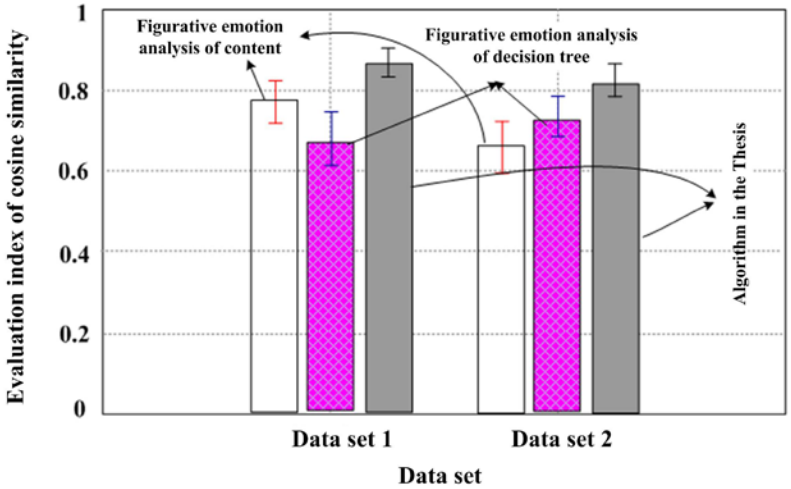

Fig. 3 Contrast of Algorithm Based on Cosine Similarity

According to comparison shown in Fig. 3, it can be seen that in cosine similarity index, this algorithm is superior to the cases that figurative emotion analyses of content-based micro-blog short text and decision are selected separately. In the data set 1, for figurative language, effect on figurative emotion analysis of content-based micro-blog short text is better than that of decision; in data set 2, for mixed short text, effect on figurative emotion analysis of decision is better than that of content-based micro-blog short text. At the same time, in stability of the algorithm, this algorithm is better than the two algorithms selected; in data set 1, for figurative language, stabilization effect on figurative emotion analysis of content-based micro-blog short text is better than that of decision; in data set 2, for mixed short text, stabilization effect on figurative emotion analysis of decision is superior to that of content-based micro-blog short text.

Comparison between results of terms in data set 1 and data set 2 and actual score is shown in Fig. 4-5. The algorithm in Literature[4] is selected.

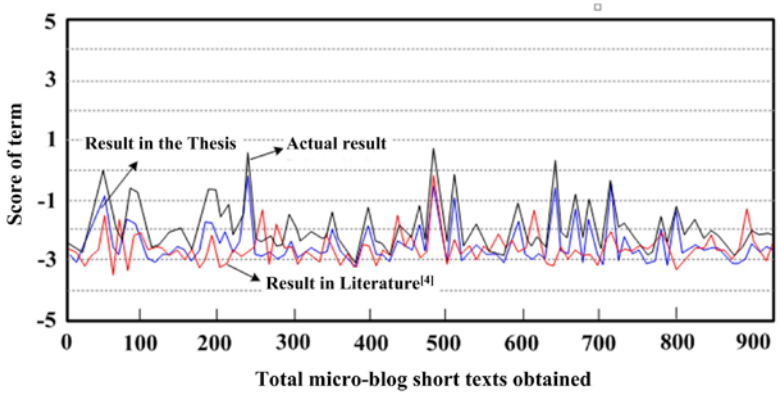

Fig. 4 Comparison of Term S

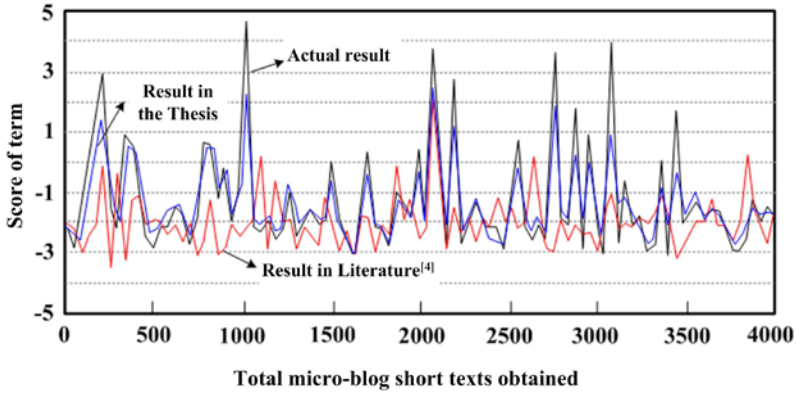

Fig. 5 Comparison of Term Scoring in Data Set 2 coring in Data Set 1

It can be seen from Fig. 4 and Fig. 5 that in comparison result of term scoring, the results of term scoring in this Thesis are closer to the real score of the term than that in Literature[4], which embodies advantage of proposed algorithm in term scoring result.

5. Conclusion

Problems to be further solved in this research are: (1) there is no nonfigurative micro-blog short text in training set. Performance of the algorithm can be improved by adding more nonfigurative micro-blog short text during learning process. (2) Here attention should only be paid to analysis of emotion mode of word (unigram model). In the next research, analysis of all pattern emotional patterns on micro-blog short text should be considered, so it is more difficult. (3) Structure of training set is still rough, and there are strong noise and disturbance, so some redundant information should be removed to improve training accuracy.

Acknowledgement

Guangdong Provincial Department of science and technology in the province of 2014 frontier and key technology innovation (2014B010117002); Guangdong Provincial Academy of Sciences comprehensive strategic cooperation (2013B091500060); Guangdong professional town small and micro enterprise service platform construction project (2013B040500010).

Reference

[1] Xin M, Wu H, Li W, et al. A public opinion classification algorithm based on micro-blog text sentiment intensity: Design and implementation[J]. International Journal of Computer Network & Information Security, 2011, 3(3).

[2] Yin C, Xiang J, Zhang H, et al. A New SVM Method for Short Text Classification Based on Semi-Supervised Learning[C]// International Conference on Advanced Information Technology and Sensor Application. 2015:100-103.

[3] Xu Y. Application of knowledge gain on multi-type feature space in microblog user classification[C]// IEEE International Conference on Granular Computing. IEEE, 2014:340-345.

[4] Wu K, Ma J, Chen Z, et al. Analysis of Subjective City Happiness Index Based on Large Scale Microblog Data[M]// Web Technologies and Applications. Springer International Publishing, 2015:365-376.

[5] Zhao S, Yao H, Zhao S, et al. Multi-modal microblog classification via multi-task learning[J]. Multimedia Tools and Applications, 2016, 75(15):8921-8938.