Acta Technica *62* (2017), No. 6B, 433–444

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

Yindi Dong[1]

Abstract. In order to solve the problem that traditional emotion analysis method for microblog information adopting vocabulary target identification is unsatisfactory for identifying concrete concept application, a recommended system mixing behavioral habits of users and microblog user labels is proposed. Firstly, use content-based method to build statistic model for emotion analysis of microblog information and carry out experimental analysis for cluster coefficients of emotion analysis so as to acquire the best selected value of cluster coefficients; secondly, use decision classifier to build emotion-based concrete method model for microblog information so as to realize the highest similar mode classification of microblog information and also use content and method model to rectify it; finally, verify the performance advantage of proposed method in concrete statistic analysis through experiments.

Key words. Behavioral habit, Content rectification, microblog information, Decision-making tree, Concrete emotion, Statistic model

.

1. Introduction

In public feelings analysis, emotion analysis is a hard mission for it aims to find a corpus emotion which does not have direct communication; it can also be called opinion excavation or emotion extraction [1, 2]. A lot of useful information can be obtained through emotion analysis of microblog information [3, 4]: for example, in e-commerce aspect, companies can popularize their products through website, blog or social network. Every trading activity is carried out on the Internet and each time when new product information is published, people will check these pieces of information immediately and leave comments to show their opinions. Thus, emotion analysis is playing a more and more important role in network information mining [5].

[1]College of Information Engineering, Chongqing City Management College, Chongqing 401331, China

The above researches emphasize on solving problem interest and focus on vocabulary level. Therefore, in this research, the purpose is to find a new method to identify image concept. One statistical method is used to provide a common model which can be easily expanded to description of graphic devices of other types.

2. Method model

2.1. Content-based method model

Basically, content-based method is based on condition statistics and it uses training set to predict score of given microblog. The output of this model is microblog scores within score range of [-5, 5].

Carry out term extraction for microblog information t_k which needs to be noted firstly; the form is:

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

Where, T_k is term set extracted from microblog information t_k; w_i is a term of microblog information t_k; m_k refers to the number of terms extracted from microblog information t_k.

By using microblog information t_k, all possible combinations can be created. Each combination represents one kind of co-occurrence term. Specific implication can be expressed based on co-occurrence term set. All combination terms can be obtained through the following method: (1) consider all conditions that co-occurrence term may occur in T_k; (2) calculate microblog score distribution in training set based on given terms. In this case, each combination is deemed as one set. Possible cluster number can be obtained through the following formula computing:

$$C_k = \left\{ (\delta_i)_{i=1}^{\gamma_k} | \gamma_k = \sum_{j=1}^{m_k} \left(\begin{array}{c} m_k \\ j \end{array} \right) \right\} . \tag{2}$$

Where, C_k is cluster set; δ is one cluster; each cluster is expressed as one feature vector; γ_k is the possible combination number created by T_k; the definition of m_k is the same as above.

Example 1: for the given term set $T_k = \{A, B, C\}$ which is extracted from microblog information t_k, based on T_k, all clusters owned by it are:

$$C_k = \{A, B, C\}; \{A, B\}; \{A, C\}; \{B, C\}; \{A\}; \{B\}; \{C\} . \tag{3}$$

Each cluster in C_k can be expressed as one feature vector and its dimensionality is equal to term number in T_k.

$$\boldsymbol{\delta} = \{\delta_1, \delta_2, \delta_3, \cdots, \delta_{m_k}\} . \tag{4}$$

Example 2: based on cluster set C_k, the following feature vector lists can be got: $(A, B, C) = \{1, 1, 1\}$; $\{A, B\} = \{1, 1, 0\}$; $\{A, C\} = \{1, 0, 1\}$; $\{B, C\} = \{0, 1, 1\}$;

$\{A\} = \{1, 0, 0\}$; $\{B\} = \{0, 1, 0\}$; $\{C\} = \{0, 0, 1\}$.

Each microblog information in Z can be expressed as one vector and they will assemble into corresponding cluster in C_k. After being allocated in one cluster, microblog information shall meet the following conditions: (1) the distance between one microblog information and one cluster shall be the smallest distance between this microblog information and other microblog information. (2) This distance shall be smaller than the given threshold value. The distance between microblog information T_k and cluster can be calculated through the following equation:

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

Where, $dis(t_k, \delta)$ refers to the distance between microblog information t_k and cluster δ; the definition of m_k is the same as above.

Each cluster has one cluster coefficient which can be calculated based on the number of characteristic items in this cluster. Cluster coefficient is used to indicate the similarity between microblog information in this cluster and the given microblog information to be analyzed. The higher the similarity between terms is, the bigger the obtained cluster coefficient value will be. For this, the following equation is defined hereby to calculate cluster coefficient value:

$$C_\delta = \lambda^{\delta_k}. \tag{6}$$

Where, C_δ is cluster coefficient value; δ_k is the number of characteristic items in given cluster. In order to obtain the best λ value in equation (6), dataset experiment is used here to analyze λ value, as shown in Fig.1.

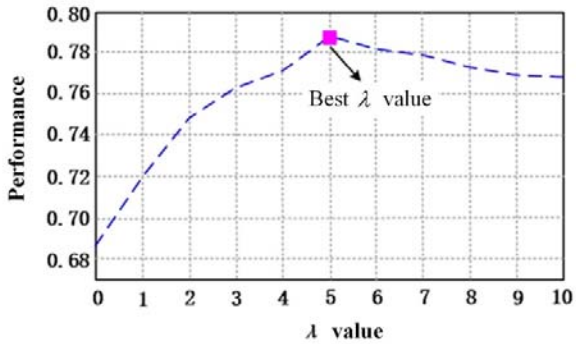

Fig. 1. Content-based performance analysis

According to Fig.1, content-based method module will have the best performance when $\lambda = 5$, then its performance will reduce gradually. This coefficient is reasonable for microblog information for it has the following characteristics: (1) since microblog information has limited length, the difference among cluster coefficients is not great. (2) C_δ is a nonlinear function which represents the importance of clusters through considering the number of characteristic items in clusters.

Example 3: the following cluster lists are given in Table 1: $\{A, B, C\}, \{A, B\}$, $\{A, C\}$, $\{B, C\}$, $\{A\}$, $\{B\}$, $\{C\}$; besides, their corresponding cluster coefficient values are also given.

Table 1. Clusters and their cluster coefficients

Cluster set	Number of characteristic terms	Cluster coefficient
{ABC}	3	$C_\delta = 5^3 = 125$
{AB }	2	$C_\delta = 5^2 = 25$
{AC}	2	$C_\delta = 5^2 = 25$
{ BC}	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, establish column diagram according to score of microblog information in clusters and cluster coefficients to express score distribution in training set. The peak value in column diagram refers to the possible optimal emotion score of microblog information.

Example 4: There are 6 non-empty clusters. Cluster $\{A, B, C\}$ includes one microblog information and its score $(< t_1, -4.0 >)$; cluster $\{A, B\}$ includes two pieces of microblog information and their scores $(< t_5, -2.0 >)$ and $(< t_6, 0.0 >)$; cluster $\{A\}$ includes one microblog information and its score $(< t_7, -1.0 >)$; cluster $\{B\}$ includes two pieces of microblog information and their scores $(< t_8, -0.5 >)$ and $(< t_9, -1.5 >)$; cluster $\{C\}$ includes three pieces of microblog information and their scores $(< t_{10}, 0.5 >)$, $(< t_{11}, 1.0 >)$, and $(< t_{12}, -4.0 >)$. Table 2 shows the above data and their respective coefficient; then data in Table 2 is shown in column diagram in Fig.2.

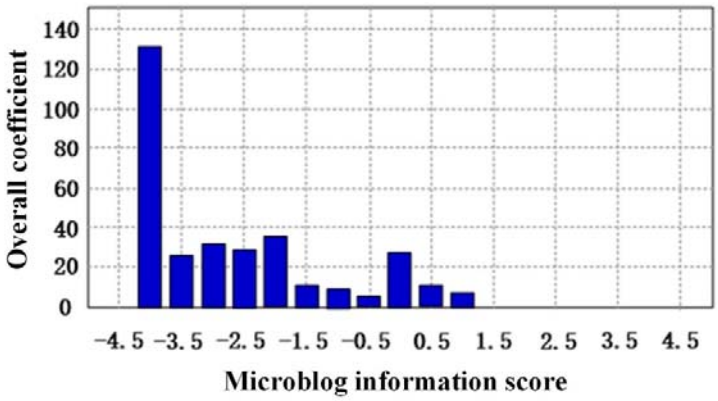

Fig. 2. Column diagram of score distribution

Table 2. Microblog information and their coefficients

Microblog information	Cluster set	Score	Cluster coefficient
$tw1$	{ABC}	-4.0	125
$tw2$	{AB}	-3.5	25
$tw3$	{AB}	-3.0	25
$tw4$	{BC}	-2.5	25
$tw5$	{BC}	-2.0	25
$tw6$	{BC}	0.0	25
$tw7$	{A}	-1.0	5
$tw8$	{B}	-0.5	5
$tw9$	{B}	-1.5	5
$tw10$	{C}	0.5	5
$tw11$	{C}	1.0	5
$tw12$	{C}	-4.0	5

According to column diagram in Fig.2, it can be known that microblog information score of content method is -4.0.

2.2. Emotion-based method model

Each term in training set has one score which indicates the emotion such as positive, negative or neutral emotion of that term. The score of this microblog information is calculated through using the following two properties: (1) occurrence probability of other terms in microblog information; (2) the score of all microblog information that include this term. The final score of this term is within $[-5 \sim 5]$; according to each microblog information, the following equation can be used to calculate term score. During initial period $\left(i^0\right)$, the score of this term can be obtained through calculating probability $P\left(S\,|w\right)$ which is the highest score probability of terms [12, 13]]:

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

Where, S_w^i refers to term score in the number i step; S_t is microblog information score; $P(S_t|w)$ refers to score probability of given microblog information; n refers to the number of terms in microblog information. Repeat the above Step until scores in the number i step and number $i-1$ step are smaller than the given threshold value.

Suppose that all pieces of microblog information are of same importance and the score of terms is the average value of all scores, namely:

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

Where, S_w refers to score; value taking interval is $[-5, 5]$. n refers to the number of microblog information that contain this term.

3. Decision-making tree classifier-based combination prediction for microblog message

3.1. Decision analysis for emotion score

For given microblog information dataset D, property space of this information is R^n; n refers to the number of information properties. Decision classifier DT can disintegrate R^n into Q different zones; the classification of different zone r_m is marked as $r_m.cl$. Decision classifier DT is equivalent to constant subsection mapping $f_{DT} : x \to r_m \cdot cl$ which can realize the establishment of mapping relation between microblog information sample $x \in D$ and corresponding zone r_m and also obtain label value $r_m \cdot cl$ under this situation.

There are two expression means for corresponding prediction zone r, constituent, and path structure of decision classifier; where, the path structure $r.p$ of prediction zone r can be described as follows:

$$r.p = \{\cap d(a_v), v = 1, 2, \cdots, K_r\} \ . \tag{9}$$

In equation (9), $d(a_v)$ refers to value taking interval of microblog information property a_v in prediction zone r; K_r refers to the number of involved points with prediction path among root nodes in zone r; operator $\cap$ refers to intersection and correlation conditions of quality supervision for different properties. Path structure $r.p$ can correlate decision classifier DT with prediction zone r and also describe rule property between zone r and root nodes.

In order to describe contents contained in microblog information set D which is contained in zone r, the constituent structure form of r is given here:

$$r.c = \{num(k_1), num(k_2), \cdots, num(k_J)\} . \tag{10}$$

In equation (10), J refers to the number of categories contained in microblog information set; $num(k_1), num(k_2), \cdots, num(k_J)$ are the numbers of samples in zone r which respectively belong to $k_1, k_2, \cdots, k_J$ and other different categories. Constituent structure $r.c$ indicates constituent relation between microblog information set D and prediction zone r.

For decision classifier DT_1 and DT_2 which are different in structure but still have relevance, it is appropriate to carry out similarity description based on affinity prediction probability of microblog information; for prediction probability $P(r)$ of microblog information, according to visit condition, it can be divided into prediction constituent probability $P(r.c)$ and prediction path probability $P(r.p)$; prediction

probability component can be expressed as:

$$P(r_m.p) = V(r_m.p) \Big/ \sum_{l=1}^{Q} V\left(r_l.p\right). \tag{11}$$

$$V(r_m.p) = \prod_{v=1}^{K_{r_m}} \left(|d(a_v)|/|dom(a_v)|\right). \tag{12}$$

$$P(r_m.c) = |r_m.c| \Big/ \sum_{l=1}^{Q} |r_l.c|. \tag{13}$$

In the above equation, equation (12) refers to hypervolume acquired after normalization operation for prediction zone r_m in property space R^n; $|dom(a_v)| = \max\left(a_v\right) - \min\left(a_v\right)$ refers to value taking interval of property a_v; $|d(a_v)| = \max(a_v) - \min(a_v)$ refers to value taking interval of microblog information property a_v in zone r_m; in equation (13), $|r_m.c| = \sum_{\rho=1}^{J} num\left(k_\rho\right)$ refers to the sum of all microblog information samples in zone r_m.

The obtained prediction value $P(r.p)$ of path probability by using equation (11) conforms to distribution consistency of properties. Therefore, under the condition that training set D is accessible, it is required to solve predicted probability value $P(r)$ according to equation (13). At the same time, equation (11) and equation (13) have only listed the component of prediction probability; the calculation form for all probability predictions is $P(r) = \{P\left(r_m\right) | m = 1, 2, \cdots, Q\}$.

After getting predicted probability value $P(r)$, the similarity expression form for decision classifiers of all types can be obtained based on the following equation:

$$\begin{aligned} S(DT_1, DT_2) &= s(P_{DT1}(r), P_{DT2}(r)) \\ &= \sum_{m=1}^{Q} \left[P_{DT1}(r_m) \cdot P_{DT2}(r_m)\right]. \end{aligned} \tag{14}$$

In equation (14), $s\left(\cdot,\cdot\right)$ refers to probability affinity expression which can represent similarity among different probability distribution and also meet the condition that $0 < s\left(\cdot,\cdot\right) < 1$. Therefore, value taking interval of $S(DT_1, DT_2)$ is also $(0, 1]$; the higher the similarity level of prediction probability between DT_1 and DT_2 is, the closer that $S(DT_1, DT_2)$ will be to upper limit 1; otherwise, the closer that $S(DT_1, DT_2)$ will be to lower limit 0. In case $P_{DT1}(r) = P_{DT2}(r)$, $S(DT_1, DT_2) = 1$.

Calculation process for decision analysis of emotion score is as follows:

Step1: (priori knowledge) train decision classifier DT_i by using dataset S_i in original data field of microblog information and use microblog information set T in target domain to train decision classifier DT_T.

Step2: successively determine the similarity between decision classifier DT_T and each microblog target information DT_i to obtain $S(DT_T, DT_i)$; in case dataset S_i in original data field of microblog information is accessible, then it will be feasible to predict constituent structure $r.c$ in microblog information zone r based on equation

(10) and to acquire probability value of prediction constituent by combining equation (13); otherwise, path structure $r.p$ can be got based on equation (13) and probability value of prediction path which is $P(r.p)$ can be obtained by combining equation (15).

Step3: carry out normalization operation for similarity $S_i(DT_T, DT_i)$ of different microblog objects and weight value ω_i can then be obatained; then allocate decision classifiers.

Step4: based on linear combination, decision classifier $DT_T = \sum_{i=1}^{N} \omega_i DT_i$ for emotion score can be obtained; output decision value of emotion score.

3.2. Content-based emotion score rectification

From each piece of microblog information in training set, one term list can be extracted; these terms can represent one model. Since the number of terms in each piece of microblog information is different, for each model, interpolation function can be used to resize the most possibility of microblog information in training set. Then, establish a vector space expression in which each dimensionality represents one matched model. Use decision-making tree classifier shown in section 3.1 in score prediction of emotion method model to carry out training of score prediction for microblog information.

Example 5: give microblog information t_k: "Some instagram photos are just so funny #sarcasm". In this example, we extract term lists and their respective scores: <some, -0.20>, <instagram, -0.20>, <photos, 0.05>, <just, -0.77>, <so, -0.60>, <funny, -0.19>, and <# sarcasm, -2.35>. Fig.3 shows the mode of the above data. In our training set, the biggest amount of terms which can be extracted from microblog information is 24.

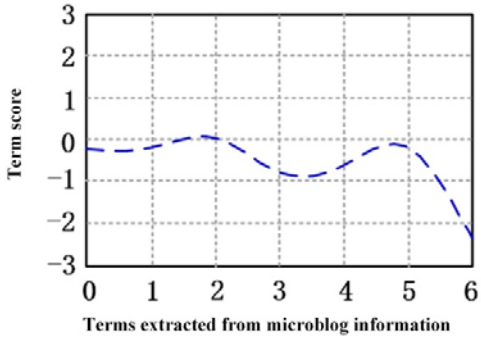

Fig. 3. Microblog information mode

Evaluation method for combined microblog information of content-based method and emotion mode-based modules is as follows:

$$S = \alpha \times SC + \beta \times SE\,. \tag{15}$$

Where, S refers to the final score of microblog information. SC is the calculated microblog information score of content-based method module; SE is the calculated

microblog information score of emotion-based method module; α and β refer to weight coefficients determined according to training errors of classification models of each method; and $\alpha + \beta = 1$.

4. Experiment comparison and analysis

4.1. Experiment settings

Generally, microblog information contains text, photo or video; this information is limited and it can contain 140 characters at most. In this Thesis, emphasis is laid on analysis of text emotion of microblog information. HTTP demand is used to obtain 8000 pieces of microblog information; ID lists are taken as training sets which have three types: 5000 sets of ironic information, 1000 sets of information with ironic meaning, and 2000 sets of metaphorical information. Due to the property of information, most ironic information, information with ironic meaning, and metaphorical information are negative. Microblog information can be expressed as the following model:

$$Z = \{< t, s > \mid s \in [-5, 5]\} . \tag{16}$$

Where, Z refers to one microblog information set in training set; t refers to microblog information; s refers to microblog information score. Term set extracted from Z can be expressed as:

$$T_z = \bigcup_{i=1}^{n} t_i = \bigcup_{i=1}^{n} \{\omega_j \,|\omega_j \in t_i \}_{j=1}^{m} . \tag{17}$$

Where, T_z is the term set extracted from Z; n refers to the number of microblog information in training set; ω_j is term; m refers to the number of terms extracted from Z.

Use cosine similarity as index to estimate the performance of proposed system. The value taking interval of scores measured from cosine similarity is [0,1], showing the similarity between our results and the predicted results. Firstly, express the above two results as vector form:

$$\begin{cases} R = \{r_1, r_2, \cdots, r_n\} , \\ E = \{e_1, e_2, \cdots, e_n\} . \end{cases} \tag{18}$$

Where, R refers to the result from algorithm in this Thesis; E is the expected result; n is the number of terms to be estimated. And the similarity index can be defined as:

$$sim\,(R, E) = \frac{\sum_{i=1}^{n} (R_i \times E_i)}{\sqrt{\sum_{i=1}^{n} (R_i^2) \times \sum_{i=1}^{n} (E_i^2)}} . \tag{19}$$

4.2. Performance evaluation

Divide microblog information data acquired in the above; due to the problem of privacy, some microblog information can not be downloaded; there are 4927 pieces of microblog information in total which have been divided into two parts; dataset 1 includes 927 pieces of microblog information, and dataset 2 includes 4000 pieces of microblog information. Dataset 1 only contains concrete microblog information which is used to evaluate concrete speech recognition capability. Dataset 2 contains both concrete and abstract microblog information. Comparing algorithms are selected as content-based image emotion analysis for microblog information, image emotion analysis for decision-making tree, and algorithm in this Thesis; comparison results are shown in Fig.4.

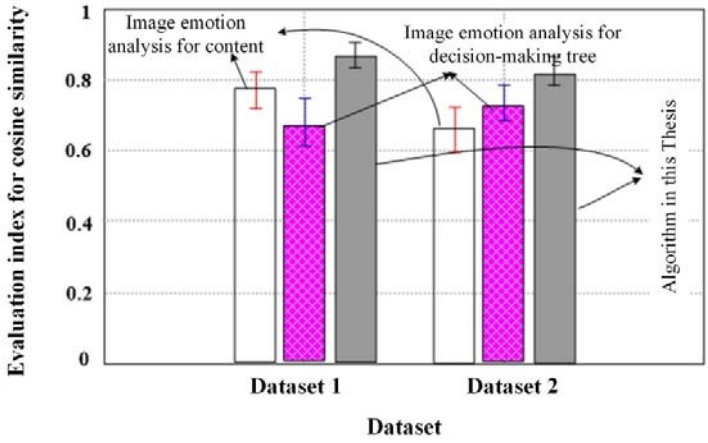

Fig. 4. Algorithm comparison based on cosine similarity index

According to comparison condition in Fig.4, it is known that the algorithm in this Thesis is better than solely selected content-based image emotion analysis for microblog information and image emotion analysis for decision-making tree with respect to cosine similarity index; in dataset 1, for concrete language, content-based image emotion analysis for microblog information is better than image emotion analysis method for decision-making tree in effect; while in dataset 2, for mixed information, the effect of image emotion analysis method for decision-making tree is better that of content-based image emotion analysis method for microblog information. At the same time, with respect to algorithm stability, the algorithm in this Thesis is better than the selected two comparing algorithms; in dataset 1, for concrete language, the stability of content-based image emotion analysis method for microblog information is better than that of image emotion analysis method for decision-making tree; in dataset 2, for mixed information, the stability of image emotion analysis method for decision-making tree is better that of content-based image emotion analysis method for microblog information.

The comparison result between score results and actual score results of terms in

dataset 1 and dataset 2 is shown in Fig.5 to Fig.6; algorithm in Literature [4] is selected as comparing algorithm.

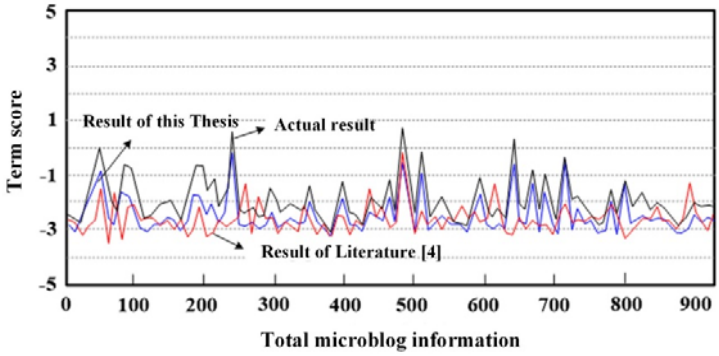

Fig. 5. Comparison among scores of terms in dataset 1

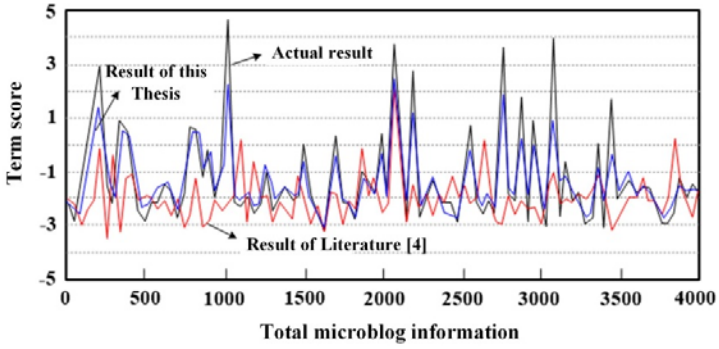

Fig. 6. Comparison among scores of terms in dataset 2

According to Fig.5 and Fig.6, it is known that in term score comparison result, term score result of the algorithm in this Thesis is more close to actual score result of terms than algorithm in Literature [4], which showing the advantage of the proposed algorithm in term score results.

5. Conclusion

In this Thesis, a statistical analysis method of concrete image emotion based on content rectification for rectification information and decision-making tree is proposed; on the basis of carrying out content-based emotion analysis and statistic for microblog information, build emotion-based concrete image method model for microblog information by using decision-making tree and also rectify by combining content-based statistic model; finally, the performance advantage of proposed algorithm in cosine similarity index is verified through experiments. However, there are still problems to be further researched and solved: (1) there is no non-concrete

microblog information in training set. By adding more non-concrete microblog information during learning process, the performance of algorithms can be improved. (2) Emphasis is only laid on the emotion mode analysis of words (unigram model). In the following research, emotion mode analysis for all images in microblog information will be of great difficulty. (3) The structure of training set is rough; there is relatively big noise interference, and some redundant information need to be deleted so as to increase training accuracy.

References

[1] Olley G S, Pakes A: (1992) *The Dynamics of Productivity in the Telecommunications Equipment Industry*[J]. Econometrica, 64(6):1263-1297.

[2] Srigley J R, Delahunt B, Eble J N, et al.: (2013) *The International Society of Urological Pathology (ISUP) Vancouver Classification of Renal Neoplasia.*[J]. American Journal of Surgical Pathology, 37(10):1469-89.

[3] Walsh S J, Pizzitutti F, Mena C F :(2014) Journal of Artificial Societies and Social Simulation[J]. Journal of Artificial Societies & Social Simulation, 17(1):116-117.

[4] Zhang M L, Zhou Z H: (2014) *A Review on Multi-Label Learning Algorithms*[J]. Knowledge & Data Engineering IEEE Transactions on, 26(8):1819-1837.

[5] Dabbish L, Stuart C, Tsay J, et al.: (2012) *Social coding in GitHub:transparency and collaboration in an open software repository*[C]// Cscw 12 Computer Supported Cooperative Work, Seattle, Wa, Usa, February. DBLP, 2012:1277-1286.

[6] Modi C, Patel D, Borisaniya B, et al.: (2013) *Review: A survey of intrusion detection techniques in Cloud*[J]. Journal of Network & Computer Applications, 36(1):42–57.

[7] Kowal P, Chatterji S, Naidoo N, et al.: (2012) *Data Resource Profile: The World Health Organization Study on global AGEing and adult health (SAGE)*[J]. International Journal of Epidemiology, 41(6):1639-1649.

[8] Hotta N, Kawamori R, Fukuda M, et al.: (2012) *Long-term clinical effects of epalrestat, an aldose reductase inhibitor, on progression of diabetic neuropathy and other microvascular complications: multivariate epidemiological analysis based on patient background factors and severity of diabetic neuropa*[J]. Diabetic Medicine, 29(12):1529-1533.

[9] Lehmann J, Cattuto C: (2012) *Dynamical classes of collective attention in twitter[C]// International Conference on World Wide Web*. ACM, 2012:251-260.

[10] Kim D G, Vargas R, Bondlamberty B, et al.: (2012) *Effects of soil rewetting and thawing on soil gas fluxes: a review of current literature and suggestions for future research*[J]. Biogeosciences, 9(7):2459-2483.

[11] Glaab E, Baudot A, Krasnogor N, et al.: *EnrichNet: network-based gene set enrichment analysis*[J]. Bioinformatics, 2012, 28(18):i451.

[12] Aiello L M, Barrat A, Schifanella R, et al.: (2012) *Friendship prediction and homophily in social media*[J]. Acm Transactions on the Web, 6(2):1-33.

[13] Zhu D C, Zhao Z D, Niu Y, et al.: (2012) *Cambrian bimodal volcanism in the Lhasa Terrane, southern Tibet: Record of an early Paleozoic Andean-type magmatic arc in the Australian proto-Tethyan margin*[J]. Chemical Geology, 328(11):290–308.

[14] Giles D M, Holben B N, Eck T F, et al.: (2012) *An analysis of AERONET aerosol absorption properties and classifications representative of aerosol source regions[J]. Journal of Geophysical Research Atmospheres*, 117(D17):127-135.

Received May 7, 2017