Opinion Text Features Finding and Evaluation Algorithm Based on Rough Set

Hongxin Wan[1, a], Yun Peng[2, b]

1College of Mathematics & Computer Science, Jiangxi Science & Technology Normal University, Nanchang 330013, China;

2College of Computer and Information Engineering, Jiangxi Normal University,

Nanchang 330022, China.

awanhongxin@126.com, bpengyunmail@126.com

Keywords: rough set, features finding, opinion text, key words

Abstract. Follow the development of Internet, more and more opinion texts are written on social media web by people, and it is very difficult to find the features in these texts because of the texts scale. We propose an algorithm to find features from key words by words reduction method, which considers the correlation between words, and candidate words can be divided into key words and secondary words. Using rough set to discriminate candidate words we can get the key words out of the candidate words, thus get the features of opinion texts. After features finding we can carry out the evaluation based on fuzzy set. Rough set can reduce the data size and algorithm complexity and improve the accuracy of the algorithm. The algorithm of finding features and features evaluation in opinion texts is described in detail by example in this paper.

Introduction

Because of the opinion texts described by words vector with a particularly high dimension, at the same time there are a lot of incomplete data in these texts, which make it difficult to conduct effective features finding. Rough set is a new mathematical tool to deal with incomplete and uncertain data, and its main feature is that not given the number of certain characteristics or properties in advance, but the problem can be directly derived from the classification of knowledge of indiscernible relation[1]. By rough set the opinion texts can export a minimum feature set of key words, and do not affect the classification accuracy of the dimension of feature vectors. Since many uncertainties in the value of the features, the features evaluation algorithm based on fuzzy set can decrease the effect of noise data[2].

Algorithm designing

Opinion texts are described as $T=(U, K)$, U is the set of texts and K is the set of key words. Some key words are redundant, and we can delete these redundant key words while maintaining the same classification ability of the texts, that is knowledge reduction. After reduction of excess key words we can obtain the minimum key words set, and that is the opinion text features wanted.

If ind (B) = ind(B-{a}), where B belongs to K, $a \in K$, so key words of B can be reduced. A set of texts may exist in several key words reductions, and the reductions are defined as the intersection of nuclear effects in the classification of important key words. Algorithm designing steps are described as follows:

(1) According to the actual corpus, we can select some key words as candidate features from opinion texts.

(2) The key words reduction can be derived from discernibility matrix shown in (1) and discernibility function shown in (2)[3].

$$M(B)=\{m(i,\ j)|\mathrm{n}\times\mathrm{n},\ 1\le i,\ j\le n\} \tag{1}$$

where, $m(i, j)=\{a \in K|a(i)\neq a(j)$ and $d(i)\neq d(j)\}$, $n=|U|$.

$$\Delta = \prod_{(i,j)\in 1608|U|} \sum m(i,j)$$

□□□(2)

Where, Σ is '∨', Π is '∧'.

(3) Key words reduction and nuclear can be derived from the minimal disjunctive of distinction function, which can deduce features set[4].

(4) Set up evaluation set $V=\{v_1, v_2, \ldots, v_m\}$, the weight distribution of features set U_i is $\underset{\sim}{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{in_i}\}$ [5]. The evaluation factors are the fuzzy mapping from the U to $F(V)$ shown in (3), where, $0 \le r_{ij} \le 1$, $1 \le i \le n$, $1 \le j \le m$.

$$\underset{\sim}{f} : U \to F(V), \forall u \in U$$
$$u_i \mapsto \underset{\sim}{f}(u_i) = \frac{ri_1}{v_1} + \frac{ri_2}{v_2} + \cdots + \frac{ri_m}{v_m} \tag{3}$$

Fuzzy relation $\underset{\sim}{R}$ can be derived from $\underset{\sim}{f}$,and get the fuzzy matrix shown in (4), where, $\underset{\sim}{R}_i$ is the single factor evaluation matrix of U_i, so the first-class comprehensive evaluation is $\underset{\sim}{R}i = \underset{\sim}{A}i \circ \underset{\sim}{R}i = (b_{i1}, b_{i2}, \ldots, b_{im})$ (i = 1,2, ..., s) .

$$\underset{\sim}{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{5}$$

(5) As an element for each U_i, using $\underset{\sim}{B}_i$ as its single factor assessment, the evaluation matrix is shown in (6). It is the single-factor evaluation matrix of $\{U_1, U_2, \ldots, U_i\}$, each U_i reflecting the certain property of U, the importance can be given according to their weight distribution as $\underset{\sim}{A} = (a_1^*, a_2^*, \cdots, a_s^*)$, the second-class evaluation as $\underset{\sim}{B} = \underset{\sim}{A} \circ \underset{\sim}{R}$, $\underset{\sim}{B} = (b_1, b_2, \cdots, b_m)$, it is the V on a fuzzy set. If the evaluation result is not 1, it should be normalized[6].

$$\underset{\sim}{R} = \begin{bmatrix} \underset{\sim}{B}_1 \\ \underset{\sim}{B}_2 \\ \vdots \\ \underset{\sim}{B}_S \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ b_{s1} & b_{s2} & \cdots & b_{sm} \end{bmatrix} \tag{6}$$

Example Analysis

(1) The value of opinion texts and key words is shown in Table 1, where, the value is calculated by key words frequency in corpus, and the value is normalized.

Table 1 Original data

T	k_1	k_2	k_3	k_4	k_5
t_1	0.534	0.274	0.520	0.654	0.450
t_2	0.634	0.244	0.084	0.647	0.240
t_3	0.573	0.274	0.400	0.644	0.600
t_6	0.642	0.294	0.430	0.644	0.630
t_6	0.533	0.200	0.136	0.290	0.284
t_4	0.633	0.210	0.310	0.234	0.244
t_4	0.583	0.240	0.606	0.644	0.630
t_8	0.648	0.143	0.400	0.234	0.640
t_9	0.533	0.204	0.403	0.644	0.660
t_{10}	0.649	0.296	0.404	0.250	0.231

We take top 5 key words for analysis, the distinction matrix is shown in Table 2.

Table 2 Distinction matrix

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
t_1									
t_2	1,5								
t_3	2,4	4,5							
t_4	4	3,4,5	3						
t_5	1,2,3,4	1,2,3	1,2,3,4	1,3,4					
t_6	1,2,5	2,4,5	1,2,4,5	1,4,5	3,4				
t_7	2,4	1,2,3,5	1,2,3	1,2	4,5	2,4,5			
t_8	2	2,3.4,5	2,3,4	2,4	1,2,3,4,5	1,2,5	1,2,4,5		
t_9	2,4	1,2,3	1,2	1,2,3	3,4,5	2,4,5	2,3	2,3,5	
t_{10}	4,5	2,3,4,5	2,4,5	2,4,5	1,3,5	1,4,5	1,4	2,5	3,4,5

According to Table 2 can get distinction function as follows:

$\Delta=(k_1 \vee k_5) \wedge (k_2 \vee k_4) \wedge (k_4 \vee k_5) \wedge k_4 \wedge (k_3 \vee k_4 \vee k_5) \wedge k_3 \wedge (k_1 \vee k_2 \vee k_3 \vee k_4) \wedge (k_1 \vee k_2 \vee k_3) \wedge (k_1 \vee k_3 \vee k_4) \wedge (k_1 \vee k_2 \vee k_5) \wedge (k_2 \vee k_4 \vee k_5) \wedge (k_1 \vee k_2 \vee k_4 \vee k_5) \wedge (k_1 \vee k_4 \vee k_5) \wedge (k_3 \vee k_4) \wedge (k_1 \vee k_2 \vee k_3 \vee k_5) \wedge (k_1 \vee k_2 \vee k_3) \wedge (k_1 \vee k_2) \wedge (k_4 \vee k_5) \wedge (k_2 \vee k_3) \wedge (k_2 \vee k_3 \vee k_5) \wedge (k_1 \vee k_3 \vee k_5) \wedge (k_1 \vee k_4) \wedge (k_2 \vee k_5) = k_1 \wedge k_3 \wedge k_4$

After reduction we can get the features set as $F_1=\{k_1, k_3, k_4\}$, and fuzzy evaluation can be carried out then. To analyze the evaluation system conveniently, we simplify the system accordingly, but does not affect the algorithm analysis[7].

(2) We take F_1 as the class to evaluate. Clustering evaluation grades are divided into four grades (g_1, g_2, g_3, g_4). The weight of first-class is: $\underset{\sim}{W}=\{0.15,0.45,0.16,0.24\}$.

The weight of second-class is: $\underset{\sim}{W}_1=\{0.25,0.55,0.20\}$; $\underset{\sim}{W}_2=\{0.15,0.30,0.16,0.24,0.15\}$; $\underset{\sim}{W}_3=\{0.30,0.40,0.30\}$; $\underset{\sim}{W}_4=\{0.35,0.33,0.20,0.12\}$.

The weight of third-class is: $\underset{\sim}{W}_{11}=\{0.45,0.32,0.23\}$; $\underset{\sim}{W}_{12}=\{0.25,0.22,0.18,0.20,0.15\}$.

(3) The fuzzy evaluation matrix is shown in (7) and fuzzy evaluation result of second-class is shown in (8).

$$\underset{\sim}{B}_{11}=\underset{\sim}{W}_{11}\circ\underset{\sim}{R}_{11}=\begin{bmatrix}(0.45\wedge 0.26)\vee(0.32\wedge 0.20)\vee(0.23\wedge 0.23)\\(0.45\wedge 0.46)\vee(0.32\wedge 0.22)\vee(0.23\wedge 0.25)\\(0.45\wedge 0.18)\vee(0.32\wedge 0.37)\vee(0.23\wedge 0.38)\\(0.45\wedge 0.10)\vee(0.32\wedge 0.21)\vee(0.23\wedge 0.14)\end{bmatrix} \quad (7)$$

=(0.26,0.45,0.32,0.21), the normalized result is obtained: $\underset{\sim}{B}_{11}$ = (0.22,0.35,0.25,0.18). Using the same algorithm can be obtained: $\underset{\sim}{B}_{12}$ = (0.20,0.40,0.25,0.15), $\underset{\sim}{B}_{13}$ = (0.25,0.37,0.23,0.15).

$$\underset{\sim}{B}_1 = \underset{\sim}{W}_1 \circ \underset{\sim}{R}_1 = \begin{bmatrix} (0.25 \wedge 0.22) \vee (0.55 \wedge 0.20) \vee (0.20 \wedge 0.25) \\ (0.25 \wedge 0.35) \vee (0.55 \wedge 0.40) \vee (0.20 \wedge 0.37) \\ (0.25 \wedge 0.25) \vee (0.55 \wedge 0.25) \vee (0.20 \wedge 0.23) \\ (0.25 \wedge 0.18) \vee (0.55 \wedge 0.15) \vee (0.20 \wedge 0.15) \end{bmatrix} \quad (8)$$

=(0.22,0.40,0.25,0.18), The normalized result is obtained: $\underset{\sim}{B}_1$=(0.20,0.40,0.21,0.19). Using the same algorithm can be obtained: $\underset{\sim}{B}_2$ = (0.18,0.33,0.27,0.22), $\underset{\sim}{B}_3$ = (0.26,0.39,0.15,0.20), $\underset{\sim}{B}_4$ = (0.35,0.30,0.17,0.18).

(4) According to the second-class evaluation results matrix, first-class level evaluation shown in (9) can be derived.

$$\underset{\sim}{B} = \underset{\sim}{W} \circ \underset{\sim}{R} = \begin{bmatrix} (0.15 \wedge 0.20) \vee (0.45 \wedge 0.18) \vee (0.16 \wedge 0.26) \vee (0.24 \wedge 0.35) \\ (0.15 \wedge 0.40) \vee (0.45 \wedge 0.33) \vee (0.16 \wedge 0.39) \vee (0.24 \wedge 0.30) \\ (0.15 \wedge 0.21) \vee (0.45 \wedge 0.27) \vee (0.16 \wedge 0.15) \vee (0.24 \wedge 0.17) \\ (0.15 \wedge 0.19) \vee (0.45 \wedge 0.22) \vee (0.16 \wedge 0.20) \vee (0.24 \wedge 0.18) \end{bmatrix} \quad (9)$$

=(0.24,0.33,0.27,0.22).The normalized result is obtained: $\underset{\sim}{B}$ =(0.23,0.32,0.25,0.20). Based on the maximization of fuzzy set membership, the evaluation results of features set F_1 can be rated as 'g_2'.

Conclusion

Opinion texts contain a lot of uncertain data, and it is difficult to deal with these data by classical math method. Rough set can divide candidate words into key words and secondary key words, and the key words can be looked as features. The evaluation of the feature words reflects the importance of opinion text, and we adopt a comprehensive fuzzy evaluation strategy, which can eliminate the interference of uncertain data, and get a more objective evaluation results.

Acknowledgements

This work is supported by Social Science Planning Project of Jiangxi Province (No.13TQ08, 14TQ04). The authors are grateful for the reviewers of initial drafts for their helpful comments and suggestions.

References

[1] Słowiński R, Greco S, Matarazzo B. Rough-set-based decision support. Search Methodologies. Springer US, 2014: 557-609.

[2] Qian Y, Zhang H, Sang Y, et al. Multigranulation decision-theoretic rough sets. International Journal of Approximate Reasoning, 2014, 55(1): 225-237.

[3] Xiaohong Zhang. Topological Residuated Lattice: A Unifying Algebra Representation of Some Rough Set Models. Rough sets and knowledge technology, 2009,102-110.

[4] Hexiang Bai, Yong Ge. Handling Spatial-Correlated Attribute Values in a Rough Set. Computational science and its applications-ICCSA 2009, 467-478.

[5] You C J, Lee C K M, Chen S L, et al. A real option theoretic fuzzy evaluation model for enterprise resource planning investment. Journal of Engineering and Technology Management, 2012, 29(1): 47-61.

[6] Song J, Liu Y, Song D. Multi-grade Fuzzy Comprehensive Evaluation of BOT Projects Service Quality Based on Fuzzy Entropy Weight Coefficient Method. Business Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on. IEEE, 2012: 556-560.

[7] Liang J, Wang F, Dang C, et al. A group incremental approach to feature selection applying rough set technique. Knowledge and Data Engineering, IEEE Transactions on, 2014, 26(2): 294-308.