Opinion Text Features Finding and Evaluation Algorithm Based on Rough Set

Hongxin Wan¹,a, Yun Peng²,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 \( \text{ind} (B) = \text{ind}(B-\{a\}) \), 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].

\[
M(B)=\{m(i, j)|n×n, 1\leq i, j\leq n\}
\]

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

\[
A = \prod_{(i, j)\in M(B)} \sum m(i, j)
\]
Where, $\Sigma$ is ‘$\lor$’, $\Pi$ is ‘$\land$’.

(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 $A = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$ [5]. The evaluation factors are the fuzzy mapping from the $U$ to $F(V)$ shown in (3), where, $0 \leq r_{ij} \leq 1$, $1 \leq i \leq n$, $1 \leq j \leq m$.

$$ u_i \mapsto f(u_i) = \frac{r_{i1}}{v_1} + \frac{r_{i2}}{v_2} + \cdots + \frac{r_{im}}{v_m} $$

(3)

Fuzzy relation $R$ can be derived from $f$, and get the fuzzy matrix shown in (4), where, $R_i$ is the single factor evaluation matrix of $U_i$, so the first-class comprehensive evaluation is $\overline{R} = \bigoplus R_i = (b_{11}, b_{12}, \ldots, b_{nn})$ ($i = 1, 2, \ldots, s$).

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

(5)

(5) As an element for each $U_i$, using $B_i$ 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 $A = (a_1, a_2, \ldots, a_n)$, the second-class evaluation as $\overline{B} = A \circ \overline{R}$, $\overline{B} = (b_1, b_2, \ldots, b_m)$, it is the $V$ on a fuzzy set. If the evaluation result is not 1, it should be normalized[6].

$$ R = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_s \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{s1} & b_{s2} & \cdots & b_{sn} \end{bmatrix} $$

(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.
shown in (8).

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: \( W = \{0.15,0.45,0.16,0.24\} \).

The weight of second-class is: \( W_1 = \{0.25,0.55,0.20\} \); \( W_2 = \{0.15,0.30,0.16,0.24,0.15\} \); \( W_3 = \{0.30,0.40,0.30\} \); \( W_4 = \{0.35,0.33,0.20,0.12\} \).

The weight of third-class is: \( W_{11} = \{0.45,0.32,0.23\} \); \( W_{12} = \{0.25,0.22,0.18,0.20,0.15\} \).

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

\[
B_{11} = \overline{W_{11}} \circ R_{11} = \\
\begin{bmatrix}
(0.45 \land 0.26) \lor (0.32 \land 0.20) \lor (0.23 \land 0.23) \\
(0.45 \land 0.46) \lor (0.32 \land 0.22) \lor (0.23 \land 0.25) \\
(0.45 \land 0.18) \lor (0.32 \land 0.37) \lor (0.23 \land 0.38) \\
(0.45 \land 0.10) \lor (0.32 \land 0.21) \lor (0.23 \land 0.14)
\end{bmatrix}
\]

=(0.26,0.45,0.32,0.21), the normalized result is obtained: \( B_{11} = (0.22,0.35,0.25,0.19) \). Using the same algorithm can be obtained: \( B_{12} = (0.20,0.40,0.25,0.15) \), \( B_{13} = (0.25,0.37,0.23,0.15) \).
The normalized result is obtained: \( B_1 = (0.20, 0.40, 0.21, 0.19) \). Using the same algorithm can be obtained: \( B_2 = (0.18, 0.33, 0.27, 0.22), \ B_3 = (0.26, 0.39, 0.15, 0.20), \ B_4 = (0.35, 0.30, 0.17, 0.18) \).

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

\[
B = W \circ R = \left[ \begin{array}{c}
(0.15 \land 0.20) \lor (0.45 \land 0.18) \lor (0.16 \land 0.26) \lor (0.24 \land 0.35) \\
(0.15 \land 0.40) \lor (0.45 \land 0.33) \lor (0.16 \land 0.39) \lor (0.24 \land 0.30) \\
(0.15 \land 0.21) \lor (0.45 \land 0.27) \lor (0.16 \land 0.15) \lor (0.24 \land 0.17) \\
(0.15 \land 0.19) \lor (0.45 \land 0.22) \lor (0.16 \land 0.20) \lor (0.24 \land 0.18)
\end{array} \right]
\]

(9)

\( B = (0.24, 0.33, 0.27, 0.22) \). The normalized result is obtained: \( B = (0.23, 0.32, 0.25, 0.20) \). 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**


