Three Term Weighting and Classification Algorithms in Text Automatic Classification

Qian Diao
97300BA, Shanghai Jiao Tong University
qian-diao@263.net
Hainan Diao
P.O.Box 030-8, Luoyang, Henan

Abstract

In this paper, three text automatic classification algorithms are provided. They are Bayes Method based on Bayes theorem and IDF (Invert Document Frequency), VSM based on Shannon entropy, and Fuzzy Method based on the fuzzy theory. Furthermore, the way of how to combine term weighting methods with three classification algorithms is also provided in the paper.

1. Introduction

Search Engines were designed to assist user to find useful information in Internet. They have two types. One is based on the keyword query; the other is browsing type. With the development of search engines, the trend shows as from query——search to browse——search. This has led to the development of text classification according to the taxonomy. Now we bring forward three text automatic classification algorithms here.

2. Bayes Method and IDF

In the text classification system, text sample x is expressed by its keywords w_i, so $x=(w_1,\cdots,w_i,\cdots,w_m)$. Here, m is the total number of keywords. If the correlative-weight between w_i and x is IDF (Invert Document Frequency), expressed as idf_i, and the definition of idf_i is formula (2), here, n_i is the number of texts including keyword w_i, N is the total number of text samples, k is decided by experiments.

$$idf_i = k + \log\frac{N-n_i}{n_i} \tag{1}$$

Then,

$$\begin{aligned} &P(x/c_j)\\ &= P((w_1(idf_1),\cdots,w_i(idf_i),\cdots w_m(idf_m))/c_j)\\ &= \sum_{i=1}^{m} P(w_i(idf_i)/c_j)\\ &= \sum_{i=1}^{m} P(w_i/c_j)\cdot idf_i \end{aligned} \tag{2}$$

If we put formula (3) into Bayes classification rule, we obtain Bayes Method as following:

$$[\sum_{i=1}^{m} P(w_i/c_j)\cdot idf_i]P(c_j) > [\sum_{i=1}^{m} P(w_i/c_k)\cdot idf_i]P(c_k)\ ,$$

Then $(w_1,\cdots,w_i,\cdots,w_m)\in c_j$

So the correlative-weight between w_i and c_j is (3).

$$Weight(w_i/c_j) = \frac{n_{ij}}{n_i}\cdot\frac{N_j}{N}\cdot(k+\log\frac{N-n_i}{n_i}) \tag{3}$$

Here, n_{ij} is the appearance times of w_i in the texts belonging to class c_j, N_j is the total number of texts belonging to c_j.

3. VSM Based on Shannon Entropy

$V_i = (Weight_{i1}(w_1/c_1),\cdots,Weight_{ik}(w_k/c_i),\cdots,Weight_{im}(w_m/c_m))$ is the feature vector of category i, and $D_j = (Weight_j(w_1),\cdots,Weight_j(w_k),\cdots,Weight_j(w_m))$ is the

feature vector of the testing text sample j, then after vector standardization, we can make classification by computing cosine distance between two vectors by (4).

$$Sim(V_i, D_j) = \cos\theta = \frac{\sum_{k=1}^{m} v_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{m} v_{ik}^2} \cdot \sqrt{\sum_{k=1}^{m} d_{ik}^2}} \quad (4)$$

Here, we use Shannon entropy theory to compute the correlative-weight between keyword w_k and class c_i, then we obtain formula (5).

$$Weight_{ik}(w_k / c_i) = -\left(\frac{n_i(w_k)}{\sum_{i=1}^{M} n_i(w_k)} \cdot \frac{N_i}{N}\right) \log\left(\frac{n_i(w_k)}{\sum_{i=1}^{M} n_i(w_k)} \cdot \frac{N_i}{N}\right) \quad (5)$$

4. Fuzzy Method Based on Fuzzy Theory

In Fuzzy theory, if a element x is in the field U, and the number of common sets A^* which have attributes belonging to fuzzy set $\underline{A}$ is n, then the attributive value of the element x to $\underline{A}$ is obtained by formula (6).

$$\mu_{\underline{A}}(x) = \lim_{n \to \infty} \frac{\text{The times of } x \in A^*}{n} \quad (6)$$

In the text classification system, if keyword w_i is in the text set, and there are N_j text samples belonging to class c_j, then the relations between w_i and c_j can be obtained as formula (7).

$$weight(w_i / c_j) = \mu_{c_j}(w_i) = \frac{n_{ij}}{N_j} \quad (7)$$

If a testing text sample is expressed as $D = (d_1, \cdots, d_i, \cdots, d_m)$, Here, m is the total number of keywords, M is the total number of classes, then we use formula (8) to obtain classification decision-making.

$$R_j(D) = Weight^T \cdot D = \sum_{i=1}^{m} d_i weight(w_i / c_j) \quad (8)$$

$$R_j = \sum_{i=1}^{m} weight(w_i / c_j)$$

$$\mu_j(D) = \frac{R_j(D)}{R_j}$$

If $\mu_j(D) = \max\{\mu_1(D), \cdots, \mu_N(D)\}$, Then $D \in c_j$

5. Experiments Results

The experiment situation is:

Training samples: 67093 texts

Testing samples: 2032 texts

Pretreatment result: 95928 keywords

Class Number: 26 classes

In table 1: Bayes/ VSM/Fuzzy

Testing Method	Close Test	Open Test
Testing Text Number	8000/8000/8000	2000/2000/2000
Precision	93%/97%/93%	73%/79%/74%
Classification Time (s)	407/141/395	88/48/73

6. Conclusion

Although Bayes classification method and VSM are the classical classification methods in the field of text automatic classification, the old Bayes method do not consider IDF and the traditional VSM do not use Shannon entropy theory. In this paper we put forward these new text classification methods to promote them. Furthermore, we introduce fuzzy theory into text automatic classification field.

Reference

[1] Robert Trapl, Cybernetics Theory and Applications, Hemisphere Publication Corporation, U.S.A., 1983, pp9—20.

[2] Huang Dequan, Theory of Neutral Networks and Pattern Recognition Systems, Publication of Electronic Industry, Beijing, 1996, pp12—30.