Text Similarity Computing Based on Standard Deviation

Tao Liu and Jun Guo

School of Information Engineering, Beijing University of Posts and
Telecommunications, Beijing 100876, China,
sdlclt@sohu.com

Abstract. Automatic text categorization is defined as the task to assign free text documents to one or more predefined categories based on their content. Classical method for computing text similarity is to calculate the cosine value of angle between vectors. In order to improve the categorization performance, this paper puts forward a new algorithm to compute the text similarity based on standard deviation. Experiments on Chinese text documents show the validity and the feasibility of the standard deviation-based algorithm.

1 Introduction

Text categorization has recently become an active research topic in the area of information retrieval. The objective of text categorization is to assign free text documents to one or more predefined categories based on their content. Traditionally text categorization is performed manually by domain experts. Each incoming document is read and comprehended by the expert and then it is assigned a number of categories chosen from the set of prespecified categories. This process is very time-consuming and costly, thus limiting its applicability.

A promising way to deal with this problem is to learn a categorization scheme automatically from training collection. Once the categorization scheme is learned, it can be used for classifying future documents. It involves issues commonly found in machine learning problems. Since a document may be assigned to more than one category, the scheme also requires the assignment of multiple categories. There is a growing body of research addressing automatic text categorization. A number of statistical classification and machine learning techniques has been applied to text categorization, including regression models[1][2], nearest neighbor classifiers[3][4], Bayesian classifiers[5][6], decision trees[1][6][7], rule learning algorithms[8][9][10], neural networks[1], inductive learning techniques[11][12], Support Vector Machines[13], relevance feedback[14] and voted classification[15].

In order to improve the categorization performance, this paper puts forward a new algorithm to compute the text similarity based on standard deviation. Experiments show the validity and the feasibility of the standard deviation-based algorithm.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part I, LNCS 3644, pp. 456–464, 2005.

This paper contains 6 sections. In Section 2 we describe the vector space model; in Section 3 describe several typical methods that have been successfully applied to text feature selection and categorization; Section 4 introduces the proposed method applied to text similarity computing and categorization based on standard deviation and compares it with the classical method based on cosine similarity; experimental results and evaluation are given in Section 5; finally, we draw to a conclusion.

2 Vector Space Model

The most commonly used document representation is the so called vector space model (VSM)[16]. In the vector space model, each document can be represented by vector $\mathbf{v} = (\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_m})$, where w_i represents the corresponding weight of the i^{th} feature t_i of the document and denotes the importance of t_i in describing the document's content. Therefore, the expression and matching issue of text information is converted to that of the vector in VSM [17]. Experiment shows that word is a better candidate for feature than character and phrase.

At present there are several ways of determining the weight w_i, Intuitively, w_i should express the two aspects as follows:

- The more often a word occurs in a document, the more effectively it is to reflect the content of the document.
- The more often the word occurs throughout all documents in the collection, the more poorly it discriminates between documents.

A well-known approach for computing word weights is the tf*idf weighting, which assigns the weight to word in document in proportion to the number of occurrences of the word in the document, and in inverse proportion to the number of documents in the collection for which the word occurs at least once.

Among several existing tf*idf formulas, we selected a commonly used one in our system:

$$W(t, \boldsymbol{d}) = \frac{tf(t, \boldsymbol{d}) \times \log\left(\frac{N}{n_t} + 0.01\right)}{\sqrt{\sum_{t \in \boldsymbol{d}} [tf(t, \boldsymbol{d}) \times \log\left(\frac{N}{n_t} + 0.01\right)]^2}} \tag{1}$$

where $W(t, \boldsymbol{d})$ is the weight of word t in document $\boldsymbol{d}$, $tf(t, \boldsymbol{d})$ is the frequency of word t in document $\boldsymbol{d}$, N is the number of documents in the training collection and n_t is the number of documents in the whole collection for which word t occurs at least once.

The main advantage of VSM is in the fact that it simplifies the documents' content to vectors comprising features and weights, which greatly decreases the complexity of the problem.

3 Typical Text Feature Selection and Categorization Methods

A major problem in text categorization is the high dimensionality of the feature space. Generally the feature space consists of hundreds of thousands words even for a moderated-size documents collection. Standard classification techniques can hardly deal with such a large feature set since processing is extremely costly in computational terms, and overfitting can not be avoided due to the lack of sufficient training data. Hence, there is a need for a reduction of the original feature set without decreasing the categorization accuracy, which is commonly known as dimensionality reduction in the pattern recognition literature.

Feature selection attempts to remove non-informative words from documents in order to improve categorization effectiveness and reduce computational complexity. Before feature selection, word segmentation which is necessary for Chinese text categorization has to be made because there is not apparent delimiter between the character in the text. In [18] a thorough evaluation of the five known feature selection methods: Document Frequency Thresholding, Information Gain, χ^2-statistic, Mutual Information and Term Strength is given.

In Document Frequency Thresholding, we computes the document frequency for each word in the training collection and removes those words whose document frequency is less than a predetermined threshold. The basic assumption is that rare words are either non informative for category prediction, or not influential in global performance. Information Gain is frequently employed as a termgoodness criterion in the field of machine learning [19][20]. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a word in a document. The information gain of a word t is defined to be:

$$\begin{aligned} IG\left(t\right) = &-\sum_{j=1}^{n} P\left(C_j\right)\log P\left(C_j\right) + P\left(t\right)\sum_{j=1}^{n} P\left(C_j|t\right)\log P\left(C_j|t\right) \\ &+P\left(\bar{t}\right)\sum_{j=1}^{n} P\left(C_j|\bar{t}\right)\log P\left(C_j|\bar{t}\right) \end{aligned} \tag{2}$$

Wherein, n is the number of the category, $P(C_j)$ is the probability that class C_j occurs in the total collection and $P(t)$ is that of word t. $P(C_j|t)$ can be computed as the fraction of documents from class C_j that have at least one occurrence of word t and $P(C_j|\bar{t})$ as the fraction of documents from class C_j that does not contain word t. The information gain is computed for each word of the training collection, and the words whose information gain is less than some predetermined threshold are removed. The χ^2-statistic measures the lack of independence between word t and class C_j. It is given by:

$$\chi^2(t, C_j) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{3}$$

A set of tokens with the highest χ^2 measures are then selected as keyword features. A is the number of documents from class C_j that contains word t and B is the number of documents that contains t but does not belong to class C_j. C is the number of documents from class C_j that does not contain word t and D is the number of documents that belongs to class C_j nor contains word t. Mutual Information is a criterion commonly used in statistical modelling of word associations and related applications[21][22][23]. Term strength measures how informative a word is in identifying two related documents. The strength of a word t is defined as the probability of finding t in a document which is related to any document in which t occurs[24]. In our experiments, the method of χ^2-statistic is found to be the most effective.

There exists various text categorization algorithms, such as Rocchio's algorithm, Naive Bayes, K-nearest neighbor, neural network, support vector machine etc, wherein the first two are employed in most applications. The proposed method in this paper is in fact an improved version of Rocchio's algorithm. Here we only introduce the Naive Bayes while Rocchio's algorithm will be described in detail in the next section. The naive Bayes classifier estimate the probability of each class based on Bayes theory:

$$P\left(C_j|d\right) = \frac{P\left(C_j\right)P\left(d|C_j\right)}{P\left(d\right)} \tag{4}$$

$P\left(d\right)$ is same to all the caterogies and the assumption is made that the features are conditionally independent. This simplifies the computations yielding:

$$P\left(C_j|d\right) = P\left(C_j\right)\prod_{i=1}^{m} P\left(t_i|C_j\right) \tag{5}$$

$P\left(C_j\right)$ is the probability that class C_j occurs in the total collection. An estimate $\hat{P}\left(t_i|C_j\right)$ for $P\left(t_i|C_j\right)$ is given by:

$$\hat{P}\left(t_i|C_j\right) = \frac{1+N_{ij}}{m+\sum\limits_{l=1}^{m} N_{lj}} \tag{6}$$

N_{ij} denotes the number of times word t_i occurred within documents from class C_j in the training collection.

4 Proposed Method Based on Standard Deviation

Rocchio's algorithm is the classical method for document routing or filtering in information retrieval. In this method, a prototype vector $\mu_{\mathbf{j}}$= (μ_{j1},μ_{j2},...μ_{jm}) is computed as the average vector over n_j training document vectors that belong to class C_j, where the feature $\mu_{ji} = \frac{1}{n_j}\sum\limits_{k=1}^{n_j} w_{j,ki}$ is the mean of $w_{j,ki}$ and $\mathrm{w}_{\mathrm{j,ki}}$ is the weight of word i in document $\mathbf{d_k}$ of category C_j. A document $\mathbf{d_{test}}$ is classified by calculating the similarity between document vector $\mathbf{v_{test}} = \left(\mathbf{w_{test1}}, \mathbf{w_{test2}}, \ldots, \mathbf{w_{testm}}\right)$ of $\mathbf{d_{test}}$ and each of the prototype vectors $\mu_{\mathbf{j}}$. The similarity can be computed as follows [25]:

$$Sim\left(\mathbf{d_{test}}, C_j\right) = \frac{\mathbf{v_{test}} \cdot \mu_{\mathbf{j}}}{\|\mathbf{v_{test}}\| \cdot \|\mu_{\mathbf{j}}\|} \tag{7}$$

Since (7) exactly denotes the cosine function of the angle between the two vectors, the similarity defined in (7) is usually called 'cosine similarity'. Consequently, the category into which the document $\mathbf{d_{test}}$ falls is determined by the equation:

$$c = \arg\max_{j} Sim\left(\mathbf{d_{test}}, C_j\right) \tag{8}$$

As we can see from the analysis above, Rocchio's method has such advantage as simple categorization mechanism, rapid process rate while its main defect is due to the fact that it is difficult to roundly describe the characteristics of the category with the only information of samples' mean. A new method is proposed in this paper to overcome the main defect of classical method by describing the characteristics of category more precisely with not only the mean vector but also the standard deviation and classifying documents with new similarity rather than 'cosine similarity' employed. In our study, it is found that the standard deviation, which is a common used statistics reflecting the distribution of the samples in pattern recognition, of each feature in diverse category changes distinctly, while the fact is not concerned in classical Rocchio's method leading to degraded categorization result. The new method in this paper obtains better performance because of considering of the difference of the standard deviation.

Two vectors, mean vector $\mu_{\mathbf{j}}$ and standard deviation vector $\sigma_{\mathbf{j}} = (\sigma_{j1}, \sigma_{j2}, ...\sigma_{jm})$, are chosen as the prototype vectors of category C_j, wherein $\sigma_{ji} = \sqrt{\frac{1}{n_j - 1}\sum\limits_{k=1}^{n_j}\left(w_{j,ki} - \mu_{ji}\right)^2}$. A modified street distance is proposed to be a new similarity in text categorization to accommodate the two new prototype vectors,

$$Sim\left(\mathbf{d_{test}}, C_j\right) = -\sum_{i=1}^{m} \frac{\max\left\{\left|w_{testi} - \mu_{ji}\right| - \sigma_{ji}, 0\right\}}{\sigma_{ji}^2} \tag{9}$$

As is shown, any document locates within the sphere determined by μ_{ji} and σ_{ji}^2 will be classified to category C_j. It is noted that because of the probably of being zero σ_{ji}^2 should be modified before used as denominator in (9).

In categorization experiments, mean vector $\mu_{\mathbf{j}}$ and standard deviation vector $\sigma_{\mathbf{j}}$, are obtained during the training process according to the training collection. Consequently, the similarity between document $\mathbf{d_{test}}$ to be categorized and each category is computed by (9). Finally, the categorization results is obtained by (8).

Compared with classical Rocchio's method based on 'cosine similarity', the advantage of the proposed method is illustrated in Fig.1. For the convenience of illustrating, we suppose each document has two features. Since in (7) $\mathbf{d_{test}}$ and C_j have been cosine normalized, the similarity described in (7) reflects the Euclidean distance between $\mathbf{d_{test}}$ and C_j. $\mathbf{v_{test}}$ is the document vector of the document $\mathbf{d_{test}}$ to be categorized, $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$ are mean vectors of category A and B, respectively. σ_{A1} and σ_{A2} represent the two dimensional standard deviation of A, and σ_{B1} and σ_{B2} denote that of B. Since the distance D_A between $\mathbf{v_{test}}$ and $\mu_{\mathbf{A}}$ is greater than the distance D_B between $\mathbf{v_{test}}$ and $\mu_{\mathbf{B}}$, $D_A > D_B$, the categorization result is $\mathbf{d_{test}} \in B$ based on classical Rocchio's method. However, $\mathbf{d_{test}}$ locates in the field of category A not category B, so the probability of

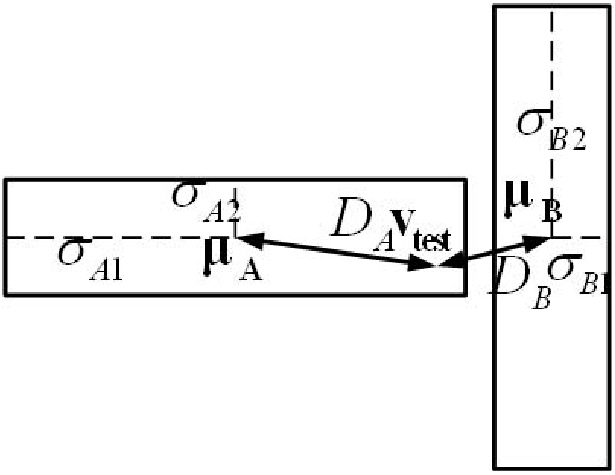

Fig. 1. Advantage of proposed method

$\mathbf{d_{test}} \in A$ is larger than that of $\mathbf{d_{test}} \in B$. It is more likely to put $\mathbf{d_{test}}$ in wrong category by classical Rocchio's method than the proposed method.

5 Experiment Results and Evaluation

Text categorization experiment results of 36 categories of Chinese text documents according to Chinese Library Classification (version 4) based on standard deviation and 'cosine similarity' are given in this section. We take Chinese Text Categorization (TC) Evaluation collection of Chinese Language Processing and Intelligent Human Machine Interface evaluation of the National High Technology Research and Development Program(HTRDP) in 2003 as the training collection and download 3134 text documents (labeled by the experts with reference to Chinese Library Classification) from the web. The HTRDP Evaluation of Chinese Language Processing and Intelligent Human Machine Interface, also called the '863' Evaluation, is a series of evaluation activities sponsored by China's National High Technology Research and Development Program (HTRDP, also called the '863' Program). The purpose of the HTRDP Evaluation is to provide infrastructural support for research and development on Chinese information processing and intelligent human-machine interface technology, to enhance interaction among industry, academia, and government, and to speed up the transfer of technology from research labs into commercial products.

Experiment results of 8 categories are shown in Table 1, where a represents the number of documents correctly assigned to this category; b represents the number of documents incorrectly assigned to this category; c represents the number of documents incorrectly rejected from this category. The category label is with reference to Chinese Library Classification.

Table 1. Experiment results of 8 categories

category	B	E	J	K	TU	G	TD	TP
a	94	69	29	55	57	27	70	77
b	7	5	28	25	38	80	18	5
c	26	50	21	20	34	3	14	50
Precision	93.07%	93.24%	50.88%	68.75%	60.00%	25.23%	79.55%	93.90%
Recall	78.33%	57.98%	58.00%	73.33%	62.64%	90.00%	83.33%	60.63%
F1	85.07%	71.50%	54.21%	70.97%	61.29%	39.42%	81.40%	73.68%
F1 Gain	2.12%	1.35%	8.83%	20.97%	22.97%	5.84%	2.91%	13.27%

Categorization effectiveness is measured in terms of the commonly used IR notions of precision and recall, adapted to the case of text categorization. Precision is defined as the probability that if a random document $\mathbf{d_{test}}$ is categorized under C_j , this decision is correct. Analogously, Recall is defined as the probability that, if a random document $\mathbf{d_{test}}$ should be categorized under C_j, this decision is taken. Another evaluation criterion that combines recall and precision is the F1 measure. The definitions of Precision, Recall and F1 are given below [26]:

$$Precision = a/(a+b) \tag{10}$$

$$Recall = a/(a+c) \tag{11}$$

$$F1 = (Precision \times Recall \times 2)/(Precision + Recall) \tag{12}$$

Table 2. Measure comparison of two methods

	Macro-averaging Precision	Macro-averaging Recall	Macro-averaging F1
cosine	70.81%	70.71%	70.76%
deviation	74.21%	74.43%	74.32%

For evaluating performance average across categories, Macro-averaging performance scores are determined by first computing the performance measures per category and then averaging these to compute the global means and are calculated as follows:

$$Macro - averaging\ Precision = \frac{1}{n}\sum_{j=1}^{n} Precision_j \tag{13}$$

$$Macro - averaging\ Recall = \frac{1}{n}\sum_{j=1}^{n} Recall_j \tag{14}$$

$$Macro - averaging\ F1 = \frac{1}{n}\sum_{j=1}^{n} F1_j \tag{15}$$

n is the number of the category. From Table 2, we can see that both the Macro-averaging Precision and Recall of proposed method is around 74%, which is greater than that of the traditional one. The validity and the feasibility of the standard deviation-based algorithm is validated by experiment results.

6 Conclusions

As Chinese text information available on the Internet continues to increase, there is a growing need for tools helping people better manage the information. Text categorization, the assignment of free text documents to one or more predefined categories based on their content, is an important component to achieve such task and attracts more and more attention.

A number of statistical classification and machine learning techniques has been applied to text categorization. In order to improve the categorization performance, this paper puts forward a new algorithm to compute the text similarity based on standard deviation. Experiments show the validity and the feasibility of the standard deviation-based algorithm.

References

[1] Fuhr, N., Hartmanna, S., Lustig, G., Schwantner, M.,Tzeras, K.: Air/x - a rule-based multistage indexing systems for large subject fields. Proceedings of RIAO'91. (1991) 606–623

[2] Yang, Y., Chute, C.G.: A Linear Least Squares Fit mapping method for information retrieval from natural language texts. Proceedings of 14th International Conference on Computational Linguistics (COLING'92).**II** (1992) 447–453

[3] Creecy, R.H., Masand, B.M., Smith, S.J., Waltz, D.L.: Trading MIPS and memory for knowledge engineering:classifying census returns on the connection machine. Comm. ACM.**35** (1992) 48–63

[4] Yang, Y., Chute, C.G.: An example-based mapping method for text classification and retrieval. ACM Transactions on Information Systems (TOIS) **12** (1994) 253–277

[5] Tzeras,K., Hartmann, S.: Automatic Indexing Based on Bayesian Inference Networks. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIDIR'93). (1993) 22–34

[6] Lewis, D.,Ringuette, M.: A comparison of two learning algorithms for text clas sification, In Third Annual Symposium on Document Analysis and Information Retrieval. (1994) 81–93

[7] Moulinier, I.: Is learning bias an issue on the text categorization problem? In Technical report, LAFORIA-LIP6, Universite Paris VI, (1997)

[8] Apte, C., Damerau, F., Weiss, S.: Towards language independent automated learning of text categorization models. In Proceedings of the Seventeenth Annual International ACM/SIGIR Conference. (1994)

[9] Wiener, E., Pedersen, J.O., Weigend, A.S.: A neural network approach to topic spotting. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval(SDAIR'95). (1995)

[10] Moulinier, I., Raskinis, G., Ganascia, J.: Text categorization: a symbolic approach. In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval. (1996)

[11] William, W. C., Singer, Y.: Context-sensitive learning methods for text classification. In SIGIR'96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1996) 307–315

[12] David, D. L., Robert E. S., Callan, J.P., Papka, R.: Training Algorithms for Linear Text Classifiers. In SIGIR '96:Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1996) 298–306

[13] Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In Proc. 10th European Conference on Machine Learning (ECML) Springer Verlag. (1998)

[14] Rocchio, J.: Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing. PrenticeHall Inc.. (1971) 313–323

[15] Weiss, S. M., Apte, C., Damerau, F. J.,Johnson, D.E., Oles, F.J., Goetz, T., Hampp, T.: Maximizing Text-Mining Performance. Intelligent Systems and Their Applications, IEEE [see also IEEE Intelligent Systems] 14 (1999) 63-69

[16] Salton, G., Lesk, M. E.: Computer evaluation of Indexing and text processing. Association for Computing Machinery **15** (1968) 8–36

[17] Salton, G., Wong, A., Yang, C. S.: A Vector Space Model for Automatic Indexing. Communications of ACM **18** (1975) 613–620

[18] Yiming, Y., Jan, P. P.: A comparative study on feature selection in text Categorization. In:Proceedings of ICML'97, 14th International Conference on Machine Learning. (1997) 412–420 Morgan Kaufmann

[19] Tom, M.: Machine Learning. McCraw Hill, 1996.

[20] Quinlan, J.: Induction of decision trees. Machine Learning, **1** (1986) 81–106

[21] Keeneth, W. C., Patric, H.: Word association norms, mutual information and lexicography. In Proceeding of ACL 27 (1989) 76–83 Vabcouver, Canada

[22] Fano, R.: Transmission of Information. MIT Press, Cambrige, MA, (1961)

[23] Wiener, E., Pedersen, J.O., Weigend, A.S.: A neural network apporach to topic spotting. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval(SDAIR'95) (1995)

[24] Yiming, Y.: Noise Reduction in a Statistical Approach to Text Categorization. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95) (1995) 256–263

[25] Salton, G.: Automatic text processing: the transformation analysis and retrieval of information by Computer. Reading, Pennsylvania: Aoldison-wesley (1989)

[26] Bin, L., Tiejun, H., Jun, C., Wen, G.: A New Statistical-based Method in Automatic Text Classification. Journal of Chinese information processing. **16** (2002) 18–24