Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (11): 2123-2126, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2123.2126

Blocking Distribution Based Hierarchical Reconstruction for Text Categorization

Wen Li, Weili Wang and Ling Chai
Information Engineering School, Nanchang University, 999Xuefu Road,
Honggutan New District, Nanchang, 330031, China

Abstract: As one of the important techniques in large-scale data organizing, text categorization has been widely investigated. But the existing hierarchical classification methods often suffer from inter-level error transmission, namely blocking. In this paper, blocking distribution based topology reconstruction method was proposed for hierarchical text categorization problem. Firstly, blocking distribution recognition technique is put forward to mining out the serious high-level misclassification class. Subsequently, original hierarchical structure are reconstructed using blocking direction information obtained ahead, which increasing the path for the blocking instance to the correct subclass. Experimental studies on Chinese text classification benchmark Tan Corp, demonstrate that the proposed algorithm performs better than the traditional hierarchical and state-of-the-art flat classification strategies.

Key words: Text categorization, blocking distribution, hierarchical reconstruction, large-scale data

INTRODUCTION

As one of the important techniques in digital data organizing, text categorization has been widely investigated. With the generation of large-scale data, much work has been focused on hierarchical classification scheme, which reply on a predefined hierarchical topology and use top-down classification process (He *et al*., 2012; Huang *et al*., 2004; Ceci and Malerba, 2007).

The main challenge of hierarchical approach is blocking, which means texts are misclassified by the higher-level classifiers and they have no chance to assign to the target category (Sun *et al*., 2003). In order to reduce blocking error, some approaches have been considered in recently years (Ruiz, 2001; Li *et al*., 2010).

The most straightforward solution is to add more channels. Yuan *et al*. (2004) considered two candidates category for each classifier, in other words, the text is assigned to the two categories which have the largest and the second largest probability value. The other solution is set a minimum acceptable threshold for each category node in the classification process (Sun *et al*., 2004). Text will be assigned to all sub-categories which the probability value is greater than a given threshold. However, such system performance greatly relies on threshold setting.

In this study, we present a blocking distribution based topology reconstruction method. The proposed algorithm is divided into two steps. Firstly, blocking distribution recognition technique is put forward to mining out the serious high-level misclassification class. Subsequently, original hierarchical structure are reconstructed using blocking direction information obtained ahead, which increasing the path for the blocking instance to the correct subclass.

The remainder of the paper is organized as follows: In the next section, we will introduce the blocking distribution based hierarchical reconstruction method in detail. Experimental results will be reported in section 3. Some conclusions to be draw in finally section.

PROPOSED ALGORITHM

Blocking distribution recognition: For hierarchical classification, given an internal class c_i, let $C = (c_{i1}, c_{i2}, ..., c_{ik})$ be a set of k target categories. Blocking distribution matrix BM is a k×k two-dimensional (2D) matrix:

$$BM(c_i) = \begin{pmatrix} 0 & \cdots & n_{in}^{il} & \cdots & n_{ik}^{il} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{il}^{im} & \cdots & n_{in}^{im} & \cdots & n_{ik}^{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{il}^{ik} & \cdots & n_{in}^{ik} & \cdots & 0 \end{pmatrix}$$

where, n_{in}^{im} is the number of blocking text which belongs to the subcategories of c_m but wrongly rejected by the classifiers and assigned to class c_n.

In our experiment hierarchical dataset, the first level contains 12 big categories (Tan *et al*., 2006): career (c_1), sport (c_2), medical (c_3), region (c_4), entertainment (c_5),

Corresponding Author: Wen Li, Nanchang University, 999Xuefu Road, Honggutan New District, Nanchang, 330031, China

Table 1: BMof experiment on TanCorp

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
c_1	0	0	0	0	0	0	3	0	0	0	0	1
c_2	0	0	0	0	0	0	0	0	7	0	0	0
c_3	0	1	0	0	0	0	2	0	2	6	0	0
c_4	0	0	1	0	0	0	0	1	0	0	2	0
c_5	0	0	1	0	0	0	0	0	1	0	1	0
c_6	0	0	0	0	0	0	0	0	0	0	0	0
c_7	9	0	2	0	0	0	0	0	6	4	3	0
c_8	0	0	0	0	0	0	0	0	0	0	0	0
c_9	1	0	0	0	0	0	1	1	0	2	0	2
c_{10}	0	1	13	0	0	0	0	1	2	0	1	0
c_{11}	0	1	0	0	35	0	2	0	2	3	0	0
c_{12}	3	0	1	0	0	0	0	1	12	3	0	0

estate (c_6), education (c_7), car (c_8), computer (c_9), science (c_{10}), art (c_{11}) and economy (c_{12}). We use SVM classifier (Joachims, 1998) with penalty cost set as 1 and the blocking distribution of first level classification results of experiment TanCorp are detailed reported in Table 1.

In order to evaluate the blocking direction of different classes, degree of blocking between class c_m and c_n denoted by DB_{Cn}^{Cm} is defined as follows:

$$S_{Cn}^{Cm} = \frac{n_{in}^{im}}{\text{num of text belongs to } c_n} \quad (1)$$

$$D_{Cn}^{Cm} = \frac{n_{in}^{im}}{\text{num of text assigned to } c_n} \quad (2)$$

$$DB_{Cn}^{Cm} = \begin{cases} 0, & \text{if } S_{Cn}^{Cm} = 0 \text{ or } D_{Cn}^{Cm} = 0 \\ \dfrac{2 \cdot S_{Cn}^{Cm} \cdot D_{Cn}^{Cm}}{S_{Cn}^{Cm} + D_{Cn}^{Cm}}, & \text{if } S_{Cn}^{Cm} > 0 \text{ and } D_{Cn}^{Cm} > 0 \end{cases} \quad (3)$$

Equation 1 and 2 calculate the blocking degree from relevance two class c_m and c_n, respectively. Eq. 3 using the harmonic mean of S_{Cn}^{Cm} and D_{Cn}^{Cm} to combined the two measure value by considering the misclassification flow of the associated two categories. By this way, it is easy to found the serious blocking direction.

Hierarchical reconstruction: The blocking degree calculated by the above method in the range 0 to 1 and the magnitude is small, so its further processed by Eq. 4:

$$Cm \overset{DB}{\rightarrow} Cn = \sqrt{\log(DB_{Cn}^{Cm} + 1)} \quad (4)$$

Rule: If:

$$(Cm \overset{DB}{\rightarrow} Cn) > \theta$$

it hypothesis that texts are obstruction from class c_m to class c_n. θ is an experience value determined by

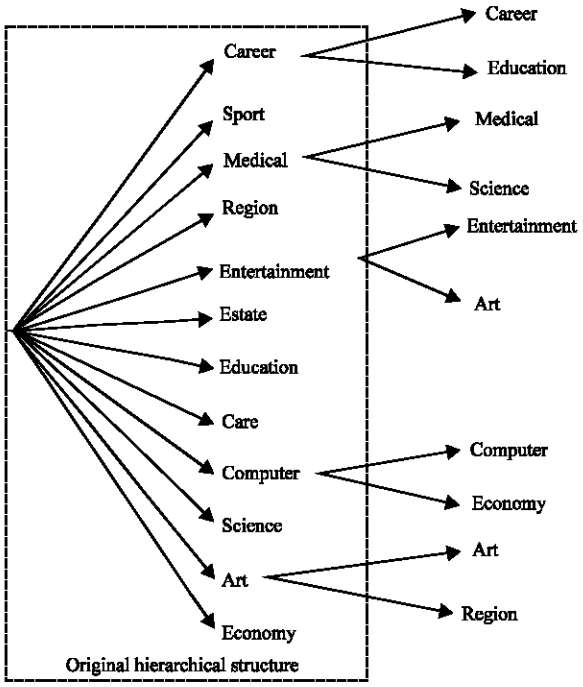

Fig. 1: Hierarchical reconstruction of experiment TanCorp

experiment. In our experiment on TanCorp, θ set to 0.11 is reasonable and the following five pairs are serious blocking direction:

$$\{art \overset{DB}{\rightarrow} entertainment, education \overset{DB}{\rightarrow} career, science \overset{DB}{\rightarrow} medical, economy \overset{DB}{\rightarrow} computer, region \overset{DB}{\rightarrow} art\}$$

For the serious blocking direction, if using the original hierarchical structure, the blocking texts have no chance to assign to the correct target category,so that the two relevant class classifier performance will be poor. This is bound to reconstruction on the predefined hierarchical topology, increasing the path for the blocking text to the correct subclass. On the basis of blocking distribution recognition, it is easy to refine the predefined hierarchy to selectively increase access to specific direction subclasses. The reconstructed hierarchical structure of experiment TanCorp is show in Fig. 1.

EXPERIMENT

Dataset and experimental setting: The Chinese hierarchical classification benchmark TanCorp has 14,150 texts (http://www.searchforum.org.cn /tansongbo/corpus.htm), collected by Songbo Tan. The corpus divided into two hierarchical levels. As mentioned in section 2, the first level contains 12 big categories and the second consists of 60 subclasses.

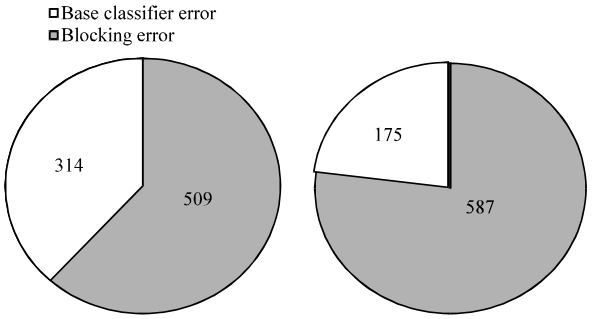

Fig. 2: Error distribution of original vs. reconstruction hierarchical structure

We adopt Vector Space Model for text representation. Features are selected according IG (Information Gain) weighting technique. To evaluate the classification performance, we present the results in both Macro-average and Micro-average F1. In our work, 70% documents are randomly sampled for training set and the remaining 30% are used for testing set.

Different from flat classification, hierarchical classification using a top-down process and the misclassification document exists following two situations:

- Blocking error: As hierarchical topology constraints, texts are misclassified by the high-level classifiers
- Base classifier error: High-level classifier has been correctly assigned texts to the target base classifier, but the base classifier misclassified text to the other leaf node

We will investigate the distribution of these two errors by experiments.

Experimental results and analysis: As mention in above section, analysis the error distribution of two types for hierarchical classification is an important and interesting work. Fig. 2 illustrates the two different error type distribution for original and reconstruction hierarchical structure. As all know, classification performance also tightly depends on feature numbers, the results was achieved when using 4000 feature which both methods can get stable and satisfactory performance.

As can be seen from Fig. 2, the proposed algorithm has 175 blocking error and it has less blocking error than original hierarchical classification which has 314 blocking error. It indicating that the refined structure accurately predicts the blocking direction and through the topology reconstruction, blocking texts have chance to arrive at the correct base class. By this way, it effectively prevents the occurrence of blocking error.

Table 2: Performance comparison among flat classification and two hierarchical classifications on TanCorp

Feature no.	Reconstruction hierarchical	Original hierarchical	Flat classification
Macro-F1 value			
500	65.58	63.76	48.78
1000	75.26	73.35	62.45
2000	75.21	73.29	62.30
3000	74.06	72.00	61.82
4000	76.09	74.10	61.78
5000	**76.26**	74.28	72.20
6000	74.90	73.78	61.92
7000	73.15	72.55	62.08
8000	73.42	72.59	62.09
Micro-F1 value			
500	74.46	73.50	66.49
1000	80.70	79.66	73.97
2000	78.84	79.64	73.83
3000	78.75	78.00	73.27
4000	**82.18**	80.74	72.87
5000	82.16	80.52	79.10
6000	80.81	80.60	72.38
7000	79.84	79.07	72.26
8000	80.06	79.12	72.00

But it also can be seen that the proposed algorithm cause a small amount of base classifier error increase. Manual analysis of these misclassified texts found the reason lie in: the refined topology adds channels for blocking text but at the same time, it will introduce more output category. The base classifier needs to compare more categories and it produces some interference to base classification procedure, resulting in a small number of text misclassification.

To evaluate the efficiency of our proposed method, Table 2 detailed report the classification performance of state-of-the-art hierarchical classification vs. the proposed hierarchical scheme. Performance of flat classification approach compared with hierarchical classification has been a major concern issue, the paper also carried out experiments on this problem. The Macro-average-F1 and Micro-average-F1 values were obtained by using 500, 1000, 2000, ... , 8000 features.

The observation of Table 2 indicates that: at most of the time, the reconstruction hierarchical scheme is outperform the original topology in both Macro and Micro F1. The best case of Macro-F1 value is 76.26% when using 5000 features and the best Micro-F1 is 82.18% when using 4000 features, which is approximately 1.98% and 1.45% higher than traditional hierarchical method.

It also easy found that the hierarchical classification model outperform flat classification. Meanwhile, the observation of Table 2 indicated that flat classification scheme showed poor stability, it obtain 72.2% Macro-F1 and 79.1% Micro-F1 which is the best performance when using 5000 features. But classification results under other feature dimension are not ideal.

CONCLUSION

In this study, blocking distribution based topology reconstruction method was proposed for hierarchical text categorization problem. Firstly, blocking distribution recognition technique was employed to mining out the serious high-level misclassification class. Then, original hierarchical structure are reconstructed using blocking direction information obtain ahead, which increasing the path for the blocking instance to the correct subclass.

Experimental studies on Chinese text classification dataset TanCorp, demonstrate that the proposed algorithm performs better than the state-of-the-art hierarchical classification method. In addition, compared with standard flat classification scheme, hierarchical classification methods show more stability.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61163023), the Science and Technology Supported Program of Jiangxi Province (Granted No. 20112BBE50045) and the Higher Education Reform Project of Jiangxi Province (Granted No.JXJG-12-1-38).

REFERENCES

Ceci, M. and D. Malerba, 2007. Classifying web documents in a hierarchy of categories: A comprehensive study. J. Intell. Inform. Syst., 28: 37-78.

He, L., Y. Jia, W. Han, S. Tan and Z. Chen, 2012. Research and development of large scale hierarchical classification problem. J. Comput., 35: 2101-2115.

Huang, C.C., S.L. Chuang and L.F. Chien, 2004. Liveclassifier: Creating hierarchical text classifiers through web corpora. Proceedings of the 13th International Conference on World Wide Web, May 17, 2004, New York, USA., pp: 184-192.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Li, W., D.Q. Miao, W. Wang and N. Zhang, 2010. Hierarchical rough decision theoretic framework for text classification. Proceedings of the 9th IEEE International Conference on Cognitive Informatics, July 7-9, 2010, Beijing, China, pp: 484-489.

Ruiz, M.E., 2001. Combining machine learning and hierarchical structures for text categorization. Ph.D. Thesis, Graduate College of University of Iowa, Ames, USA.

Sun, A., E.P. Lim and W.K. Ng, 2003. Performance measurement framework for hierarchical text classification. J. Am. Soc. Inform. Sci. Technol., 54: 1014-1028.

Sun, A., E.P. Lim, W.K. Ng and J. Srivastava, 2004. Blocking reduction strategies in hierarchical text classification. IEEE Trans. Knowl. Data Eng., 16: 1305-1308.

Tan, S., 2006. An effective refinement strategy for KNN Text classifier. Expert Syst. Appl., 30: 290-298.

Yuan, S., R. Li, S. Zhou and Y. Hu, 2004. Hierarchical Chinese document categorization. J. China Inst. Commun., 25: 55-63.