Impact of the Size of Training Set on Text Categorization

LI Jingyang, SUN Maosong

(State Key Lab of Intelligent Technology and Systems, Tsinghua University, 100084)

Abstract: Text Categorization approaches were often tested on different document collections but the pure relation between the size of training set and the classifier performance has rarely been systematically discussed. We carried out experiments on three document collections with term filtering by *Chi* and term weighting by *tfidf*, and attempt to reveal in this paper some attention-getting phenomena. A *Rocchio-style* classifier and a *k-nearest neighbor* classifier showed quite different characteristics when *training set* scales.

Keywords: Text Categorization, Training set.

1. Introduction

Statistic based Text Categorization (TC) systems commonly adopt machine learning method on document collection as the whole and only training data. So the characteristics of the training set play a key role in the performance of a classifier. Corpora variability may impact a lot on classifier performances [4]. So a particular approach was often examined on more than one document collection. But a pure size-oriented analysis is absent. The relation of F1-measure vs. category frequency was examined [3], but only on individual categories, not the whole training set. The common sense of "the larger the better" relation between corpus and performance is not for sure, despite the effectiveness of eliminating the sparseness problems [7].

Many components and their parameters participate in a TC procedure, which vary dependently or independently, such as the document collection and its size, class label set and its size, term selection and its dimension, term weighting, and classifier and its parameters and thresholds [10]. It is hard to sample them all but the result would still be meaningful and informative [4] by fixating some variable parts on some common or stable choices and focusing on the impact of size of training set.

2. Classifiers and Document Collections

To simplify the experiment settings and concentrate on the scaling problem, we only take account of single-label categorization [1]. As most classifiers are for independent binary classifying or category ranking (m-ary classifying by a cutting policy), they have the potential to deal with both single-label and multi-label categorization [4].

We chose two popular classifiers [3, 4]: Rocchio-style classifiers are commonly used as a weak baseline. For the de-emphasis of negative examples and single-labeling mode, we choose the centroid-based version (class-centroid vector, CCV) as a simple representative [1, 2]. The other one is k-nearest neighbor (kNN), which has been regarded as one of the top-performers (SVM, LLSF, etc.) [4, 6]. Main reason of the choice is that CCV is profile based, and kNN is instance based (lazy learning). They might show different sensibilities to training set scaling.

作者简介：李景阳，男，博士研究生，E-mail：lijingyang@gmail.com

is k-nearest neighbor (kNN), which has been regarded as one of the top-performers (SVM, LLSF, etc.) [4, 6]. Main reason of the choice is that CCV is profile based, and kNN is instance based (lazy learning). They might show different sensibilities to training set scaling.

We chose three corpora (Table 1), among which RCV1 [3] is large enough to be used for study on scaling problem. We took an accumulative policy to pick out documents proportional-equally from each category of (outside the test set) and add them into the training set, and started the TC system at some predefined sizes.

Table 1. Corpora

Name	Size (binary byte)	Doc Num	Cat Num	Test Doc Num
CN1[3]	$2.1*10^7$	3600	36	720
CN2[4]	$2.2*10^8$	71673	55	7140
RCV1[5]	$6.3*10^8$ (token files)	804414	101	2790

3. Settings of Classifier-free Steps

Chinese character bigrams (better than words [9]) and English words are chosen for document indexing. As kNN and Rocchio classifiers are both sensitive to irrelevant terms [3], the ChiMax criterion is chosen as the state-of-the-art best single-statistic criterion for term filtering [3, 5, 6]; the final term count is 70000 for CN1/CN2 and 5000 ([4]) for RCV1. Since ChiMax is know to be unreliable for rare terms [5, 6], we took a df cut (eliminating terms with df<3) [6]. tfidf is chosen for term weighting as a simple but effective statistic. MicroF1 is chosen for evaluation, while in single-label categorization, precision, recall, F-measure, and break-even-point have the same value under microaveraging method. The k value of kNN is set to 20 ([4]).

4. Results

Experiment results (Fig. 1) shows that kNN is a stable classifier which is not a surprise. But the analysis of the abnormal behavior of CCV is quite intractable.

For each multi-labeled document, any label with its sub-label (in the category hierarchy) also assigned on the document is removed from the doc. Docs still with multi labels are deleted.

CN1 is so small and heavily sparse, so the performance increased along with the size of training set. CN2 documents have obvious centralized topic but some category pairs are hard to distinct from each other and the class-centroids are mixed up. While RCV1 have many topic-amphibolous documents which resulted in undistinguishable class-centroids when the training set scaled larger and document set of each category

5 Conclusion

The partial comparison reminds us of the possibility that the size of training set might unusually impact on TC. More

[3] Data collection of *Chinese TC Contest* (2003) supported by National 863 Science Fund.
[4] Full-text of *Encyclopedia of China*.
[5] We used the token (not vector) files and topic categories of the RCV1-v2 version [3] For each multi-labeled document, any label with its sub-label (in the category hierarchy) also assigned on the document is removed from the doc. Docs still with multi labels are deleted.

additional comparisons are necessary in the future.

References

[27] Fabrizio Sebastiani: Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34(1). ACM Press New York (2002) 1-47

[28] Susan Domais, John Platt, David Heckerman: Inductive Learning Algorithms and Representations for Text Categorization. Proceedings of CIKM (1998) 148-155

[29] David D. Lewis, Yiming Yang, Tony G. Rose, Fan Li: RCV1: A New Benchmark Collection for Text Categorization Research. JMLR (2004) 5:361-397

[30] Yiming Yang: An Evaluation of Statistical Approaches to Text Categorization. Technical report CMU-CS-97-127, Computer Science Dept., Carnegie Mellon University (1997)

[31] Yiming Yang, Jan O. Pedersen: A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML (1997) 412-420

[32] Monica Rogati, Yiming Yang: High-performing Feature Selection for Text Classification. ACM Conference on Information and Knowledge Management (2002) 659-661

[33] James R. Curran, Miles Osborne: A Very Very Large Corpus doesn't Always Yield Reliable Estimates. Meeting of the Association for Computational Linguistics (2001)

[34] Yiming Yang, Jian Zhang, Bryan Kisiel: A Scalability Analysis of Classifiers in Text Categorization. Proceedings of the 28th International ACM SIGIR Conference (2003) 96103

[35] Jianyun Nie, Fuji Ren: Chinese Information Retrieval: Using Characters or Words? Information Processing and Management Vol. 35, (1999) 443-462

[36] Yiming Yang: A Study on Thresholding Strategies for Text Categorization. Proceedings of 24th International ACM SIGIR Conference (2001)