A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines

Man Lan*
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
lanman@i2r.a-star.edu.sg

Chew-Lim Tan
Department of Computer Science
National University of Singapore, 3 Science Drive 2, Singapore 117543
tancl@comp.nus.edu.sg

Hwee-Boon Low
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
hweeboon@i2r.a-star.edu.sg

Sam-Yuan Sung
Department of Computer Science
National University of Singapore, 3 Science Drive 2,Singapore 117543
ssung@comp.nus.edu.sg

ABSTRACT

Term weighting scheme, which has been used to convert the documents as vectors in the term space, is a vital step in automatic text categorization. In this paper, we conducted comprehensive experiments to compare various term weighting schemes with SVM on two widely-used benchmark data sets. We also presented a new term weighting scheme $tf.rf$ to improve the term's discriminating power. The controlled experimental results showed that this newly proposed $tf.rf$ scheme is significantly better than other widely-used term weighting schemes. Compared with schemes related with tf factor alone, the *idf* factor does not improve or even decrease the term's discriminating power for text categorization.

Categories and Subject Descriptors

I.7 [**Document and Text Processing**]: Document Preparation

General Terms

Performance

Keywords

term weighting schemes, text categorization, SVM

1. INTRODUCTION

Text categorization, the task of automatically assigning unlabelled documents into predefined categories, has been widely studied in the recent decades. Many researchers have studied text categorization based on different term weighting schemes and different kernel functions of SVMs [3] [4] [1]. In [1], the authors pointed out that it is the text representation schemes which dominate the performance of text categorization rather than the kernel functions of SVM. That is, choosing an appropriate term weighting scheme is more important than choosing and tuning kernel functions of SVM for text categorization.

However, even given these previous studies, we could not definitely draw a conclusion as to which term weighting scheme is better than others for SVM-based text categorization, because we know that comparisons are reliable only when based on experiments performed by the same author under carefully controlled conditions.

For this purpose, our study focused on the comparison of various term weighting schemes only. Specifically, our benchmark adopted the linear SVM algorithm and we used McNemar's significance tests [2] to validate if there is significant difference between two term weighting schemes.

2. TERM WEIGHTING SCHEMES

We adopted a tabular representation similar to that one in [5] and compared the following ten term weighting schemes listed in Table 1. Most of these term weighting schemes have been widely used in information retrieval and text categorization and/or have shown good performance in practice. Noted that other weighting schemes may exist, but these ten term weighting schemes were chosen due to their reported superior classification results or their typical representation when using support vector machines. For example, ITF representation proposed by [1] is included because the experimental results show that when combined with linear kernel of SVM it needs the minimum of support vectors (i.e. best generalization).

From this table, we can find that the first four term weighting schemes are different variants of tf factor. Then the next

*Ms. Lan is also a PhD candidate in the Department of Computer Science, National University of Singapore.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

Table 1: Summary of term weighting schemes

NAME	DESCRIPTION
binary	binary feature representation (1 for presence and 0 for absence)
tf	tf only
$logtf$	$log(1+tf)$
ITF	$1-1/(1+tf)$
idf	idf alone ($idf = log(N/n_i)$)
$tf.idf$	classic $tf.idf$
$logtf.idf$	$log(1+tf).idf$
$tf.idf$-$prob$	probabilistic idf, actually is the approximate $tf.term\ relevance$ [5]
$tf.chi$	$tf.\chi^2$
$tf.rf$	$tf.relevance\ frequency$ is our new weighting scheme (rf=$log(1+n_i/n_{i-})$)

four schemes are different variants of $tf.idf$. The $tf.chi$ scheme is a typical representation which combines tf factor with one feature selection metric (here is χ^2). The last weighting representation is our newly presented scheme in order to improve the term's discriminating power for text categorization.

3. COMPARATIVE EXPERIMENTS

3.1 Benchmark Methodology

To compare the performance between two term weighting schemes, we employed the McNemar's significance tests [2] based on the micro-averaged precision/recall *break-even point*, which is defined as the value where *recall* equals to *precision*.

The first data collection we used is from the top 10 largest categories of the Reuters-21578 corpus. One noticeable issue of Reuters corpus is the skewed category distribution problem. For the second data collection, we randomly selected 300 samples per category among 20 categories from the 20 Newsgroups corpus. Compared with the skewed category distribution in the Reuters corpus, the 20 categories in the 20 Newsgroups corpus are uniform distribution.

3.2 Results and Discussion

Figure 1 and 2 depict the micro-averaged break-even point performance on the Reuters and the 20 Newsgroups data sets by using ten term weighting schemes at different number of features, respectively. To achieve high performance in terms with break-even point, different number of vocabularies is required for the two data sets; however, both of the best break-even points are achieved by using our newly presented scheme $tf.rf$. Furthermore, these term weighting schemes have been shown consistent performance compared with the others on the two different data sets. The trends are distinctive that the $tf.rf$ scheme has always been shown significant better performance than others. It is clearly to know that all the observations are supported by the following McNemar's significance tests.

4. CONCLUSIONS

Based on the experimental results, our conclusions are:

- Our newly presented $tf.rf$ scheme shows significant better performance than other schemes based on two widely-used data sets with different category distributions
- The schemes related with term frequency alone, such as tf, $logtf$, ITF show rather good performance but still worse than the $tf.rf$ scheme
- The idf factor, taking the collection distribution into consideration, does not improve or even decrease the term's discriminating power for text categorization
- The $binary$ and $tf.chi$ representations significantly underperform the other term weighting schemes

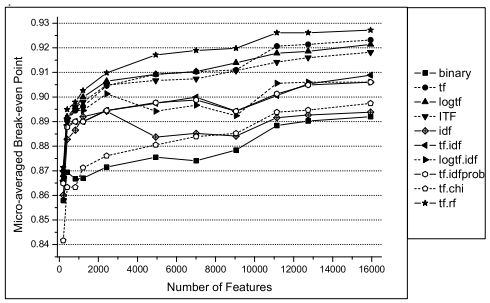

Figure 1: Results for the Reuters-21578 corpus

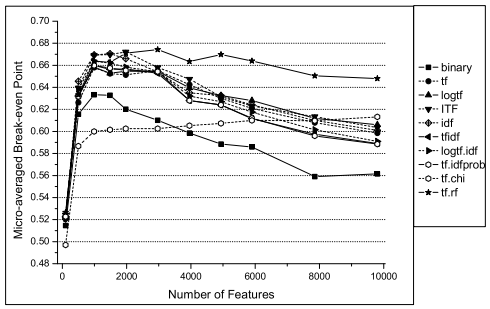

Figure 2: Results for the 20 Newsgroups corpus

5. REFERENCES

[1] E. Leopold and J. Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423 – 444, January - February - March 2002.

[2] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, 1998.

[3] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM Press, 1998.

[4] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[5] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.