Scalable Term Selection for Text Categorization

Jingyang Li
National Lab of Intelligent Tech. & Sys.
Department of Computer Sci. & Tech.
Tsinghua University, Beijing, China
lijingyang@gmail.com

Maosong Sun
National Lab of Intelligent Tech. & Sys.
Department of Computer Sci. & Tech.
Tsinghua University, Beijing, China
sms@tsinghua.edu.cn

Abstract

In text categorization, term selection is an important step for the sake of both categorization accuracy and computational efficiency. Different dimensionalities are expected under different practical resource restrictions of time or space. Traditionally in text categorization, the same scoring or ranking criterion is adopted for all target dimensionalities, which considers both the discriminability and the coverage of a term, such as χ^2 or IG. In this paper, the poor accuracy at a low dimensionality is imputed to the small average vector length of the documents. Scalable term selection is proposed to optimize the term set at a given dimensionality according to an expected average vector length. Discriminability and coverage are separately measured; by adjusting the ratio of their weights in a combined criterion, the expected average vector length can be reached, which means a good compromise between the specificity and the exhaustivity of the term subset. Experiments show that the accuracy is considerably improved at lower dimensionalities, and larger term subsets have the possibility to lower the average vector length for a lower computational cost. The interesting observations might inspire further investigations.

1 Introduction

Text categorization is a classical text information processing task which has been studied adequately (Sebastiani, 2002). A typical text categorization process usually involves these phases: document indexing, dimensionality reduction, classifier learning, classification and evaluation. The vector space model is frequently used for text representation (document indexing); dimensions of the learning space are called *terms*, or *features* in a general machine learning context. Term selection is often necessary because:

- Many *irrelevant* terms have detrimental effect on categorization accuracy due to *overfitting* (Sebastiani, 2002).
- Some text categorization tasks have many relevant but *redundant* features, which also hurt the categorization accuracy (Gabrilovich and Markovitch, 2004).
- Considerations on computational cost:
  (i) Many sophisticated learning machines are very slow at high dimensionalities, such as LLSF (Yang and Chute, 1994) and SVMs. (ii) In Asian languages, the term set is often very large and redundant, which causes the learning and the predicting to be really slow. (iii) In some practical cases the computational resources (time or space) are restricted, such as hand-held devices, real-time applications and frequently retrained systems. (iv) Some deeper analysis or feature reconstruction techniques rely on matrix factorization (e.g. LSA based on SVD), which might be computationally intractable while the dimensionality is large.

Sometimes an aggressive term selection might be needed particularly for (iii) and (iv). But it is notable that the dimensionality is not always directly

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 774–782, Prague, June 2007.

connected to the computational cost; this issue will be touched on in Section 6. Although we have many general feature selection techniques, the domain specified ones are preferred (Guyon and Elisseeff, 2003). Another reason for ad hoc *term* selection techniques is that many other pattern classification tasks has no *sparseness problem* (in this study the sparseness means a sample vector has few nonzero elements, but not the high-dimensional learning space has few training samples). As a basic motivation of this study, we hypothesize that the low accuracy at low dimensionalities is mainly due to the sparseness problem.

Many term selection techniques were presented and some of them have been experimentally tested to be high-performing, such as Information Gain, χ^2 (Yang and Pedersen, 1997; Rogati and Yang, 2002) and Bi-Normal Separation (Forman, 2003). Everyone of them adopt a criterion scoring and ranking the terms; for a target dimensionality d, the term selection is simply done by picking out the top-d terms from the ranked term set. These high performing criteria have a common characteristic — both discriminability and coverage are implicitly considered.

- *discriminability*: how unbalanced is the distribution of the term among the categories.
- *coverage*: how many documents does the term occur in.

(Borrowing the terminologies from document indexing, we can say the *specificity* of a term set corresponds to the discriminability of each term, and the *exhaustivity* of a term set corresponds to the coverage of each term.) The main difference among these criteria is to what extent the discriminability is emphasized or the coverage is emphasized. For instance, empirically *IG* prefers high frequency terms more than χ^2 does, which means *IG* emphasizes the coverage more than χ^2 does.

The problem is, these criteria are nonparametric and do the same ranking for any target dimensionality. Small term sets meet the specificity–exhaustivity dilemma. If really the sparseness is the main reason of the low performance of a small term set, the specificity should be moderately sacrificed to improve the exhaustivity for a small term set; that is to say, the term selection criterion should consider coverage more than discriminability. Contrariwise, coverage could be less considered for a large term set, because we need worry little about the sparseness problem and the computational cost might decrease.

The remainder of this paper is organized as follows: Section 2 describes the document collections used in this study, as well as other experiment settings; Section 3 investigates the relation between sparseness (measured by *average vector length*) and categorization accuracy; Section 4 explains the basic idea of scalable term selection and proposed a potential approach; Section 5 carries out experiments to evaluate the approach, during which some empirical rules are observed to complete the approach; Section 6 makes some further observations and discussions based on Section 5; Section 7 gives a concluding remark.

2 Experiment Settings

2.1 Document Collections

Two document collections are used in this study.

CE (Chinese Encyclopedia): This is from the electronic version of the Chinese Encyclopedia. We choose a Chinese corpus as the primary document collection because Chinese text (as well as other Asian languages) has a very large term set and a satisfying subset is usually not smaller than 50000 (Li et al., 2006); on the contrary, a dimensionality lower than 10000 suffices a general English text categorization (Yang and Pedersen, 1997; Rogati and Yang, 2002). For computational cost reasons mentioned in Section 1, Chinese text categorization would benefit more from an high-performing aggressive term selection. This collection contains 55 categories and 71674 documents (9:1 split to training set and test set). Each documents belongs to only one category. Each category contains 399–3374 documents. This collection was also used by Li et al. (2006).

20NG (20 Newsgroups[1]): This classical English document collection is chosen as a secondary in this study to testify the generality of the proposed approach. Some figures about this collection are not shown in this paper as the figures about CE, viz. Figure 1–4 because they are similar to CE's.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups

2.2 Other Settings

For CE collection, character bigrams are chosen to be the indexing unit for its high performance (Li et al., 2006); but the bigram term set suffers from its high dimensionality. This is exactly the case we tend to tackle. For 20NG collection, the indexing units are stemmed[2] words. Both term set are *df*-cut by the most conservative threshold ($df \geq 2$). The sizes of the two candidate term sets are $|\mathcal{T}_{\text{CE}}| = 1067717$ and $|\mathcal{T}_{\text{20NG}}| = 30220$.

Term weighting is done by $\textit{tfidf}(t_i, d_j) = \log(\textit{tf}(t_i, d_j) + 1) \cdot \log\left(\frac{df(t_i)+1}{N_d}\right)$[3], in which t_i denotes a term, d_j denotes a document, N_d denotes the total document number.

The classifiers used in this study are support vector machines (Joachims, 1998; Gabrilovich and Markovitch, 2004; Chang and Lin, 2001). The kernel type is set to linear, which is fast and enough for text categorization. Also, Brank et al. (2002) pointed out that the complexity and sophistication of the criterion itself is more important to the success of the term selection method than its compatibility in design with the classifier.

Performance is evaluated by microaveraged F_1-measure. For single-label tasks, microaveraged *precision*, *recall* and F_1 have the same value.

χ^2 is used as the term selection baseline for its popularity and high performance. (*IG* was also reported to be good. In our previous experiments, χ^2 is generally superior to *IG*.) In this study, features are always selected *globally*, which means the maximum are computed for category-specific values (Sebastiani, 2002).

3 Average Vector Length (AVL)

In this study, *vector length* (how many different terms does the document hold after term selection) is used as a straightforward sparseness measure for a document (Brank et al., 2002). Generally, document sizes have a *lognormal distribution* (Mitzenmacher, 2003). In our experiment, vector lengths are also found to be nearly lognormal distributed, as shown in Figure 1. If the correctly classified documents and the wrongly classified documents are separately investigated, they both yield a nearly lognormal distribution.

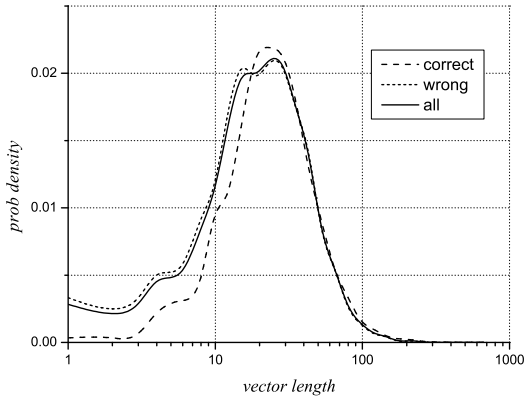

Figure 1: Vector Length Distributions (smoothed), on CE Document Collection

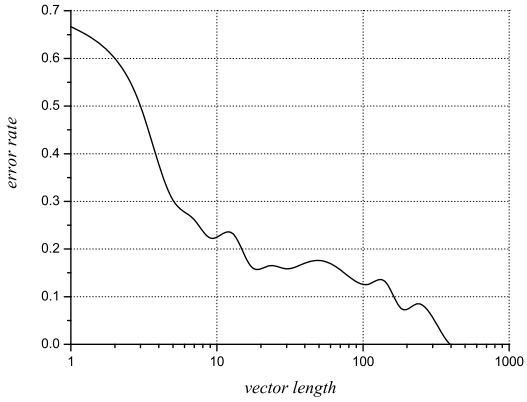

Figure 2: Error Rate vs. Vector Length (smoothed), on CE Collection, 5000 Dimensions by χ^2

Also in Figure 1, wrongly classified documents shows a relatively large proportion at low dimensionalities. Figure 2 demonstrates this with more clarity. Thus the hypothesis formed in Section 1 is confirmed: there is a strong correlation between the sparseness degree and the categorization error rate.

Therefore, it is quite straightforward a thought to measure the "sparseness of a term subset" (or more precisely, the exhaustivity) by the corresponding *average vector length* (*AVL*) of all documents.[4] In the

[2] Stemming by Porter's Stemmer (`http://www.tartarus.org/ martin/PorterStemmer/`).

[3] In our experiments this form of *tfidf* always outperforms the basic $\textit{tfidf}(t_i, d_j) = \textit{tf}(t_i, d_j) \cdot \log\left(\frac{df(t_i)+1}{N_d}\right)$ form.

[4] Due to the lognormal distribution of vector length, it seems more plausible to average the logarithmic vector length. However, for a fixed number of documents , $\log \frac{\sum |d_j|}{|\mathcal{D}|}$ should hold a nearly fixed ratio to $\frac{\sum \log |d_j|}{|\mathcal{D}|}$, in which $|\mathcal{D}|$ denotes the document number and $|d_j|$ denotes the document vector length.

remainder of this paper, (log) *AVL* is an important metric used to assess and control the sparseness of a term subset.

4 Scalable Term Selection (STS)

Since the performance droping down at low dimensionalities is attributable to low *AVL*s in the previous section, a scalable term selection criterion should *automatically* accommodate its favor of high coverage to different target dimensionalities.

4.1 Measuring Discriminability and Coverage

The first step is to separately measure the discriminability and the coverage of a term. A basic guideline is that these two metrics should not be highly (positive) correlated; intuitively, they should have a slight negative correlation. The correlation of the two metrics can be visually estimated by the joint distribution figure. A bunch of term selection metrics were explored by Forman (2003). *df* (*document frequency*) is a straightforward choice to measure coverage. Since *df* follows the Zipf's law (inverse power law), $\log(df)$ is adopted. High-performing term selection criterion themselves might not be good candidates for the discriminability metric because they take coverage into account. For example, Figure 3 shows that χ^2 is not satisfying. (For readability, the grayness is proportional to the log probability density in Figure 3, Figure 4 and Figure 12.) Relatively, *probability ratio* (Forman, 2003) is a more straight metric of discriminability.

$$PR(t_i, c) = \frac{P(t_i|c_+)}{P(t_i|c_-)} = \frac{df(t_i, c_+)/df(c_+)}{df(t_i, c_-)/df(c_-)}$$

It is a symmetric ratio, so $\log(PR)$ is likely to be more appropriate. For multi-class categorization, a global value can be assessed by $PR_{\max}(t_i) = \max_c PR(t_i, c)$, like $\chi^2_{\max}$ for χ^2 (Yang and Pedersen, 1997; Rogati and Yang, 2002; Sebastiani, 2002); for brief, *PR* denotes $PR_{\max}$ hereafter. The joint distribution of $\log(PR)$ and $\log(df)$ is shown in Figure 12. We can see that the distribution is quite even and they have a slight negative correlation.

4.2 Combined Criterion

Now we have the two metrics: $\log(PR)$ for discriminability and $\log(df)$ for coverage, and a parametric

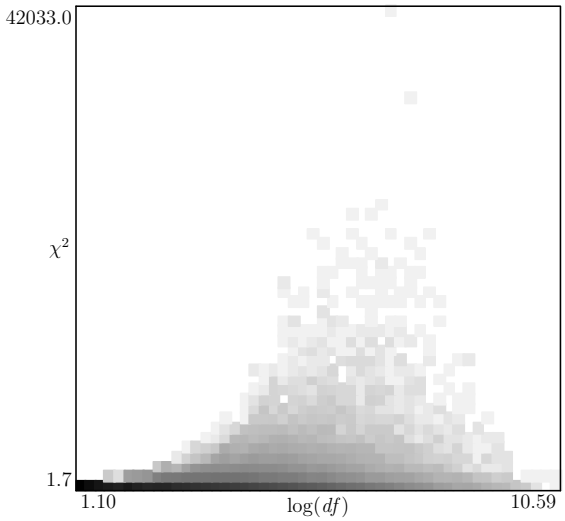

Figure 3: $\big(\log(df), \chi^2\big)$ Distribution, on CE

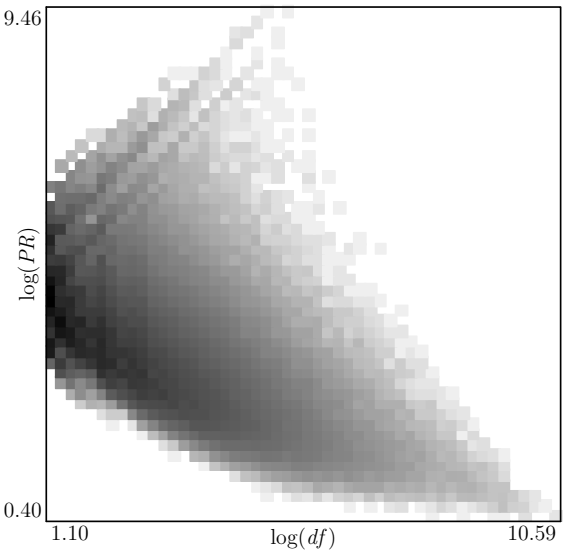

Figure 4: $(\log(df), \log(PR))$ Distribution, on CE

term selection criterion comes forth:

$$\zeta(t_i; \lambda) = \left(\frac{\lambda}{\log(PR(t_i))} + \frac{1-\lambda}{\log(df(t_i))} \right)^{-1}$$

A *weighted harmonic averaging* is adopted here because either metric's being *too small* is a severe detriment. $\lambda \in [0, 1]$ is the weight for $\log(PR)$, which denotes how much the discriminability is emphasized. When the dimensionality is fixed, a smaller λ leads to a larger *AVL* and a larger λ leads to a smaller *AVL*. The optimal λ should be a function

of the expected dimensionality (k):

$$\lambda^*(k) = \arg\max_{\lambda} F_1(\mathcal{S}_k(\lambda))$$

in which the term subset $\mathcal{S}_k(\lambda) \in \mathcal{T}$ is selected by $\zeta(\circ;\lambda)$, $|\mathcal{S}_k| = k$, and F_1 is the default evaluation criterion. Naturally, this optimal λ leads to a corresponding optimal *AVL*:

$$AVL^*(k) \longleftrightarrow \lambda^*(k)$$

For a concrete implementation, we should have an (empirical) function to estimate λ^* or AVL^*:

$$AVL^\circ(k) \doteq AVL^*(k)$$

In the next section, the values of AVL^* (as well as λ^*) for some k-s are figured out by experimental search; then an empirical formula, $AVL^\circ(k)$, comes forth. It is interesting and inspiring that by adding the "corpus *AVL*" as a parameter this formula is universal for different document collections, which makes the whole idea valuable.

5 Experiments and Implementation

5.1 Experiments

The expected dimensionalities (k) chosen for experimentation are
CE: 500, 1000, 2000, 4000, ..., 32000, 64000;
20NG: 500, 1000, 2000, ..., 16000, 30220.[5]

For a given document collection and a given target dimensionality, there is a corresponding *AVL* for a λ, and vice versa (for the possible value range of *AVL*). According to the observations in Section 5.2, *AVL* other than λ is the direct concern because it is more intrinsic, but λ is the one that can be tuned directly. So, in the experiments, we vary *AVL* by tuning λ to produce it, which means to calculate $\lambda(AVL)$.

$AVL(\lambda)$ is a monotone function and fast to calculate. For a given *AVL*, the corresponding λ can be quickly found by a Newton iteration in [0,1]. In fact, $AVL(\lambda)$ is not a continuous function, so λ is only tuned to get an acceptable match, e.g. within ± 0.1.

[5] STS is tested to the whole $\mathcal{T}$ on 20NG but not on CE, because (i) $\mathcal{T}_{\text{CE}}$ is too large and time consuming for training and testing, and (ii) χ^2 was previously tested on larger k and the performance (F_1) is not stable while $k > 64000$.

For each k, by the above way of fitting λ, we manually adjust *AVL* (only in integers) until $F_1(\mathcal{S}_k(\lambda(AVL)))$ peaks. By this way, Figure 5–11 are manually tuned best-performing results as observations for figuring out the empirical formulas.

Figure 5 shows the F_1 peaks at different dimensionalities. Comparing to χ^2, STS has a considerable potential superiority at low dimensionalities. The corresponding values of AVL^* are shown in Figure 6, along with the *AVL*s of χ^2-selected term subsets. The dotted lines show the trend of AVL^*; at the overall dimensionality, $|\mathcal{T}_{\text{CE}}| = 1067717$, they have the same $AVL = 898.5$. We can see that $\log(AVL^*)$ is almost proportional to $\log(k)$ when k is not too large. The corresponding values of λ^* are shown in Figure 7; the relation is nearly linear between λ^* and $\log(k)$.

Now it is necessary to explain why an empirical $AVL^\circ(k)$ derived from the straight line in Figure 6 can be used instead of $AVL^*(k)$ in practice. One important but not plotted property is that the performance of STS is not very sensitive to a small value change of *AVL*. For instance, at $k = 4000$, $AVL^* = 120$ and the F_1 peak is 85.8824%, and for $AVL = 110$ and 130 the corresponding F_1 are 85.8683% and 85.6583%; at the same k, the F_1 of χ^2 selection is 82.3950%. This characteristic of STS guarantee that the empirical $AVL^\circ(k)$ has a very close performance to $AVL^*(k)$; due to the limited space, the performance curve of $AVL^\circ(k)$ will not be plotted in Section 5.2.

Same experiments are done on 20NG and the results are shown in Figure 8, Figure 9 and Figure 10. The performance improvements is not as significant as on the CE collection; this will be discussed in Section 6.2. The conspicuous relations between AVL^*, λ^* and k remain the same.

5.2 Algorithm Completion

In Figure 6 and Figure 9, the ratios of $\log(AVL^*(k))$ to $\log(k)$ are not the same on CE and 20NG. Taking into account the *corpus AVL* (the *AVL* produced by the whole term set): $AVL_{\mathcal{T}_{\text{CE}}} = 898.5286$ and $AVL_{\mathcal{T}_{\text{20NG}}} = 82.1605$, we guess $\frac{\log(AVL^*(k))}{\log(AVL_{\mathcal{T}})}$ is capable of keeping the same ratio to $\log(k)$ for both CE and 20NG. This hypothesis is confirmed (not for too high dimensionalities) by Figure 11; Section 6.2

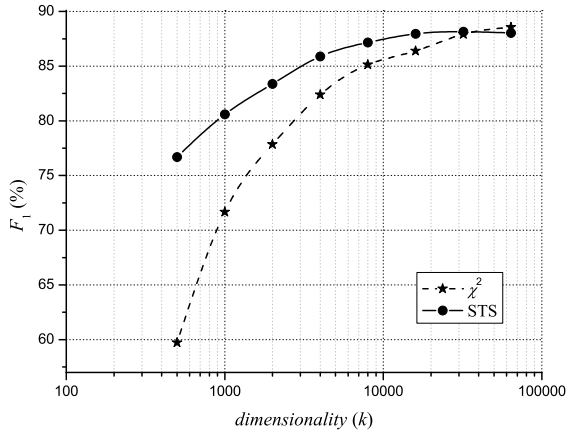

Figure 5: Performance Comparison, on CE

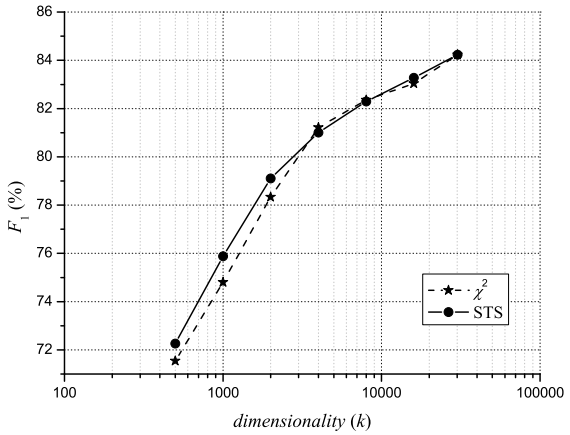

Figure 8: Performance Comparison, on 20NG

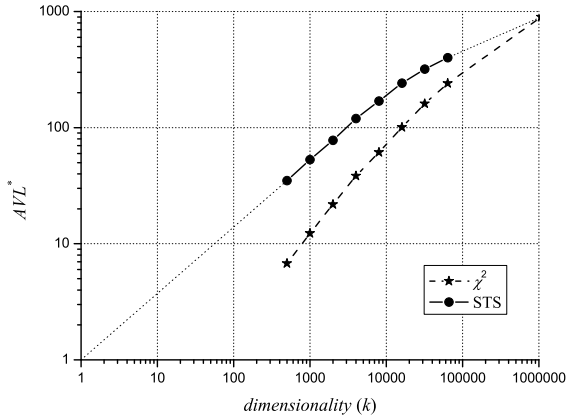

Figure 6: *AVL* Comparison, on CE

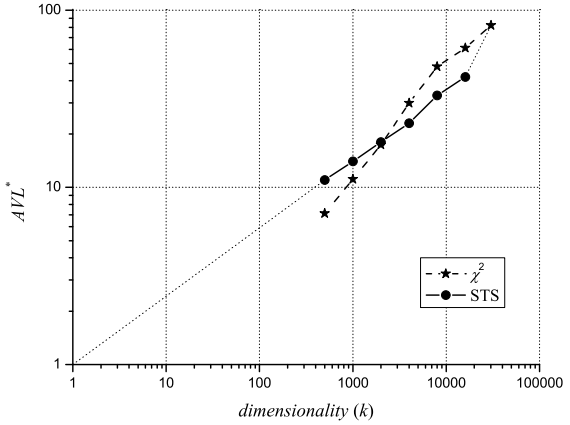

Figure 9: *AVL* Comparison, on 20NG

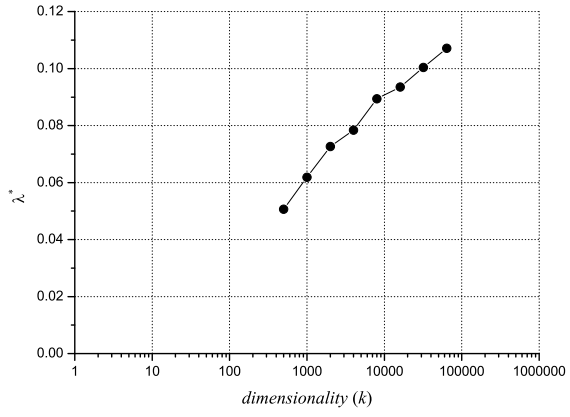

Figure 7: Optimal Weights of $\log(PR)$, on CE

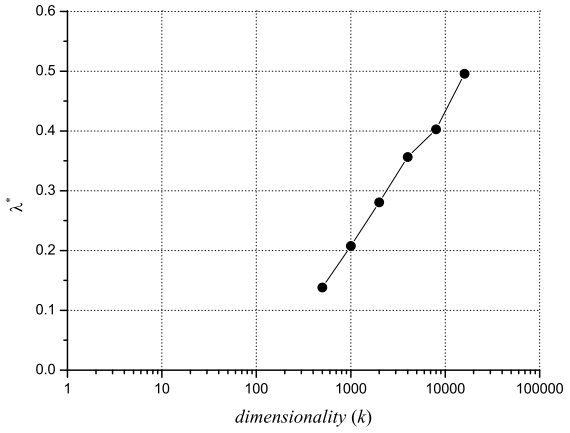

Figure 10: Optimal Weights of $\log(PR)$, on 20NG

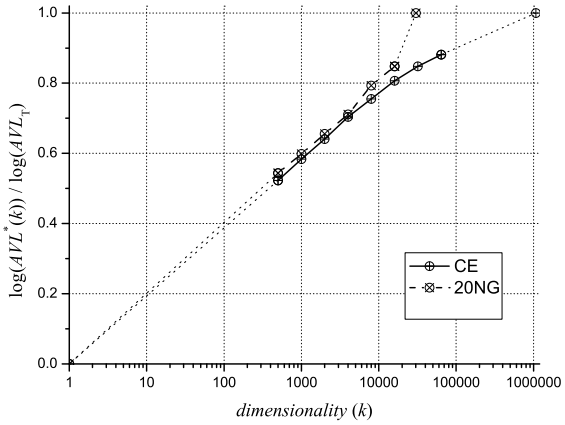

Figure 11: $\frac{\log(AVL^*(k))}{\log(AVL_{\mathcal{T}})}$, on Both CE and 20NG

contains some discussion on this.

From the figure, we get the value of this ratio (the base of log is set to e):

$$\gamma = \frac{\log(AVL^*(k))/\log(AVL_{\mathcal{T}})}{\log(k)} \cong 0.085$$

which should be a universal constant for all text categorization tasks.

So the empirical estimation of $AVL^*(k)$ is given by

$$\begin{aligned} AVL^{\circ}(k) &= \exp(\gamma \log(AVL_{\mathcal{T}}) \cdot \log(k)) \\ &= AVL_{\mathcal{T}}^{\gamma \log(k)} \end{aligned}$$

and the final STS criterion is

$$\begin{aligned} \zeta(t_i, k) &= \zeta(t_i; \lambda(AVL^{\circ}(k))) \\ &= \zeta(t_i; \lambda(AVL_{\mathcal{T}}^{\gamma \log(k)})) \end{aligned}$$

in which $\lambda(\circ)$ can be calculated as in Section 5.1. The target dimensionality, k, is involved as a parameter, so the approach is named *scalable* term selection. As stated in Section 5.1, $AVL^{\circ}(k)$ has a very close performance to $AVL^*(k)$ and its performance is not plotted here.

6 Further Observation and Discussion

6.1 Comparing the Selected Subsets

An investigation shows that for a quite large range of λ, term rankings by $\zeta(t_i; \lambda)$ and $\chi^2(t_i)$ have a strong correlation (the *Spearman's rank correlation coefficient* is bigger than 0.999). In order to compare the two criteria's preferences for discriminability and coverage, the selected subsets of different dimensionalities are shown in Figure 12 (the corresponding term density distribution was shown in Figure 4) and Figure 13. For different dimension-

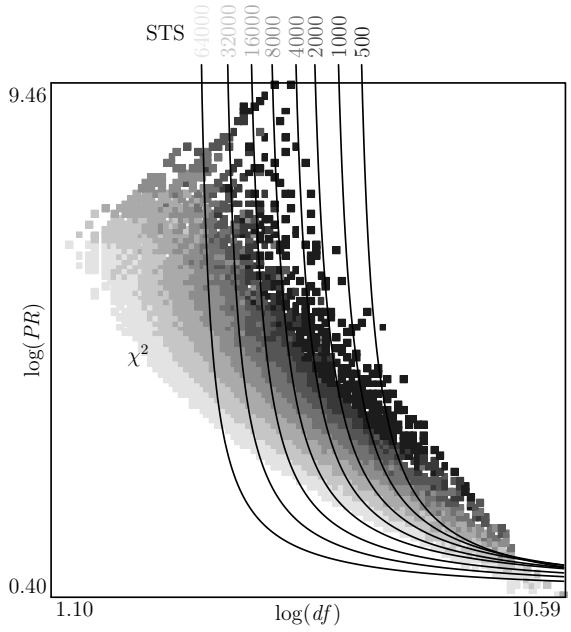

Figure 12: Selection Area Comparison of STS and χ^2 on Various Dimensionalities, on CE

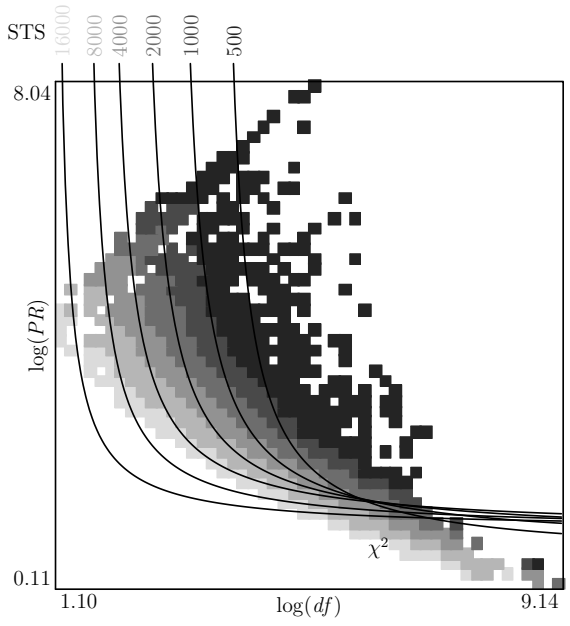

Figure 13: Selection Area Comparison of STS and χ^2 on Various Dimensionalities, on 20NG

alities, the selection areas of STS are represented by boundary lines, and the selection areas of χ^2 are represented by different grayness.

In Figure 12, STS shows its superiority at low dimensionalities by more emphasis on the coverage of terms. In Figure 13, STS shows its superiority at high dimensionalities by more emphasis on the discriminability of terms; lower coverage yields smaller index size and lower computational cost. At any dimensionality, STS yields a relatively fixed bound for either discriminability or coverage, other than a compromise between them like χ^2; this is attributable to the harmonic averaging.

6.2 Adaptability of STS

There are actually two kinds of sparseness in a (vectorized) document collection:

collection sparseness: the high-dimensional learning space contains few training samples;

document sparseness: a document vector has few nonzero dimensions.

In this study, only the document sparseness is investigated. The collection sparseness might be a backroom factor influencing the actual performance on different document collections. This might explain why the explicit characteristics of STS are not the same on CE to 20NG: (comparing with χ^2, see Figure 5, Figure 6, Figure 8 and Figure 9)

CE. The significant F_1 improvements at low dimensionalities sacrifice the short of *AVL*. In some learning process implementations, it is *AVL* other than k that determines the computational cost; in many other cases, k is the determinant. Further more, possible post-processing, like matrix factorization, might benefit from a low k.

20NG. The F_1 improvements at low dimensionalities is not quite significant, but *AVL* remains a lower level. For higher k, there is less difference in F_1, but the smaller *AVL* yield lower computational cost than χ^2.

Nevertheless, STS shows a stable behavior for various dimensionalities and quite different document collections. The existence of the universal constant γ empowers it to be adaptive and practical. As shown in Figure 11, STS draws the *relative* $\log AVL^*(k)$ to the same straight line, $\gamma \log(k)$, for different document collections. This might means that the *relative AVL* is an intrinsic demand for the term subset size k.

7 Conclusion

In this paper, Scalable Term Selection (STS) is proposed and supposed to be more adaptive than traditional high-performing criteria, viz. χ^2, *IG*, *BNS*, etc. The basic idea of STS is to separately measure discriminability and coverage, and adjust the relative importance between them to produce a optimal term subset of a given size. Empirically, the constant relation between target dimensionality and the optimal relative average vector length is found, which turned the idea into implementation.

STS showed considerable adaptivity and stability for various dimensionalities and quite different document collections. The categorization accuracy increasing at low dimensionalities and the computational cost decreasing at high dimensionalities were observed.

Some observations are notable: the loglinear relation between optimal average vector length (AVL^*) and dimensionality (k), the semi-loglinear relation between weight λ and dimensionality, and the universal constant γ. For a future work, STS needs to be conducted on more document collections to check if γ is really universal.

In addition, there could be other implementations of the general STS idea, via other metrics of discriminability and coverage, other weighted combination forms, or other term subset evaluations.

Acknowledgement

The research is supported by the National Natural Science Foundation of China under grant number 60573187, 60621062 and 60520130299.

References

Janez Brank, Marko Grobelnik, Nataša Milic-Fraylingand, and Dunjia Mladenic. 2002. Interaction of feature selection methods and linear classification models. *Workshop on Text Learning held at ICML-2002*.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Evgeniy Gabrilovich and Shaul Markovitch. 2004. Text categorization with many redundant features: using aggressive feature selection to make svms competitive with c4.5. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 41, New York, NY, USA. ACM Press.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML '98*, number 1398, pages 137–142. Springer Verlag, Heidelberg, DE.

Jingyang Li, Maosong Sun, and Xian Zhang. 2006. A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization. In *Proceedings of COLING-ACL '06*, pages 545–552. Association for Computational Linguistics, July.

Michael Mitzenmacher. 2003. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1:226–251.

Monica Rogati and Yiming Yang. 2002. High-performing feature selection for text classification. In *Proceedings of CIKM '02*, pages 659–661. ACM Press.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Yiming Yang and Christopher G. Chute. 1994. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, 12(3):252–277.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.