Two-dimensional Clustering for Text Categorization

Hiroya Takamura and **Yuji Matsumoto**
Department of Information Technology
Nara Institute of Science and Technology
8516-9, Takayama, Ikoma, Nara 630-0101, Japan
{hiroya-t,matsu}@is.aist-nara.ac.jp

Abstract

We propose a new method to improve the accuracy of Text Categorization using two-dimensional clustering. In a number of previous probabilistic approaches, texts in the same category are implicitly assumed to be generated from an identical distribution. We empirically show that this assumption is not accurate, and propose a new framework based on two-dimensional clustering to alleviate this problem. In our method, training texts are clustered so that the assumption is more likely to be true, and at the same time, features are also clustered in order to tackle the data sparseness problem. We conduct some experiments to validate the proposed two-dimensional clustering method.

1 Introduction

Text Categorization is the task of classifying texts into their most plausible category. One problem in most previous probabilistic approaches to Text Categorization is that texts in the same category are assumed to be generated by an identical distribution (we call it the i.d. assumption, in this paper). However, categories are manually defined and there is no predefined probabilistic structure behind them, as discussed in the next section. Another problem with Text Categorization is the data-sparseness problem caused by the high dimensionality of the feature space. The frequency of each word is usually so small that it is difficult to estimate reliable statistics.

In order to tackle these problems, we propose a new framework based on two-dimensional clustering. We first cluster training texts into several clusters whose elements can be thought as being generated from an identical distribution before estimating the probability model of each category. The data-sparseness problem is more critical, if the number of parameters is larger as in the text clustering approach we are adopt. So we alleviate this problem by clustering features (words). That is to say, we *cluster both texts and features simultaneously.*

Through experiments, we show that our approach works well with probabilistic classifiers.

In Natural Language Processing, several clustering applications have been proposed, for example in (Brown, 1992; Li and Abe, 1998). Among those applications, (Baker et al, 1998) applied the class-distributional clustering to Text Categorization. They theoretically proved the optimality of their clustering method in terms of Naive Bayes Score, and validated it empirically. In class-distributional clustering, occurrences of categories given a word are regarded as a probability distribution, and words are clustered according to this distribution. In (Slonim and Tishby, 2001), they use the Information Bottleneck Method (Tishby et al, 1999) for Text Categorization. Both (Baker et al, 1998) and (Slonim and Tishby, 2001), however, deal only with word clustering. Unlike those methods, our method adopts a two-dimensional clustering of words and texts.

The idea of clustering words and texts simultaneously is also pursued in (Slonim and Tishby, 2000; Dhillon, 2001). However, those are concerned exclusively with clustering and do not propose any framework applicable to Text Categorization.

This paper is organized as follows. In Section 2, we investigate the i.d. assumption in Text Categorization. In Section 3, we describe our clustering methods. In Section 4, we explain our categorization methods. Section 5 presents the experiments and results with discussions. Finally, in Section 6, we summarize our research.

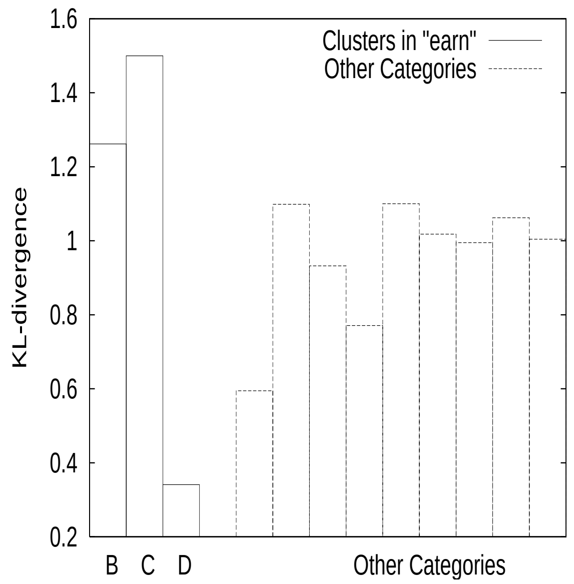

Figure 1: KL-divergence from cluster A in "earn"

2 Human-made Categories and Probabilistic Structure

As observed in the previous section, probabilistic structure does not always underlie the categories defined by humans. In order to provide empirical evidence for this observation, we conducted a small experiment using the Reuters-21578 data set. We first clustered a set of texts labeled with "earn", which is the largest category in this corpus. We obtained four clusters[1], each of which contains several hundreds documents or more. Then, regarding the occurrences of words given a cluster or a category as random events, we computed the KL (Kullback-Leibler) divergence between a cluster (A) in "earn" and the other three clusters (B, C and D) in "earn", and between the same cluster A and the (10 most frequent) categories other than "earn".

If the occurrences of words in texts in the same category are (approximately) identically distributed, the values of the KL divergence to the clusters B, C and D should be smaller than those to the other categories, because the clusters are labeled as "earn". The result is shown in Figure 1. The three boxes on the left correspond to the distances from cluster A to clusters B, C and D. The other boxes correspond to the distances from A to other categories. Divergence is larger for B and C than for other categories. The result is unexpected given the i.d. assumption.

This example suggests that the i.d. assumption does not always hold true. This fact leads to the inaccurate estimation of statistics. For example, the probability of the occurrence of a word given a category is estimated in the Naive Bayes classification. However, for some texts in the category, this word might tend to appear frequently, while for others not. In spite of that, one certain value represents the probability, in the previous approaches. So it may hold that categorization accuracy can be improved if we cluster texts in order to make the i.d. assumption more likely to be true.

3 Two-dimensional Clustering of Texts and Words

In this section, we propose a framework to overcome the problem caused by the violation of the i.d. assumption.

Our approach uses a bottom-up clustering technique. At the initial stage, each cluster has only one word or text. At each step, the most similar pair of words or texts are merged into one cluster. As a measure of similarity (or dissimilarity), we use the likelihood reduction caused by merging. This measure is related to Jensen-Shannon divergence. In the following, we explain our probability model, the clustering algorithm, the relation to Jensen-Shannon divergence and the stopping criteria for clustering.

3.1 Probability Model

The probability model that we adopt is the hard clustering model proposed in (Li and Abe, 1998), in which they dealt with the co-occurrence of two words. In the present case, however, we want the co-occurrence probability of a word and a text, which is expressed as :

$$P(w,d) = P(C_w,C_d)P(w|C_w)P(d|C_d) \quad w \in C_w, d \in C_d \tag{1}$$

where w denotes a word, d a text, and C_w and C_d are the clusters that w and d belong to respectively.

[1]The number of clusters is determined by the AIC (Akaike Information Criterion), which will be explained later.

Given a set of co-occurrence samples of words and texts:

$$S = \{(w_1, d_1), (w_2, d_2), \cdots, (w_m, d_m)\}, \quad (2)$$

its log-likelihood is calculated as :

$$\begin{aligned} & \textstyle\sum_{(w,d)\in S} \log P(w,d) \\ = & \textstyle\sum_{(w,d)\in S} \log P(C_w, C_d) P(w|C_w) P(d|C_d). \quad (3) \end{aligned}$$

The parameters in model (1) are estimated with the Maximum Likelihood Estimation :

$$P(C_w, C_d) = \frac{N(C_w, C_d)}{|S|}, \quad (4)$$

$$P(w|C_w) = \frac{N(w)}{N(C_w)}, \quad (5)$$

$$P(d|C_d) = \frac{N(d)}{N(C_d)}, \quad (6)$$

where $N(x)$ denotes the frequency of x.

3.2 Clustering Algorithms

In the algorithm described in (Li and Abe, 1998), given two positive integers k and l, merging for the first dimension is performed k times, followed by l merges for the second dimension.

We propose two different clustering algorithms. In both algorithms, a pair of words or texts are chosen and merged at each step, based on the model described in Section 3.1. The difference is the way to choose the pair of words or texts to be merged. One is what we call *text-first* clustering, in which text clustering is conducted first, followed by word clustering. The other is *greedy* clustering, in which, at each step, the pair with the least likelihood decrease is searched from the word pairs and the text pairs, and merged :

- Text-first Clustering
  1. Initialize
  2. Merge two *text clusters* with the least likelihood decrease repeatedly, while the stopping criterion is not satisfied.
  3. Merge two *word clusters* with the least likelihood decrease repeatedly, while the stopping criterion is not satisfied.
- Greedy Clustering
  1. Initialize
  2. Merge two *text clusters* or two *word clusters* with the least likelihood decrease repeatedly, while the stopping criterion is not satisfied.

We set the constraints that only texts in the same category can be merged (we call *the category-constraint*), and that only words with the same part-of-speech can be merged (*pos-constraint*). The category-constraint is indispensable in our method, because of our categorization method which is explained later. Both of these constraints decrease the computational time needed for clustering.

The text-first clustering has the advantage that word clustering can be conducted using the information given by class-distribution[2]. Class-distributional clustering is a special case of text-first clustering. If the stopping criterion of the text clustering step is set as "no two clusters can be merged without violating the category-constraint", then text-first clustering is identical to the class-distributional clustering.

3.3 The Relation to Jensen-Shannon Divergence

Here we show that using the criterion of the least likelihood decrease is equivalent to selecting the closest pair of clusters in terms of a certain distance as a probability distribution. Let ΔL denote the decrease of the log-likelihood (3) caused by merging word-clusters i and j. Let $|S|$ denote the number of the whole training examples. Using $P(C_{ij}, C_d) = P(C_i, C_d) + P(C_j, C_d)$, ΔL divided by $|S|$ is transformed as :

$$\begin{aligned} & \tfrac{1}{|S|}\Delta L \\ = & \textstyle\sum_{C_d} -P(C_{ij}, C_d) \log \frac{P(C_{ij},C_d)}{P(C_{ij})P(C_d)} \\ & + \textstyle\sum_{C_d} P(C_i, C_d) \log \frac{P(C_i,C_d)}{P(C_i)P(C_d)} \\ & + \textstyle\sum_{C_d} P(C_j, C_d) \log \frac{P(C_j,C_d)}{P(C_j)P(C_d)} \\ = & \textstyle\sum_{C_d} P(C_i, C_d) \{\log \frac{P(C_i,C_d)}{P(C_i)P(C_d)} \\ & \qquad - \log \frac{P(C_{ij},C_d)}{P(C_{ij})P(C_d)}\} \\ & + \textstyle\sum_{C_d} P(C_j, C_d) \{\log \frac{P(C_j,C_d)}{P(C_j)P(C_d)} \end{aligned}$$

[2] More precisely, the information used in the text-first clustering is different from the information given by class-distribution, but as the clustering proceeds, these two types of information become more similar.

$$
\begin{aligned}
& \quad - \log \frac{P(C_{ij}, C_d)}{P(C_{ij})P(C_d)}\} \\
= & \quad P(C_i) \sum_{C_d} P(C_d|C_i) \log \frac{P(C_d|C_i)}{P(C_d|C_{ij})} \\
& + P(C_j) \sum_{C_d} P(C_d|C_j) \log \frac{P(C_d|C_j)}{P(C_d|C_{ij})} \\
= & \quad P(C_i) D_{KL}(P(\cdot|C_i)||P(\cdot|C_{ij})) \\
& + P(C_j) D_{KL}(P(\cdot|C_j)||P(\cdot|C_{ij})), \qquad (7)
\end{aligned}
$$

where $D_{KL}(p||q)$ is the KL-divergence between the probability distribution p and q. The last line of (7) is the Jensen-Shannon divergence, which is also known as "KL divergence to the mean". That is, in our method, the closest pair of clusters in terms of the Jensen-Shannon divergence is merged at each step. In other words, the clustering method used in (Baker et al, 1998) is valid in terms of the likelihood.

3.4 AIC-based Stopping Criterion

As the stopping criterion in the clustering algorithm, we adopt AIC (Akaike Information Criterion) (Akaike, 1974). In (Li and Abe, 1998), they use MDL (Minimum Description Length) Principle (Rissanen, 1987). We do not use MDL Principle, because it tends to predict too small numbers of clusters in preliminary experiments (for text clustering, it predicted a smaller number of clusters than the number of categories, which is not suitable for our method because of the category-constraints).

AIC is realized as follows. The decrease of the number of parameters caused by merging a pair of clusters is :

$$
\Delta N_p = \begin{cases} |C(\mathbf{D})| - 1, & \text{(word-merge)} \\ |C(\mathbf{W})| - 1, & \text{(text-merge)} \end{cases} \qquad (8)
$$

where,

$$
\begin{aligned}
|C(\mathbf{D})| = & \quad \text{Number of clusters of words,} \\
|C(\mathbf{W})| = & \quad \text{Number of clusters of texts.}
\end{aligned}
$$

According to AIC, the stopping criterion should be

$$
-\Delta L + \Delta N_p > 0. \qquad (9)
$$

The first term ΔL denotes the decrease of log-likelihood (3) caused by merging.

Note that, in the text-first clustering, there are two possible points where AIC is applied. One is the point when the text clustering is finished, and the other is when the word clustering is finished.

4 Categorization

Probabilistic classifiers are expected to yield good results combined with our clustering method, but the performance of non-probabilistic classifiers with our method is unpredictable. We evaluate our clustering method using NB (Naive Bayes) classifiers (Mitchell, 1997), which is a probabilistic classifier, and SVMs (Support Vector Machines) (Vapnik, 1995), which is a non-probabilistic classifier.

In our method, the texts are clustered beforehand. So we first categorize the test texts and predict which cluster each test text belongs to. Then, we assign to each text the category that the predicted cluster belongs to (in our clustering method, all the training texts in each cluster are supposed to have the same category tag).

For the NB classifier, we use the Multinomial Model (McCallum and Nigam, 1998), but ignore the concern of document length.

SVM is a binary classifier based on Structural Risk Minimization (Vapnik, 1995). It has a high generalization performance and has been successfully applied to Text Categorization, for example in (Joachims, 1998). In order to apply SVMs to multi-class classification, we use the one-versus-rest method. However, when constructing a hyperplane for one cluster, the training texts belonging to the other clusters in the same category are removed from the training set.

5 Experiments

5.1 Experimental Settings

The data corpus used in this research is Reuters-21578[3]. We removed the texts whose body was meaningless, after applying ModApte-split, which is a standard way to split the corpus into training texts and test texts. This procedure yielded 8815 training texts, 3023 test texts and 116 categories. We used as features only nouns, verbs, proper nouns, adjectives and adverbs that occur five times or more in the whole training data. Stemming was also done using TreeTagger (Schmid, 1994).

Some of the texts in Reuters-21578 have multiple category-tags. In the clustering phase, we introduced multiple copies of those texts and label each text with one of its tags, so that every

[3]Available at http://www.research.att.com/~lewis/

text has one tag (otherwise the texts with multiple tags can never be merged according to the category-constraint). After clustering, we treat those texts belonging to multiple clusters in the categorization phase.

We compared our method with the method based on the class-distribution clustering, which shows good performance (Baker et al, 1998) (actually, we also tried the feature selection methods based on Information Gain and Mutual Information (Mitchell, 1997; Church, 1990), but they only deteriorated accuracy).

For the categorization with SVM, the package, TinySVM[4], was used. The kernel function used is the linear kernel.

The performance of each method is evaluated in terms of accuracy. Accuracy is computed for various compression rates of words (in this paper, we define the compression rate of words as the number of word-clusters divided by the number of all words. The compression rate of texts is defined similarly).

5.2 Results

The accuracies without clustering are 0.863 and 0.890 for NB classifiers and SVMs, respectively. According to AIC, 141 was selected as the number of text clusters, which is slightly more than the number of original categories. The categories that have multiple clusters even after the clustering are “earn (7 clusters)”, “acq (6 clusters)”, “others (8 clusters)”, “crude (3 clusters)”, “money-fx (3 clusters)”, “grain (2 clusters)”, “interest (2 clusters)” and “trade (2 clusters)”.

The results of categorization experiments are shown in Figures 2 and 3. Figure 2 shows the performance of NB classification combined with the text-first clustering and the class-distributional clustering, with various compression rates of words. As for the text-first clustering, the accuracies after the text-clustering step are displayed, because here we want to clarify the influence of the clustering of texts.

“Text-first Clustering” corresponds to the proposed method. “Class-Dist Clustering” corresponds to the class-distributional clustering method (this abbreviation is used in other figures and tables as well).

[4]Available at http://cl.aist-nara.ac.jp/˜taku-ku/software/TinySVM/ .

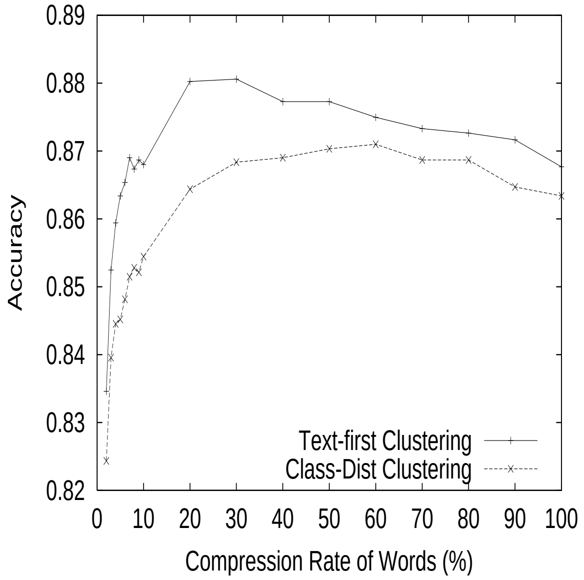

Figure 2: Categorization Accuracy (NB with text-first clustering and class-distributional clustering)

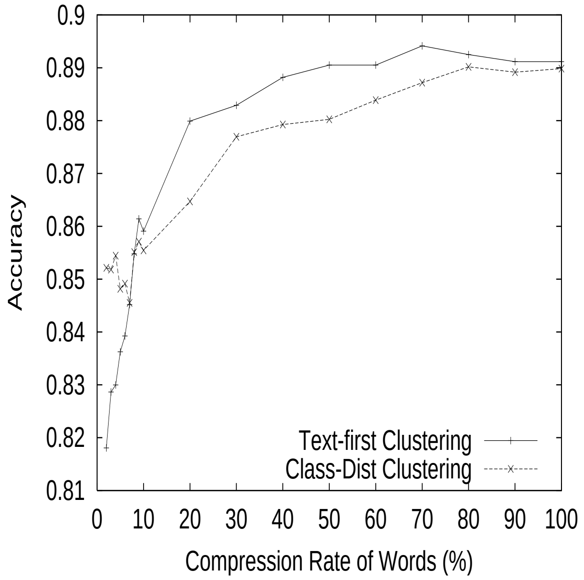

Figure 3: Categorization Accuracy (SVM with text-first clustering and class-distributional clustering)

Figure 3 shows the performance of the SVM combined with the text-first clustering and the class-distributional clustering, with various compression rates of words.

Table 1 shows the AIC-predicted numbers of clusters and the corresponding accuracies, to-

Clustering Method	Compress. Rate (%)	Accuracy
Class-Dist(AIC)	13.4	0.859
(Actual)	60	0.871
Text-first(AIC)	16.6	0.880
(Actual)	30	0.881

Table 1: Prediction of Number of Word-clusters

gether with the actual best compression rates and their accuracies.

Table 2 shows the performance of NB classification combined with greedy clustering. In the case of greedy clustering, it is necessary to display both word compression rates and text compression rates, so we didn't include the results of the greedy clustering into Figure 2. In Table 2, the compression rates predicted by AIC, 9.7% for words and 6.8% for texts, are also displayed.

5.3 Discussion

At the point of 100% word compression rate (i.e. no comression) in Figure 2, text-first clustering performs better than class-distributional clustering, although the difference is small (at this point, texts have been clustered in the text-first clustering). As the word compression rate decreases, the difference of the accuracy increases. This means that the combination of text clustering and word clustering works well.

Figure 3 shows that, also for SVMs, text-first clustering outperforms class-distributional clustering. However, the performance is worse for smaller compression rates. This means, in terms of accuracy, word-clustering is not effective for SVMs. The clustering of texts is still effective.

Predicted compression rates in Table 1 are not close to the actual best compression rates, although the corresponding accuracy is not so different for the text-first clustering. The difference of two AIC-predicted accuracies is significant in the sign-test (with 1% significance-level). The difference of the AIC-predicted accuracy of our method and the accuracy without clustering is also siginificant in the same test with 5% siginificance-level.

Table 2 shows that the greedy-clustering does not work well. The reason would be that word-clustering in the early stage cannot use the information of class-distribution.

6 Conclusion

We proposed a new method to improve the accuracy of Text Categorization using the two-dimensional clustering. In our method, both training texts and features are clustered before estimating the probability model.

Our approach is motivated by the fact that, in most previous probabilistic approaches, one category is assumed to have one identical probabilistic distribution, but this assumption is not always true, as discussed in this paper. Our two-dimensional clustering approach alleviates this problem, and at the same time, it can avoid the data-sparseness problem.

Through experiments, we showed that two-dimensional clustering worked well with Naive Bayes Classifiers and that, for the SVMs, two-dimensional clustering outperformed class-distributional clustering.

Future work includes the following.

First, in this research, we conducted experiments with only one data set. It would be desirable to confirm our conclusions with further experiments using different data sets. We used AIC as a stopping criterion of the text clustering step in the text-first clustering. But we haven't investigated whether AIC was valid as the turning criterion, because it needs experiments over two-dimensional parameter space. This point has to be investigated. As a stopping criterion, AIC does not always work well enough. Better criteria should be pursued. In our framework, AIC is actually targetting the joint probability of words and texts. But, in order to obtain a better stopping criterion, AIC should be incorporated in a more sophisticated way, such that it aims at the categorization.

We used an agglomerative clustering, but a divisive clustering method might be better in terms of computational time.

One of the possible extensions of this model is the soft version, as discussed in (Hofmann, 1998), in which the Expectation-Maximization algorithm is used with the soft version of this model.

References

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control*, vol. AC-19, pp. 716–723.

Table 2: Categorization Accuracy (NB with Greedy Clustering)

Word Compres.(%)	100.0	90.0	80.0	70.0	60.0	50.0	40.0	30.0	20.0
Text Compres.(%)	100.0	94.8	94.3	93.8	93.1	91.8	43.1	29.9	17.0
Acc.	0.717	0.718	0.719	0.722	0.731	0.739	0.807	0.834	0.836

10.0	9.7(AIC)	9.0	8.0	7.0	6.0	5.0	4.0	3.0	2.0
7.1	6.8	6.2	5.5	4.9	4.1	3.5	2.8	2.3	1.8
0.848	0.848	0.847	0.848	0.849	0.846	0.846	0.843	0.837	0.841

Baker, D. and McCallum, A. 1998. Distributional Clustering of Words for Text Classification. *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pp. 96–103.

Brown, P., Pietra, V.J., deSouza, P.V., Lai, J.C. and Mercer, R.L. 1992. Class-based N-gram Models of Natural Language. *Computational Linguistics*, 18(4), pp. 467–479.

Church, K. and Hanks, P. 1990. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics* 16(1), pp. 22–29.

Dhillon, I. 2001. Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning. Technical Report 2001-05, UT Austin CS Dept.

Hofmann, T. and Puzicha, J. 1998. Statistical Models for Cooccurrence Data. AI–MEMO 1625, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning.*

Li, H. and Abe, N. 1998. Word Clustering and Disambiguation Based on Co-occurrence Data. *Proceedings of COLING-ACL 98*, pp. 749–755.

McCallum, A. and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48.

Mitchell, T. 1997. *Machine Learning*, McGraw Hill.

Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3). pp. 103–134.

Rissanen, J. 1987. Stochastic Complexity. *Journal of Royal Statistical Society*, Series B, 49(3), pp. 223–239.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pp. 44–49, Manchester.

Slonim, N. and Tishby, N. 2000. Document Clustering using Word Clusters via the Information Bottleneck Method. *Research and Development in Information Retrieval*, pp. 208–215.

Slonim, N. and Tishby, N. 2001. The Power of Word Clusters for Text Classification. *23rd European Colloquium on Information Retrieval Research.*

Tishby, N., Pereira, F. and Bialek, W. 1999. The Information Bottleneck Method. *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* Springer.