A Method of Cluster-Based Indexing of Textual Data

Akiko Aizawa
National Institute of Informatics
akiko@nii.ac.jp

Abstract

This paper presents a framework for clustering in text-based information retrieval systems. The prominent feature of the proposed method is that documents, terms, and other related elements of textual information are clustered simultaneously into small overlapping clusters. In the paper, the mathematical formulation and implementation of the clustering method are briefly introduced, together with some experimental results.

1 Introduction

This paper is an attempt to provide a view of *indexing* as a process of generating many small *clusters* overlapping with each other. Individual clusters, referred to as micro-clusters in this paper, contain multiple subsets of associated elements, such as documents, terms, authors, keywords, and other related attribute sets. For example, a cluster in Figure 1 represents 'a set of documents written by a specific community of authors related to a subject represented by a set of terms'.

Our motivations for considering such clusters are that (i) the universal properties of text-based information spaces, namely large scale, sparseness, and local redundancy (Joachims, 2001), may be better manipulated by focusing on only limited sub-regions of the space; and also that (ii) the multiple viewpoints of information contents, which a conventional retrieval system provides, can be better utilized by considering not only the relations between 'documents' and 'terms' but also associations between other attributes such as 'authors' within the same unified framework.

Based on the background, this paper presents a framework of micro-clustering, within which we adopt a probabilistic formulation of co-occurrences of textual elements. For simplicity, we focus primarily on the co-occurrences between 'documents' and 'terms' in our explanation, but the presented framework is directly applicable to more general cases with more than two attributes.

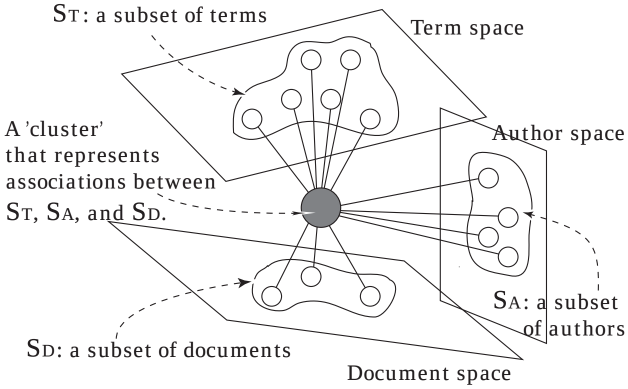

Figure 1: Cluster-based indexing of information spaces.

2 Background Issues

A view from indexing

In information retrieval research, matrix transformation-based indexing methods such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990) have recently become quite common. These methods can be viewed as an established basis for exposing hidden associations between documents and terms. However, their objective is to generate a compact representation of the original information space, and it is likely in consequence that the resulting orthogonal vectors are dense with many non-zero elements (Dhillon and Modha, 1999). In addition, because the reduction process is globally optimized, matrix transformation-based methods become computationally infeasible when dealing with high-dimensional data.

A view from clustering

The document-clustering problem has also been extensively studied in the past (Iwayama and Tokunaga, 1995; Steinbach et al., 2000). The majority of the previous approaches to clustering construct either a partition or a hierarchy of target documents, where the generated clusters are either exclusive or nested. However, generating mutually exclusive or tree-structured clusters in general is a hard-constrained problem and thus is likely to suffer high computational costs when dealing with large-scale data. Also, such a constraint is not necessarily required in actual applications, because 'topics' of documents, or rather 'indices' in our context, are arbitrarily overlapped in nature (Zamir and Etzioni, 1998).

Basic Strategy:

Based on the above observations, our basic strategy is as follows:

- Instead of generating component vectors with many non-zero elements, produce only limited subsets of elements, i.e., micro-clusters, with significance weights.
- Instead of transforming the entire co-occurrence matrix into a different feature space, extract tightly associated sub-structures of the elements on the graphical representation of the matrix.
- Use entropy-based criteria for cluster evaluation so that the sizes of the generated clusters can be determined independently of other existing clusters.
- Allow the generated clusters to overlap with each other. By assuming that each element can be categorized into multiple clusters, we can reduce the problem to a feasible level where the clusters are processed individually.

Related studies:

Another important aspect of the proposed micro-clustering scheme is that the method employs simultaneous clustering of its composing elements. This not only enables us to combine issues in *term* indexing and *document* clustering, as mentioned above, but also is useful for connecting matrix-based and graph-based notions of clustering; the latter is based on the association networks of the elements extracted from the original co-occurrence matrices. Some recent topics dealing with this sort of duality and/or graphical views include: the Information Bottleneck Method (Slonim and Tishby, 2000), Conceptual Indexing (Dhillon and Modha, 1999; Karypis and Han, 2000), and Bipartite Spectral Graph Partitioning (Dhillon, 2001), although each of these follows its own mathematical formulation.

3 The Clustering Method

3.1 Definition of Micro-Clusters

Let $D = \{d_1, \cdots, d_N\}$ be a collection of N target documents, and let S_D be a subset of documents such that $S_D \subseteq D$. Likewise, let $T = \{t_1, \cdots, t_M\}$ be a set of M distinct terms that appear in the target document collection, and let S_T be a subset of terms such that $S_T \subseteq T$. A *cluster*, denoted as c, is defined as a combination of S_T and S_D:

$$c = (S_T, S_D). \tag{1}$$

The co-occurrences of terms and documents can be expressed as a matrix of size $M \times N$ in which the (i, j)-th cell indicates that t_i $(\in T)$ appears in d_j $(\in D)$. We make the value of the (i, j)-th cell equal to $freq(t_i, d_j)$. Although we primarily assume the value is either '1' (exist) or '0' (not exist) in this paper, our formulation could easily be extended to the cases where $freq(t_i, d_j)$ represents the actual number of times that t_i appears in d_j.

The observed total frequency of t_i over all the documents in D is denoted as $freq(t_i, D)$. Similarly, the observed total frequency of d_j, i.e. the total number of terms contained in d_j, is denoted as $freq(T, d_j)$. These values correspond to summations of the columns and the rows of the co-occurrence matrix. The total frequency of all the documents is denoted as $freq(T, D)$. Thus,

$$\begin{aligned} freq(T, D) &= \sum_{t_i \in T} freq(t_i, D) = \sum_{d_j \in D} freq(T, d_j) \\ &= \sum_{t_i \in T} \sum_{d_j \in D} freq(t_i, d_j). \end{aligned} \tag{2}$$

We sometimes use $freq(t_i)$ for $freq(t_i, D)$, $freq(d_j)$ for $freq(T, d_j)$ and F for $freq(T, D)$.

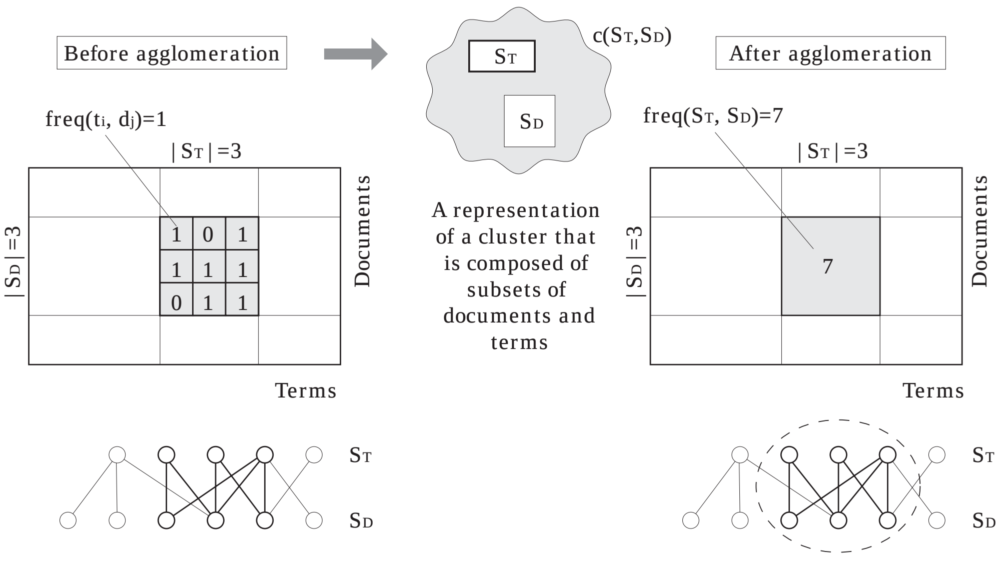

Figure 2: Example of a cluster defined on a co-occurrence matrix.

When a cluster c is being considered, T and D in the above definitions are changed to S_T and S_D. In this case, $freq(t_i, S_D)$ and $freq(S_T, d_j)$ represent the frequencies of t_i and d_j within $c = (S_T, S_D)$, respectively. In the co-occurrence matrix, a cluster is expressed as a 'rectangular' region if terms and documents are so permuted (Figure 2).

3.2 Probabilistic Formulation

The view of the co-occurrence matrix can be further extended by assigning probabilities to each cell. With the probabilistic formulation, t_i and d_j are considered as independently observed events, and their combination as a single co-occurrence event (t_i, d_j). Then, a cluster $c = (S_T, S_D)$ is also considered as a single co-occurrence event of observing one of $t_i \in S_T$ within one of $d_j \in S_D$.

In estimating the probability of each event, we use a simple discounting method similar to the *absolute discounting* in probabilistic language modeling studies (Baayen, 2001). The method subtracts a constant value δ, called a discounting coefficient, from all the observed term frequencies and estimates the probability of t_i as:

$$P(t_i) = \frac{freq(t_i) - \delta}{F}. \quad (3)$$

Note that the discounting effect is stronger for low-frequency terms. For high-frequency terms, $P(t_i) \approx freq(t_i)/F$. In the original definition, the value of δ was uniquely determined, for example as $\delta = \frac{m(1)}{M}$ with $m(1)$ being the number of terms that appear exactly once in the text. However, we experimentally vary the value of δ in our study, because it is an essential factor for controlling the size and quality of the generated clusters.

Assuming that the probabilities assigned to documents are not affected by the discounting, $P(d_j|t_i) = freq(t_i, d_j) \ / \ freq(t_i)$. Then, applying $P(t_i, d_j) = P(d_j|t_i)P(t_i)$, the co-occurrence probability of t_i and d_j is given as:

$$P(t_i, d_j) = \frac{freq(t_i) - \delta}{freq(t_i)} \cdot \frac{freq(t_i, d_j)}{F}. \quad (4)$$

Similarly, the co-occurence probability of S_T and S_D is given as:

$$P(S_T, S_D) = \frac{freq(S_T) - \delta}{freq(S_T)} \cdot \frac{freq(S_T, S_D)}{F}. \quad (5)$$

3.3 Criteria for Cluster Evaluation

The evaluation is based on the information theoretic view of the retrieval systems (Aizawa, 2000). Let $\mathcal{T}$ and $\mathcal{D}$ be two random variables corresponding to the events of observing a term and a document, respectively. Denote their occurrence probabilities as $P(\mathcal{T})$ and $P(\mathcal{D})$, and their co-occurrence probability as a joint distribution $P(\mathcal{T}, \mathcal{D})$. By the general definition of traditional information theory, the mutual information between $\mathcal{T}$ and $\mathcal{D}$, denoted as $\mathcal{I}(\mathcal{T}, \mathcal{D})$, is calculated as:

$$\mathcal{I}(\mathcal{T}, \mathcal{D}) = \sum_{t_i \in T} \sum_{d_j \in D} P(t_i, d_j) log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}, \quad (6)$$

where the values of $P(t_i, d_j)$ and $P(t_i)$ are calculated using Eqs. (3) and (4). $P(d_j)$ is determined by $P(d_j) = \sum_{t_i \in T} P(t_i, d_j)$, or approximated simply by $P(d_j) = freq(d_j)/F$. Next, the mutual information after agglomerating S_T and S_D into a single cluster (Figure 2) is calculated as:

$$\begin{aligned} \mathcal{I}'(\mathcal{T}, \mathcal{D}) \quad = \quad & \sum_{t_i \notin S_T} \sum_{d_j \notin S_D} P(t_i, d_j) log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \\ & + P(S_T, S_D) log \frac{P(S_T, S_D)}{P(S_T)P(S_D)}, \quad (7) \end{aligned}$$

where $P(S_T) = \sum_{t_i \in S_T} P(t_i)$ and $P(S_D) = \sum_{d_j \in S_D} P(d_j)$.

The fitness of a cluster, denoted as $\delta\mathcal{I}(S_T, S_D)$, is defined as the difference of the two information values given by Eqs.(6) and (7):

$$\begin{aligned} \delta\mathcal{I}(S_T, S_D) &= \mathcal{I}'(\mathcal{T}, \mathcal{D}) - \mathcal{I}(\mathcal{T}, \mathcal{D}) \\ &= P(S_T, S_D) log \frac{P(S_T, S_D)}{P(S_T)P(S_D)} \\ &\quad - \sum_{t_i \in S_T} \sum_{d_j \in S_D} P(t_i, d_j) log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (8) \end{aligned}$$

Without discounting, the value of $\delta\mathcal{I}(S_T, S_D)$ in the above equation is always negative or zero. However, with discounting, the value becomes positive for uniformly dense clusters, because the frequencies of individual cells are always smaller than their agglomeration and so the discounting effect is stronger for the former.

Using the same formula, we calculated the significance weights t_i in $c = (S_T, S_D)$ as:

$$\delta\mathcal{I}(t_i, S_D) = \sum_{d_j \in S_D} P(t_i, d_j) log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}, \quad (9)$$

and the significance weights of d_j as:

$$\delta\mathcal{I}(S_T, d_j) = \sum_{t_i \in S_T} P(t_i, d_j) log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (10)$$

In other words, all the terms and documents in a cluster can be jointly ordered according to their contribution in the entropy calculation given by Eq. (7).

To summarize, the proposed probabilistic formulation has the following two major features. First, *clustering is generally defined as an operation of agglomerating a group of cells in the contingency table.* Such an interpretation is unique because existing probabilistic approaches, including those with a duality view, agglomerate entire rows or columns of the contingency table all at once. Second, *the estimation of the occurrence probability is not simply in proportion to the observed frequency.* The discounting scheme enables us to trade off (i) the loss of averaging probabilities in the agglomerated clusters, and (ii) the improvement of probability estimations by using larger samples sizes after agglomeration.

It should be noted that although we have restricted our focus to one-to-one correspondences between terms and documents, the proposed framework can be directly applicable to more general cases with $k(\geq 2)$ attributes. Namely, given k random variables $X_1, \cdots, X_k$, Eq. (8) can be extended as:

$$\begin{aligned} & \delta\mathcal{I}(S_{X_1}, \cdots, S_{X_k}) \\ & \quad = P(S_{X_1}, \cdots, S_{X_k}) log \frac{P(S_{X_1}, \cdots, S_{X_k})}{P(S_{X_1}) \cdots P(S_{X_k})} \quad (11) \\ & \quad - \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_k \in S_{X_k}} P(x_1, \cdots, x_k) log \frac{P(x_1, \cdots, x_k)}{P(x_1) \cdots P(x_k)}. \end{aligned}$$

3.4 Cluster Generation Procedure

The cluster generation process is defined as the repeated iterations of cluster initiation and cluster improvement steps (Aizawa, 2002).

First, in the cluster initiation step, a single term t_i is selected, and an initial cluster is then formulated by collecting documents that contain t_i and terms that co-occur with t_i within the same document. The collected subsets, respectively, become S_D and S_T of the initiated cluster. On the bipartite graph of terms and documents (Figure 2), the process can be viewed as a two-step expansion starting from t_i.

Next, in the cluster improvement step, all the terms and documents in the initial cluster are tested for elimination in the order of increasing significance weights given by Eqs. (9) and

(10). If the performance of the target cluster is improved after the elimination, then the corresponding term or document is removed. When finished with all the terms and documents in the cluster, the newly generated cluster is tested to see whether the evaluation value given by Eq. (8) is positive. Clusters that do not satisfy this condition are discarded. Note that the resulting cluster is only locally optimized, as the improvement depends on the order of examining terms and documents for elimination.

At the initiation step, instead of randomly selecting an initiating term, our current implementation enumerates all the existing terms $t_i \in T$. We also limit the sizes of S_T and S_D to $k_{max} = 50$ to avoid explosive computation caused by high frequency terms. Except for k_{max}, the discounting coefficient δ is the only parameter that controls the sizes of the generated clusters. The effect of δ is examined in detail in the following experiments.

4 Experimental Results

4.1 The Data Set

In our experiments, we used NTCIR-J1[1], a Japanese text collection for retrieval tasks that is composed of abstracts of conference papers organized by Japanese academic societies. In preparing the data for the experiments, we first selected 52,867 papers from five different societies: 23,105 from the Society of Polymer Science, Japan (SPSJ), 20,482 from the Japan Society of Civil Engineers (JSCE), 4,832 from the Japan Society for Precision Engineering (JSPE), 2,434 from the Ecological Society of Japan (ESJ), and 2,014 from the Japanese Society for Artificial Intelligence (JSAI).

The papers were then analyzed by the morphological analyzer ChaSen Ver.2.02 (Matsumoto et al., 1999) to extract nouns and compound nouns using the Part-Of-Speech tags. Next, the co-occurrence frequencies between documents and terms were collected. After preprocessing, the number of distinctive terms was 772,852 for the 52,867 documents.

4.2 Clustering Results

In our first experiments, we used a framework of unsupervised text categorization, where the quality of the generated clusters was evaluated by the goodness of the separation between different societies. To investigate the effect of the discounting parameter, it was given the values $\delta = 0.1, 0.3, 0.5, 0.7, 0.9, 0.95$.

Table 1 compares the total number of generated clusters (c), the average number of documents per cluster (s_d), and the average number of terms per cluster (s_t), for different values of δ. We also examined the ratio of unique clusters that consist only of documents from a single society (r_s), and an inside-cluster ratio that is defined as the average relative weight of the dominant society for each cluster (r_i). Here, the weight of each society within a cluster was calculated as the sum of the significance weights of its component documents given by Eq. (10).

The results shown in Table 1 indicate that reducing the value of δ improves the quality of the generated clusters: with smaller δ, the single society ratio and the inside-cluster ratio becomes higher, while the number of generated clusters becomes smaller.

Table 1: Summary of clustering results.

δ	c	s_d	s_t	r_s	r_i
0.10	136,832	3.25	9.3	0.953	0.983
0.30	187,079	3.94	29.4	0.896	0.960
0.50	196,208	4.81	39.7	0.866	0.951
0.70	196,911	5.39	44.4	0.851	0.948
0.90	197,164	5.81	46.3	0.841	0.945
0.95	197,193	5.89	46.6	0.839	0.944

4.3 Categorization Results

In our second experiment, we used a framework of supervised text categorization, where the generated clusters were used as indices for classifying documents between the existing societies, and the categorization performance was examined.

For this purpose, the documents were first divided into a training set of 50,182 documents and a test set of 2,641 documents. Then, assuming that the originating societies of the training documents are known, the significance weights of the five societies were calculated for each cluster generated in the previous experiments. Next, the test documents were assigned to one of the five societies based on the membership of the multiple clusters to which they belong.

[1] http://research.nii.ac.jp/ntcir/

For comparison, two supervised text categorization methods, naive Bayes and Support Vector Machine (SVM), were also applied to the same training and test sets.

The results are shown in Table 2. In this case, the performance was better for larger δ, indicating that the major factor determining the categorization performance was the number of clusters rather than their quality. For $\delta = 0.5 \sim 0.95$, each tested document appeared in at least one of the generated clusters, and the performance was almost comparable to the performance of standard text categorization methods: slightly better than naive Bayes, but not so good as SVM. We also compared the performance for varied sizes of training sets and also using different combination of societies, but the tendency remained the same.

Table 2: Summary of categorization results.

δ	correct	judge	F-value
0.10	2,370	2,446	0.932
0.30	2,520	2,623	0.957
0.50	2,575	2,641	0.975
0.70	2,583	2,641	0.978
0.90	2,584	2,641	0.978
0.95	2,583	2,641	0.978
naive Bayes	2,579	2,641	0.977
SVM	2,602	2,641	0.985

4.4 Further Analysis

Analysis of categorization errors

Table 3 compares the patterns of misclassification, where the columns and rows represent the classified and the real categories, respectively. It can be seen that as far as minor categories such as ESJ and JSAI are concerned, the proposed micro-clustering method performed slightly better than SVM. The reason may be that the former method is based on locally conformed clusters and less affected by the skew of the distribution of category sizes. However, the details are left for further investigation.

In addition, by manually analyzing the individual misclassified documents, it can be confirmed that most of them dealt with inter-domain topics. For example, nine out of the ten JSCE documents misclassified as ESJ were related to environmental issues; six out of the 14 JSPE documents misclassified as JSCE, as well as all seven JSPE documents misclassified as JSAI, were related to the application of artificial intelligence techniques. These were the major causes of the performance difference of the two methods.

Table 3: Analysis of miss-classification.

(a) *Micro-clustering results*

	judge				
	SPSJ	JSCE	JSPE	ESJ	JSAI
r SPSJ	1146	7	2	0	0
e JSCE	5	1007	1	10	1
a JSPE	3	14	216	1	7
l ESJ	0	1	0	120	0
JSAI	0	3	1	1	95

(b) *Text categorization results*

	judge				
	SPSJ	JSCE	JSPE	ESJ	JSAI
r SPSJ	1150	2	3	0	0
e JSCE	2	1017	1	2	2
a JSPE	5	9	226	1	0
l ESJ	0	2	0	119	0
JSAI	1	3	6	0	90

Effect of local improvement:

We also tested the categorization performance without local improvement where the top 50 terms at most survive unconditionally after forming the initial clusters. In this case, the clustering works similarly to the automatic relevance feedback in information retrieval. Using the same data set, the result was 2,564 correct judgments (F-value 0.971), which shows the effectiveness of local improvement in reducing noise in automatic relevance feedback.

Effect of cluster duplication check:

Because we do not apply any duplication check in our generation step, the same cluster may appear repeatedly in the resulting cluster set. We have also tested the other case where clusters with terms or document sets identical to existing better-performing clusters were eliminated. The obtained categorization performance was slightly worse than the one without elimination. For example, the best performance obtained for $\delta = 0.9$ was 2,582 correct judgments (F-value 0.978) with 137,867 (30% reduced) clusters.

The results indicate that the system does not necessarily require expensive redundancy checks for the generated clusters as a whole. Such consideration becomes necessary when the formulated clusters are presented to users, in which case, the duplication check can be applied only locally.

5 Discussion

In this paper, we reported a method of generating overlapping micro-clusters in which documents, terms, and other related elements of text-based information are grouped together.

Comparing the proposed micro-clustering method with existing text categorization methods, the distinctive feature of the former is that the documents on borders are readily viewed and examined. In addition, the terms in the cluster can be further utilized in digesting the descriptions of the clustered documents. Such properties of micro-clustering may be particularly important when the system actually interacts with its users.

For comparison purposes, we have used only the conventional documents-and- terms feature space in our experiments. However, the proposed micro-clustering framework can be applied more flexibly to other cases as well. For example, we have also generated clusters using the co-occurrences of the triple of documents, terms, and authors. Although the performance was not much different in terms of text categorization (2,584 correct judgments out of 2,639 judgments, the precision slightly improved), we can confirm that many of the highly ranked clusters contain documents produced by the same group of authors, emphasizing the characteristics of such generated clusters.

Future issues include: (i) enhancing the probabilistic models considering other discounting techniques in linguistic studies; (ii) developing a strategy for initiating clusters by combining different attribute sets, such as documents or authors; and also (iii) establishing a method of evaluating overlapping clusters. We are also looking into the possibility of applying the proposed framework to Web document clustering problems.

References

A. Aizawa. 2000. The feature quantity: An information theoretic perspective of tfidf-like measures. In *Proc. of ACM SIGIR 2000*, pages 104–111.

A. Aizawa. 2002. An approach to microscopic clustering of terms and documents. In *Proc. of the 7th Pacific Rim Conference on Artificial Intelligence (to appear).*

R. H. Baayen. 2001. *Word frequency distributions.* Kluwer Academic Publishers.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of American Society of Information Science*, 41:391–407.

I. S. Dhillon and D. S. Modha. 1999. Concept decomposition for large sparse text data using clustering. Technical Report Research Report RJ 10147, IBM Almaden Research Center.

I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. Technical Report 2001-05, UT Austin CS Dept.

M. Iwayama and T. Tokunaga. 1995. Cluster-based text categorization: a comparison of category search strategies. In *Proc. of ACM SIGIR'95*, pages 273–281.

T. Joachims. 2001. A statistical learning model of text classification for support vector machines. In *Proc. of ACM SIGIR 2001*, pages 128–136.

G. Karypis and E.-H. Han. 2000. Fast supervised dimensionality reduction algorithm with applications to document categorization and retrieval. In *Proc. of the 9th ACM International Conference on Information and Knowledge Management*, pages 12–19.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, K. Matsuda, and M. Asahara. 1999. Morphological analysis system chasen 2.0.2 users manual. NAIST Technical Report NAIST-IS-TR99012, Nara Institute of Science and Technology.

N. Slonim and N. Tishby. 2000. Document clustering using word clusters via the information bottleneck method. In *Proc. of ACM SIGIR 2000*, pages 2008–2015.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining.*

O. Zamir and O. Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Proc. of ACM SIGIR'98*, pages 46–54.