

On Combining Multiple Clusterings

Tao Li
School of Computer Science
Florida International University
Miami, FL 33199
taoli@cs.fiu.edu

Mitsunori Ogihara
Computer Science Dept.
University of Rochester
Rochester, NY 14627-0226
ogihara@cs.rochester.edu

Sheng Ma
IBM T.J. Watson Research
Center
Hawthorne, NY 10532
shengma@us.ibm.com

ABSTRACT

Many problems can be reduced to the problem of combining multiple clusterings. In this paper, we first summarize different application scenarios of combining multiple clusterings and provide a new perspective of viewing the problem as a categorical clustering problem. We then show the connections between various consensus and clustering criteria and discuss the complexity results of the problem. Finally we propose a new method to determine the final clustering. Experiments on kinship terms and clustering popular music from heterogeneous feature sets show the effectiveness of combining multiple clusterings.

Categories and Subject Descriptors

I.2 [Artificial Intelligence]: Learning; I.5.3 [Pattern Recognition]: Clustering

General Terms

Algorithms, Experimentation, Theory

Keywords

multiple clusterings, combining, categorical

1. INTRODUCTION

1.1 Problem Overview

Generally the problem of combining multiple clusterings is: given multiple clusterings of the dataset, find a combined clustering which would provide better cluster results. Many problems can be reduced to the problem of combining multiple clusterings:

- Ensemble clustering: Clustering is an inherently ill-posed problem due to the lack of label information. Different clustering algorithms and even multiple replications of the same algorithm result in different solutions due to random initializations and stochastic learning methods. In supervised learning area, ensemble methods, by combining multiple classifiers, have shown to be a popular way to overcome

instability in classification problems [3]. The success of ensemble methods in classification provides the main motivation for applying ensemble methods in clustering. The problem of ensemble clustering is to find a combined clustering result based on multiple clusterings of the dataset. There are many ways to obtain multiple clusterings such as applying different clustering algorithms; using resampling to get subsamples of the dataset [30], utilizing feature selection methods such as random projection to get different feature spaces [13], and exploiting the randomness of the clustering algorithm.

- Clustering with Multiple Criteria: In many applications, especially in the social sciences, clustering problems often require optimization over more than one criterion [11] or clustering needs to meet additional constraints which are not included in the clustering criterion [12]. A typical example is *second world war politicians* problem [10], where the data were obtained by asking many subjects to rate the dissimilarities of second world war politicians. Each subject corresponds to a optimization criterion and hence clustering the politicians needs to optimize multiple criteria. For clustering with multiple criteria, solutions optimal according to each particular criterion are not identical. The core problem is then how to find the best solution so as to satisfy as much as possible all the criteria considered. A typical approach is to combine multiple clusterings obtained via single criterion clustering algorithms based on each criterion [8].
- Distributed clustering: Over the years, data set sizes have grown rapidly with the advances in technology and the increasingly automated business processes. Many data sets are, in nature, geographically distributed across multiple sites. In a distributed environment, data sites may be *homogeneous*, i.e., different sites containing data for exactly the same set of features, or *heterogeneous*, i.e., different sites storing data for different set of features, possibly with some common features among sites. To cluster the distributed datasets, one way is to first cluster them locally and then combine the clustering obtained at each site [26, 20].
- Three-way clustering: When a multivariate phenomenon is observed on different occasions, the datasets collected are identified according to three modes: units (rows), variables (columns) and occasions (layers) and then arranged into a three-way data matrix $X = (x_{ijh})$ [38]. Typical examples include macroeconomics performance data of different regions over different time period or medical examinations of different plants over different locations. One way to perform clustering on three-way datasets is to cluster the units based

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'04, November 8–13, 2004, Washington, DC, USA.
Copyright 2004 ACM 1-58113-874-1/04/0011 ...\$5.00.

on the variables at each occasion and then combine the clustering results [1].

In addition, combining multiple clusterings also provides a framework for knowledge reuse and can be used to exploit existing knowledge implicit in legacy clusterings [35].

1.2 Content of The Paper

The contribution of this paper is four-fold: we first provide a new perspective of viewing the problem of combining multiple clustering as a categorical clustering problem, we then show the connections between various consensus and clustering criteria. Third, we discuss the complexity results of the problem, and finally we propose a new method to determine the final clustering. Our experimental results on kinship terms and clustering popular music from heterogeneous feature sets show the effectiveness of combining multiple clusterings.

The rest of the paper is organized as follows: Section 2 formally defines the problem of combining multiple clusterings and presents two different perspectives: consensus partition and categorical clustering. Section 3 introduces various criteria for combining multiple clusterings. Section 4 illustrates the connections among them. In particular, the equivalence relation between consensus partition and categorical clustering is established. Section 5 discusses the complexity results of the problem and Section 6 presents a method for determining final clusterings. Section 7 presents our experiments on kinship terms and clustering popular music, and finally section 8 concludes.

2. PROBLEM DEFINITION

Formally the problem of combining multiple clusterings can be described as follows: let $D = \{d_1, d_2, \dots, d_n\}$ be a set of n data points. Suppose we are given a set of T partitions $\mathcal{P} = \{P_1, P_2, \dots, P_T\}$ of the data points in D . Each partition $P_i, i = 1, \dots, T$ consists of a set of clusters $P_i = \{C_i^1, C_i^2, \dots, C_i^{k_i}\}$ where k_i is the number of clusters for partition P_i and $D = \bigcup_{j=1}^{k_i} C_j^i$. Our goal is to find a final clustering $P = \{C_1, \dots, C_K\}$ of D such that the points inside each C_i are ‘‘similar’’ to each other.

Basically the problem of combining multiple clusterings can be solved from two perspectives: one is using consensus classification methods to find a target partition for optimizing consensus functions [8, 18], and the other one, as we will show later, is performing categorical (or binary) clustering in the space induced by the multiple partitions.

A consensus function maps a given set of partitions $\mathcal{P} = \{P_1, P_2, \dots, P_T\}$ to a final partition P . There are many ways to define the consensus function [8]. On the other hand, it is convenient to characterize the consensus problem as a binary clustering problem where the attributes are induced by the partitions. Let $p = \sum_{i=1}^T k_i$, then each point $d_i \in D$ can be represented as a p -dimensional vector

$$d_i = (d_{i11}, \dots, d_{i1k_1}, \dots, d_{ij1}, \dots, d_{ijk_j}, \dots, d_{iT1}, \dots, d_{iT k_T}) \quad (1)$$

$$d_{ijl} = \begin{cases} 1 & d_i \in C_j^l \\ 0 & \text{Otherwise} \end{cases} \quad 1 \leq j \leq T, 1 \leq l \leq k_j$$

After inducing new binary representation, finding the final clustering can be achieved with various categorical clustering algorithms. The notations used for problem definition are listed in the Table 1.

¹In this paper, we interchangeably use partition and clustering. By partition, we mean a set of mutually exclusive and collectively exhaustive classes such that each point is in one and only one cluster. Note that partition is an equivalence relation.

$D = \{d_1, d_2, \dots, d_n\}$	The set of n data points
$\mathcal{P} = \{P_1, P_2, \dots, P_T\}$	The set of partitions
$P_t = \{C_t^1, C_t^2, \dots, C_t^{k_t}\}$	The clusters in a partition
k_t	Number of clusters for partition P_t
$P = \{C_1, \dots, C_K\}$	Final partition
K	Number of clusters for final partition P
$d_i = (d_{ijl})$	Induced vector representation for d_i

Table 1: Notations For Problem Definition

3. VARIOUS CRITERIA

There are many different ways to define the consensus function such as co-associations between data points or based on pairwise agreements between partitions. Some of the criteria are based on the similarity between data points and some of them are based on the estimates of similarity between partitions. In what follows, we survey various criteria for combining multiple clusterings. The connections among various criteria are elaborated in Section 4.

Note that each partition P_t defines an associated $n \times n$ matrix that stores the information, for each pair of points, whether they are in the same cluster. The entries of the matrix are defined as follows:

$$M_t(i, j) = \begin{cases} 1 & i, j \text{ belongs to the same cluster} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

- Partition Difference:** Given two partitions A and B ², a common measure of their difference is³:

$$\Delta(A, B) = \sum_{i,j} (A(i, j) - B(i, j))^2$$

Intuitively, $\Delta(A, B)$ computes the number of pairs of points that belong to different clusters in A and B . A target partition P thus should minimize $\sum_{t=1}^T \Delta(P, P_t)$.

- Consensus Matrix:** Another measure based on the associated matrices is the normalized consensus matrix, defined by

$$S(i, j) = \frac{1}{T} \sum_{t=1}^T M_t(i, j)$$

Basically, S indicates, for each pair of points, the proportion of times in which they are clustered together. $S(i, j)$ is a measure of co-association indicating the ‘‘similarity’’ between i and j .

- Katz & Powell Index:** Consensus measures can also be defined based on the indexes of pairwise agreements among partitions. The first index for comparing two partitions was constructed by Katz and Powell [21] as follows: given two partition matrices A, B , the index of agreement between them is

$$\Gamma(A, B) = \frac{n^2 N_{AB} - N_A N_B}{\sqrt{N_A(n^2 - N_A)N_B(n^2 - N_B)}}$$

where N_A and N_B are respectively the number of 1’s in A and B , and N_{AB} is the number of entries in A and B both

²We also use A and B to denote the matrices associated with the partitions.

³Note that since $A(i, j)$ and $B(i, j)$ are either 0 or 1, thus $\Delta(A, B) = \sum_{i,j} (A(i, j) - B(i, j))^2 = \sum_{i,j} |A(i, j) - B(i, j)|$.

defined by a 1. Using the classical contingency table notation, we have

$$N_{AB} = \sum_{i=1}^n \sum_{j=1}^n A_{ij} B_{ij} = \sum_{u=1}^p \sum_{v=1}^q n_{uv}^2$$

$$N_A = \sum_{ij} A_{ij} = \sum_{u=1}^p n_{u.}^2, N_B = \sum_{ij} B_{ij} = \sum_{v=1}^q n_{.v}^2$$

where n_{uv} is the number of points in the cluster u of A and the cluster v of B ; $n_{u.}$ and $n_{.v}$ refer respectively to the number of points in clusters u and v . It has been shown that the above index is equivalent to the ordinary product moment correlation coefficient between the off-diagonal cells of A and B ordered identically [19]. The best partition should maximize the value of the overall agreement

$$P_{best} = \operatorname{argmax}_P \sum_{t=1}^T \Gamma(P, P_t). \quad (3)$$

4. **Chance-Corrected Measures:** A well-known approach to define chance-corrected measures of an association or agreement is

$$l_N = \frac{l - \tau}{l_{max} - \tau}$$

where l_N is the chance-corrected measure, τ represents a structure in which $l_N = 0$ and l_{max} is the maximum value of the coefficients regardless of the fixed margins [27]. Cohen's Kappa [6], defined by

$$K(A, B) = \frac{N_{AB} - N_A N_B}{n^2 - N_A N_B},$$

is an example of chance-corrected measures. A family of chance-corrected rand index measures were introduced in [18]. Similarly, the best partition should maximize the overall agreement defined by the chance-corrected measure.

5. **Chi-squared Measure:** Chi-squared measure for comparing two partitions A and B is defined as follows:

$$\chi^2(A, B) = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{uv} - E_{uv})^2}{E_{uv}} \quad (4)$$

where $E_{uv} = \frac{n_{u.} n_{.v}}{n}$. Assuming the sizes of clusters in each clustering are fixed, we have

$$\chi^2(A, B) = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{uv} - E_{uv})^2}{E_{uv}} = \sum_{i=1}^p \sum_{j=1}^q \frac{n_{uv}^2}{E_{uv}} - n$$

Similarly, the best partition should maximize the overall agreement defined by the chi-squared measure.

6. **Category Utility Functions:** Given the partition $P = \{C_1, \dots, C_K\}$, the category utility function $U(P, P_t)$ is defined as follows [15, 28]:

$$U(P, P_t) = \sum_{r=1}^K p(C_r) \sum_{j=1}^{k_t} P(C_t^j | C_r)^2 - \sum_{j=1}^{k_t} p(C_t^j)^2 \quad (5)$$

where $p(C_r) = \frac{|C_r|}{n}$, $p(C_t^j | C_r) = \frac{|C_t^j \cap C_r|}{|C_r|}$, $p(C_t^j) = \frac{|C_t^j|}{n}$. In other words, the function $U(P, P_t)$ assesses the agreement between two partitions as the difference between the expected number of classes of partition P_t that can be

correctly predicted both with the knowledge of clustering P or without it. The overall utility of the partition P with respect to $\mathcal{P} = \{P_1, P_2, \dots, P_T\}$ can then be defined as

$$U(P, \mathcal{P}) = \sum_{t=1}^T U(P, P_t)$$

and thus the best partition should maximize the value of the overall utility.

7. **Information-Theoretic Measures:** Strehl and Ghosh [35] suggested an information theoretic function based on mutual information

$$P_{best} = \operatorname{argmax}_P I'(P, \mathcal{P})$$

where $I'(P, \mathcal{P}) = \frac{1}{T} \sum_{t=1}^T I'(P, P_t)$. $I'(P, P_t)$ is the normalized mutual information between P and P_t and is calculated by

$$I'(P, P_t) = \frac{I(P, P_t)}{\sqrt{H(P)H(P_t)}}$$

where

$$p(C_r, C_t^j) = \frac{|C_t^j \cap C_r|}{n}$$

$$I(P, P_t) = \sum_{r=1}^K \sum_{j=1}^{k_t} p(C_r, C_t^j) \log \left(\frac{p(C_r, C_t^j)}{p(C_r)p(C_t^j)} \right)$$

$$H(P) = - \sum_{r=1}^K p(C_r) \log p(C_r), H(P_t) = - \sum_{j=1}^{k_t} p(C_t^j) \log p(C_t^j)$$

8. **Within/Between Cluster Distances:** Another collection of criteria is based on clustering criteria for categorical data. For clustering, a well-known criterion is to find the final partition minimizing either the intra-cluster distance

$$\sum_k \sum_{x, y \in C_k} \operatorname{dist}(x, y) \text{ or } \sum_k \sum_{x \in C_k} \operatorname{dist}(x, \bar{C}_k)$$

or maximizing the between-cluster distance $\sum_k \operatorname{dist}(\bar{C}_k, \bar{C})$ where \bar{C}_k is the centroid for cluster C_k and \bar{C} is the overall centroid. dist can be defined by L_p distances or other dissimilarity measures.

4. RELATIONS AMONG DIFFERENT CONSENSUS FUNCTIONS

We have derived the connections among various criteria and they can be briefly summarized in Figure 1. In particular, the equivalence relation between the two perspectives: consensus partition and categorical clustering, is established. The mathematical details of derivation on the connections are presented in the rest of this section.

First, the consensus matrix defines similarity between data points and the similarity is defined by the fraction of the number of clusters shared by them across all the partitions. Define $\operatorname{dist}(i, j) = \alpha S(i, j) + \beta$ for some $\alpha < 0, \beta$, then maximizing $\sum_{r=1}^K \sum_{i, j \in C_r} S(i, j)$ is equivalent to minimize $\sum_{r=1}^K \sum_{i, j \in C_r} \operatorname{dist}(i, j)$. A popular choice of $\alpha = -1, \beta = 1$ defines $\operatorname{dist}(i, j) = 1 - S(i, j) = \frac{1}{T} |d_i - d_j|$, where d_i is the induced vector representation for point i as in Equation 1.

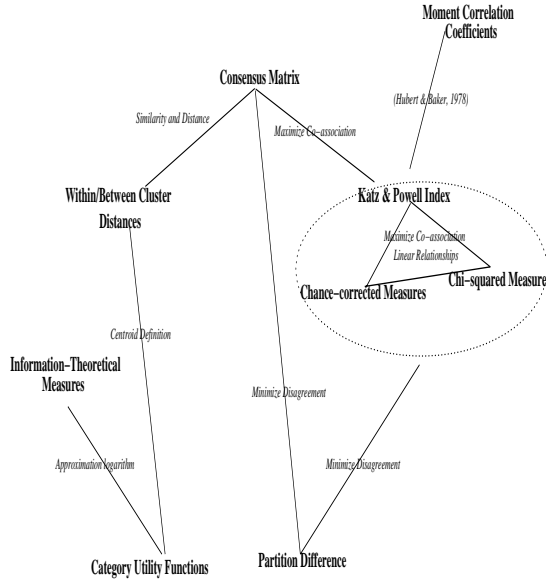


Figure 1: Summary of Relations for Various Consensus Criteria. The words beside the links describe connections between the criteria.

For consensus criteria based on Katz & Powell index, the chance-corrected and chi-squared measures, if the sizes of clusters in each clustering are fixed, they all reduce to maximizing $\sum_t N_{PP_t}$. In particular, given two clusterings A and B , if the sizes of each cluster within each clustering are the same, Katz & Powell index, Kappa and chi-squared measures are linearly related to each other. Mathematically, we have

$$n_{u.} = \frac{n}{p}, n_{.v} = \frac{n}{q}, 1 \leq u \leq p, 1 \leq v \leq q.$$

Then

$$\chi^2(A, B) = \sum_{i=1}^p \sum_{j=1}^q \frac{n_{uv}^2}{E_{uv}} - n = \frac{pq}{n} \sum_{i=1}^p \sum_{j=1}^q n_{uv}^2 - n$$

$$\begin{aligned} \Gamma(A, B) &= \frac{n^2 N_{AB} - N_A N_B}{\sqrt{N_A(n^2 - N_A)N_B(n^2 - N_B)}} \\ &= \frac{pq \sum_{i=1}^p \sum_{j=1}^q n_{uv}^2}{n^2 \sqrt{(p-1)(q-1)}} - \frac{1}{\sqrt{(p-1)(q-1)}} \\ &= \frac{1}{n \sqrt{(p-1)(q-1)}} \chi^2(A, B) \end{aligned}$$

$$\begin{aligned} K(A, B) &= \frac{N_{AB} - N_A N_B}{n^2 - N_A N_B} = \frac{pq \sum_{i=1}^p \sum_{j=1}^q n_{uv}^2 - n^4}{pq n^2 - n^4} \\ &= \frac{1}{n(pq - n^2)} \chi^2(A, B) - \frac{1 - n^2}{pq - n^2} \end{aligned}$$

Moreover, observe that

$$\begin{aligned} \sum_{r=1}^K \sum_{i,j \in C_r} S(i, j) &= \sum_k \sum_{i,j \in C_r} \frac{1}{T} \sum_{t=1}^T M_t(i, j) \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{u=1}^K \sum_{v=1}^{k_t} n_{uv}^2 = \frac{1}{T} \sum_{t=1}^T N_{PP_t} \end{aligned}$$

Hence maximizing $\sum_k \sum_{i,j \in C_k} S(i, j)$ is equivalent to maximiz-

ing $\sum_t N_{PP_t}$.

Next we show the relations between the partition difference and other measures. Denote M as the associated matrix for partition P ,

$$\begin{aligned} \sum_{t=1}^T \Delta(P, P_t) &= \sum_{t=1}^T \sum_{i,j} (M(i, j) - M_t(i, j))^2 \\ &= \sum_{t=1}^T \sum_{u=1}^K \sum_{v=1}^{k_t} n^2 - \sum_{t=1}^T \sum_{u=1}^K \sum_{v=1}^{k_t} n_{uv}^2 \\ &= \text{Constant} - T \sum_{r=1}^K \sum_{i,j \in C_r} S(i, j) \end{aligned}$$

So minimizing $\sum_{t=1}^T \Delta(P, P_t)$ is consistent with maximizing $\sum_k \sum_{i,j \in C_k} S(i, j)$ and $\sum_t N_{PP_t}$. On the other hand,

$$\begin{aligned} \sum_{t=1}^T \Delta(P, P_t) &= \sum_{t=1}^T \sum_{i,j} (M(i, j) - M_t(i, j))^2 \\ &= \text{Constant} - \sum_{i,j} M(i, j) [2 \sum_{t=1}^T M_t(i, j) - T] \\ &= \text{Constant} + T \sum_{i,j} M(i, j) [-2S(i, j) + 1] \quad (6) \end{aligned}$$

Define $dist(i, j) = -2S(i, j) + 1$, hence minimizing $\sum_{t=1}^T \Delta(P, P_t)$ is equivalent to minimizing $\sum_{i,j} M(i, j) dist(i, j) = \sum_{r=1}^K \sum_{i,j \in C_r} dist(i, j)$, i.e., the popular clustering criterion.

Now let's turn to the category utility function. We can show that the category utility function is equivalent to the square-error criterion defined for the induced categorical clustering problem. As we mentioned above, we can transform the partition P_t assuming k_t values by k_t binary features and the solution of the partition problem can be approached by the clustering algorithm operating in the induced space. Note that

$$\begin{aligned} U(P, P_t) &= \sum_{r=1}^K p(C_r) \sum_{j=1}^{k_t} p(C_t^j | C_r)^2 - \sum_{j=1}^{k_t} p(C_t^j)^2 \\ &= \sum_{r=1}^K \sum_{j=1}^{k_t} \frac{[p(C_r)p(C_t^j | C_r) - p(C_t^j)p(C_r)]^2}{p(C_r)} \\ &= \sum_{r=1}^K p(C_r) \sum_{j=1}^{k_t} \left(\frac{p(C_r)p(C_t^j | C_r)}{p(C_r)} - p(C_t^j) \right)^2 \end{aligned}$$

The overall mean for the tj -th attribute is $p(C_t^j)$ and the mean for the cluster k in the final partition is $\frac{p(C_r)p(C_t^j | C_r)}{p(C_r)}$.

$$\begin{aligned} \sum_{t=1}^T U(P, P_t) &= \sum_{r=1}^K p(C_r) \sum_{t=1}^T \sum_{j=1}^{k_t} \left(\frac{p(C_r)p(C_t^j | C_r)}{p(C_r)} - p(C_t^j) \right)^2 \\ &= \sum_{r=1}^K p(C_r) dist(\bar{C}_r, \bar{C}) \end{aligned}$$

Where \bar{C}_r and \bar{C} are the cluster representative for cluster r and the overall cluster center respectively. Hence maximizing the category utility function is equivalent to maximizing the within-cluster distance weighted by cluster sizes. This suggests the relations between the category utility function (using the between-attribute (partition) similarity measures) and the square-error clustering criterion (using distances between points and prototypes).

Finally let's take a look at the information theoretical measures.

$$\begin{aligned}
I(P, P_t) &= \sum_{r=1}^K \sum_{j=1}^{k_t} p(C_r, C_t^j) \log \left(\frac{p(C_r, C_t^j)}{p(C_r)p(C_t^j)} \right) \\
&= \sum_{r=1}^K p(C_r) \sum_{j=1}^{k_t} p(C_t^j | C_r) \log p(C_t^j | C_r) \\
&\quad - \sum_{j=1}^{k_t} p(C_t^j) \log p(C_t^j)
\end{aligned} \tag{7}$$

In fact, if we use $x - 1$ to approximate $\log x$, we can see that Equation 7 and Equation 5 are equivalent.

5. COMPLEXITY RESULTS ON MULTIPLE CLUSTERING

In this section, we summarize complexity results on the problem of combining multiple clusterings. There are two perspectives on the problem: consensus partition and categorical clustering.

5.1 Consensus Partition

From the consensus partition perspective, the problem of combining multiple clusterings is to find a target partition which optimizes the consensus criteria. To investigate the complexity of problem, let's look at the partition difference measure. From Equation 6, maximizing $\sum_{t=1}^T \Delta(P, P_t)$ is equivalent to minimizing $\sum_{i,j} M(i, j) \text{dist}(i, j)$. Note that M is the associated matrix for partition P , so it should satisfy the equivalence conditions in Table 2:

$M(i, j) \in \{0, 1\}$	Boolean matrix
$M(i, i) = 1, i = 1, \dots, n$	Reflexivity
$M(i, j) = M(j, i)$	Symmetry
$M^2(i, j) \leq M(i, j)$	Transitivity

Table 2: Conditions on the Associated Matrix

The reflexivity and symmetry conditions reduce the size of the problem and we only need to compute $M(i, j), i = 1, \dots, n; j = i + 1, \dots, n$. $M^2(i, j) \leq M(i, j)$ can be expressed as $M(i, k) + M(j, k) - M(i, j) \leq 1, \forall i, j, k$. With these conditions, the problem of finding the consensus partition is equivalent to an integer programming problem

$$\begin{cases} \text{Minimize } \sum_{i=1}^n \sum_{j=i+1}^n M(i, j) \text{dist}(i, j) \text{ where} \\ M(i, j) \in \{0, 1\} \\ M(i, k) + M(j, k) - M(i, j) \leq 1, \forall i, j, k \end{cases} \tag{8}$$

It has been shown in general that the 0 - 1 integer programming problem is NP-hard [31].

5.2 Clustering Complexity

Another way to look at the problem is from the clustering perspective. The number of the points to be clustered is n , and we need to find a clustering into k -classes. Then the multinomial number $\binom{n}{n_1, \dots, n_k}$ is the number of clusterings whose cluster sizes are n_1, \dots, n_k . Thus the number of all the clusterings of k clusters is $\sum_{n_1 + \dots + n_k = n} \binom{n}{n_1, \dots, n_k}$. Taking the ordering of clusters into consideration, the number of ways to par-

ition the set of n points into k -disjoint clusters is $N(n, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^{k-i} \binom{k}{i} i^n$ [9].

It is easily seen that this number is growing exponentially. Generally when $k \geq 3$ and the number of features is greater than 1, the clustering problem is proved to be NP-complete [5, 31]. There are few cases, when the number of clusters is less than 3 or the data only contain one feature, the clustering problem is polynomial solvable [5].

6. FINDING THE FINAL CLUSTERING

In previous section, we have shown the relations between various criteria and hence the problem of combining multiple clusterings can be solved via binary clustering. A critical problem in clustering is to determine the number of clusters. In this section, we present a method to determine the number of clusters for final clustering. The similar idea has been appeared in our previous work on document clustering [23].

LEMMA 1. Let $X = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$, i.e., all entries in the

matrix $X \in R^{n \times n}$ are 1. Then the only nonzero eigenvalue of X is n .

LEMMA 2. Let $L = \begin{pmatrix} X_1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & X_k \end{pmatrix}$. That is, L

is a block diagonal matrix, with each block matrix formed as in Lemma 1. Let n_i be the size of the matrix X_i , for $i = 1, \dots, k$. Then the only nonzero eigenvalues of L are n_1, n_2, \dots, n_k .

LEMMA 3. Let A and E be two symmetric matrices with the same dimensions. Then

$$|\lambda_i(A) - \lambda_i(A + E)| \leq \|E\|_2, \text{ for } i = 1, \dots, n$$

where $\lambda_i(A)$ denotes the i -th largest eigenvalue of the matrix A , similarly, $\lambda_i(A + E)$ for matrix $(A + E)$.

Proofs for Lemma 1 and Lemma 2 are immediate and Lemma 3 follows from the standard results in matrix perturbation theory [14].

THEOREM 4. Let $M = L + E$ where L has the form as in Lemma 2 and E is a matrix with a small value in each entry. Then M has k dominant eigenvalues, which are close to n_1, n_2, \dots, n_k .

Theorem 4 follows directly from Lemma 2 and Lemma 3. Given a binary dataset D , DD^T is a $n \times n$ matrix and can be thought as a similarity matrix between data points. The similarity of data point i and j is simply the inner product of the i -th row and the j -th row of D . If D is normalized, then each entry of DD^T shows the cosine similarity between corresponding data points. Since the permutation of the matrices does not change the spectral properties, we can order the points in D according to which cluster they are in and hence DD^T can be regarded as the addition of two matrices: $DD^T = L + E$ as described above. Hence the number of clusters can then be decided from the eigenvalues of DD^T .

7. EXPERIMENTS

7.1 Kinship Terms

The method described in previous sections is first applied to the analysis of 15 kinship terms data⁴ provided by 85 female undergraduates at Rutgers University [33]. The 15 terms were *Grandfather(GrF)*, *Grandmother(GrM)*, *Grandson(GrS)*, *Granddaughter(GrD)*, *Brother(Bro)*, *Sister(Sis)*, *Father(Fat)*, *Mother(Mot)*, *Son*, *Daughter(Dau)*, *Nephew(Nep)*, *Niece(Nie)*, *Uncle(Unc)*, *Aunt(Aun)* and *Cousin(Cou)*. Each student was instructed to provide a grouping of the terms on the basis of some aspect of meaning. The number of groups that the students pick is from 2 to 8. We first characterize the consensus problem as a binary clustering problem using Equation 1 and we then solve the problem using clustering approach. After transformation, each term is then represented as a 421-dimension binary vector. Denote the transformed dataset as D , then using the approach described in previous section, we can decide the number of final clusters. Figure 2 shows the top eigenvalues of the normalized DD^T . Note that the eigenvalues are close to the cluster sizes and there is a big drop from the third largest eigenvalue to the fourth largest eigenvalue. Hence we choose the final number of clusters as 3. Figure 3 lists the clustering results. These results are consistent with previous analysis of these data reported in [16, 17].

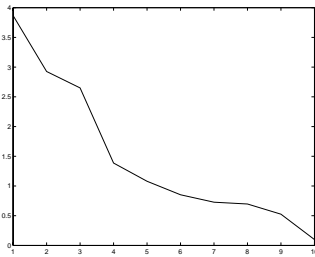


Figure 2: Top eigenvalues of DD^T . The Y-axis indicates the eigenvalues and the X-axis indicates the order of the eigenvalues.

Index	Members
I	{ <i>GrF</i> , <i>GrM</i> , <i>GrS</i> , <i>GrD</i> }
II	{ <i>Sis</i> , <i>Fat</i> , <i>Mot</i> , <i>Son</i> , <i>Dau</i> }
III	{ <i>Nep</i> , <i>Nie</i> , <i>Unc</i> , <i>Aun</i> , <i>Cou</i> }

Table 3: Clustering Results

7.2 Clustering popular music using both lyrics and acoustic data

This section addresses the issue of clustering popular music, i.e., clustering the music songs into groups denoted by the artists. As a fundamental and effective tool for efficient organization, summarization, navigation and retrieval of large amount of music data, clustering has been very active and enjoying a growing amount of attention. Ellis et al. [32] point out that similarity between music songs reflects personal tastes and suggest that different measures have to be combined together so as to achieve reasonable results in similarity retrieval. Our previous work has explored the idea of

⁴The dataset can be downloaded from <http://www.solar.mcs.st-and.ac.uk/~allan/KinshipTerms.data>.

music artist style identification by semi-supervised learning from both lyrics and content [24]. In this section, we take the approach of combining multiple clusterings: first perform clustering on each separate feature sets and then combine the clustering results.

7.2.1 Heterogeneous Feature Sets

In this section, we describe the feature sets extracted from the lyrics and the acoustic content.

Text-Based Style Features: Previous study on stylometric analysis has shown that statistical analysis on text properties could be used for text genre identification and authorship attribution [34, 2] and over one thousand stylometric features (style makers) have been proposed in variety research disciplines [36]. To choose features for analyzing lyrics, one should be aware of some characteristics of popular song lyrics. For example, song lyrics are usually brief and are often built from a very small vocabulary. In song lyrics, words are uttered with melody, so the sound they make plays an important in determination of words.

We divided the text-based style features into three different feature sets: *Bag-of-words*, *Part-of-Speech Statistics* and *Lexical/orthographic features*. Each of these three feature sets has been applied in previous study on stylometric analysis [36]. The text-based features are summarized in Table 4.

- **Bag-of-words:** We compute the TF-IDF measure for each words and select the top 200 words as our features. We do not apply stemming operations.
- **Part-of-Speech statistics:** We use the output of Brill’s part-of-speech(POS) tagger [4] as the basis for feature extraction. POS statistics usually reflect the characteristics of writing. There are 36 POS features extracted for each document, one for each POS tag expressed as a percentage of the total number of words for the document.
- **Lexical/Orthographic Features:** By lexical features, we mean features of individual word-tokens in the text. The most basic lexical features are lists of 303 generic function words taken from [29]⁵, which generally serve as proxies for choice in syntactic (e.g., preposition phrase modifiers vs. adjectives or adverbs), semantic (e.g., usage of passive voice indicated by axillary verbs), and pragmatic (e.g., first-person pronouns indicating personalization of a text) planes. We also use orthographic features of lexical items, such as capitalization, word placement, word length distribution. Word orders and lengths are very useful since the writing of lyrics usually follows a certain rhythm.

Content-Based Features: There has been a considerable amount of work in extracting descriptive features from music signals for music genre classification and artist identification [37, 25]. In our study, we use timbral features along with wavelet coefficient histograms. The feature set consists of the following three parts and totals 35 features.

- **Mel-Frequency Cepstral Coefficients (MFCC):** MFCC is designed to capture short-term spectral-based features. After taking the logarithm of the amplitude spectrum based on short-term Fourier transform for each frame, the frequency bins are grouped and smoothed according to Mel-frequency scaling, which is designed to agree with perception. MFCC features are generated by decorrelating the Mel-spectral vectors using discrete cosine transform.

⁵Available on line at <http://www.cse.unsw.edu.au/~min/ILLDATA/Function.word.htm>

Type	Number
Function Words (FW)	303
Token Place	5
Capitalization	10
Start of ...	9
Word Length	6
Line Length	6
Average Word Length	1
Average Sentence Length	1
POS features	36
Bag-Of-Words	200

Table 4: Summary of Feature Sets for Lyric Styles.

- Other Timbral Features: Timbral features consist of *Spectral Centroid*, *Spectral Rolloff*, *Spectral Flux*, *Zero Crossings* and *Low Energy*. *Spectral Centroid* is the centroid of the magnitude spectrum of short-term Fourier transform and is a measure of spectral brightness. *Spectral Rolloff* is the frequency below which 85% of the magnitude distribution is concentrated. It measures the spectral shape. *Spectral Flux* is the squared difference between the normalized magnitudes of successive spectral distributions. It measures the amount of local spectral change. *Zero Crossings* is the number of time domain zero crossings of the signal. It measures noisiness of the signal. *Low Energy* is the percentage of frames that have energy less than the average energy over the whole signal. It measures amplitude distribution of the signal.
- DWCH(Daubechies Wavelet Coefficients Histogram): To extract DWCH features, the Db8 filter with seven levels of decomposition is applied to three seconds of sound signals. After the decomposition, the histogram of the wavelet coefficients is computed at each subband. Then the first three moments of a histogram is used [7] to approximate the probability distribution at each subband. In addition, the subband energy is computed at each subband, which is defined as the mean of the absolute value of coefficients, for each subband. More details on DWCH feature extraction can be found in [25].

7.2.2 Experimental Results on Three Artists

The first experiment is performed on a dataset consisting of 106 songs from 11 albums by three artists (4 albums from Elton John, 4 albums from Joni Mitchell and 3 albums from Led Zeppelin). The sound recording and the lyrics from them are obtained and we use the method described in Section 7.2.1 to extract the text-based and content-based features. Four different feature sets: three text-based feature sets and one content-based feature set are obtained, and we label them as Feature set 1(*Bag-of-words*), 2(*Part-of-Speech Statistics*), 3(*Lexical/orthographic features*) and 4(*content-based features*) respectively. The experiment is to cluster the music. The artists' names are used as labels to evaluate the performance of clustering.

We use *purity* [39], which measures the extent to which each cluster contained data points from primarily one class, as our metric for cluster validity. The purity of a clustering solution is obtained as a weighted sum of individual cluster purities and is given by $Purity = \sum_{i=1}^K \frac{n_i}{n} P(S_i)$, $P(S_i) = \frac{1}{n_i} \max_j (n_i^j)$ where S_i is a particular cluster of size n_i , n_i^j is the number of documents of the i -th input class that were assigned to the j -th cluster, K is the

number of clusters and n is the total number of points⁶. In general, the larger the values of purity, the better the clustering solution is.

Table 5 presents the experimental results on three artists including both the clustering performances on different feature combinations and the result of combining multiple clusterings. For multiple clusterings, we first performed clustering on each of the four feature sets and then combine the clusterings results. The clustering algorithms were implemented using K -means approaches [22]. It can be observed that the clustering performance increases as more features are added in. On the combination of all the four feature sets, the clustering purity is the highest among all the feature combinations. This indicates that all the feature sets seem to be useful for identifying the music styles. The result of combining clusterings is slightly better than that on the combination of all four feature sets. The improvement is partially due to the ability of the multiple clusterings algorithm to learn the correlation structure among different feature sets. In addition, combining multiple clustering can also overcome the instability inherent in the clustering problem.

Feature Set(s)	Purity
1	0.53
2	0.48
3	0.52
4	0.54
1+2	0.64
1+3	0.61
1+4	0.56
2+3	0.58
2+4	0.55
3+4	0.58
1+2+3	0.68
1+2+4	0.66
2+3+4	0.60
1+2+3+4	0.688
Multiple Clusterings	0.698

Table 5: Experimental Results on Three Artists. The clustering performance was tested using various combinations of feature sets. Multiple Clusterings row shows the purity result of combining multiple clusterings and all the other rows show the results of clustering on the corresponding feature combinations. The results were obtained by averaging ten trials.

7.3 Experiments on Artist Similarity

The second experiment is performed on the dataset consisting of 570 songs from 56 albums of a total of 43 artists. The sound recording and the lyrics from them were obtained. In this experiment, we try to identify the artist similarities.

7.3.1 Similarity Ground Truth

Although we believe that the degree at which a listener finds a piece of music similar to another is influenced by the listener's cultural and music backgrounds and by the listener's state of mind, to make our investigation more plausible we choose to use similarity information available at All Music Guide (www.allmusic.com) as the ground truth, assuming that this information is reflection of multiple individual listeners. By examining All Music Guide artist pages, if the name of an artist X appears on the list of artists similar to Y, it is considered that X is similar to Y. The similarity graph of these nodes are shown in Figure 3. We did not agree completely

⁶ $P(S_i)$ is also called the individual cluster purity.

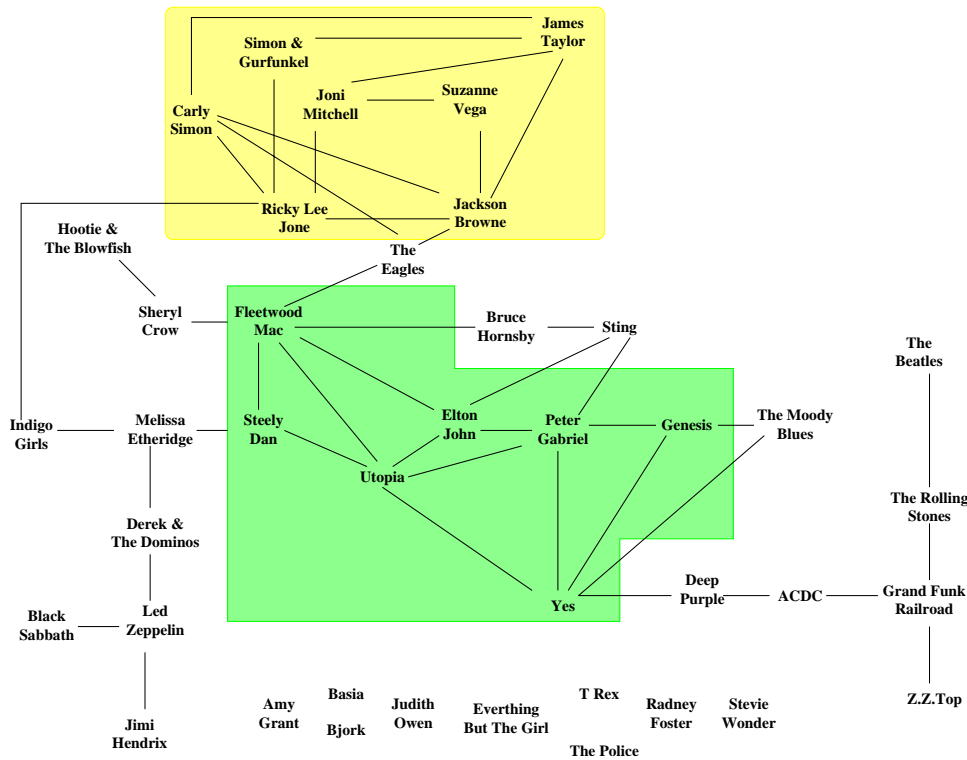


Figure 3: Artist Similarity Graph.

with the artist similarity thus obtained but nonetheless used it as the ground truth.

7.3.2 Results Analysis

We first group songs into different clusters using both lyrics and content data via the approach of combining multiple clusterings (as described in Section 7.2.2). Then we define the similarity between two artists as the ratio of the number of songs in the same cluster to the total number of songs. A hierarchical clustering scheme is then used to generate a similarity dendrogram of all the artists. Figure 4 shows the generated dendrogram from our experiment.

We can derive three clusters from the dendrogram as listed in Table 6. In the similarity graph of Figure 3, if the name of an artist X appears on the list of artists similar to Y, it is considered that X is similar to Y. The cluster structures listed in Table 6 indicate that our approach of combining multiple clusterings correctly learned relationships revealed in the artist similarity graph. For example, Fleetwood Mac and Hootie (as well as Sheryl Crow, Hootie) were correctly clustered into the same class. AC/DC and Deep Purple were classified as the same class with Elton John. However, there were inconsistencies between the cluster structures generated by the dendrogram and the similarity graph. Using analytical similarity measures to obtain the ground truth about artist similarity, thereby improving upon the data provided by web information resources, will be our future goal. In conclusion, We can conclude from these experiments that artist similarity can be efficiently learned using multiple data sources.

8. CONCLUSIONS

Many application problems can be reduced to the problem of combining multiple clusterings. This paper provides a unified view

Clusters	Members
No.1	{ Jackson Browne, Genesis, Suzanne Vega, Melsissa Etheridgem The Rolling Stones, Carly Simon, Utopia, Hootie }
No. 2	{ ZZ Top, Elton John, James Taylor, Led Zeppelin, Sheryl Crow, Fleetwood Mac, AC/DC, Ricky Lee Jones, Black Sabbath, Sting, Deep Purple }
No. 3	{ Jim Hendrix, Yes, Joni Mitchell, Eagles, Steely Dan, Peter Garbiel Derek & The Dominos }

Table 6: Cluster Memberships.

of the problem of combining multiple clusterings by exploring the connections among various criteria. In addition, it shows the equivalence between the two different perspectives for combining multiple clusterings: consensus partition and categorical clustering. A novel method based on the matrix theory for determining the final clustering is also presented. Experiment results show the effectiveness of combining multiple clusterings.

9. REFERENCES

- [1] Arabie, P., Carroll, J. D., & Desarbo, W. (1987). *Three-way scaling and clustering*. Newbury Park, CA: Sage publications.
- [2] Argamon, S., Saric, M., & Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 475–480). Washington, D.C.: ACM Press.

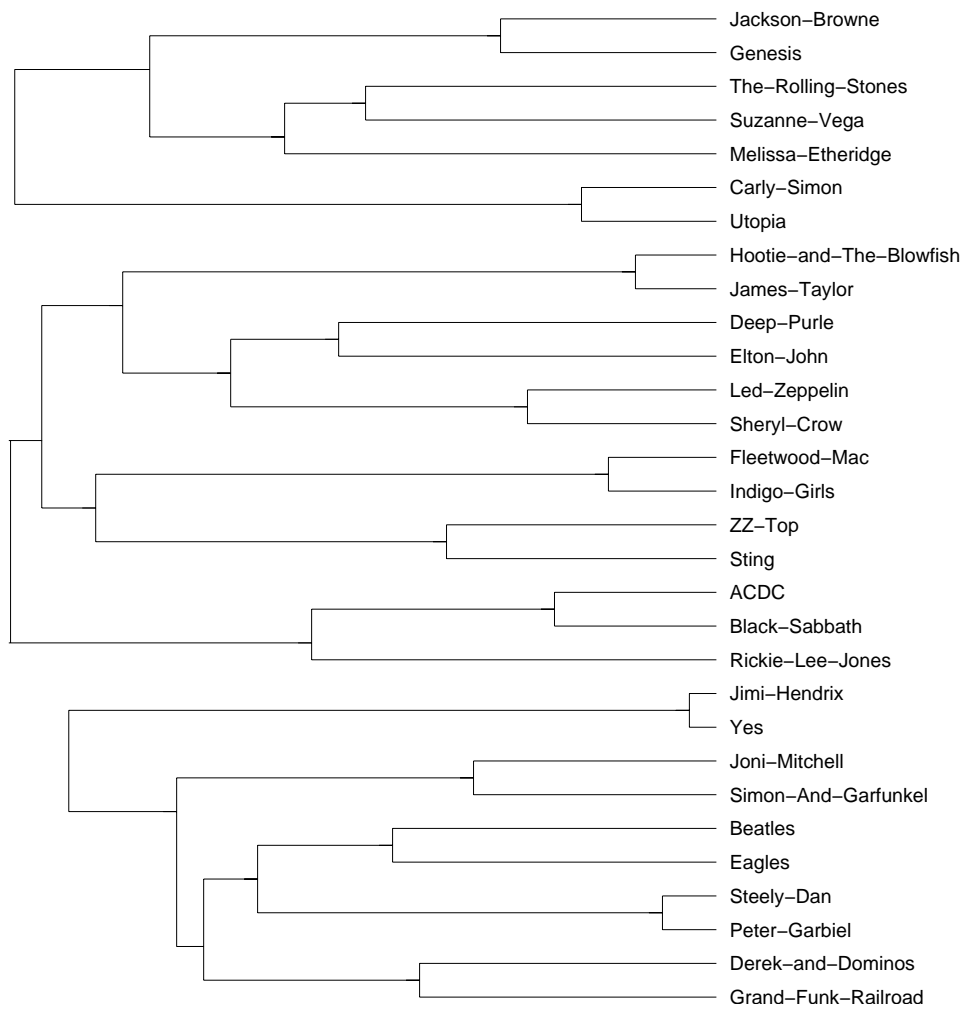


Figure 4: Artist Similarity Dendrogram.

- [3] Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36, 105–139.
- [4] Bill, E. (1994). Some advances in transformation-based parts of speech tagging. *Proceedings of the twelfth national conference on Artificial intelligence (vol. 1)* (pp. 722–727). American Association for Artificial Intelligence.
- [5] Brucker, P. (1977). On the complexity of clustering problems. *Optimization and Operations Research* (pp. 45–54). Springer-Verlag.
- [6] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- [7] David, A., & Panchanathan, S. (2000). Wavelet-histogram method for face recognition. *Journal of Electronic Imaging*, 9, 217–225.
- [8] Day, W. H. E. (1986). Foreword: Comparison and consensus of classifications. *Journal of Classification*, 3, 183–185.
- [9] Duran, B. S., & Odell, P. L. *Cluster analysis: a survey*. New York, NY: Springer.
- [10] Everitt, B. S. (1987). *Introduction to optimization methods and their application in statistics*. Chapman and Hall.
- [11] Ferligoj, A. (1992). Direct multicriteria clustering algorithm. *Journal of Classification*, 9, 43–61.
- [12] Ferligoj, A., & Batagelj, V. (1983). Some types of clustering with relational constraints. *Psychometrika*, 48, 541–552.
- [13] Fern, X. Z., & Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)* (pp. 186–193). Morgan Kaufmann Publishers.
- [14] Golub, G. H., & Loan, C. F. V. (1991). *Matrix computations*. The Johns Hopkins University Press.
- [15] Goodman, L. A., & Kruskal, W. H. (1954). Measures of associations for cross classification. *Journal of the American Statistical Association*, 49, 732–764.
- [16] Gordan, A. D., & Vichi, M. (1998). Partitions of partitions. *journal of classification*, 15, 265–285.
- [17] Gordan, A. D., & Vichi, M. (2002). Obtaining partitions of a set of hard or fuzzy partitions. *Classification, Clustering and Data Analysis: recent advances and applications* (pp. 75–79). Springer.
- [18] Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *journal of classification*, 2, 193–218.
- [19] Hubert, L. J., & Baker, F. B. (1978). Evaluating the conformity of sociometric measurements. *Psychometrika*, 43, 31–41.
- [20] Kargupta, H., Huang, W., Sivakumar, K., & Johnson, E. L. (2001). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems*, 3, 422–448.
- [21] Katz, L., & Powell, J. H. (1953). A proposed index of the conformity of one sociometric measurement to another. *Psychometrika*, 18, 249–256.
- [22] Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley.
- [23] Li, T., Ma, S., & Ogihara, M. (2004). Document clustering via adaptive subspace iteration. *Proceedings of Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*. To appear.
- [24] Li, T., & Ogihara, M. (2004). Music artist style identification by semisupervised learning from both lyrics and content. *Proceedings of the ACM Conference on Multimedia*. To appear.
- [25] Li, T., Ogihara, M., & Li, Q. (2003a). A comparative study on content-based music genre classification. *SIGIR'03* (pp. 282–289). ACM Press.
- [26] Li, T., Zhu, S., & Ogihara, M. (2003b). Algorithms for clustering high dimensional and distributed data. *Intelligent Data Analysis Journal*, 7, 305–326.
- [27] Messatfa, H. (1992). An algorithm to maximize the agreement. *Journal of Classification*, 9, 5–15.
- [28] Mirkin, B. (2000). Reinterpreting the category utility function. *Machine Learning*, 45, 219–228.
- [29] Mitton, R. (1987). Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management*, 23, 103–209.
- [30] Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning Journal*, 52, 91–118.
- [31] Moret, B. M. (1998). *The theory of computation*. Addison-Wesley.
- [32] P.W.Ellis, D., Whitman, B., Berenzweig, A., & Lawrence, S. (2002). The quest for ground truth in musical artist similarity. *Proceedings of 3rd International Conference on Music Information Retrieval* (pp. 170–177).
- [33] Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.
- [34] Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26, 471–496.
- [35] Strehl, A., & Ghosh, J. (2003). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3, 583–617.
- [36] Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measure of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.
- [37] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10.
- [38] Vichi, M. (1999). One-mode classification of a three-way data matrix. *journal of classification*, 16, 27–44.
- [39] Zhao, Y., & Karypis, G. (2001). *Criterion functions for document clustering: Experiments and analysis* (Technical Report). Department of Computer Science, University of Minnesota.

Acknowledgments

The authors want to thank Mr. Chengliang Zhang for helping with the experiments in Section 7.3. We are also grateful to the conference reviewers for their helpful comments and suggestions. The first and the second authors are supported in part by NSF grants EIA-0080124 and EIA-0205061 and in part by NIH grant P30-AG18254.