Geometric Clustering Models for Multimedia Databases *

Mary Inaba† and Hiroshi Imai†

1 Introduction

Recently, in the fields of information retrieval, Data Mining, or Knowledge Discovery in Databases (KDD), is intensively studied to extract implicit useful information from large amount of data. One of the important objectives of KDD is to obtain generalizations by grouping similar objects via clustering. In the case of multi-media databases such as full text database and image database, geometric structure of the feature space of the objects may be used, where computational geometry may play an important role.

As for the geometric structure, in the full text databases, a geometric approach, called the vector-space method, has been developed for advanced information retrieval of full text databases (Salton et al. [7, 6]). Roughly, this method first fixes a set of d terms, and maps each text to a point in the d-dimensional space such that the i-th coordinate value is the frequency of term i. In image databases, it is also natural to adopt geometric approaches in querying images by their content, as discussed in [2]. By counting the frequencies of d colors, each image is mapped to a point in the k-dimensional space, called the feature space of images.

This paper investigates geometric structures for geometric clustering in these databases. We first describe our formulation of the geometric clustering problems, and summarize some of our algorithmic results. We also mention using the Kullback-Leibler divergence as a measure. Then, we clarify underlying geometric structures for the vector-space model for texts and the quadratic form measure for images.

2 Geometric Clustering Problem

Now we proceed to defining our clustering problem. For a set S of n vectors $\boldsymbol{t}_i$ in the d-dimensional space, a k-clustering $(S_1, S_2, \ldots, S_k)$ of S is a partition of S into k subsets S_l $(l = 1 \ldots, k)$, i.e.,

$$S_l \cap S_{l'} = \emptyset \ (l \neq l'), \quad \bigcup_{l=1}^{k} S_l = S.$$

In the Euclidean space, when we use a criteria, "minimize the sum of squared distances", this problem can be considered as an optimization problem

$$\min\{\sum_{j=1}^{k} \mathrm{Cost}(S_j) \mid k\text{-clustering}(S_1, ..., S_k) \ \text{ of } S\}$$

and, the cost of each cluster S_j is

$$\mathrm{Cost}(S_j) = \sum_{i \in S_j} \|\boldsymbol{t}_i - \boldsymbol{r}_j\|^2$$

where $\boldsymbol{r}_j$ is the representative point of each cluster S_j, and is set to the centroid of points in S_j.

Similarly, in the statistical parametric space (e.g., see [1]), which can be regarded as the extension of the Euclidean space, we can consider the criteria "minimize the sum of divergence", where the squared Euclidean distance is the simplest form of divergence, and, the most well-known divergence is the Kullback-Leibler divergence, or the relative entropy, in information theory. The Kullback-Leibler divergence $D(\boldsymbol{p}\|\boldsymbol{q})$ for d-dimensional vectors $\boldsymbol{p} = (p_i)$, $\boldsymbol{q} = (q_i)$ such that each coordinate value is positive and the sum of all coordinate values is 1,

$$D(\boldsymbol{p}\|\boldsymbol{q}) = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}.$$

Then, the cost of cluster S_j is defined for this divergence by

$$\mathrm{Cost}(S_j) = \sum_{i \in S_j} D(\boldsymbol{t}_i\|\boldsymbol{r}_j)$$

where r_j is the centroid with respect to the divergence. For the divergence, the Voronoi diagram can be defined and it has similar nice properties like the Euclidean Voronoi diagram. A brief exposition, with a video, is found in [5].

Theorem 1 *In the Euclidean Space (statistical parametric space), an optimal solution to the k-clustering problem with the criteria "minimize the sum of squared distances" ("minimize the sum of divergence") can be characterized by a partition of n points induced by the Voronoi diagram generated by the centroids (centroids by divergence) of each cluster.*

Based on this characterization, all the possible partitions generated by the Voronoi Diagrams may be enumerated to find an optimal clustering. For the Euclidean case, the number of all possible partitions is analyzed in [4]. Since this number is quite large even

*This research was supported in part by the Grant-in-Aid of the Ministry of Education, Science, Sports and Culture, Japan.

†Department of Information Science, University of Tokyo, 113-0033, Japan, E-mail: {mary,imai}@is.s.u-tokyo.ac.jp

for moderate k, d, n, a randomized approximate algorithm is presented to overcome that difficulty [4] and some experimental results are given in [3]. For the general divergence, similar arguments holds, which will be discussed in a full version.

3 Vector Space Model for Texts

In the vector-space model, all information items of stored texts are represented by vectors, or points, of terms, or keywords, in the space whose dimension is the number of terms. A term is typically a word. In automatic processing of various texts, the terms are derived directly from the texts under consideration.

Since the terms does not equally represent the contents of texts, it is important to use a term-weighting system which assigns high weights to terms deemed important and lower weights to the less important terms. There are many term-weighting systems, and a typical one described in [6] is given by the equation $f_t \times 1/f_c$(term frequency times inverse collection frequency), which favors terms with a high frequency (f_t) in particular documents but with a low frequency overall in the collection (f_c).

Then, all texts and text queries are represented by weighted term vectors $\boldsymbol{t}_i$ in the d-dimensional space where d is the number of terms. Of course, the l-th element in $\boldsymbol{t}_i$ is the weight assigned to the l-th term in the document i. The similarity $\mathrm{sim}(\boldsymbol{t}_i, \boldsymbol{t}_j)$ between two vectors $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ of given two documents is defined by

$$\mathrm{sim}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \frac{\boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{t}_j}{\|\boldsymbol{t}_i\| \|\boldsymbol{t}_j\|} = \cos\theta$$

where $\| \ \|$ denotes the L_2 norm, and θ is the angle between two vectors $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$. The similarity value ranges from 0 (low similarity) to 1 (high similarity).

Here, we define dissimilarity instead of similarity as follows:

Definition 1 (Dissimilarity of two vectors) *The dissimilarity* $\mathrm{dis}(\boldsymbol{t}_i, \boldsymbol{t}_j)$ *between two vectors* $\boldsymbol{t}_i$ *and* $\boldsymbol{t}_j$ *is defined to be*

$$\mathrm{dis}(\boldsymbol{t}_i, \boldsymbol{t}_j) = 1 - \mathrm{sim}(\boldsymbol{t}_i, \boldsymbol{t}_j).$$

From the geometrical viewpoint, text vector $\boldsymbol{t}_i$ is mapped to a normalized vector $\tilde{\boldsymbol{t}}_i$ by the projection on the sphere with radius 1, then, the dissimilarity between two texts $\boldsymbol{t}_i, \boldsymbol{t}_j$ can be considered as the squared Euclidean distance between these normalized vector $\tilde{t}_i$, $\tilde{t}_j$. That is,

$$\mathrm{dis}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \frac{1}{2} \|\tilde{\boldsymbol{t}}_i - \tilde{\boldsymbol{t}}_j\|^2.$$

Similarly, if we use the hyperplane $x_1 + x_2 + \cdots + x_d = 1$ instead of sphere $x_1^2 + x_2^2 + \cdots + x_d^2 = 1$, each text can be regarded as the probability distribution in statistical parametric space, and the Kullback-Leibler divergence may be used as dissimilarity.

4 Quadratic Form Model for Images

In the Query-by-Image-Content (QBIC) system [2] the following measure is used to estimate the closeness of two images by their color frequency vectors. First, find $\tilde{d}$ representative colors for the original color images, and quantize the images by using the $\tilde{d}$ colors. Let $B = (b_{ij})$ be a matrix of order $\tilde{d}$ such that b_{ij} is the negative of some dissimilarity between color i and color j. Let $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ be two normalized frequency vector of two images with respect to the $\tilde{d}$ colors, where normalization is done by dividing the frequency vector by the total count so that the sum becomes 1. Then, the distance-like function between two images is defined to be $\mathrm{dis} = \Delta\boldsymbol{\xi}^{\mathrm{T}} B \Delta\boldsymbol{\xi}$ with $\Delta\boldsymbol{\xi} = \boldsymbol{\xi}_1 - \boldsymbol{\xi}_2$.

When b_{ij} is defined to be the negative of squared Euclidean distance between two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the d-dimensional space, i.e., $b_{ij} = -\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, which is the case in the original framework of the QBIC system based on the Luv color system with $\tilde{d} = 64$ or 256 and $d = 3$, then the distance dis is expressed as

$$\mathrm{dis} = 2\|A\Delta\boldsymbol{\xi}\|^2$$

where A is a matrix whose i-th column is $\boldsymbol{x}_i$. Hence, simply applying the singular value decomposition of A, we can reduce the clustering problem by the distance dis to the geometric clustering problem with our objective function in the d-dimensional space.

References

[1] S. Amari: *Differential-Geometric Methods in Statistics*. Lecture Notes in Statistics, Vol.28, Springer-Verlag, 1985.

[2] C. Faloutsos, R. Barber, M. Flickner, W. Niblack, D. Petkovic and W. Equitz: Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, Vol.3 (1994), pp.231–262.

[3] M. Inaba, H. Imai and N. Katoh: Experimental Results of Randomized Clustering Algorithm. *Proceedings of the 12th Annual ACM Symposium on Computational Geometry*, 1996, pp.C1–C2.

[4] M. Inaba, N. Katoh and H. Imai: Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering. *Proceedings of the 10th ACM Symposium on Computational Geometry*, 1994, pp.332–339.

[5] K. Sadakane, H. Imai, K. Onishi, M. Inaba, F. Takeuchi, and K. Imai: Voronoi Diagrams by Divergences with Additive Weights. *Proceedings of the 14th ACM Symposium on Computational Geometry*, Video descriptions, 1998, to appear.

[6] G. Salton, J. Allan, C. Buckley and A. Singhal: Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. *Science*, Vol.264, 1994, pp.1421–1426.

[7] G. Salton, A. Wong, C. S. Yang: A Vector Space Model for Automatic Indexing. *Communications of the ACM*, Vol.18, No.11, 1975, pp.613–620.