情報検索・全文データベースでの文書クラスタリングでの幾何構造活用

稲葉真理　今井浩

東京大学理学系研究科情報科学専攻

アブストラクト：情報検索や全文データベースでの質問処理に関連して，文書をキーワードの出現頻度情報にしたがってその特徴空間の点に写像し，その特徴空間での幾何構造を活用することが行なわれている．しかし，従来法では，幾何構造は単に距離を導くために使われている場合が多く，その次のクラスタリングによる汎化や高速処理化のために幾何構造が十分に活用されていない．本稿では，この文書のクラスタリング問題に幾何構造が役に立つことを示し，本研究グループが従来から示してきた手法が適用できることを述べる．

Geometry Helps Clustering Texts in Information Retrieval and Text Databases

Mary Inaba and Hiroshi Imai

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

Abstract: In the field of information retrieval of full text databases, the vector-space model has been developed to process texts efficiently over 20 years. In this model, each text is mapped to a point in the feature space by considering frequencies of keywords appearing in the text. So far, this geometric structure is mainly used to derive the distance between two texts only. This paper describes that the geometric structure of this feature space can further be employed in clustering texts in databases. We propose using geometric clustering algorithms for variance-based clustering developed by our groups. This paper then surveys such a geometric clustering approach to analyze texts by making full use of their geometric structures. Thus, besides near neighbor search algorithms, existing geometric clustering algorithms can be made use of to automatically analyze texts in text databases.

1 Introduction

A geometric approach, called the vector-space method, has been developed for advanced information retrieval of full text databases (Salton et al. [15, 14]). Recent trends, that an enormous amount of machine-readable texts become available very easily via Internet, CD, etc., shed strong light on this method, and investigations from theoretical sides has also emerged (e.g., see Indyk, Motwani, Raghavan, Vempala [11]. However, so far, geometric structure in the space of weighted term vectors is only used in deriving the similarity value between two texts, and further processing such as text analysis, text theme identification, etc., is done simply using such values without returning to the geometric structure.

In this paper, we show a geometric clustering approach which fully utilizes the geometric structure in clustering texts in the space of weighted term vectors. Here, as a distance measure between two points in the space, we use the squared Euclidean distance, which is justified as follows.

In the vector-space model, the similarity is defined as

$$\mathrm{sim}(t_i, t_j) \cos\theta$$

for angle θ between two vectors t_i, t_j corresponding to the two texts. By defining the dissimilarity, or distance, $\mathrm{dis}(t_i, t_j)$ to be $1 - \cos\theta$, we show that this model corresponds to a geometric model on the normalized vectors $\tilde{t}_i$ such that the distance between two points is measured by the square of their Euclidean distance.

In view of this fundamental geometric observations, this paper considers a clustering problem of finding a k-clustering $(S_1, S_2, \ldots, S_k)$ of a set S of n normalized vectors $\tilde{t}_i$ in the d-dimensional space minimizing

$$\sum_{l=1}^{k} \sum_{\tilde{t}_i, \tilde{t}_j \in S_l} \mathrm{dis}(\tilde{t}_i, \tilde{t}_j)$$

which is equivalent to minimizing

$$\sum_{l=1}^{k} \sum_{\tilde{t}_i, \tilde{t}_j \in S_l} \|\tilde{t}_i - \tilde{t}_j\|^2.$$

(It should be noted that this problem is strongly connected to a clustering problem considered in Salton et al. [15] to design a nice space of weighted term vectors for retrieval.) Furthermore, when the cost of a cluster is defined to be the average distance for each point to points, including itself, in the cluster, the problem becomes finding a k-clustering minimizing

$$\sum_{l=1}^{k} \left(\sum_{\tilde{t}_i \in S_l} \frac{1}{|S_l|} \left(\sum_{\tilde{t}_j \in S_l} \mathrm{dis}(\tilde{t}_i, \tilde{t}_j) \right) \right)$$
$$= \sum_{l=1}^{k} \frac{1}{|S_l|} \sum_{\tilde{t}_i, \tilde{t}_j \in S_l} \mathrm{dis}(\tilde{t}_i, \tilde{t}_j)$$
$$= \sum_{l=1}^{k} \frac{1}{|S_l|} \sum_{\tilde{t}_i, \tilde{t}_j \in S_l} \|\tilde{t}_i - \tilde{t}_j\|^2.$$

This problem is known as the clustering problem minimizing the sum of squared distances [6, 10].

In Hasegawa, Imai, Inaba, Katoh, Nakano [6] and Inaba, Katoh, Imai [10] (see also [8, 9, 13]), these problems are considered as a geometric clustering problem minimizing the all-pair sum of squared distances and that minimizing the sum of squared distances, and geometric analyses are performed. We here summarize their implications in this space of weighted term vectors.

For automatic analysis, theme generation of texts, so far only clustering methods using the similarity matrix are used (Salton, Allan, Buckley, Singhal [14], Kitagawa, Mizuuchi, Tajima, Tanaka [12]). In general, the clustering problem based on the similarity matrix is more general than the geometric clustering, since in the latter problem the similarity matrix is basically induced from the coordinates of objects and hence is less general (or, has more properties induced by geometry). However, for texts, as summarized above, the similarity and dissimilarity between objects are defined utilizing geometric structures, and then the clustering problem for text databases becomes a special case of the geometric clustering, as depicted in Figure 1. Hence, what have been done for clustering texts can be done in geometric setting. Then, it is more powerful to apply geometric techniques, as discussed in this paper, to the problem. For example, in the geometric setting, the concept of representative points such as centroids (see [4]), medoids are well-defined and computationally easy to find, while such concepts are difficult to define or hard to compute in the simple similarity/dissimilarity matrix setting.

We expect that our geometric clustering approach would be useful enough to tackle these problems as the geometric clustering paradigms such as the ordinary k-means has demonstrated their power. Computational results will be reported in subsequent reports for this geometric clustering approach which is theoretically founded in this paper.

2 Vector-Space Model

To explain the vector-space model [15, 14], we are here based on the descriptions in [14].

In the vector-space model, all information items, of stored texts as well as information queries, 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-

Clustering texts in the vector space model (latent semantic indexing space) $\subseteq$ **Geometric Clustering** $\subseteq$ Clustering by similarity/dissimilarity matrix

Graph Clustering

Figure 1: Relations among clustering problems: As far as distances induced from geometry of the corresponding space in clustering texts, the clustering problem is a geometric clustering, and hence by utilizing the corresponding geometric structures more powerful clustering results may be obtained!

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 [14] 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, texts are represented by weighted term vectors t_i in the d-dimensional space where d is the number of terms. Of course, the l-th element in t_i is the weight' assigned to the l-th term in the document i. The similarity $\mathrm{sim}(t_i, t_j)$ between two vectors t_i and t_j of given two documents is defined by

$$\mathrm{sim}(t_i, t_j) = \frac{t_i \cdot t_j}{\|t_i\| \|t_j\|} = \cos\theta$$

where $\cdot$ is the vector inner product operation, $\| \; \|$ denotes the L_2 norm and θ is the angle between two vectors t_i and t_j. The similarity value ranges from 0 (low similarity) to 1 (high similarity).

The latent semantic indexing applies the singular value decomposition to a matrix of the above set of vectors (see Berry, Dumais [3]). When the dimension is reduced to two, the results can be presented visually, and clusters may often be found in this planar configuration of points. In this sense, the latent semantic indexing approach is sometimes viewed also as a clustering method. This dimension reduction may be considered as a way of removing noisy and/or useless parts in the original vectors. However, by this dimension reduction, much information is lost, and it is often the case that even after the dimension reduction we still have to treat vectors in dimensions, say 50 to 100. Then, again, some clustering methods based on the geometric of this reduced space are needed.

3 Dissimilarity and Clustering Criterion

For information retrieval, using the similarity is more natural. For example, it would elucidate quantitatively users or analyzers of text databases to answer, 'texts t_i and t_j matches 98%' for very similar texts when the similarity is 0.98.

For clustering purposes, it is more convenient to have an appropriate definition of dissimilarity. The dissimilarity may be defined as the minus, inverse or minus logarithm of similarity. Since the value of the similarity defined by the cosine of angle has been recognized as useful, the minus of similarity is more suitable in this case. Furthermore, to treat the problem in a geometric setting without losing any meaningful information, we here define the dissimilarity as follows:

Definition 1 (Dissimilarity of two vectors) *The dissimilarity* $\mathrm{dis}(t_i, t_j)$ *between two vectors* t_i *and* t_j *is defined to be* $1 - \mathrm{sim}(t_i, t_j)$.

By this definition, the all-pair sum of similarities of two vectors in a cluster of l vectors is exactly $\binom{l}{2}$ minus the all-pair sum of dissimilarities of two vectors in the cluster. Again, the dissimilarity value ranges from 0 (high similarity) to 1 (low similarity).

Now we proceed to defining our clustering problem. For a set S of n vectors 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.,

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

Since the geometric structure of this space of weighted term vectors has not yet fully been understood (this is left as a future problem), we here adopt the most simple all-pair criterion for our cluster cost:

Definition 2 (Intracluster cost: all-pair case) *The all-pair intracluster cost of a cluster S_l is defined to be*

$$\sum_{t_i,t_j \in S_l} \mathrm{dis}(t_i, t_j).$$

Note that $\mathrm{dis}(t_i, t_j) = 0$ when $t_i = t_j$.

Now, the average intracluster cost of a cluster S_l is defined to be

Definition 3 (Intracluster cost: average case) *The average intracluster cost of a cluster S_l is defined to be*

$$\frac{1}{|S_l|} \sum_{t_i,t_j \in S_l} \mathrm{dis}(t_i, t_j) = \frac{1}{|S_l|} \sum_{t_i \in S_l} \left(\sum_{t_j \in S_l} \mathrm{dis}(t_i, t_j) \right).$$

The intercluster cost is defined, as usual, by the summation of all intracluster costs $\mathrm{Intracost}(S_l)$:

Definition 4 (Intercluster cost) *The intercluster cost of a k-clustering $(S_1, S_2, \ldots, S_k)$ is defined to be*

$$\mathrm{Intercost}(S_1, \ldots, S_k) = \sum_{l=1}^{k} \mathrm{Intracost}(S_l).$$

Now, our clustering problem is stated as follows:

Definition 5 *The clustering problem is to find a k-clustering $(S_1, S_2, \ldots, S_k)$ of S that minimizes* $\mathrm{Intercost}(S_1, \ldots, S_k)$ *in each of all-pair and average intracluster costs.*

4 Normalizing Weighted Term Vectors

The domain of weighted term vectors may not be bounded in general. Since the similarity of two vectors is defined by the cosine of their angle, normalizing each vector t_i to

$$\tilde{t}_i = \frac{t_i}{\|t_i\|},$$

where $\|\cdot\|$ denotes the L_2 norm, does not change the similarity value. For normalized vectors, we have the following.

Lemma 1

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

Proof: For an isosceles triangle with edge length 1, 1 and d and angle θ between two edges of length 1, we have

$$d^2 = 2(1 - \cos\theta),$$

which implies the lemma. □

Thus, the clustering problem in the all-pair case is to find a k-clustering minimizing

$$\sum_{l=1}^{k} \sum_{\tilde{t}_i,\tilde{t}_j \in S_l} \|\tilde{t}_i - \tilde{t}_j\|^2,$$

and that in the average case is to find a k-clustering minimizing

$$\sum_{l=1}^{k} \frac{1}{|S_l|} \sum_{\tilde{t}_i,\tilde{t}_j \in S_l} \|\tilde{t}_i - \tilde{t}_j\|^2,$$

both of which are considered in [6, 10]. It should be noted that minimizing the average cost is equivalent to minimizing the sum of $\mathrm{dis}(\bar{t}, t_i)$ from the centroid $\bar{t}$ of the cluster to each point t_i.

5 Survey of Our Previous Results to the Clustering Problem and Their Implications

5.1 Clustering into k clusters

The above clustering problem is studied in [6, 10] as the 'all-pair sum of squared errors' problem and 'sum of squared errors' problem. Optimum clusterings in both cases are induced by the Voronoi diagrams [10], which are powerful tools in computational geometry. See Figure 2 and 3. Applying the theorems in [10], we obtain the following.

Theorem 1 ((Inaba, Katoh, Imai [10])) *The k-clustering problem for n vectors of texts in the space of weighted term vectors can be solved in time $O(n^{(d+2)k+1})$ and $O(n^{dk+1})$ in the case of all-pair and average cases, respectively.* □

The bounds in this theorem are polynomial in n but exponential in d and k. In the application to text databases, d is not so small (in fact, reducing d by using the principal component decomposition, etc., is another research issue for the vector-space method). In this regard, this theorem is not so useful, although it can be

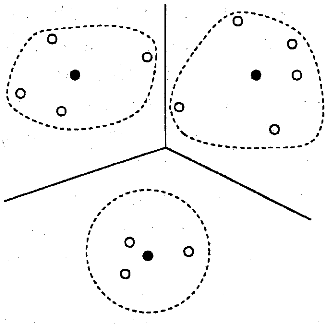

Figure 2: 3-clustering of 12 points (case of the sum of squared distances)

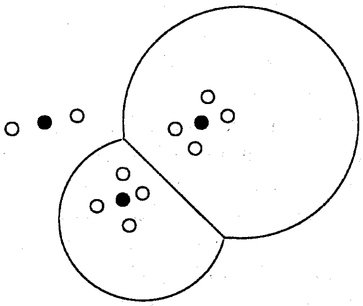

Figure 3: 3-clustering of 10 points (case of the all-pairs sum of squared distances)

used to derive a nontrivial bound for the following practical clustering algorithm.

For the average case, the following well-known k-means algorithm works well in practice (see [5, 7, 16])

```
Standard $k$-Means Algorithm:
Find an initial $k$-clustering $S_j$ $(j = 1, \ldots, k)$ for
$n$ vectors $\tilde{t}_i$ $(i = 1, \ldots, n)$;
repeat
    Compute the centroid of each cluster:
        $\bar{t}(S_j) = \frac{1}{|S_j|} \sum_{\tilde{t}_j \in S_j} \tilde{t}_j;$
    Update the $k$-clustering to the Voronoi partition induced by the Euclidean Voronoi diagram of $\bar{t}(S_j)$;
until a local minimum is found.
```

Thus, for the average case, this well-established algorithm can be fully utilized.

Concerning the implementations of the k-means algorithm, many computational-geometric techniques such as near neighbor search can be applied. The nearest neighbor search problem becomes harder for higher dimensional problems, and recently efficient algorithms for approximate nearest neighbor have been proposed (e.g., see Indyk, Motwani, Raghavan, Vempala [11]).

Besides these things, computational geometry can be used to find good initial solutions. Randomized algorithm for the case of minimizing the sum of squared distances in Inaba, Katoh, Imai [10] may be used. Also, a simple clustering technique of dividing points by only hyperplanes perpendicular to some axis (e.g., see Wan, Wong, Prusinkiewicz [17]) may be effective in the high dimensional case. In fact, when this clustering technique is considered from the view point of decision functions, this has connection with a decision tree algorithm T2 (Auer, Holte, Maass [2]). Since initial solutions strongly affects local minima obtained by the k-means algorithm, this approach needs further investigations.

5.2 Finding a most related cluster

So far, we have discussed grouping object in the target space into k similar groups. For example, find a group of texts, among huge collections of texts, having the most similarity in some measure to a given text. In such problems, finding one group of texts which are closely related to one another is important.

As such a type of geometric clustering with respect to the measure used above, the problem of finding a subset of k points, for given k, that minimizes the all-pair sum of squared Euclidean distances among them is considered in Aggarwal, Imai, Katoh, Suri [1]. By changing k appropriately, we can find a most closely related cluster among the target objects. Applications of such results should also be investigated.

6 Concluding remarks

In this paper we have presented a theoretical framework for making full use of geometry in the space of weighted term vectors in the vector space model. Checking the efficiency of this framework is definitely necessary for real applications, and this will be done in a subsequent work. We do hope that our theoretically founded framework work well in practice.

Acknowledgment

The work of the authors was supported in part by the Grant-in-Aid on Priority Areas, 'Advanced Databases,' of Ministry of Education, Science, Sports and Culture of Japan.

References

[1] A. Aggarwal, H. Imai, N. Katoh and S. Suri: Finding k Points with Minimum Diameter and Related Problems. *Journal of Algorithms*, Vol.12 (1991), pp.38–56.

[2] P. Auer, R. C. Holte and W. Maass: Theory and Applications of Agnostic PAC-learning with Small Decision Trees. *Proceedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann, 1995, pp.21–29.

[3] M. Berry and S. Dumais: Using Linear Algebra for Intelligent Information Retrieval *SIAM Review*, Vol.37, No.4 (1995), pp.573–595.

[4] M. Charikar, C. Chekuri, T. Feder and R. Motwani: Incremental Clustering and Dynamic Information Retrieval. *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC'97)*, 1997, pp.626–634.

[5] A. Gersho and R. M. Gray: *Vector Quantization and Signal Compression*. Kluwer Academic, 1992.

[6] S. Hasegawa, H. Imai, M. Inaba, N. Katoh and J. Nakano: Efficient algorithms for variance-based k-clustering. *Proceedings of the First Pacific Conference on Computer Graphics and Applications*, World Scientific, 1993, pp.75–89.

[7] H. Imai and M. Inaba: Geometric Clustering with Applications. *Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM)*, Vol.76, Suppl. (1996), pp.183–186.

[8] 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.

[9] M. Inaba, H. Imai, M. Nakade, and T. Sekiguchi: Application of an Effective Geometric Clustering Method to the Color Quantization Problem. *Proceedings of the 13th ACM Symposium on Computational Geometry*, 1997, pp.477–478.

[10] 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.

[11] P. Indyk, R. Motwani, P. Raghavan and S. Vempala: Locality-Preserving hashing in Multidimensional Spaces. *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC'97)*, 1997, pp.618–625.

[12] M. Kitagawa, Y. Mizuuchi, K. Tajima and K. Tanaka: Clustering and Cut Detecting for Information Organization of Mailing Lists (in Japanese). *Proceedings of the 8th Workshop of Data Engineering (DEWS'97)*, 1997, pp.221–226.

[13] T. Ono, Y. Kyoda, T. Masada, K. Hayase, T. Shibuya, M. Nakade, M. Inaba, H. Imai, K. Imai and D. Avis: A Package for Triangulations. *Proceedings of the 12th Annual ACM Symposium on Computational Geometry*, Video descriptions, 1996, pp.V17-V18.

[14] 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.

[15] 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.

[16] S. Z. Selim and M. A. Ismail: K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.PAMI-6 (1984), pp.81–87.

[17] S. J. Wan, S. K. M. Wong and P. Prusinkiewicz: An algorithm for multidimensional data clustering. *ACM Transactions on Mathematical Software*, Vol.14, No.2 (1988), pp.153–162.