Spectral Clustering for a Large Data Set by Reducing the Similarity Matrix Size

Hiroyuki Shinnou, Minoru Sasaki

Ibaraki University,
4-12-1 Nakanarusawa, Hitachi, Ibaraki, Japan 316-8511
{shinnou, msasaki}@mx.ibaraki.ac.jp

Abstract

Spectral clustering is a powerful clustering method for document data set. However, spectral clustering needs to solve an eigenvalue problem of the matrix converted from the similarity matrix corresponding to the data set. Therefore, it is not practical to use spectral clustering for a large data set. To overcome this problem, we propose the method to reduce the similarity matrix size. First, using k-means, we obtain a clustering result for the given data set. From each cluster, we pick up some data, which are near to the central of the cluster. We take these data as one data. We call these data set as "committee." Data except for committees remain one data. For these data, we construct the similarity matrix. Definitely, the size of this similarity matrix is reduced so much that we can perform spectral clustering using the reduced similarity matrix

1. Introduction

In this paper, we proposed the method to reduce the similarity matrix size in order to use spectral clustering for a large data set.

Document clustering is the task of dividing a document's data set into groups based on document similarity. This is the basic intelligent procedure, and is important in text mining systems (Michael W. Berry, 2003). As the specific application, relevant feedback in IR, where retrieved documents are clustered, is actively researched (Hearst and Pedersen, 1996) (Kummamuru et al., 2004) etc.

Spectral clustering is a powerful clustering method using partitioning of a graph (Ding et al., 2001). It uses an object function to find the optimal partition of the graph. The optimal solution of the object function corresponds to a solution of an eigenvalue problem. Using this solution, spectral clustering generates the final clustering result for the given data set.

Spectral clustering is actually powerful, but needs to solve the eigenvalue problem of the Laplacian matrix converted from the similarity matrix corresponding to the given data set. Therefore, we cannot use spectral clustering for a large document data set (Dhillon et al., 2005)(Liu et al., 2007). In this paper, we propose the method to reduce the similarity matrix size.

First, using k-means, we obtain a clustering result for the given data set. From each cluster, we pick up some data, which are near to the central of the cluster. We take these data as one data. We call these data set as "committee" according to the paper (Pantel and Lin, 2002). Data except for committees remain one data. For these data, we construct the similarity matrix. Definitely, the size of this similarity matrix is reduced so much that we can perform spectral clustering using the reduced similarity matrix

Note that our method needs a reasonably accurate clustering result to reduce the similarity matrix size. Therefore, the final clustering result generated by spectral clustering must be improved from the initial clustering result. That is, our method is regarded as the method to improve the given clustering result.

In the experiment, we used seven document data sets to evaluate our method. We compared our method with k-means and spectral clustering, Mcut. The experiment showed our method improves the clustering result generated by k-means.

In future we will investigate the proper reduction degree, and improve the similarity definition.

2. Spectral clustering

In spectral clustering the data set is represented as a graph. Each data point is represented as a vertex in the graph. If the similarity between data x and y is non-zero, the edge between x and y is drawn and the similarity is used as the weight of the edge. From this graph, clustering can be seen to correspond to the segmentation of the graph into a number of subgraphs by cutting the edges. The preferable cutting is such that the sum of the weights of the edges in the subgraph is large and the sum of weights of the cut edges is small. To find the ideal cut, the object function is used. The spectral clustering method finds the desirable cut by using the fact that an optimum solution of the object function corresponds to the solution of an eigenvalue problem. Different object functions are proposed. In this paper, we use the object function of Mcut (Ding et al., 2001).

First, we define the similarity $cut(A, B)$ between the subgraph $A$and Bas follows:

$$cut(A, B) = W(A, B).$$

The function $W(A, B)$ is the sum of the weights of the edges between A and B. We define $W(A)$ as $W(A, A)$.

The object function of Mcut is the following:

$$Mcut = \frac{cut(A, B)}{W(A)} + \frac{cut(A, B)}{W(B)} \qquad (1)$$

The clustering task is to find A and B to minimize the above equation.

Note that the spectral clustering method divides the data set into two groups. If the number of clusters is larger than two, the above procedure is iterated recursively.

The minimization problem of Eq. 1 is equivalent to the problem of finding the n dimensional discrete vector y to minimize the following equation:

$$J_m = \frac{y^T(D-W)y}{y^T W y} \tag{2}$$

where W is the similarity matrix of data, $D = diag(We)$ and $e = (1, 1, \cdots, 1)^T$. Each element in the vector y is a or $-b$, where

$$a = \sqrt{\frac{d_B}{d_A d}},$$
$$b = \sqrt{\frac{d_A}{d_B d}},$$
$$d_X = \sum_{i \in X} d_i$$

and $d = d_A + d_B$. If the i-th element of the vector y is a (or $-b$), the i-th data element belongs to the cluster A (or B). We can solve Eq. 2 by converting the discrete vector y to the continuous vector y. Finally, we can obtain an approximate solution to Eq. 2 by solving the following eigenvalue problem:

$$(I - D^{-1/2} W D^{-1/2})z = \lambda z \tag{3}$$

We obtain the eigenvector z, that is, the Fielder vector, corresponding to the second minimum eigenvalue by solving the eigenvalue problem represented by Eq. 3. We can obtain the solution y to Eq. 2 from $y = D^{-1/2}z$. By the sign of the i-th value of y, we can judge whether the i-th data element belongs to cluster A or B.

Note that Eq. 1 is the object function when the number of clusters is two. The object function used in NMF is the following general object function for k clusters.

$$Mcut_K = \frac{cut(G_1, \bar{G_1})}{W(G_1)} + \frac{cut(G_2, \bar{G_2})}{W(G_2)} + \cdots + \frac{cut(G_k, \bar{G_k})}{W(G_k)} \tag{4}$$

where is $\bar{G_i}$ the complement of G_i. The smaller $Mcut_k$ is, the better it is.

3. Reduction of the similarity matrix size

Suppose X is divided into cluster A and B by a spectral clustering method. Note that the value of Eq. 1 for this segmentation is minimum.

Now, we translate a subset $A' \subset A$ into one data a'.

$$A' = \{a_1, a_2, \cdots, a_m\} \subset A$$

First, we define the similarity between a' and data $c \in {A'}^C$[1] as follows:

$$sim(a', c) = \sum_{i=1}^{m} sim(a_i, c) \tag{5}$$

Next, we define the similarity between a' and a' as follows:

$$sim(a', a') = \sum_{i=1}^{m} \sum_{j=1}^{m} sim(a_i, a_j) \tag{6}$$

Using Eq. 5 and Eq. 6, we can translate the set A' into one data a'. As the result, we obtain the new data set $X' = a' \cup {A'}^C$. Let W' the similarity matrix of X'. The size of W' is $(N - m + 1) \times (N - m + 1)$.

It is clear that the above similarity definition is exact. It can be confirmed by applying the clustering result of X to X'. That is, if data $x \in X - A'$ belongs to the cluster C according to the clustering result of X, we set the cluster of data x as C in the clustering result of X', And we set the cluster of data a' as A. Computing the value of Eq. 1 for this segmentation using W', it is same to the value of Eq. 1 for the clustering result of X.

The value of Eq. 1 for the clustering result of X is minimum. Therefore, spectral clustering for W' generates the same clustering result.

3.1. Construction of Committee by k-means

We need the collect clustering result of data set X to reduce the similarity matrix size by the above method. However, it is impossible to obtain the collect clustering result.

Here, we take notice that A' is a subset of A, and need not to be A. What we requires for A' is that A' does not include wrong data. We call this set A' as "committee" according to (Pantel and Lin, 2002).

In this paper, we perform k-means first. From each cluster, we pick up some data, which are near to the central of the cluster. These picked up data is the committee. Specifically, committee is made by picking up 80% data of the cluster in this paper. However, this value is changeable.

3.2. Improved similarity definition using central point

After making committees, we construct the small similarity matrix using Eq. 5 and Eq. 6. However, Eq. 5 and Eq. 6 cannot actually be used because Eq. 5 and Eq. 6 assume that A' does not include wrong data, but this assumption is wrong.

Therefore, instead of Eq. 5 and Eq. 6, we must use the more practical similarity definition. Thus, we define them as follows:

$$sim(a', c) \approx m \cdot sim(\bar{a}, c) \tag{7}$$

where $\bar{a}$ means the central point of A'.

$$sim(a', a') \approx m/2 \tag{8}$$

4. Experiment

To confirm the effectiveness of our method, we performs our method for seven document data sets (tr12, tr31, mm, la12, sports, ohscal, cacmcisi) provided in CLUTO site [2]. Table 1 shows information of these data set: the number of data, the number of dimension, the number of non-zero data and the number of clusters.

[1] ${A'}^C$ means the complement set of A'.

[2] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

Table 1: Data sets

data	# of data	# of dimension	# of clusters
tr12	313	5804	8
tr31	927	10128	7
mm	2521	126373	2
cacmcisi	4663	41681	2
la12	6279	31472	6
sports	8580	126373	7
ohscal	11162	11465	10

First we perform k-means for the data set. Using distance between the central point of the cluster and data, we make a committee from each cluster. Next, using Eq. 7 and Eq. 8, we make the reduced similarity matrix. Finally, we perform spectral clustering using the reduced similarity matrix to obtain the final clustering result.

We evaluate our method by comparing it with k-means and Mcut. In this paper,we used entropy and purity for clustering evaluation. It is natural that our method is worse than Mcut. However, it is remarkable that our method is better than k-means. The size of four data set (cacmcisi、la12、sports、ohscal) are big, so we cannot perform the spectral clustering, Mcut for these data sets.

This experiment showed that our method can improve the clustering result generated by k-means.

Table 2: Result (Entropy)

data	Mcut	k-means	our method
tr12	**0.3800**	0.4366	0.3840
tr31	**0.2946**	0.3419	0.3414
mm	**0.9715**	0.9847	0.9837
cacmcisi	—	0.6768	0.6744
la12	—	0.4523	0.4575
sports	—	0.3142	0.3049
ohscal	—	0.5678	0.5722

Table 3: Result (Purity)

data	Mcut	k-means	our method
tr12	**0.7061**	0.6550	0.6741
tr31	**0.8037**	0.7605	0.7702
mm	**0.5799**	0.5601	0.5688
cacmcisi	—	0.6869	0.6869
la12	—	0.7015	0.7019
sports	—	0.7735	0.7871
ohscal	—	0.5434	0.5440

5. Discussions

In the paper, our method sets the reduction degree to be 80%. However, this value is changeable.

For the data set ‘mm’, we varied the reduction degree from 100% to 0%, in steps of 10%. The result is shown in Fig. 1 and Fig. 2.

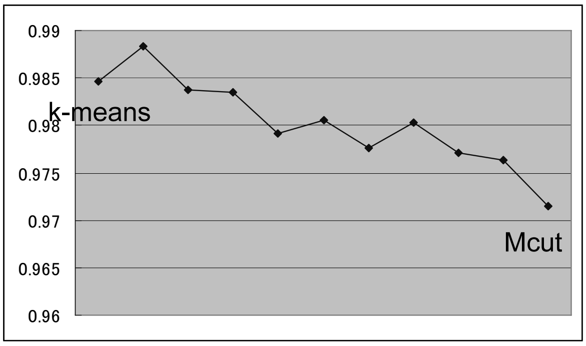

Figure 1: Entropy for the reduction degree

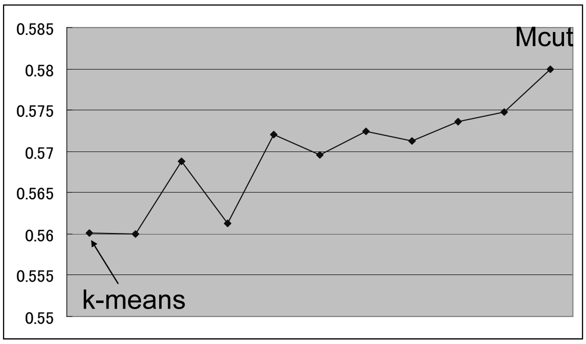

Figure 2: Purity for the reduction degree

Theoretically, the larger the reduction degree is, the nearer the result is to result of k-means. And the smaller the reduction degree is, the nearer the result is to result of Mcut. Moreover, the change must be monotonous.

Curves of Fig. 1 and Fig. 2 are not exactly monotonous. However they are likely to be monotonic. We can consider two reasons that curves are not exactly monotonous. The one is that our committee includes error, and another is that the similarity definition is not proper.

In future we will investigate the proper reduction degree, and improve the similarity definition.

Lastly, we note that viewing our method as method to improve the given clustering result, “first validation” of (Dhillon et al., 2002) and “Link based refinement” of (Ding et al., 2001) are informative to refine our method.

6. Conclusion

In this paper, we proposed the method to reduce the similarity matrix size in order to use spectral clustering for a large data set.

Our method needs a reasonably correct clustering result to reduce the similarity matrix size. Thus, our method is regarded as the method to improve the given clustering result. The experiment using seven data sets showed that our method improves the clustering result generated by k-means.

In future we will investigate the proper reduction degree, and improve the similarity definition.

Acknowledgements

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas, “ Japanese Corpus ”, 19011001, 2007.

7. References

Inderjit S. Dhillon, Yuqiang Guan, and J. Kogan. 2002. Iterative Clustering of High Dimentional Text Data Augmented by Local Search. In *The 2002 IEEE International Conference on Data Mining*, pages 131–138.

Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. 2005. A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts. In *The University of Texas at Austin, Department of Computer Sciences. Technical Report TR-04-25*.

Chris Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst Simon. 2001. Spectral Min-max Cut for Graph Partitioning and Data Clustering. In *Lawrence Berkeley National Lab. Tech. report 47848*.

Marti A. Hearst and Jan O. Pedersen. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of SIGIR-96*, pages 76–84.

Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. 2004. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. In *Proceedings of WWW-04*, pages 658–665.

Tie-Yan Liu, Huai-Yuan Yang, Xin Zheng, Tao Qin, and Wei-Ying Ma. 2007. Fast Large-Scale Spectral Clustering by Sequential Shrinkage Optimization. In *ECIR*, pages 319–330.

Michael W. Berry, editor. 2003. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer.

P. Pantel and D. Lin. 2002. Document clustering with committees. In *Proceedings of SIGIR-02*.