

Spectral Analysis of Text Collection for Similarity-based Clustering

Wenyuan Li, Wee-Keong Ng and Ee-Peng Lim
 Center for Advanced Information Systems
 Nanyang Technological University
 Nanyang Avenue, Singapore 639798, Singapore
 liwy@pmail.ntu.edu.sg {awkng, aseplim}@ntu.edu.sg

Clustering of text collections is generally difficult due to its high dimensionality, heterogeneity, and large size. These characteristics compound the problem of determining the appropriate similarity space for clustering algorithms. In this paper, we propose to use the spectral analysis of the similarity space of a text collection to predict clustering behavior before actual clustering is performed. Spectral analysis is a technique that has been adopted across different domains to analyze the key encoding information of a system. Using spectral analysis for prediction is useful in first determining the quality of the similarity space and discovering any possible problems the selected feature set may present.

The similarity matrix $\mathbf{S} = (s_{ij})_{n \times n}$ has an associate graph $\mathcal{G}(\mathbf{S})$. Due to the different scale of the spectrums of \mathbf{S} 's, it is difficult to analyze and compare them. Thus, we transform \mathbf{S} to the *weighted Laplacian* \mathbf{L} , which has “normalized” eigenvalues with the same scale ($0 \leq \text{eig}(\mathbf{L}) \leq 2$) [1]. For the convenience of computation, we use \mathbf{L} , the variant of the weighted Laplacian, instead of \mathbf{L} , which is defined to be: $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{S} - \mathbf{I})\mathbf{D}^{-1/2}$, where $\mathbf{D} = \text{diag}(d_i)$ ($d_i = \sum_j s_{ij}$) denotes the diagonal matrix. We have $\text{eig}(\mathbf{L}) = \{1 - \lambda | \lambda \in \text{eig}(\mathbf{S})\}$. There are three observations of the $\mathcal{G}(\mathbf{S})$ spectrum for the clustering behavior of \mathbf{S} , which can be accounted for by spectral graph theory. Suppose $1 = \lambda_1 \geq \dots \geq \lambda_n$ be eigenvalues of $\mathcal{G}(\mathbf{S})$,

1. If λ_2 is higher, there exists a better bipartition for \mathbf{S} .
2. For the sequence $\alpha_i = \frac{\lambda_i}{\lambda_2}$ ($i \geq 2$), $\exists k \geq 2$, it has $\alpha_i \rightarrow 1$ and $\alpha_i - \alpha_{i+1} > \delta$ ($0 < \delta < 1$), then k indicates the cluster number of the text collection.

3. If the curve of the spectrum is closer to the x-axis, the clustering behavior of \mathbf{S} is worse. (It is measured by $\sum \lambda_i^2$)

Two text collections are used in experiments: **(TC1)** the classic text collection – CACM, CISI, CRAN and MEDLINE; **(TC2)** web pages from categories in Yahoo! Directory [3]. The first experiment investigates the impact of feature sets on text clustering in **TC1** by Observation 3. We use two different feature sets: (T) terms; (TP) the combination of terms and phrases. The similarity measure is cosine. Three clustering algorithms of “CLUTO” toolkit [2]

	bisecting k -means		graph-based		hierarchical	
	T	TP	T	TP	T	TP
F -measure	0.822	0.817	0.795	0.846	0.563	0.753
Purity	0.853	0.892	0.845	0.869	0.455	0.679
Entropy	0.299	0.262	0.285	0.259	0.654	0.419

Table 1. Results of Clustering Algorithms

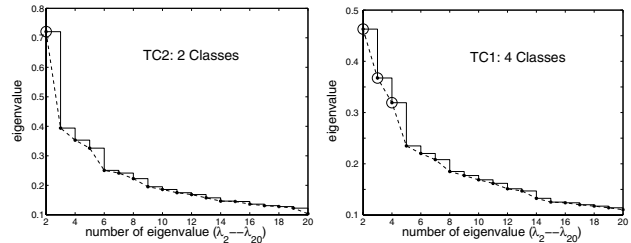


Figure 1. Cluster Number on Text Collections

and three external measures are used. The \mathbf{S} 's $\sum \lambda_i^2$ of T and TP feature sets are 2.39 and 2.66, respectively. It shows the same result as Table 1: among feature sets, TP has higher clustering quality than T.

In the second experiment, we estimate the number of clusters by Observation 2. Two selected text collections are used: two classes of **TC2**; four classes of **TC1**. In Figure 1, eigenvalue curve is drawn from λ_2 , as λ_1 is always 1. Circled eigenvalues (plus λ_1) indicate the number of clusters.

References

- [1] F. R. K. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, Providence, Rhode Island, 1997.
- [2] G. Karypis. CLUTO – a clustering toolkit. Technical Report #02-017, University of Minnesota, 2002.
- [3] M. P. Sinka and D. W. Corne. A large benchmark dataset for web document clustering. In *Proceedings of the Second International Conference on Hybrid Intelligent Systems*, Chile, December 2002.