FAÇADE: A Fast and Effective Approach to the Discovery of Dense Clusters in Noisy Spatial Data

Yu Qian Gang Zhang Kang Zhang
Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75083-0688, USA
{yxq012100, gxz014000, kzhang}@utdallas.edu

ABSTRACT

FAÇADE (Fast and Automatic Clustering Approach to Data Engineering) is a spatial clustering tool that can discover clusters of different sizes, shapes, and densities in noisy spatial data. Compared with the existing clustering methods, FAÇADE has several advantages: first, it separates true data and noise more effectively. Second, most steps of FAÇADE are automatic. Third, it requires only *O(nlogn)* time. 2D and 3D visualizations are used in FAÇADE to assist parameter selection and result evaluation. More information on FAÇADE is available at: http://viscomp.utdallas.edu/FACADE.

1. INTRODUCTION

Many clustering methods have been proposed for finding dense clusters in spatial databases. The solutions to three challenging issues, however, remain unsatisfactory: recognizing and handling noise, minimizing the number of input parameters, and scalability. Few existing clustering algorithms can meet all these requirements at the same time, i.e., to efficiently and automatically discover clusters of all shapes, sizes, and densities as well as the noise. Most clustering methods contain inherent limitations: partitioning methods are usually efficient and insensitive to noise, but lack the ability of finding arbitrarily shaped clusters. Hierarchical clustering methods are usually slower than partitioning ones and not very robust to noise although they have advantage at finding natural clusters. Density-based methods can discover clusters of arbitrary shapes and filter out noise, yet usually rely on a set of user-input parameters and inefficient for high-dimensional data.

This paper presents a novel hierarchical clustering approach, called FAÇADE, which can handle huge amount of spatial data with heavy noise and find natural clusters correctly. The whole process of FAÇADE is illustrated in Fig. 1 and composed of six steps: 1) Modeling the spatial data set with *k*-mutual neighborhood graph; 2) Applying *k*-core algorithm [1] to remove noise and outliers according to the structural information; 3) Applying GraphZip [3] to compress the data set produced by the *k*-core algorithm; 4) Constructing a *1*-mutual neighborhood graph on the compressed data set, in which each sub-graph (connected component) is regarded as a group; 5) Mapping back the grouping information of the compressed data to the original data; 6) Merging the groups hierarchically according to the connections between two groups. All steps can be completed in *O(nlogn)* time.

SIGMOD 2004, June 13–18, 2004, Paris, France.

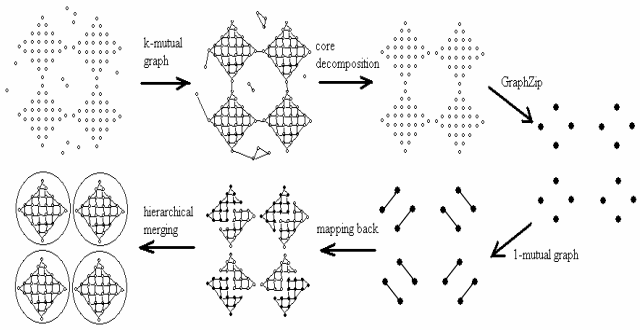


Fig. 1. The six steps of FAÇADE

2. Main Clustering Phases

There are two main phases in FAÇADE: Step 2 uses *k*-core algorithm [1] to recognize noise and Step 3 applies GraphZip [3] to compress the data set to improve the efficiency of hierarchical combination. We assume a corresponding *k*-mutual graph has been constructed for the given data set.

2.1 Noise Removal

The *k*-core algorithm is used in FAÇADE to separate data points into different layers according to connections between graph vertices. The notion of core [1] is: let $G = (V, E)$ be a graph. V is the set of vertices and E is the set of edges. A sub-graph $H_k = (W, E \mid W)$ induced by the set W is a *k*-core or a core of order k iff $\forall v$ in W: *degree (v)* $\geq k$ and H_k is the maximum sub-graph with this property. The core of maximum order is also called the *main core*.

The algorithm for determining the core hierarchy is simple: from a given graph $G=(V, E)$, recursively delete all vertices of degrees less than k and the lines incident with them, the remaining graph is the *k*-core. The *k*-core algorithm costs only *O(m)* time, where m is the number of graph edges. For a *k*-mutual graph with n vertices and m edges, we have $m \leq kn/2$ if n is the number of vertices, so applying *k*-core algorithm to a *k*-mutual graph requires only linear time. FAÇADE uses core decomposition as an effective noise removing method for spatial data sets.

Algorithm *GraphZip (Data Set D)*
begin
Construct 1-nearest neighbor graph G for D;
Create an empty data set D';
For each connected-component C of G:
Generate a point p that is located at the center of C;
Add p to D' and update the mapping file;
if $O(|D'|) \leq O(\sqrt{n})$ **return** *D';***else** *GraphZip(D');*
end

Fig. 2. The *GraphZip* Algorithm

2.2 Spatial Data Compression

FAÇADE compresses the input data using a method called GraphZip [3] with four characteristics: a) preserving original spatial patterns with smaller data points, b) its result is deterministic and reproducible for different executions or data input orders, c) the compression process is data-driven and parameter-free, and d) it requires only *O(nlogn)* time for *n* data points. The process of GraphZip is described in Fig. 2.

3. Visualization of Experimental Results

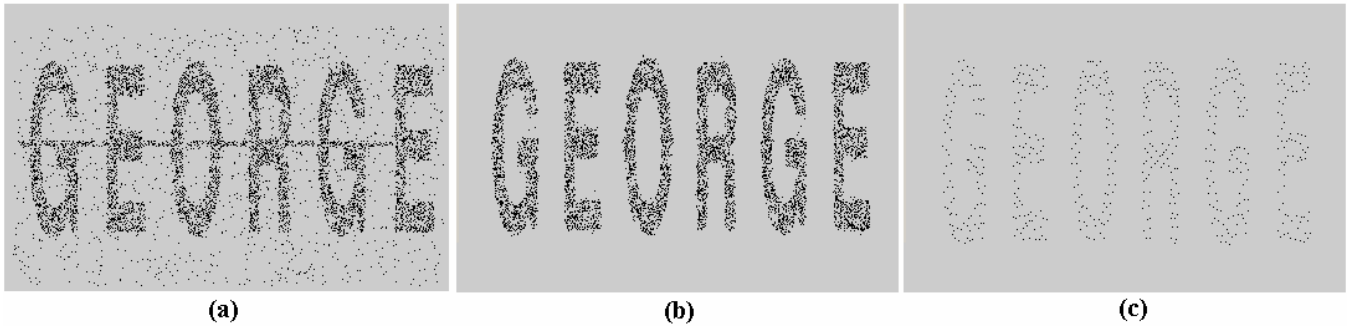
FAÇADE provides 2D and 3D visualization for each step of the clustering process. This section provides visualizations for one benchmark dataset used in CHAMELEON [2]. More examples can be found at our website. Fig. 3 is the 2D visualization of the results of noise removal and data compression. Fig. 4 shows 3D visualization of noise removal and final clustering results. In Fig. 4 (a), the data points are separated into different layers/cores by *k*-core algorithm while different layers are represented with different gray levels. True data are clearly separated from noise, as they belong to the cores at higher layers while noise belongs to those at lower layers. Fig. 4 (b) illustrates the clustering results after applying FAÇADE where different clusters are represented with different gray levels and in different layers. Fig. 5 provides 2D visualization of final clustering results with different clusters in different gray levels. FAÇADE demo with user customization is publicly available at http://viscomp.utdallas.edu/FACADE.

References:

[1] Seidman, S. B. Network structure and minimum degree. *Social Networks*, 5, 1983, pp. 269-287.

[2] Karypis, G., Han, E., and Kumar, V. CHAMELEON, A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, Vol.32, 1999, pp. 68-75.

[3] Qian, Y. and Zhang, K., GraphZip: a fast and automatic compression method for spatial data clustering. In *Proc. of the 2004 ACM Symposium on Applied Computing (SAC'04)*, pp. 571-575.



(a) **(b)** **(c)**

Fig. 3 (a) The original data sets **(b)** The resulting data sets after applying core decomposition **(c)** after applying GraphZip.

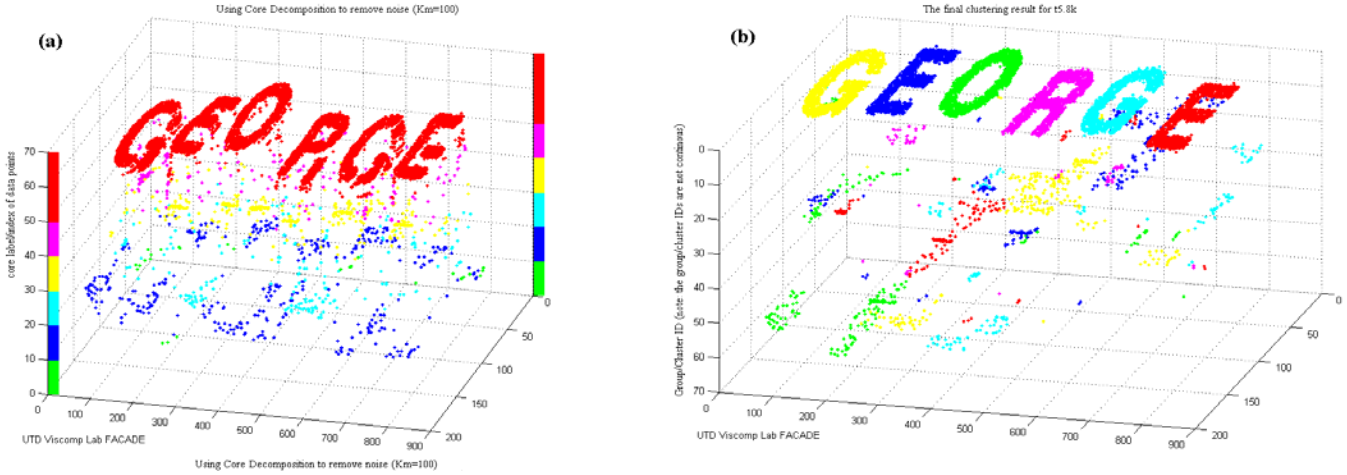


Fig. 4 (a) Data points separated into layers and noise located at the bottom; **(b)** final clustering results.

(a) **(b)** **(c)**

Fig. 5. The clustering results of **(a)** true data only; **(b)** adding back noise; **(c)** the whole data set without noise removal.