Computers, Environment and Urban Systems 35 (2011) 320–332

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

An adaptive spatial clustering algorithm based on delaunay triangulation

Min Deng [a,*], Qiliang Liu [a], Tao Cheng [b], Yan Shi [a]

[a] Department of Surveying and Geo-informatics, Central South University, Changsha, China
[b] Department of Civil, Environmental and Geomatic Engineering, University College London, Gower St., WC1E 6BT, London, United Kingdom

ARTICLE INFO

Article history:
Received 26 August 2010
Received in revised form 11 February 2011
Accepted 14 February 2011
Available online 12 March 2011

Keywords:
Spatial clustering
Adaptive
Delaunay triangulation
Spatial data mining

ABSTRACT

In this paper, an adaptive spatial clustering algorithm based on Delaunay triangulation (ASCDT for short) is proposed. The ASCDT algorithm employs both statistical features of the edges of Delaunay triangulation and a novel spatial proximity definition based upon Delaunay triangulation to detect spatial clusters. Normally, this algorithm can automatically discover clusters of complicated shapes, and non-homogeneous densities in a spatial database, without the need to set parameters or prior knowledge. The user can also modify the parameter to fit with special applications. In addition, the algorithm is robust to noise. Experiments on both simulated and real-world spatial databases (i.e. an earthquake dataset in China) are utilized to demonstrate the effectiveness and advantages of the ASCDT algorithm.

1. Introduction

More attention has been paid to spatial data mining in the last decade (Miller & Han, 2009), in order to discover valuable information or knowledge from spatial databases. Spatial clustering has played an important role in the process of spatial data mining. It aims to classify a spatial database into several clusters without any prior knowledge (e.g., probability distribution and the number of clusters). In general, spatial points in the same cluster are similar to one another and dissimilar to the points in different clusters. Spatial clustering can be generally employed to segment geographic regions with different characteristics and extract meaningful spatial patterns. Spatial clustering also can help to generalize the aggregate point features and find the optimal positions of the public facilities.

There are mainly two types of spatial clustering methods. One considers only the spatial attributes of spatial points (i.e. geometric coordinates), and the other considers both spatial and non-spatial attributes of spatial points. Currently, spatial clustering has a wide range of applications, such as crime hot-spot analysis (Estivill-Castro & Lee, 2002a), land use detection (Ester, Kriegel, Sander, & Xu, 1996; Sander, Ester, Kriegel, & Xu, 1998), earthquake analysis (Pei, Zhu, Zhou, Li, & Qin, 2009; Xu, Ester, Kriegel, & Sander, 1998), vehicle crashes (Yamada & Thill, 2007) and agricultural environments (Schoier & Borruso, 2004, 2007). In this paper, the spatial clustering of geo-referenced 2-D point data is emphasized, where only the spatial attribute of point data is considered. As a matter of fact, this is the most fundamental clustering task for exploratory spatial analysis, which plays a critical role in further investigations (Estivill-Castro & Lee, 2002a; Ng & Han, 1994).

A number of clustering algorithms have been developed for spatial databases. They can be classified roughly into partitioning algorithms (MacQueen, 1967; Kaufman & Rousseeuw, 2005; Ng & Han, 1994), hierarchical algorithms (Estivill-Castro & Lee, 2002a; Guha, Rastogi, & Shim, 1998; Karypis, Han, & Kumar, 1999; Xu & Wunsh, 2009; Zhang, Ramakrishnan, & Livny, 1996), density-based algorithms (Ankerst, Breunig, Kriegel, & Sander, 1999; Ester et al., 1996; Hinneburg & Keim, 1998; Nosovskiy, Liu, & Sourina, 2008), graph-based algorithms (Estivill-Castro & Lee, 2002b; Liu, Nosovskiy, & Sourina, 2008; Zahn, 1971; Zhong, Miao, & Wang, 2010), grid-based algorithms (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998; Sheikholeslami, Chatterjee, & Zhang, 1998; Wang, Yang, & Muntz, 1997), model-based algorithms (McLachlan & Krishnan, 1996; Xu et al., 1998) and various combinational algorithms (Lin & Chen, 2005; Tsai & Yen, 2007). All the above-mentioned algorithms have proved able to handle some specific applications.

Currently, the applications on complicated spatial databases bring new demands for clustering algorithms. A complicated spatial database may contain clusters adjacent to each other, with arbitrary geometrical shapes and/or different densities. In addition, a large amount of noise possibly exists. Here noise refers to spatial points which do not belong to any clusters, and can be classified into two kinds, i.e., background noise and chains. The background noise is some isolated points which often distribute randomly in the database; the chains are formed of noises which connect two separated clusters. New clustering algorithms should be developed in order to

* Corresponding author. Address: Dept. of Surveying and Geo-informatics, Central South University, Changsha, Hunan, China.
E-mail address: dengmin028@yahoo.com (M. Deng).

0198-9715/$ - see front matter Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2011.02.003

be adaptive to different cases. In the light of this, an adaptive spatial clustering algorithm based on Delaunay triangulation (ASCDT for short) is developed in this paper. The development of this algorithm will be expected to enhance the applications in the fields as earthquake analysis (Pei et al., 2009; Xu et al., 1998), cartographic generalization (Li, 2007), geographic customer segmentation (Miller & Han, 2009) and other exploratory data analysis in geographic information systems (Estivill-Castro & Lee, 2002a).

The rest of the paper is organized as follows. In Section 2, related work and the strategy on adaptive spatial clustering are described in detail. The ASCDT algorithm is performed fully in Section 3. Experiments on both simulated and real databases are shown in Section 4. Section 5 summarizes the main findings and highlights the directions for future research.

2. Related work and our strategy on discovering spatial clusters

2.1. Related work

As mentioned above, spatial clustering algorithms are designed to discover the clusters of certain structures. In this section, the performance of some classical clustering algorithms will be examined in detail, according to the requirements of the spatial clustering algorithms. This examination will be made based upon several comparative experiments and further used to substantiate the analysis of the ASCDT algorithm proposed in this paper (to be discussed in Section 4).

There are several challenges in spatial clustering. The first is to discover clusters of arbitrary shapes. Partitioning methods, such as K-means (MacQueen, 1967), PAM (Kaufman & Rousseeuw, 2005), and CLARANS (Ng & Han, 1994), are suitable only for detecting clusters of spherical shape and similar sizes. Traditional hierarchical algorithms, such as single-link and complete-link, have difficulty in discovering clusters with different shapes. Improved hierarchical algorithms, such as BIRCH (Zhang et al., 1996) and CURE (Guha et al., 1998), can deal with clusters of more complex shapes, but they are unable to discover very complicated clusters. Density-based methods, such as DBSCAN (Ester et al., 1996), DENCLUE (Hinneburg & Keim, 1998), and OPTICS (Ankerst et al., 1999) can discover arbitrary-shaped clusters, but they perform well only in cases in which the clusters are well separated and have similar densities.

The second challenge is to handle spatial data with mixed density. There are two cases. One case is that the density of one cluster is even but different from the other cluster, as shown in Fig. 1a; the other case is that the density of the cluster is not even, as shown in Fig. 1b. Density-based algorithms, such as DBSCAN, DENCLUE, and OPTICS, are all influenced by the density problem. Several improved graph-based algorithms, such as AMOEBA (Estivill-Castro & Lee, 2002a), CHAMELEON (Karypis et al., 1999), and AUTOCLUST (Estivill-Castro & Lee, 2002b can be utilized to discover clusters with different densities in a spatial database (the case in Fig. 1a). However they cannot detect clusters with uneven internal densities (the case in Fig. 1b).

The third challenge is that spatial clustering algorithms should be robust to noise and the "touching problems" (Zhong et al., 2010). When the noise is isolated from the spatial points, the density-based and graph-based algorithms, such as DBSCAN and AUTOCLUST, can deal with the noise well. However if the noise from one or more chains connects two clusters, which is also called the "chaining problem" (Fig. 2a), this has been proven to be a very challenging problem. Touching problems can be classified into two types. One is the famous "neck problem" (Fig. 2b) (Zahn, 1971); the other is the "adjacent problem" (Fig. 2c) (Estivill-Castro & Lee, 2002b), in which spare clusters are adjacent to high-density clusters. The MST algorithm (Zahn, 1971) can resolve the single chain problem and the touching problem. However, this algorithm on the one hand relies heavily on the prior information in the structure of the database; and on the other hand it cannot deal with multi-chains. The AUTOCLUST algorithm can be used to remove single or multi-chains, but it does not always work well. The AUTOCLUST algorithm still suffers from the neck problem. The 2-MSTClus algorithm (Zhong et al., 2010) can handle the chaining effect and the touching problem, but it is seriously affected by noise.

Lastly, spatial clustering algorithms should not rely on too much prior knowledge to obtain satisfactory results, because prior knowledge of a spatial database is usually unavailable for many applications. Partitioning algorithms, such as K-means, PAM and CLARANS, require the users to set the number of clusters. Hierarchical algorithms are involved in the setting of the merge or split conditions. Among the density-based algorithms, which include DBSCAN, DENCLUE, and OPTICS, it is indeed hard to determine the threshold of the density. Model-based algorithms, such as EM (McLachlan & Krishnan, 1996) and DBCLASD (Xu et al., 1998) require users to make an assumption about the probability distribution of the data. Graph-based algorithms, such as TRICLUST (Liu et al., 2008) require domain knowledge to set the weights of the different statistical features.

The analysis of the classical spatial clustering algorithms mentioned above is summarized in Table 1. It shows that the existing clustering algorithms have difficulty in satisfying various challenges of spatial clustering, and thus may encounter difficulties when dealing with practical applications. As a result, a new strategy needs to be developed.

2.2. A new strategy on discovering spatial clusters in spatial database

It is well known that geographic phenomena are the combination of global effect (large scale effect) and local effect (small scale effect) (Haining, 2003). The global effect influences the geographic phenomena first, and later the local effect (Bailey & Gatrell, 1995). Spatial clusters can be viewed as certain geographic phenomena that are merged into a spatial database. To detect spatial clusters more accurately, it is reasonable to remove the global effect first, and then the local effect. Therefore, a two-level strategy is employed to detect the clusters in a spatial database.

In the process of spatial clustering, the construction of the spatial proximity relationship is a critical issue. Delaunay triangulation has been proven to be a powerful tool for capturing spatial proximity in spatial analysis and spatial clustering (Estivill-Castro

(a) (b)

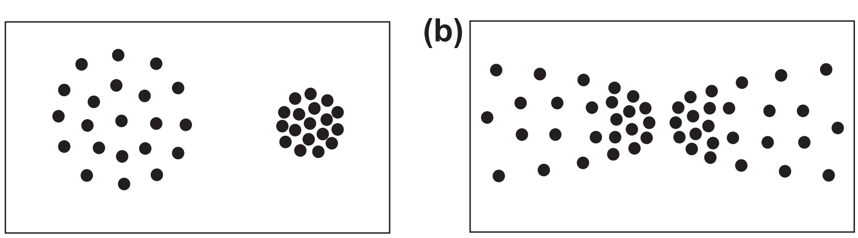

Fig. 1. Examples of clusters with uneven densities. (a) represent a database with clusters of various densities; (b) shows clusters with uneven densities.