Expert Systems with Applications 39 (2012) 11135–11148

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Map segmentation for geospatial data mining through generalized higher-order Voronoi diagrams with sequential scan algorithms

Ickjai Lee *, Christopher Torpelund-Bruin, Kyungmi Lee

School of Business (IT), James Cook University, Cairns, QLD 4870, Australia

ARTICLE INFO

Keywords:
Map segmentation
Higher order Voronoi diagrams
Geospatial data mining
Sequential-scan algorithms

ABSTRACT

Segmentation is one popular method for geospatial data mining. We propose efficient and effective sequential-scan algorithms for higher-order Voronoi diagram districting. We extend the distance transform algorithm to include complex primitives (point, line, and area), Minkowski metrics, different weights and obstacles for higher-order Voronoi diagrams. The algorithm implementation is explained along with efficiencies and error. Finally, a case study based on trade area modeling is described to demonstrate the advantages of our proposed algorithms.

1. Introduction

Segmentation is one approach to build topology (Chen, Wang, & Feng, 2010; Hanafizadeh & Mirzazadeh, 2011; Lee, Qu, & Lee, 2012; Seng & Lai, 2010) and it is one popular method for geospatial data mining. Generalized Voronoi Diagrams (GVDs) are generalizations of the ordinary Voronoi diagram to various metrics, different weights, in the presence of obstacles, complex data types (point, line and area), and higher order. Recently, Lee and Torpelund-Bruin (2012) proposed an efficient and effective GVD algorithms for use in geospatial data mining. However, the model is limited to the first order and is not able to capture higher order scenarios. In many real business settings, people are more interested in the second or third nearest object. For instance, patients are interested in the second nearest hospital when the first nearest hospital is fully occupied or closed. This article extends (Lee & Torpelund-Bruin, 2012) to implement truly flexible higher-order Voronoi diagrams (HOVD) for map segmentation and geospatial data mining.

HOVD have been studied by many researchers and found useful for a variety of applications when k number of points are considered for partitioning (Boots & South, 1997; Lee & Gahegan, 2002; Lee & Lee, 2007; Lin & Kung, 2001; Xie, Wang, & Cao, 2007). Recently, interest has been growing into exploiting HOVD for the partitioning of geospatial information for GIS applications (Pinliang, 2008). However, current HOVD modeling techniques are limited because of the complexity of the traditional vector-based algorithms used (Lee & Lee, 2009). This encompasses generators being typically limited to points in the absence of obstacles, the underlying metric limited to the Euclidean metric and weights of generators assumed to be invariant. For accurate geospatial analysis, a robust and versatile method is required to model the diverse and various components associated with real world geospatial analysis. This has lead to the exploration of efficient raster-based methods for GVD.

Due to advances in Web 2.0 technologies and the prevalence of Web Map Service (WMS) such as Google Map (http://maps.google.com), Google Earth (http://earth.google.com), NASA World Wind (http://worldwind.arc.nasa.gov), and Open Street Map (http://www.openstreetmap.org), districting Web maps through raster-based GVD is of emergence. In this article, we propose efficient and effective sequential-scan algorithms for HOVD districting. We extend the distance transform algorithm (Shih & Wu, 2004) to include complex primitives (point, line, and area), Minkowski metrics, different weights and obstacles for HOVD. The algorithm implementation is explained along with efficiencies and error. These new algorithms can be used to enhance accuracy for order-k and ordered order-k queries in various geospatial applications. To demonstrate the advantages of our proposed algorithms within an application, a case study based on trade area modeling is described.

Section 2 begins with a brief explanation of the properties of Voronoi diagrams. Section 3 describes current raster-based techniques for distance transforms and their application to Voronoi diagrams. Section 4 describes the reasoning of the sequential-scan algorithm for higher-order analysis. Sections 4.2.2 and 4.3.2 describe the implementation of the generalized higher-order algorithms for no obstacles and obstacles in the plane respectively. Sections 6 and 7 reflect the efficiency and error of the described algorithms. Section 8 describes a case study demonstrating how the algorithms can be applied for various applications and studies. Finally, Section 9 reflects on the study and future work.

* Corresponding author.
E-mail addresses: Ickjai.Lee@jcu.edu.au (I. Lee), christopher.torpelund@my.jcu.edu.au (C. Torpelund-Bruin), Joanne.Lee@jcu.edu.au (K. Lee).

0957-4174/$ - see front matter Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.042

2. Generalized Voronoi Diagrams

Let $G = \{g_1, g_2, \ldots, g_n\}$ be a set of *generator* points of interest in the plane P in R^m space. Every location p in P can be assigned to the closest generator $g \in G$ with a certain distance metric defined by $dist(p,g)$. The assignment of generators G over the plane gives the set of Voronoi regions $\mathcal{V} = \{V(g_1), V(g_2), \ldots, V(g_n)\}$. If a generator is equally close to two points in G then the location becomes part of a Voronoi boundary, whilst if it is equally close to more than two points then the location becomes a Voronoi vertex. The Voronoi boundaries between g_i and g_j can be defined as $e(g_i,g_j) = V(g_i) \cap V(g_j)$. The complete boundary of g_i over g_j gives the dominance region $Dom(g_i,g_j)$. The dominance region defined by the subsequent Voronoi regions is generalized by the type of generator, such as points, lines or polygons, various weights, plane constraints, and the metric space used. The result of the combination of generalizations generates a districted plane polygons and arcs made up of lines and Bezier curves. The way GVD can be represented is virtually limitless. For a further detailed explanation of the properties of Voronoi diagrams readers should consider (Okabe, Boots, Sugihara, & Chiu, 2000). This article focuses on higher-order generalizations of Voronoi diagrams produced with the Minkowski metric so these properties are briefly examined in the next sections.

2.1. Minkowski metric

The Minkowski metric, also called the Minkowski tensor or pseudo-Riemannian metric, defines the setting for space–time. The three ordinary dimensions of space are combined with a single dimension of time to form a four-dimensional manifold. The Minkowski metric is defined as:

$$d_{L_p}(g,g_i) = \left[\sum_{j=1}^{m} |x_j - x_{ij}|^p\right]^{1/p} \quad (1 \leqslant p \leqslant \infty) \tag{1}$$

where $(x_1, x_2, \ldots, x_m)$ and $(x_{i1}, x_{i2}, \ldots, x_{im})$ are the Cartesian coordinates of g and g_i, respectively. The parameter p can be in the range of $1 \leqslant p \leqslant \infty$. When $p = 1$, then $d_{L_1}(g,g_i) = \sum_{j=1}^{m} |x_j - x_{ij}|$ is the Manhattan metric. The Minkowski metric becomes the Euclidean metric when $p = 2$. If $p = \infty$, then the Minkowski metric becomes $d_{L_\infty}(g,g_i) = max_j|x_j - x_{ij}|$, which is called the chessboard metric. The most popular distance metric d is the Euclidean metric. Being rotationally invariant makes it most desirable for depicting two- and three-dimensional characteristics of the Earth's surface, subsurface, and atmosphere. However, in urban geography the Manhattan distance metric better approximates real world situations because of its constraints on diagonal movement (Krause, 1975). Due to this attribute, the Manhattan distance can be derived generally in a very efficient linear time $O(n)$ (Kolountzakis & Kutulakos, 1992). However, a drawback is that diagonal distances are over-estimated because a diagonal connection counts as 2 steps rather than $\sqrt{2}$ for the Euclidean distance.

2.2. Higher-order Voronoi diagrams

HOVD are natural and useful generalizations of Voronoi diagrams for more than one generator (Okabe et al., 2000). They provide tessellations where each region has the same unordered k closest sites for a given k. Let us consider the set of generators G in a plane. Considering the higher-order Voronoi diagram of the k closest unordered (order-k) points can be defined as $\mathcal{V}^k = \left\{V\left(G_1^k\right), \ldots, V\left(G_l^k\right)\right\}$, where the order-$k$ Voronoi region $V\left(G_i^k\right)$ for a random subset G_i^k consists of k generators from the set G and G_l^k represents all of the possible subsets. The set of points in the plane assigned to the order-k Voronoi region $V\left(G_i^k\right)$ can be defined as:

$$V\left(G_i^k\right) = \left\{p|\max_{g_h}\left\{d(p,g_h)|g_h \in G_i^k\right\} \leqslant \min_{g_j}\left\{d(p,g_j)|g_j \in G/G_i^k\right\}\right\} \tag{2}$$

Because HOVD are a set of k Generalized Voronoi Diagrams, the same generalizations apply. Order-k Voronoi diagrams have been combined with weights to generate *Order-k Multiplicatively Weighted Voronoi Diagram* (OKMWVD). If the order of k generators is considered then this generates the *Ordered, Order-k, Multiplicatively Weighted Voronoi diagram* (OOKMWVD) (Boots & South, 1997).

3. Raster based algorithms and distance transforms

The general idea of the raster based method is to expand the Voronoi diagram incrementally by adding a point at a time for all points in the image by considering the *number of neighbors* or *direction of connection* (Fortune, 1987; Lee & Drysdale, 1981; Li, Chen, & Li, 1999). The Voronoi diagram is represented by a discrete grid lattice of size $N \times N$ which gives N^2 points in the plane. Unlike vector districting methods, which generally have efficient time complexities at the expense of a lack of diversity, raster based methods are robust and versatile and are able to support accurate modeling of the diverse and various components associated with real world geospatial analysis. Despite the potential advantages, there is little literature based purely on this method (Li et al., 1999). The methods of generating raster based Voronoi diagrams are based heavily on the body of work related to *distance transforms* (DTs). DTs are characterized by the mappings of each image pixel (foreground) into its smallest distance to regions of interest (background) (Rosenfeld & Pfaltz, 1966). Generally, these algorithms can be divided into *iterative algorithms* and *sequential algorithms* based on the order used to scan the pixels (Cuisenaire & Macq, 1999).

Iterative algorithms use *wave* scans over points in the background image B which are seen as sources from which distance values are calculated for the foreground image F. The waves propagate out from the sources like how a grass fire would burn from the source of ignition outwards. Time complexity for such parallel processing is typically $O(N)$. However, the computing architecture required for this parallel processing requires a separate processor for each generator in the plane. This results in a *time-processor complexity* of $O(N^3)$ (Danielsson & Tanimoto, 1983). The same propagation wave scan is possible with a single processor with a time-complexity of $O(N^2)$ (Rosenfeld & Pfaltz, 1966). However, most iterative algorithms based on this single processor architecture are inefficient as many pixels are often unnecessarily updated. Alternatively, sequential algorithms have better time-processor complexities for single processors.

Sequential-scan algorithms are characterized by the *number of neighbors* and *plane-sweeps* that are considered. The number of neighbors range from four-neighborhoods for the Manhattan metric, to eight or more arbitrarily large neighborhoods. The shape of neighbors being considered also differs between algorithms. There is a trade-off between the size and shape of the neighborhood with the number of iterations needed to achieve a solution and the quality of the approximation (Breu, Gil, Kirkpatrick, & Werman, 1995). Chamfer DTs include individual weights in various sized masks to minimize the deviation from the Euclidean DT (EDT) (Borgefors, 1984). The downside with this method is that the usage of weighted masks can only allow approximate distance values. Even large Chamfer masks can only approximate to within 2% error (Cuisenaire, 1999). Recent methods to reduce the error include