基于相交关系的 GML空间线对象离群检测算法

朱 娟[1], 吉根林[2]

(1.南京师范大学计算机科学与技术学院,江苏 南京 210097)
(2.江苏省信息安全保密技术工程研究中心,江苏 南京 210097)

[摘要] 提出了一种基于相交关系的 GML空间线对象离群检测算法 DOL_IR,该算法首先计算 GML线对象与其他空间对象的相交关系,定义基于相交关系的相异度,将其作为空间线对象之间距离的度量准则,利用 DBSCAN聚类算法检测离群的基于空间相交关系的线对象.实验结果表明,算法 DOL_IR能准确地检测出离群的基于空间相交关系的线对象,并具有较高的效率.

[关键词] GML,线对象,相交关系,离群检测
[中图分类号] TP391 [文献标识码] A [文章编号] 1001-4616(2010)03-0127-04

An Algorithm for Detecting Outlier Lines Based on Intersection Relationship for GML Data

Zhu Juan[1], Ji Genlin[2]

(1. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097, China)
(2. Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210097, China)

Abstract A new algorithm DOL_IR is presented for detecting outlier lines based on intersection relationship for GML data. Intersection relations between spatial lines and other spatial objects are computed. The difference degree between one line and another line is defined as the standard of the distance between one line and another line. Algorithm DBSCAN is used to detect outlier lines based on intersection relationship. The experimental results show that algorithm DOL_IR can detect outlier lines based on intersection relationship accurately and effectively.

Key words GML, lines, intersection relationship, outlier detection

离群检测是数据挖掘中的一个重要方面.现有的离群检测方法有:基于统计的离群检测算法[1],基于距离的离群检测算法[2],基于密度的离群检测算法[3],基于深度的离群检测算法[4],基于偏移的离群检测算法[5,6],基于聚类的离群检测算法[7]等等.在现实生活中,离群检测有着广泛的应用,例如:信用卡的恶意透支监测、灾难天气预测、网络入侵检测等.现在,空间离群检测算法的研究已经引起了研究者的兴趣,空间离群检测在地理信息系统和空间数据分析中有很多应用,包括公共安全、公共健康、生态环境、交通运输和基于位置的服务等领域.但是,目前的空间离群检测算法很少考虑空间对象在拓扑关系上的差异性.

GML(Geography Mark-up Language)是一种用于描述现实世界中地理对象的标识语言[8].目前,针对GML数据对象的空间拓扑关系的离群检测的研究较少,而且很少涉及线、面对象.文献[9]提出了基于空间相邻关系的 GML点对象的离群检测算法;文献[10]提出了一种基于面包含关系的 GML空间离群面检测算法.本文研究 GML线对象与其他空间对象的相交关系,定义相交关系的相异度,将其作为空间线对象之间距离的度量准则,利用 DBSCAN聚类算法检测离群的基于空间相交关系的线对象.实验结果表明,算法 DOL_IR能够有效地挖掘离群的基于空间相交关系的线对象,并具有较高的效率.

收稿日期: 2010-06-10
基金项目:国家自然科学基金(40871176).
通讯联系人:吉根林,博士,教授,博士生导师,研究方向:数据挖掘技术及其应用. E-mail: glj@njnu.edu.cn

1　相关概念

GML空间对象用点、线、面对象来表示. 点对象 P_i 表示为: $P_i = \{PId_i, (x_i, y_i), PLabel_i\}$, 其中, PId_i 是点对象 P_i 的 id号, (x_i, y_i) 是 P_i 的坐标, $PLabel_i$ 是 P_i 的类别号; 线对象是由若干点对象组成, 线对象 L_i 表示为: $L_i = \{LId_i, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \cdots, (x_{is}, y_{is}), LLabel_i\}$, 其中, LId_i 是线对象 L_i 的 id号, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}), …, (x_{is}, y_{is}) 是组成线对象 L_i 的所有点对象的坐标, $LLabel_i$ 是 L_i 的类别号; 面对象是由若干首尾相连的线段组成, 面对象 R_i 表示为: $R_i = \{RId_i, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \cdots, (x_{it}, y_{it}), RLabel_i\}$, 其中, RId_i 是面对象 R_i 的 id号, (x_{i1}, y_{i1}), (x_{i2}, y_{i2}), …, (x_{it}, y_{it}) 是组成面对象 R_i 的所有线段的端点的坐标, $RLabel_i$ 是 R_i 的类别号.

假设线对象的集合 $L = \{L_1, L_2, \cdots, L_m\}$, 根据实际情况, 线对象只可能与线对象、面对象相交. 对每种类型的相交对象数量进行归一化处理, 将其值规范化到 $[0, 1]$, 并将线对象 L_p $(p = 1, 2, \cdots, m)$ 的相交关系表示为: $IS(L_p) = \{\langle LLabel_1, LCount_{p1}\rangle, \langle LLabel_2, LCount_{p2}\rangle, \cdots, \langle LLabel_k, LCount_{pk}\rangle, \langle RLabel_1, RCount_{p1}\rangle, \langle RLabel_2, RCount_{p2}\rangle, \cdots, \langle RLabel_r, RCount_{pr}\rangle\}$, 其中, k、r 分别是线对象、面对象的类别总数, $LCount_{pi}$ 是与 L_p 相交类别为 $LLabel_i$ 的归一化数量 $(i = 1, 2, \cdots, k)$, $RCount_{pj}$ 是与 L_p 相交类别为 $RLabel_j$ 的归一化数量 $(j = 1, 2, \cdots, r)$.

定义 **1**　为了比较两个相同类别的线对象在相交关系上的差异, 本文定义线对象 L_i 与 L_j 在相交关系上的相异度 $d(L_i, L_j)$ 如下:

$$d(L_i, L_j) = \sum_{u=1}^{k} \delta(LCount_{iu}, LCount_{ju}) + \sum_{v=1}^{r} \delta(RCount_{iv}, RCount_{jv}),$$

其中, $\delta(LCount_{iu}, LCount_{ju}) = \begin{cases} |LCount_{iu} - LCount_{ju}|, & LCount_{iu} \neq 0 \text{且} LCount_{ju} \neq 0 \\ 1, & LCount_{iu} = 0 \text{或} LCount_{ju} = 0 \end{cases}$,

$$\delta(RCount_{iv}, RCount_{jv}) = \begin{cases} |RCount_{iv} - RCount_{jv}|, & RCount_{iv} \neq 0 \text{且} RCount_{jv} \neq 0 \\ 1, & RCount_{iv} = 0 \text{或} RCount_{jv} = 0 \end{cases},$$

将 $d(L_i, L_j)$ 定义为线对象 L_i 与 L_j 之间基于相交关系的距离.

2　*GML* 空间线对象相交关系的计算

GML线对象、面对象都是由若干个连续的线段组成. 因此, 要判断 2个空间对象是否相交就是要判断 2个空间对象中是否存在线段相交.

文献[11]提出了一种判断线段相交算法, 该算法首先判断 2条线段的外接矩形是否相交, 如果 2条线段的外接矩形不相交(如图 1所示), 则可以断定 2条线段不相交; 如果 2条线段的外接矩形相交(如图 2、图 3所示), 则利用向量几何中混合积的性质, 通过由端点所构成的向量与辅助向量的混合积的正负来判断两线段是否相交.

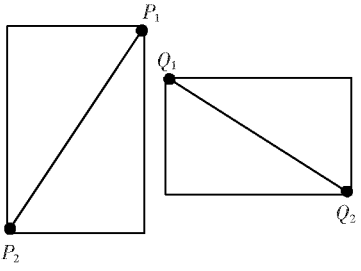

图 1　2条线段的外接矩形不相交

Fig.1　Non-intersection of the rectangles of two segments

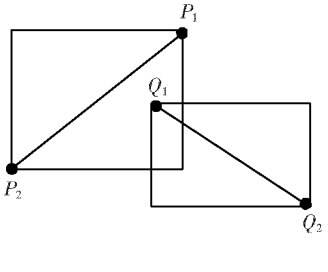

图 2　2条线段的外接矩形相交但两条线段不相交

Fig.2　The intersection of the rectangles of two segments and non-intersection of two segments

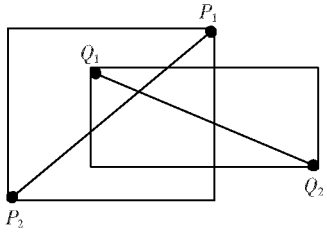

图 3　2条线段相交

Fig.3　The intersection of two segments

算法描述如下:

输入: 线段 P_1P_2 的端点坐标 $P_1(x, y)$、$P_2(x, y)$, Q_1Q_2 的端点坐标 $Q_1(x, y)$、$Q_2(x, y)$

输出: 2条线段是否相交

步骤:

(1) Rect1 = external_rectangle(P_1P_2); //求 P_1P_2的外接矩形 Rect1

(2) Rect2 = external_rectangle(Q_1Q_2); //求 Q_1Q_2的外接矩形 Rect2

(3) if(! rectangle_intersection(Rect1, Rect2)) return false;

//判断 Rect1与 Rect2是否相交, 如果不相交, 则 P_1P_2与 Q_1Q_2不相交

(4) else{ //如果 Rect1与 Rect2不相交, 则通过混合积的方法判断 P_1P_2与 Q_1Q_2是否相交

(5) $d1 = (P_2.x - P_1.x)*(Q_1.y - P_1.y) - (P_2.y - P_1.y)*(Q_1.x - P_1.x)$;

(6) $d3 = (P_2.x - P_1.x)*(Q_2.y - P_1.y) - (P_2.y - P_1.y)*(Q_2.x - P_1.x)$;

(7) $d2 = (Q_2.x - Q_1.x)*(P_1.y - Q_1.y) - (Q_2.y - Q_1.y)*(P_1.x - Q_1.x)$;

(8) $d4 = (Q_2.x - Q_1.x)*(P_2.y - Q_1.y) - (Q_2.y - Q_1.y)*(P_2.x - Q_1.x)$;

(9) $d1 = d1*d3$; $d2 = d2*d4$

(10) if($d1 <= 0 \&\& d2 <= 0$)

(11) return true

(12) else return false

(13) }

在判断线对象 L与另一空间对象 O (可以是线对象, 也可以是面对象) 是否相交时, 首先判断 L的外接矩形与 O的外接矩形是否相交, 如果 2个外接矩形不相交, 则可以断定线对象 L与对象 O不相交; 如果 2个对象的外接矩形相交, 则依次判断线对象 L与对象 O中是否存在线段相交, 如果存在线段相交, 则线对象 L与对象 O相交, 否则线对象 L与对象 O不相交.

3　空间线对象离群检测算法 DOL_IR

算法 DOL_IR首先对 GML文档进行预处理, 获得空间对象集合, 然后计算线对象的相交关系, 将 2个线对象的相异度作为 DBSCAN算法[12]中 2个线对象之间的距离, 利用 DBSCAN算法进行聚类, 发现其中的噪声即为离群的基于空间相交关系的线对象.

输入: GML文档 D, 邻域半径 eps, 最小密度 minpts

输出: 离群对象的集合 Out

步骤:

(1) Object= ReadGml(D); //解析 GML, 获得空间对象集合

(2) Intersection= Get_Intersection(Object); //计算线对象的相交关系

(3) C= DBSCAN(Object, Intersection, eps, minpts); //用 DBSCAN算法聚类

(4) Out= Get_Outlier(C); //得出离群的基于空间相交关系的线对象

4　实验结果及分析

为了验证算法 DOL_IR的有效性, 在 CPU2.11GHz, 480MB内存的微机上, 用 VC++实现了本算法. 实验的 GML数据为合成数据, 数据源如图 4所示. 图中细线代表公路, 带箭头的线代表河流, 粗线代表铁路, 三角形代表湖海, 矩形代表住宅区. 为了便于计算, 实验中分别用 0, 1, 2, 3, 4来表示上述的 5种类型. 实验的目的是找出图中离群的公路, 计算每条公路与其他空间对象的相交关系, 如表 1所示. 例如, L_1〈0, 3〉〈1, 1〉〈2, 2〉〈3, 3〉〈4, 2〉表示公路 L_1与 3条公路, 1条河

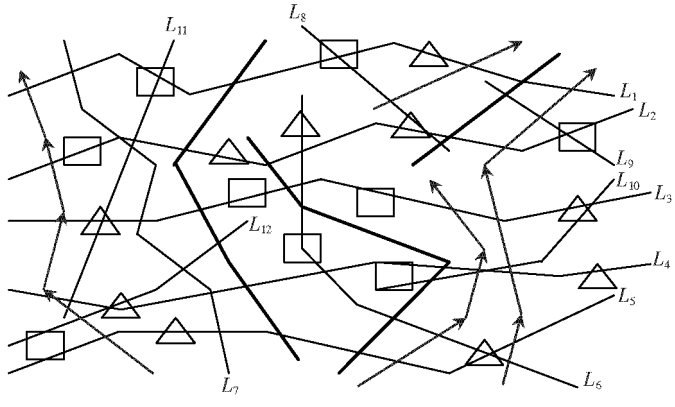

图 4　GML合成数据示意图

Fig.4　Samples of GML data

流, 2条铁路, 3个湖海, 2个住宅区相交. DOL_IR算法执行结果显示公路 L_7 为离群的公路.

为了测试算法 DOL_IR执行的时间效率, 采用合成数据源进行测试, 该数据源中包含 50%的线对象和 50%的面对象, 由于基于相交关系的线对象离群检测算法尚无文献报道, 故无法将算法 DOL_IR与其他算法进行比较, 算法 DOL_IR的运行效率如图 5所示. 实验表明, 算法 DOL_IR能够有效地挖掘离群的线对象, 且具有较好的性能.

表 1　公路的相交关系

Table 1　Intersection relations of roads

公路	相交对象 1	相交对象 2	相交对象 3	相交对象 4	相交对象 5
L_1	〈0, 3〉	〈1, 1〉	〈2, 2〉	〈3, 3〉	〈4, 2〉
L_2	〈0, 5〉	〈1, 2〉	〈2, 2〉	〈3, 2〉	〈4, 3〉
L_3	〈0, 4〉	〈1, 2〉	〈2, 2〉	〈3, 3〉	〈4, 2〉
L_4	〈0, 5〉	〈1, 2〉	〈2, 1〉	〈3, 3〉	〈4, 2〉
L_5	〈0, 2〉	〈1, 3〉	〈2, 1〉	〈3, 3〉	〈4, 2〉
L_6	〈0, 4〉	〈1, 2〉	〈2, 1〉	〈3, 2〉	〈4, 1〉
L_7	〈0, 5〉				
L_8	〈0, 2〉	〈2, 1〉	〈3, 1〉	〈4, 1〉	
L_9	〈0, 1〉	〈2, 1〉	〈3, 1〉	〈4, 1〉	
L_{10}	〈0, 2〉	〈1, 1〉	〈2, 1〉	〈3, 2〉	〈4, 1〉
L_{11}	〈0, 4〉	〈1, 1〉	〈2, 1〉	〈3, 1〉	
L_{12}	〈0, 1〉	〈1, 1〉	〈2, 1〉	〈3, 1〉	〈4, 1〉

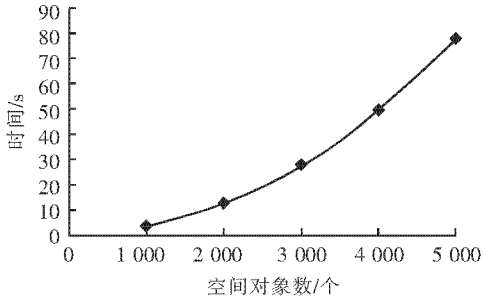

图 5　算法 DOL_IR的运行效率

Fig.5　Runtime of algorithm DOL_IR

5　结束语

近年来, 空间离群检测已成为数据挖掘的热点, 本文考虑了空间对象在相交关系上的差异, 提出了一种基于相交关系的 GML空间线对象离群检测算法 DOL_IR. 实验结果表明, 该算法能够有效地检测出离群的线对象, 并具有较高的效率. 算法中判断 2个空间对象是否相交的算法耗时较大还有待改进; 另外, 如何选取参数还需进一步的研究.

[参考文献]

[1] Barnett V, Lewis T. Outliers in Statistical Data[M]. New York: John Wiley & Sons, 1994.

[2] Knorr E, Ng R. Finding intensional knowledge of distance-based outliers[C]// Proc of the 25th Very Large Databases Conference. Edinburgh: Morgan Kaufmann Publishers, 1999: 211-222.

[3] Breunig M M, Kriegel H P, Ng R T, et al. Optics of identifying density-based local outliers[C]// Proc of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, Lecture Notes in Computer Science 1704. Prague: Springer, 1999: 262-270.

[4] Prearata F, Shamos M. Computational Geometry: An Introduction[M]. Berlin: Springer-Verlag, 1988.

[5] Jagadish H V, Koudas N, Muthukrishnan S. Mining deviants in a time series databases[C]// Proc of the 25th Conference on Very Large Databases. Edinburgh: Morgan Kaufmann Publishers, 1999: 102-113.

[6] 郑建国, 焦李成. 偏差检测挖掘方法研究[J]. 计算机工程, 2001, 27(8): 33-35.

[7] Sheng Yijiang, Qing Boan. Clustering-based outlier detection method[C]// 5th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway: IEEE Computer Society, 2008: 429-433.

[8] 张书亮, 闾国年, 龚健雅, 等. 地理标示语言——Geo-Web基础[M]. 北京: 科学出版社, 2008: 3-4.

[9] 陈佳春, 吉根林. 基于空间相邻关系的 GML点对象离群检测算法[J]. 南京师范大学学报: 工程技术版, 2009, 9(1): 61-63.

[10] 李尼格, 鲍培明, 沙露. 一种基于面包含关系的 GML空间离群面检测算法[J]. 广西师范大学学报: 自然科学版, 2009, 27(3): 118-121.

[11] 张宏, 温永宁, 刘爱利, 等. 地理信息系统算法基础[M]. 北京: 科学出版社, 2006: 23-24.

[12] Ester M, Kriegel H P, Jorg S, et al. A density-based algorithm for discovering clustering in large spatial databases with noise[C]// Proceeding of 2nd Conference on Knowledge Discovery in Databases. Piscataway: IEEE Press, 1996: 226-231.

[责任编辑: 顾晓天]