

doi:10.3799/dqkx.2010.044

基于成分关联区域相似度的面实体模糊匹配算法

叶亚琴^{1,2}, 万波^{1,2}, 陈波³

1. 中国地质大学信息工程学院, 湖北武汉 430074
2. 地理信息系统软件及其应用教育部工程研究中心, 湖北武汉 430074
3. 武汉中地数码科技有限公司, 湖北武汉 430074

摘要: 空间目标匹配是空间数据库增量更新的第一步,也是关键一步。研究了基于空间目标匹配的变化信息的获取算法。通过研究空间数据中存在的 uncertainty 问题,提出将模糊理论引入到空间目标匹配算法中。重点研究如何用模糊的方法解决空间目标匹配问题,并以面实体为例说明了具体匹配过程,提出了基于成分关联区域相似度的面实体模糊匹配算法。该算法利用成分关联区域的度量因子,确定模糊拓扑关系隶属度矩阵,进而量化隶属度矩阵,最终确定模糊拓扑关系分类。算法综合利用了图幅索引、成分关联因子等进行优化,简化计算复杂度,提高了算法效率。

关键词: 实体匹配;模糊拓扑关系;成分关联区域。

中图分类号: TP214

文章编号: 1000-2383(2010)03-0385-06

收稿日期: 2010-01-15

The Fuzzy Match Algorithm between Area Object Considering Associated Area Similarities

YE Ya-qin^{1,2}, WAN Bo^{1,2}, Chen Bo³

1. Faculty of Information Engineering, China University of Geosciences, Wuhan 430074, China
2. Engineering Research Center of GIS Software and Applications, Ministry of Education, Wuhan 430074, China
3. Wuhan Zondycyber Co., Ltd., Wuhan 430074, China

Abstract: Spatial objects' matching is the first step and also the key step of incremental updating for spatial database. In this paper, a spatial objects' matching algorithm for finding the changed information is studied. On the basis of the research on the uncertainty problem existing in spatial data, this paper suggests fuzzy theory is introduced to spatial objects' matching algorithm. The paper focuses on how to solve spatial objects' matching problem by using fuzzy methods. And taking the region entities as examples for studying matching process, the author proposes the region objects' matching algorithm considering associated area similarities. This algorithm firstly uses associated area similarity's measuring genes to confirm the fuzzy topological relationship matrix, then quantifies the degree of membership matrix, and finally determines the relationship between fuzzy topological classifications. By using frame index, associated area similarity's measuring genes, the algorithm optimizes and simplifies computational complexity, and improves the algorithm efficiency as well.

Key words: spatial objects' matching; fuzzy topologic relation; associated area.

空间目标匹配是重要的多时态空间数据变化信息获取方式。所谓空间目标匹配,即通过分析空间实体的差异和相似性,识别出不同来源图中表达现实世界同一地物或地物集(即同名实体)的过程(张桥平等,2004)。空间目标匹配是基于增量信息提取的空间数据库增量更新的第一步,也是关键一步,该技

术的完成质量直接影响着更新的效率与准确度,对多时态空间目标匹配算法的研究具有很强的现实意义。

空间目标匹配作为基础研究内容,其成果不仅应用于地图比较,还在空间数据库的增量更新、空间认知、数据融合等诸多方面广泛应用,是目前学者们

研究的热点问题. 它也属于地理空间推理的研究范畴, 是空间认知过程中非常重要的基本活动.

空间目标匹配存在两个层次: 第一层是相同比例尺下的多时态数据或者多来源数据的相同目标匹配; 第二层是不同比例尺下多尺度数据的目标匹配. 两个层次体现出了空间目标匹配的多时态、多尺度的特性, 并且第二层次比第一个层次的匹配更为复杂, 因为存在不同的比例尺, 涉及到比例尺的转换问题. 另一方面, 从匹配方式上讲, 空间目标匹配可分为几何匹配、拓扑关系匹配和语义匹配(图 1). 几何匹配和拓扑关系匹配分别通过计算待匹配实体间的几何相似度(similarity)和拓扑相似度值来确定匹配目标. 它们都是通过度量地物的空间形态特征来获取匹配结果的, 同属于空间匹配; 而语义匹配则是通过计算实体属性的相似度进行匹配. 本文将重点讨论以空间匹配的方式解决第一个层次的目标匹配的具体方案, 即相同比例尺下多时态数据目标匹配的空间匹配.

作为研究热点, 学者们在实体匹配方法上进行了热烈地探讨. 以面实体匹配为例, Wentz(1997)提出了若干个面实体的形状度量指标, 包括面的紧致度(面积周长比)、边界的描述和面的构成成分; 张桥平等(2004)提出了基于模糊拓扑关系分类的面实体匹配方法; 刘志勇(2006)提出采用面质心结合多种检验规则的几何匹配方法, 用点在面内规则进行粗匹配, 再结合多边形的面积和面密度进行匹配检验; Foley(1997)提出通过计算边界线间的面积以求得线实体间的形状相似度; 傅仲良和吴建华(2007)考虑了重叠部分面积(或长度)占源要素的面积(或长度)的比重以及源要素与目标要素的面积(或长度)的比值两个因素, 提出了基于权重的相似性计算模型; 章莉萍等(2008)以居民地为研究对象, 提出了增量式凸壳匹配方法; 郝燕玲等(2008)利用形状中心点确定位置相似度, 用形状描述函数表示形状相似度, 采用加权平均法来综合各空间匹配指标, 以确定

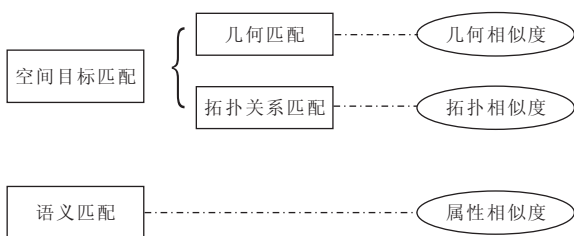


图 1 空间目标匹配的分类

Fig. 1 The kinds of spatial objects match

匹配结果.

尽管, 国内外在矢量数据的空间目标匹配方面做了很多工作, 也取得了一定的进展, 但是目前还没有比较理想地解决矢量数据的匹配问题, 仍存在诸多难点、重点问题需要解决. 一方面是因为实体匹配过程本身所需的智能化程度较高、算法复杂性较大所致; 另一方面促使笔者思考适宜于空间实体匹配的数据模型到底应该是什么样的.

李德仁等(2004)、叶亚琴等(2006)提出基于模糊理论的模糊空间要素模型较矢量数据模型能够更好地解决上述问题: (1)更接近不确定性自然界的真实表达, 更能包容数据测量时带来的误差; (2)具备更多的背景知识和语义信息, 这些信息可辅助于模拟人对地物特征的认知过程, 有助于形成更加智能化的匹配算法; (3)更利于运用模糊要素模型、模糊拓扑关系理论及系统的模糊数学理论进行空间目标的匹配.

此前也有学者将模糊集的概念引入到空间目标匹配中, 提出了基于模糊拓扑关系分类的面实体匹配方法. 但其方法中形态距离的计算公式较为复杂, 计算量较大. 本文将尝试采用基于点集的定量方法计算模糊拓扑关系矩阵, 通过计算可信度值寻找匹配实体的方法来解决面实体匹配问题. 将在引入模糊理论和方法的基础上, 采用模糊集合的截集和卵黄模型相结合的方法来描述模糊区域, 并运用模糊集理论通过模糊相似度指标度量、模糊拓扑关系及模糊推理等模糊数学理论和方法处理空间目标匹配问题.

本文提出的匹配方法称为基于成分关联区域相似度的面实体模糊匹配算法(fuzzy match arithmetic between area object considering associated area similarity), 简称为 CAAS-FMA 算法.

1 基于成分关联区域相似度的面实体模糊匹配算法

1.1 候选匹配实体集的确定

候选匹配实体集的确定是地图数据库同名实体中的一个重要过程, 它将不仅决定实体匹配的效率, 还决定匹配的正确性和完整性.

首先利用图幅索引缩小候选匹配实体集的范围. 图幅索引正是根据基础数据库的数据特点——“纵向分层, 横向分幅”而设计的索引, 它能有效提高数据查询的效率. 利用它能快速查找可能与待匹配实体的最小外包矩形(max bound rectangle),

MBR)相交的实体,以此滤去大部分无关数据.图幅索引的查询高效在于使用每个图幅的范围与待匹配实体的 MBR 是否相交,若不相交则淘汰属于该图幅的所有实体,否则将所有属于该图幅的实体都列为候选匹配集.

然后,再在上一步产生的候选匹配实体集中逐个判断候选实体的 MBR 是否确实与待匹配实体的 MBR 相交,并求出两个矩形的相交面积,若面积比大于 0.6,则进入下一个筛选步骤.最后,形成最终的候选实体集.

由于实体的匹配是一个非常耗时的过程,因此应尽量避免大范围的搜索.本文对候选匹配集进行了较细致的筛选,且效率较高.

1.2 成分关联区域的度量因子

获得了候选匹配集合后,下一步是确定实体间空间拓扑关系相似度的度量方法.拓扑相同部分多少就是拓扑关系的相似程度.

实体间的拓扑关系是非常复杂的,在很多情况下,很难对其进行非常详细的描述,这时就需要对这些拓扑关系进行抽象.空间拓扑关系抽象是将详细描述的关系根据一定的要求概括为更粗粒度的拓扑关系,这种变化事实上就包含复杂程度的减少;因此,为了详细描述拓扑关系,必须结合度量方法对拓扑关系进行精炼.例如,对区域之间的拓扑关系的抽象,可以分别使用长度和面积两种度量.换句话说,现在问题是面实体间的拓扑关系的主要度量因子有哪些,拓扑关系如何公式化.在弄清楚这个问题之前,先来了解一下“成分关联区域”的概念.

任何两个边界交成分之间的两个区域的边界,在两个区域的并集内都会形成一个小区域,称之为“成分关联区域”(郭庆胜等,2006)(图 2a).对成分关联区域的度量研究是描述空间拓扑关系的相似性的常用方法,也是拓扑关系等价性抽象的基础内容.之前学者们对成分关联区域的度量作了一定的研究,但未见将其引入空间目标匹配.当将成分关联

区域应用于空间目标匹配时,成分关联区域的重要性决定了匹配实体可能匹配的程度.

成分关联区域的重要性跟这些区域的面积大小有直接联系:如果成分关联区域的面积越大,相关的成分就越重要.因此,考虑成分关联区域的面积,采用相交面积比率来描述 a 、 b 两个面成分关联区域的形状,定义为:

$$A_{ab} = \frac{\text{Area}(a \cap b)}{\text{Area}(a)},$$

$$A_{ba} = \frac{\text{Area}(a \cap b)}{\text{Area}(b)}, \quad (1)$$

式(1)中, A_{ab} 表示成分关联区域对 a 的相交面积比率, A_{ba} 表示成分关联区域对 b 的相交面积比率. A_{ab} 和 A_{ba} 表示的侧重点不一样,现实中存在很多的例子.若 b 包含 a ,则 $A_{ba} = 1$,并且 $A_{ab} < A_{ba}$.所以考虑双向因子值,有助于解决空间目标匹配中的 $M:N$ 的匹配问题.

最理想的匹配方法是对人类认知的模拟.多时态地图在消除整体和局部坐标偏差后,同一面实体总是有较大的重叠面积;面实体之间的重叠面积大小反映了两个面之间的距离差异以及形状差异等,而且它反映了两个面实体之间的整体相似性,符合人眼的视觉效果.

只考虑成分关联区域的面积比率还不够,还不能区分出狭长区域.例如在图 2b 中,区域 b 是狭长区域, $A_{ba} \cong 1$,但事实上该区域的形态与 a 区域的形态存在较大差别.因此,引入相交周长比率,其定义为:

$$L_{ab} = \frac{\text{Len}(a \cap b)}{\text{Len}(a)}, \quad (2)$$

式(2)中, $\text{Len}()$ 表示相交边界长度, L_{ba} 表示成分关联区域对 b 的相交周长比率.该因子可以辅助相交面积比率,判断狭长区域相交情况.

因此,双向的相交面积比率和相交周长比率可以确认为成分关联区域的度量因子,能够表示成分关联区域的重要性.并且两个因子的取值范围均为 $[0,1]$,有助于隶属度矩阵分量的量化.当然,成分关联区域的度量因子还不只这两个因子,但考虑相对重要性和在空间目标匹配的实用性问题,本文最终采用上述两个因子.

1.3 面实体之间匹配关系的确定

将成分关联区域的相似度因子进行融合,形成判断面实体间的相似度公式为:

$$\text{sim}(A, B) = \frac{\text{Area}(A \cap B)}{\text{Area}(A)} \times \alpha + \frac{\text{Len}(A \cap B)}{\text{Len}(A)} \times \beta,$$

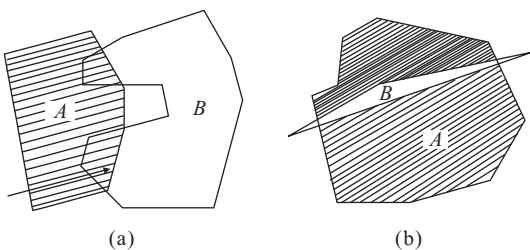


图2 成分关联区域(a)和其形状(b)

Fig. 2 Associated area (a) and its shape (b)

$$\text{sim}(B, A) = \frac{\text{Area}(A \cap B)}{\text{Area}(B)} \times \alpha + \frac{\text{Len}(A \cap B)}{\text{Len}(B)} \times \beta, \quad (3)$$

式(3)中, α 和 β 是两个控制系数, 且 $\alpha + \beta = 1$, 这里暂时使用 $\alpha = 0.7, \beta = 0.3$.

根据面实体相似度的值进一步确认面实体之间的匹配类型. 当 $\text{sim}(A, B)$ 接近 1 时, 表明 A 是整体与 B 匹配的; 当 $\text{sim}(B, A)$ 接近 1 时, 表明 B 是整体匹配 A 的. (1) 当两个值都接近于 1 时, 表明这两个面实体之间的匹配关系是 1:1 的; (2) 将其其他情况分为 1:N 和 M:1 的情况, 取最大相似度对象. 当存在 M:N 的情况时, 可将其拆分成 1:N 和 M:1 的情况; (3) 对 1:N、M:1 的情况中多的一方实体进行重组, 以多面的形式组合成一个整体.

通过上述确认与重组过程, 将所有的匹配变成了 1:1 的对应关系, 有助于模糊拓扑关系的确定.

1.4 模糊拓扑关系隶属度矩阵的形成

确定了成分关联区域的度量因子, 接下来便是解决如何形成模糊拓扑关系的隶属度矩阵问题了. 在基于 4 交模型的模糊空间拓扑关系描述和成分关联区域的度量因子的基础上, 模糊区域之间的拓扑关系可以描述为:

$$R = \begin{bmatrix} m(\tilde{A}^0 \cap \tilde{B}^0) & m(\tilde{A}^0 \cap \partial\tilde{B}) \\ m(\partial\tilde{A} \cap \tilde{B}^0) & m(\partial\tilde{A} \cap \partial\tilde{B}) \end{bmatrix}, \quad (4)$$

式(4)中, $m(\tilde{A}^0 \cap \tilde{B}^0)$ 为模糊区域间的拓扑关系的分量量化值, 即拓扑关系相似度的定量值. 这里, 结合成分关联区域的度量因子, 将该分量确定为:

$$\begin{aligned} m(\tilde{A}^0 \cap \tilde{B}^0) &= \frac{\text{Area}(\tilde{A}^0 \cap \tilde{B}^0)}{\min[\text{Area}(\tilde{A}^0), \text{Area}(\tilde{B}^0)]}, \\ m(\tilde{A}^0 \cap \partial\tilde{B}) &= \frac{\text{Area}(\tilde{A}^0 \cap \partial\tilde{B})}{\min[\text{Area}(\tilde{A}^0), \text{Area}(\partial\tilde{B})]}, \\ m(\partial\tilde{A} \cap \tilde{B}^0) &= \frac{\text{Area}(\partial\tilde{A} \cap \tilde{B}^0)}{\min[\text{Area}(\partial\tilde{A}), \text{Area}(\tilde{B}^0)]}, \\ m(\partial\tilde{A} \cap \partial\tilde{B}) &= \frac{\text{Area}(\partial\tilde{A} \cap \partial\tilde{B})}{\min[\text{Area}(\partial\tilde{A}), \text{Area}(\partial\tilde{B})]}. \end{aligned} \quad (5)$$

式(5)中, $m(\cdot)$ 表示两个模糊集 \tilde{A} 、 \tilde{B} 的内域 (α 截集 \tilde{A}^0 和 \tilde{B}^0) 和宽边界 ($\partial\tilde{A}$ 和 $\partial\tilde{B}$) 分别相交的程度. $\min[\cdot]$ 表示取两者中较小值. 由此定义可知, 这种空间关系向量中的每一个分量的值都在 $[0, 1]$ 之间, 可以很好地表示两个模糊区域的内部和宽边界两两相交的程度. 这样, 实体间的模糊隶属度矩阵便形成并

确定了.

1.5 模糊拓扑关系的定量化与分类

为了把这些模糊的空间拓扑关系转化成明确的空间拓扑关系, 还需要将隶属度矩阵进行量化. 前面定义的 8 种空间拓扑关系的“参考拓扑空间向量”包括相离(d)、相等(e)、相切(t)、相交(o)、覆盖(c)、被覆盖(cb)、包含(ct)和被包含(ctb). 隶属度矩阵量化就是找出模糊拓扑关系的隶属度矩阵与前面定义的 8 种“参考拓扑空间向量”之间的关系, 将模糊空间拓扑关系归为某一种明确的空间拓扑关系. 这一过程中, 可以使用空间向量相关度的定义. 设 S 、 T 是两个 n 维空间向量, 空间向量 S 、 T 的相关度为:

$$r(S, T) = 1 - \frac{1}{n} \sum_{i=1}^n |X_{si} - X_{ti}|, \quad (6)$$

式(6)中, X_{si} 和 X_{ti} 分别表示空间向量 S 、 T 的第 i 个分量; $|X_{si} - X_{ti}|$ 表示这两个分量的欧式距离.

根据空间向量的相关度定义可知, 相关度越大, 说明两个向量相关程度越大, 即两个向量越类似. 这样, 分别计算一个被考察的模糊拓扑关系与几种作为标准的参考空间向量相关度, 然后根据比较计算结果, 就可以将这个空间向量进行分类. 因此, 下一步需要分别计算被考察空间向量与 8 种参考拓扑空间向量 ($R_d, R_e, R_t, R_o, R_c, R_{cb}, R_{ct}, R_{ctb}$) 间的相关度的值, 依次为 $r(R_M, R_d), r(R_M, R_e), r(R_M, R_t), r(R_M, R_o), r(R_M, R_c), r(R_M, R_{cb}), r(R_M, R_{ct}), r(R_M, R_{ctb})$. 空间向量的拓扑关系分类为:

$$X =$$

$$\left\{ X \mid r(R_M, R_X) = \max(r), 0 \leq r(R_M, R_X) \leq 1 \right\}. \quad (7)$$

即比较相关度的值, 将空间向量归为相关度最大值所属的参考拓扑空间向量类.

2 基于 CAAS-FMA 分类的匹配算法小结

对上述的匹配算法进行总结, 得出给予 CAAS-FMA 分类的匹配算法的逻辑流程. 设有待匹配实体为 T , 最小外包矩形为 $\text{MBR}(T)$, 基于 CAAS-FMA 分类的面实体匹配算法的步骤如图 3 所示. 具体步骤为:

(1) 通过对比最小外包矩形的重叠面积比, 确定候选匹配实体集; (2) 将成分关联区域的度量因子——双向相交面积比率和双向相交周长比率加入

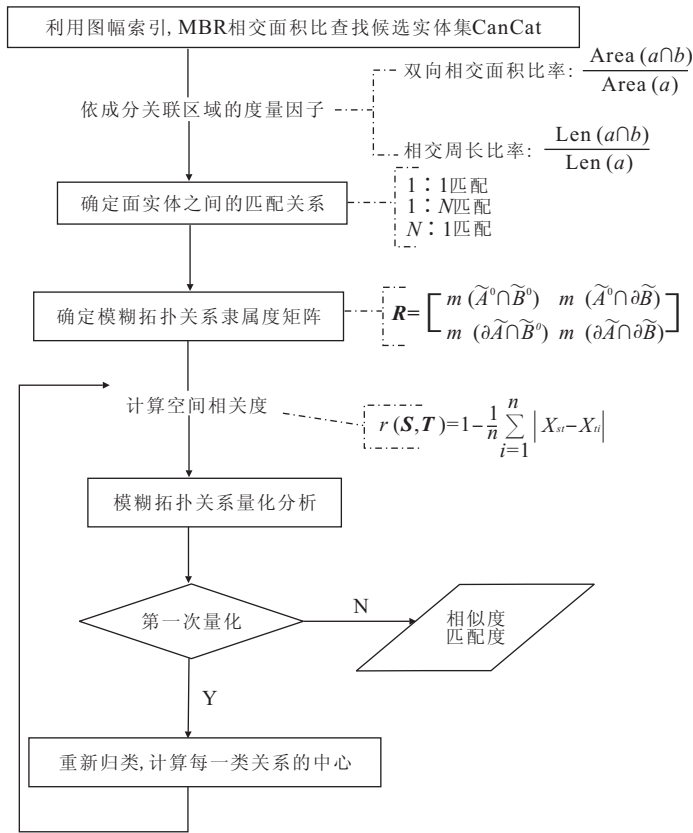


图 3 基于 CAAS-FMA 分类的面实体匹配算法的步骤

Fig. 3 The steps of the match algorithm between area object considering associated area similarities

到实体匹配中;(3)确定面实体之间的匹配关系,并将非一比一匹配进行转换和重组,均归为一比一匹配关系;(4)量化隶属度矩阵的分量值,确定并形成模糊拓扑关系的隶属度矩阵;(5)将模糊拓扑关系的隶属度矩阵定量化.利用分量相关度公式,分别计算矩阵与 8 种“参照拓扑空间向量”的空间相关度,取最大值,将模糊拓扑关系归为某一种特定的空间拓扑关系;(6)将初步结果进行归类,然后计算每一类的中心,以此产生新的参考空间拓扑关系向量;(7)重新计算相关度,再次分类,得出结果.

本文提出的基于成分关联区域相似度的多时态空间目标匹配算法具有如下优点:

(1)将成分关联区域的相似度度量因子引入空间目标匹配.对于数据更新问题,要素匹配时不仅要考虑源要素与目标要素相交面积的大小或比值,而且要顾及到源要素、目标要素以及两者交集之间的相关性.因此,笔者在空间匹配过程中重点考虑了成分关联区域的相似度.并综合考虑相交面积比率和相交周长比率两种因子,减轻了单一因子表达能力的不足,且计算复杂度也不大;(2)利用图幅索引,提

高候选实体集的选取效率,减少参与匹配的实体数量,避免了大范围搜索;(3)考虑双向匹配,可以处理非一比一的复杂匹配问题;(4)匹配结果中包含了所属关系的隶属度,为进一步分析确定面实体间的非增量匹配类型提供了依据.

3 检验实验

为检验上述算法的正确性,笔者以 1996 年和 1997 年土地利用的地类图斑数据为例,进行了空间目标匹配实验.数据说明如表 1 所示,实验结果如图 4.

实验表明,本文提出的基于成分关联区域相似度的面实体模糊匹配算法是正确可行的,并且对于非一比一匹配情况的处理比较令人满意.

表 1 实体匹配检验数据说明(1996—1997)

Table 1 The introduction to the data used in the test of objects match

源类名	比较类名	实体总数	匹配耗时	差异实体个数
DLTB1996	DLTB1997	1 728	5 s	20

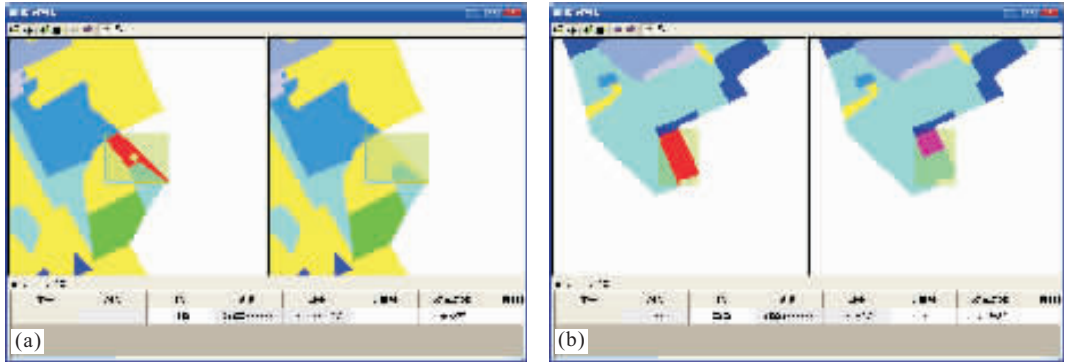


图 4 因实体合并产生的 $N:1$ (a)和因实体分割产生的 $1:N$ (b)匹配结果

Fig. 4 The result of $N:1$ match which caused by objects' combination (a) and $1:N$ match which caused by objects'split (b)

4 结论

本文利用成分关联区域相似因子,采用模糊数学的方法进行了空间匹配,下一步还将研究成分关联区域的其他相似因子.另外在将上述匹配结果应用到增量更新的过程中,还需要以模糊拓扑关系及匹配指标的计算结果为基础,进一步分析模糊拓扑关系的变化检测与分类,探讨地理实体的变化类型的自动推断方法.

References

- Foley, H., 1997. A multiple criteria based approach to performing conflation in geographical information systems. Tulane University, New Orleans.
- Fu, Z. L., Wu, J. H., 2007. Update technologies for multi-scale spatial database. *Geomatics and Information Science of Wuhan University*, 32(12): 1115—1118, 1148 (in Chinese with English abstract).
- Guo, Q. S., Du, X. C., Yan, W. Y., 2006. Geo-spatial reasoning. Science Press, Beijing (in Chinese).
- Hao, Y. L., Tang, W. J., Zhao, Y. X., et al., 2008. Area feature matching algorithm based on spatial similarity. *Acta Geodaetica et Cartographica Sinica*, 37(4): 501—506 (in Chinese with English abstract).
- Li, D. R., Gong, J. Y., Zhang, Q. P., 2004. Conflation of geographic databases. *Science of Surveying and Mapping*, 29(1): 1—4 (in Chinese with English abstract).
- Liu, Z. Y., 2006. The research on areal feature matching among the conflation of urban geographic databases (Dissertation). Hehai University, Nanjing (in Chinese).
- Wentz, E. A., 1997. Shape analysis in GIS. Proc. of ACSM/ASPRS. Seattle Washington, 204—213.
- Ye, Y. Q., Zuo, Z. J., Chen, B., 2006. Orient-entity spatial

data model. *Earth Science—Journal of Chinese University of Geosciences*, 31(5): 595—599 (in Chinese with English abstract).

- Zhang, L. P., Guo, Q. S., Sun, Y., 2008. The method of matching residential features in topographic maps at neighboring scales. *Geomatics and Information Science of Wuhan University*, 33(6): 604—607 (in Chinese with English abstract).
- Zhang, Q. P., Li, D. R., Gong, J. Y., 2004. Areal feature matching among urban geographic databases. *Journal of Remote Sensing*, 8(2): 107—112 (in Chinese with English abstract).

附中文参考文献

- 傅仲良, 吴建华, 2007. 多比例尺空间数据库更新技术研究. *武汉大学学报(信息科学版)*, 32(12): 1115—1118, 1148.
- 郭庆胜, 杜晓初, 闫卫阳, 2006. 地理空间推理. 北京: 科学出版社.
- 郝燕玲, 唐文静, 赵玉新, 等, 2008. 基于空间相似性的面实体匹配算法研究. *测绘学报*, 37(4): 501—506.
- 李德仁, 龚健雅, 张桥平, 2004. 论地图数据库合并技术. *测绘科学*, 29(1): 1—4.
- 刘志勇, 2006. 城市地图数据库合并中的面实体匹配研究(硕士学位论文). 南京: 河海大学.
- 叶亚琴, 左泽均, 陈波, 2006. 面向实体的空间数据模型. *地球科学——中国地质大学学报*, 31(5): 595—599.
- 章莉萍, 郭庆胜, 孙艳, 2008. 相邻比例尺地形图之间居民地要素匹配方法研究. *武汉大学学报(信息科学版)*, 33(6): 604—607.
- 张桥平, 李德仁, 龚健雅, 2004. 城市地图数据库面实体匹配技术. *遥感学报*, 8(2): 107—112.