第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·软件技术与数据库· 文章编号：1000—3428(2007)18—0051—02 文献标识码：A 中图分类号：TP311.5

空间数据融合技术的研究

彭煜玮，彭智勇
(武汉大学软件工程国家重点实验室，武汉 430072)

摘 要：空间数据融合技术是地理信息系统集成的重要组成部分，在多表现、多分辨率空间数据库中，该技术占据了重要的地位。该文介绍了几种主要的空间数据融合技术，并进行了比较和分析。结果表明了空间数据融合技术的有效性，并描述了空间数据融合技术发展前景。
关键词：空间数据库；空间数据融合；本体；地理信息系统集成

Research on Spatial Data Fusion Techniques

PENG Yu-wei, PENG Zhi-yong
(State Key Lab of Software Engineering, Wuhan University, Wuhan 430072)

【Abstract】 Spatial data fusion technique is an important part of GIS integration. In research area of multi-representation or multi-resolution spatial databases, spatial data fusion technique is important. Some techniques are proposed. But most of them stay on theoretical and experimental stages. There is no practical technique now. This paper introduces several spatial data fusion techniques and compares them. The results show the technique is effective. Spatial data fusion technique development in the future is described.
【Key words】 spatial database; spatial data fusion; ontology; GIS integration

在地理信息系统和空间数据库中，由于不同组织所采用的数据采集方式和存储空间的数据模型不同，因此即使表示同一组客观实体，各个单位所采用的数据格式和精度都不同。如何将这些不同数据格式和精度的空间数据整合起来，是一个很重要的问题，这也是空间数据融合技术所要解决的目标。另外，在表现空间数据库的研究中，将现有的、多个“单一表现”的数据库联系在一起，建立表示同一实体的不同空间对象联系，来维持一致性，这个过程依赖于空间数据融合技术。

空间数据融合技术的目标就是：将2个或多个空间数据集中，表示同一客观世界实体的空间对象标识出来，并将它们放在“融合集”中，一组表示同一客观实体的对象形成一个融合集。在信息集成技术中，数据融合的主要对象是结构化数据(如关系型数据)或者半结构化数据(如XML)。

在这2种情况下，对象具有一个全局标识符，也就是表示同一个客观实体的对象一定会具有相同的全局标识符，在这种条件下，数据融合显然是比较容易的。在空间数据的情况下，由于缺乏全局标识符，因此空间数据的融合也显得更加困难。

为了解决“空间数据中缺乏全局标识符”问题，研究者们提出了多种解决方案，大致可以分成以下3类：(1)基于Ontology的技术[1]；(2)基于非空间属性的技术[2]；(3)基于空间位置的技术[3]。

1 基于属性的空间数据融合技术

基于属性的空间数据融合技术是一种比较常见的技术。该技术的基本思想为：通过对空间对象的部分或者全部特征属性进行匹配，来决定2个或者多个空间对象的匹配程度。文献[2]提出了一种基于属性的方法，它可以从多个数据源中对空间对象进行匹配。匹配的过程分为3步，见图1。

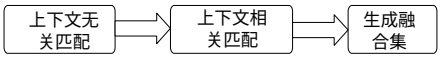

图1 匹配过程

1.1 上下文无关匹配

对于空间信息来说，上下文关系是非常重要的(也就是说2个对应的对象除了本身相似之外，其周边环境也应该非常相似)，但是从效率的角度来考虑，直接对数据集进行上下文相关的匹配，十分消耗时间和空间。因此，上下文无关匹配的作用就是：进行空间对象本身的相似性计算工作，将明显不可能对应的空间对象过滤掉，剩下的空间对象比原始数据集中的对象少得多，在此基础上，可方便地进行上下文相关匹配。由于计算上下文无关相似度只需要对空间对象的各属性进行简单的相似度计算，因此很多常用的相似度计量方法都可以被采用，例如字符串、标量、位置、形状(机器视觉领域里提供了丰富的比较形状的成果[4])等相似度计量方法。对象之间的上下文无关相似度通过计算各属性的相似度的加权平均值(权值表示属性的重要度)获得。但是，如果其中任意一个属性有明显的差异，那么仍然称这2个对象不能匹配。

1.2 上下文相关相似度

在实际应用中，可能存在2个(或多个)对象的位置、形状的相似度非常高，但是它们并不是相互对应的。因此，需要使用上下文来进一步进行相似度匹配[2]。上下文为：一个

基金项目：国家自然科学基金资助项目(60573095)；教育部新世纪优秀人才计划基金资助项目(NCET-04-0675)；教育部博士点基金资助项目(20050486024)；软件工程国家重点实验室开放基金资助项目(SKLSE05-01)
作者简介：彭煜玮(1980－)，男，博士研究生，主研方向：数据库系统，地理信息系统；彭智勇，教授、博士
收稿日期：2006-11-01 **E-mail：**sandsman@126.com

区域中空间对象之间的关系，通常有拓扑、距离和方位等。但是点状对象缺乏拓扑关系，Samal的研究使用距离和角度来定义对象的地理上下文。

在 Samal 的方法中，用邻接图来表示对象的上下文。邻接图是一个星形图，它表示居于中心的对象的上下文关系。所有在同一个数据集中的其他对象都通过一条边连接到中心。每条边通过其两端的节点间的距离以及角度来标记其权重，即每条边都是一个向量。

为了进一步减小开销，可以使用地标(所谓地标就是在同一块地理区域的不同表示中都能够被找到的对象)来引导匹配过程。将 2 个进行匹配的对象的邻接图中不是地标的其他对象及其对应的边删除，然后计算这 2 个邻接图之间的总的向量偏移量，其结果就是 2 个对象的上下文相关相似度。

1.3 2 个数据源的对象匹配

在得到上下文无关相似度和上下文相关相似度之后，可以将它们综合在一起来得到融合集。最简单的方法就是将二者平均，但相关的研究[2]表明，这会对地标的作用带来很大的影响。为了解决这个问题，文献[2]提出了一个宽松的迭代算法来处理源对象集。然后再使用上下文无关相似度和相关相似度来平均得到综合相似度。将综合相似度大于某个阈值(由用户指定)的对象放在一个融合集中。

在这一类方法中，需要融合的源数据集之间具有比较统一的模式，但实际中这种情况是很少出现的，由于不同数据源面向的是不同的应用需求，它们的数据模式很可能大相径庭，在这种环境下，这一类方法就很难能适用。

2 基于 Ontology 的空间数据融合技术

本体(Ontology)[3]的概念起源于哲学领域，是人类对自然界“存在论”的一种哲学观点，它意味着知识和知晓(knowing)。从知识共享的角度考虑，本体是对概念和关系的描述。

空间数据融合(乃至于更广泛的数据融合)的难点是：各组织对同一客观实体的理解和表示方式的不同，那么引入本体来统一这些不同的概念和认识，应该是一个不错的选择，如 F T Fonseca 等提出的本体驱动的地理信息系统的概念、H Uitermark 等提出的基于本体的地理数据集整合方法。

2.1 本体驱动的地理信息系统

Fonseca 和 Egenhofer 提出了本体驱动的地理信息系统(ontology-driven geographic information systems, ODGIS)的概念。ODGIS 在于从根本上解决不同组织对同一实体(概念)的认知不同这一问题，它基于的假设为：所有的有待于整合的系统都是基于本体来建立的。

ODGIS 基于从多种本体得来的软组件建立起来。这些软组件都是可以用来构建新应用的类。由于它们是从本体得来的，因此这些类中嵌入了从本体中抽取的知识。

参与融合的系统都是基于本体的，这样很容易判断来自于不同系统的对象是否表示同一客观实体(概念)。在这种假设之下，空间数据融合的问题将变得简单化。但是，ODGIS 的重点是：在知识生成的阶段，知识的生成、本体的定义都是需要领域专家参与的，并且需要进行大量的工作。

2.2 基于本体的地理数据集整合

H. Uitermark 等提出的基于本体的地理数据集整合方法的主要思想就是：制定一套在 GIS 领域里适用的领域本体，并在数据采集和存储的时候，将每个对象都和相应的本体相关联，这样，在数据融合的时候，与同一本体相关联的对象则可以视为是相互对应的，可以形成一个融合集。

基于本体的空间数据融合技术从原理上就是要为每个对象创建一个全局标识符，这样就使得融合过程变得非常容易。但是由于目前并没有比较成熟的本体库可用，而且要使目前已经建立的空间信息系统能够有效地融合在一起，需要大量的改造工作，因此基于本体的技术不能有效地缓解对空间数据融合的需求。

3 基于位置的空间数据融合技术

基于空间位置的技术是一类新颖的方法，只利用对象的位置来寻找到正确的融合集，成功率很高。C. Beeri等提出了几种基于位置的空间数据融合技术[4]。

3.1 片面最近邻居连接方法

片面最近邻居连接方法(one-sided nearest-neighbor join method)是商业地理信息系统中常用的方法。对于一个对象来说，在其他数据集中与之距离最近的对象就是它的对应对象，对应对象组成一个融合集。片面最近邻居连接方法是非对称的，即B与A的连接和A与B的连接得到的结果集，可能是不同的。C. Beeri等修改了片面最近邻居连接方法[4]，在结果集中加入了单一融合集的概念。如果一个对象找不到任何对应对象，那么它自身单独构成一个融合集(单一融合集)。

总之，片面最近邻居连接算法在一个数据集覆盖另一个时会达到比较好的效果。

3.2 相互最近方法

相互最近方法(mutually-nearest method)对片面最近邻居连接方法进行了改进。如果 2 个对象分别是另外一个对象的片面最近邻居，则称它们是一对相互最近对象。在相互最近方法中，为每一对相互最近对象都生成一个二元融合集。而为每一个不出现在任何二元融合集中的对象生成一个单一融合集。C Beeri 等还给出了融合集的可信度的定义，融合集的可信度是根据其中两个对象之间的距离以及它们和各自次近邻居之间距离计算得来。通过设置可信度阈值，可以过滤掉那些可信度偏小的融合集。

相互最近方法对于传统片面最近邻居连接方法的主要优势为：对于 2 个数据集的重叠程度不敏感。特别是，它在一个数据集覆盖另一个数据集的情况下，表现良好。

3.3 概率方法

在概率方法(probabilistic method)中，一个融合集的可信度依赖于集合中的对象是对应对象的概率，而这个概率则依赖于对象之间的距离。概率函数可以根据对象之间的距离、距离衰退参数 α 和距离上界来计算。在这种方法下，所有含有同一对象的融合集的可信度之和应该等于 1。与相互最近方法相似，通过可信度可以过滤融合集。

概率方法主要用于 2 个数据集的重叠较大的情况。

3.4 标准化权重方法

标准化权重方法(normalized-weights method)是概率方法的一个变种，它在重叠较小时效果比较好。在标准化权重方法中，使用概率方法定义的概率将权重(即可信度值)赋予每一个融合集。标准化权重方法将使用一个迭代算法标准化这些初始权重。该算法将在融合集之间产生相互的影响。

标准化权重方法在数据集之间的重叠较小时，能得到比较好的结果。但是，这些结果还是没有重叠较大情况下，使用概率方法得到的结果好，因为指派给单一融合集(即最后一行和最后一列)的权重并没有被标准化。所以，这些权重即使在它们应该几乎为 0 时，重叠也较大。 (下转第 55 页)