第16卷 第8期
2006年8月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.8
Aug. 2006

范例推理在地理信息系统中的应用研究

俞 磊,王 亮,王淑静,贾瑞玉
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:范例推理(CBR)是一种用先前求解问题的经验和方法,通过类比和联想来解决当前相似问题的推理技术,它是动态决策环境下求解不良结构问题的常用方法。GIS系统作为一种新兴的地学工具,具有很强的空间分析能力,但由于地学问题的复杂性,一些地学现象很难用确切的模型进行模拟和预测。考虑到范例推理系统在处理半结构化和非结构化问题方面的出色能力,文中探讨了一个基于范例推理的GIS系统结构,并给出了地理范例的构建方法和表达模型。

关键词:范例推理;地理信息系统;地理案例
中图分类号:TP39;P208 **文献标识码**:A **文章编号**:1673-629X(2006)08-0173-03

Case-Based Reasoning's Application in GIS

YU Lei, WANG Liang, WANG Shu-jing, JIA Rui-yu
(School of Computer Sci. and Tech. of Anhui Univ., Hefei 230039, China)

Abstract: Cass-based reasoning system is a very important machine-learning method. To explain a new object case, CBR system will search original case in the case-base, find an old case which is the most similar to the new object base. GIS, as a new geography tool, is famous for its powerful space analysis ability. But because of the complexity of the geography problems, it is difficult to construct an assured model to predict geography phenomena. Considering the fineness ability to solve the half-structured and non-structured problem of the CBR system, the paper discussed the structure of the CBR-based GIS system. It presents how to construct and how to explain a geo-case.

Key words: case-based reasoning, GIS; geo-case

0 引 言

范例推理(CBR)是一种由目标范例的提示而得到历史记忆中的源范例,并由源范例来指导目标范例求解的一种策略,它是一种重要的机器学习方法[1]。基于范例的问题求解方法非常适用于那些没有很强的理论模型,领域知识不完全、难以定义或定义不一致,需要依赖丰富经验的问题[2]。

目前关于CBR的研究主要集中在以下几个方面:范例的索引及检索技术;范例修正技术及其修正规则的获取方法;范例库的维护技术及其性能的研究;范例工程的自动化;范例推理的理论基础;范例推理与其它方法(包括学习技术、多Agent技术、推理方法、数据挖掘、数据仓库技术)的集成技术;范例推理的应用;研制CBR开发平台;CBR融合进大规模并行处理;基于Web的分布式CBR系统;可视化CBR技术及对话式CBR模型等。

地理信息系统是1963年由Roger F. Tomlinson提出来的,并于20世纪80年代开始走向成熟[3]。它是一种采集、存储、管理、分析、显示与应用地理信息的计算机系统,是分析与处理海量地理数据的通用技术。更严格地说,GIS是以数字化的形式反映人类社会赖以生存的现实世界的现势和变迁的各种数据,以及描述这些空间数据特征的属性;以模型化的方法来模拟地理系统空间研究对象的行为;在计算机软件、硬件支持下,以特定的格式支持输入/输出、存储、显示,以及进行地理信息查询、综合地学分析、辅助决策的有效的工具。但是在解决较为复杂的空间问题,尤其是半结构化和非结构化的空间问题的时候,GIS遇到了困难。一些模糊的、不确定的因素很难在系统中表述,也很难用一个非常准确的数学模型去描述它。

CBR系统为人们提供了一种解决空间问题的新思路。在地理学系统中,环境非常复杂,影响结果的条件繁多,甚至有些现象到现在还无法知道其本质和原理。在这种情况下,想利用一个原理模型来作出非常准确的预测和推理几乎是不可能的。而CBR系统作为一种经验式的问题处理系统不需要知道原理,只要有足够多的历史数据并从中抽取出源范例就可以得到较为准确的结果。而这一问题就求解的过程也是人类认识事物的最初方式,是符合人类认知学和心理学的。

下面主要论述CBR应用于地理信息系统的优势,并

收稿日期:2005-11-12
基金项目:安徽省教育厅科学研究资助项目(05010703,05050701)
作者简介:俞 磊(1982-),男,安徽定远人,硕士研究生,研究方向为决策支持在智能软件中的应用;贾瑞玉,硕士研究生导师,研究方向为智能计算与计算机图形学。

给出基于 CBR 的地理信息系统结构以及地理范例的构建方法和表达模型。

1 基于 CBR 的地理信息系统结构

1.1 CBR 应用于 GIS 的几点优势

单一的 GIS 虽然具有较好的空间处理和空间分析能力,但在决策支持方面,尤其是半结构化和非结构化问题上效果一般。

CBR 系统虽然善于引用已有经验解决问题,但在空间数据的组织上却不太合理,在空间数据的管理上显得凌乱,对空间数据的分析处理更显困难。

由此可见,CBR 和 GIS 两者在处理地学问题上是优势互补的,具体地说,将 CBR 应用于 GIS 的优势体现在以下几点:

(1)使用 CBR 可以避免知识理解和知识提取方面的困难[4]。

由于地学问题通常都是复杂的,有时候甚至呈现出极强的随机性。因此很难用一个确定的数学模型对一些地学现象进行模拟,或者即使有,也会由于实际情况比理想状况要复杂而使得模型的准确性大打折扣。这种情况下利用 CBR 就是最好的选择,因为它是一种先验式的问题处理方法,它解决新问题的时候依靠的是源范例(即经验),无需对现象的本质有太深的了解,只要根据领域专家的建议提取出若干个特征属性并对原始范例的特征属性及最终的问题解决方案进行索引并记录就可以建立起一个 CBR 系统的范例库。每当出现新问题的时候就从范例库中寻找与之相似的范例进行修正就可以得到一个理论上较为合理的解。

(2)使用 CBR 可以很方便地处理边界情况和特殊信息。

GIS 系统对于边界情况的处理不是很好。以气象为例,有如下两条规则:当空气湿度 $\geqslant w$,气温 t 到 T 摄氏度,且空气中有足够多的凝结核,会出现降雨;当空气湿度 $\geqslant w$,气温低于 t 摄氏度,且空气中有足够多的凝结核,会出现降雪。如果用 *GIS* 来处理这两条规则,它会非常严格地按照规则给出结果:气温 $\geqslant t$ 时候它给出的结果就是"降雨",气温 $< t$ 的时候就是"降雪"。但实际情况中在某个地域内,可能由于某些特殊的原因(如海拔高度)导致在气温 $= t$ 的时候也出现降雪,这就造成了预测结果与实际结果的不一致。而一旦引入 CBR,它虽然不知道有"海拔高度"这一特殊信息,但却可以根据以往的经验很好地处理这一边界情况,给出与实际情况相同的结果。其成功的原因就在于 CBR 隐含地处理了"海拔高度"这一特殊信息。

(3)使用 CBR 可以很好地处理模糊信息。

在地学问题中经常会出现模糊信息,这在单纯的地理信息系统中是很难处理的,比如领域专家在描述下雨的时候可能会用:在城市 A,3 月到 4 月间,当天温度低于 20 度的时候会出现小雨天气。这里"小雨"这个概念在没有统一标准的情况下就很难在 GIS 中界定究竟降雨量多少才算是小雨,而在 CBR 系统中就很好表示,只需添加一个专用的特征属性即可。

1.2 基于 CBR 的 GIS 的系统结构

CBR 与 GIS 的集成系统如图 1 所示,它包括以下几个模块:用户界面;CBR 中的范例推理模块和范例库;GIS 空间数据分析处理模块和空间数据库模块。其中用户界面和 GIS 的数据处理和分析模块直接联系,用于将结果呈现给用户,并将用户的修改意见等反馈给系统以维护范例库。GIS 的数据处理和分析模块与范例库和 GIS 的空间数据库连接,用于提取范例进行处理,并根据空间数据库的数据通过用户界面将系统建议的问题处理结果以最直观的方式呈现给用户。范例推理模块与范例库直连,用于维护范例库,同时又与 GIS 的数据处理和分析模块连接,两者结合得出问题处理结果。

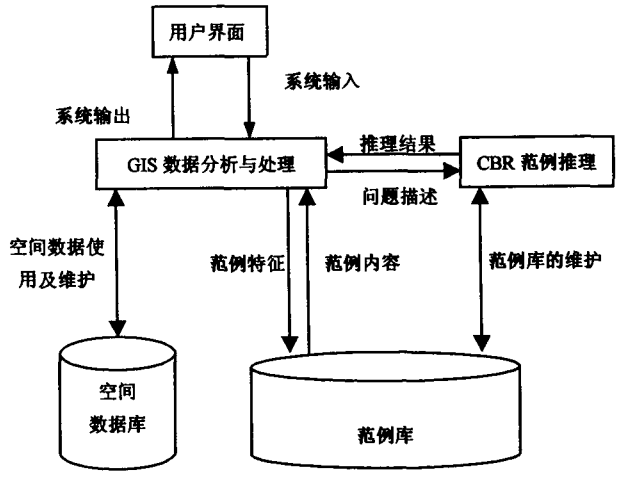

图 1 基于 CBR 的地理信息系统结构

2 地理范例的构建与表达

上文论述了基于 CBR 的 GIS 的系统结构,可以看出范例库是非常重要的一个部件。整个系统成功与否很大一部分就要看范例库构建的是否合理。但是这里的范例与一般的范例又有所不同,因为它牵涉到空间位置的表达。下面就重点讨论一下地理范例的构建和表达。

2.1 地理范例的定义

地理范例从字面上理解,就是包括地理信息的范例,这是一个不精确的定义。一般采用如下定义:"地理范例是具有空间定位的范例的统称,其反映或表达了某种或某类地理现象或事件"[5]。从上面的定义可以看出,包含空间定位信息的范例不一定就是地理范例。例如,某城市要建一个广场,这个范例里肯定包含有广场的选址这一空间定位信息,但这不能称为地理范例,因为这个广场的选址在哪里并不是由地理因素来决定的,它的选定更多地是依赖于诸如文化、经济等方面的因素。

而在地矿探索问题中,某种矿物的分布的决定因素就

是地理环境,这才是典型的地理范例。

2.2　地理范例的构建

根据上面的地理范例的定义,就可以着手建立地理范例。从问题—方法—结果这三方面入手进行范例的构建,在这样的总体框架下,给出了地理范例的构建流程,如图 2 所示。

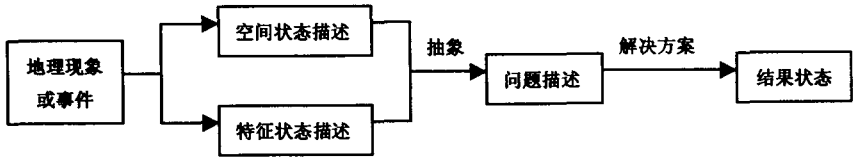

图 2　地理范例的构建流程

由图可以看出:一个地理现象或事件发生之后,应该先对其空间状态和特征状态进行提取、抽象,随之得到一个最有效、最简洁的问题描述。然后再处理解决方案和结果状态(在这里就是将原始范例的解决方案和结果状态记录到范例中)。至此一个范例就建立完毕,重复以上的过程就可以建立起一个范例库。

2.3　地理范例的表达

上文的地理范例的构建仅仅讨论了一个范例的逻辑结构,并没有给出一个计算机能够接受的范例表达。传统的范例表达是利用由领域专家提供的一组特征的定量描述,并没有体现出空间的信息。一个完整的地理范例表达应该包括空间信息和属性信息,范例的表达模型应该包括:

(1)空间信息的表达。

假设范例发生的真实空间是一个 n 维的空间,每一维都有界,最大值分别为 Max1,Max2,Max3,…,Max$n$,对应于各最大的位数分别为 $k1,k2,k3,\cdots,kn$,则 n 维空间中某个单元(Val1,Val2,Val3,…,Val$n$)的空间编码可以定义为:Space - address = Val1′Val2′Val3′…Val$n$′。

Val1′Val2′Val3′…Val$n$′ 分别为 $k1,k2,k3,\cdots,kn$ 位的整型数值,是补 0 补够 kn 位的 Val1,Val2,Val3,…,Val$n$ 的值[5]。举例如下:

某范例发生在 2 维空间中,x,y 方向的最大值分别为 500,200,对应的位数分别为 3,3,则对于空间任意一点有空间编码如下:

$$\text{Space - address} = x'y' = \begin{cases} 00x00y & 0 < x < 10, 0 < y < 10 \\ 0x0y & 10 < x < 100, 10 < y < 100 \\ xy & 100 \leqslant x \leqslant 500, 100 \leqslant y \leqslant 200 \\ 500200 & x = 500, y = 200 \\ 000000 & x = 0\ y = 0 \end{cases}$$

依次类推可以得到日常生活中的三维空间的编码。如果再加上时间就是一个真实的四维世界的精确定位。

空间编码在地理范例中是最为重要的一环,一个好的地理范例编码可以很方便地取到解决问题所需要的范例空间,减少范例检索的范围,提高范例检索的效率。

(2)属性信息的表达。

这里提出一种基于 Tesseral 的属性信息表达方法。在地理范例系统中,要考虑到范例之间的依赖关系和各属性对结果的影响的大小不同即各属性的权重不同。有些学者认为可以在范例抽取算法中将权重体现出来。但笔者认为权重最好还是放在特征属性中比较好,因为这样可以使得系统有较好的可移植性。不必每换一个环境就要修改算法。具体的表达如下:

假设地理范例的属性为 $A_1,A_2,\cdots,A_n$。它们的取值类型各不相同,通过一定的转换将每个属性的取值范围都映射到区间 $\{-Vmax,\cdots,0,\cdots,Vmax\}$ 上[6]。比如,值域为 $\{Vmin,\cdots,Vmax\}$,经过变换可以将值域映射到 $\{-(Vmax-Vmin),\cdots,0,\cdots,(Vmax-Vmin)\}$ 上,如果属性值是枚举型,即 $\{V_1,V_2,\cdots,V_n\}$,则将值域映射到 $\{-n,\cdots,0,\cdots,n\}$ 上。这样每个属性都看作是 n 维范例空间的一维[7],然后采用公式 $D_{An} = \log_2 Vmax$ 给每个属性分配一个整型类型位数。处理完这 n 维空间之后再加一个维数 D_{n+1},这第 $n+1$ 个维用于记录前面 n 维的权重,表达为:$D_{n+1} = \{W_1,W_2,\cdots,W_n\}$。这其中的 $W_1,W_2,\cdots,W_n$ 由领域专家给出意见确定权重。例如:

设有一个 n 维的属性空间 $D_1,D_2,\cdots,D_n$ 将被定义为:

$D_1: =<-Vmax1, Vmax1>$

$D_2: =<-Vmax2, Vmax2>$

…

$D_n: =<-Vmaxn, Vmaxn>$

$D_{n+1}: =<W_1,W_2,\cdots,W_n>$

对应的属性空间的位模式如下:

$D_{An}(\log_2 V\min n\text{bits})$ … $D_{A2}(\log_2 V\min 2\text{bits})$ $D_{A1}(\log_2 V\min 1\text{bits})$

采用这种模式,实际上表达了一个 n 维空间。地址计算如下:

$$\text{address} = \sum_{i=1}^{n} V_i * \text{Base}_i$$

其中 V_i 为每一个属性的取值,$\text{Base}_i = 2^{\sum_{i=0}^{i-1} b_i}$,$b_j$ 是分配给属性 A_j 的位数。

利用上面的公式就可以根据属性的值计算出一个唯一的 n 维范例空间的地址。然后根据这个地址抽取其权重就可以进行范例推理过程。

3　结束语

文中论述了 CBR 系统原理及其在与地理信息系统结合上的优势,给出了基于 CBR 的地理信息系统的系统结构。论述了地理范例的构建和表达,提出了在地理范例的表达中采用基于 Tesseral 的方法来表达其属性信息并主张在属性信息中加入权重的字段,以提高系统的可移植性。但是在权重确定上是依赖于领域专家的建议,在以后

(下转第 178 页)

应该与原来的八元树一样。对于形式一,在恢复时根据 status=0 或 1 判断是否为中间节点或叶节点,若为中间节点,则开辟 8 个孩子节点,直至叶节点,即按照它的存入的顺序来构造八元树。对于形式二,根据每个存入叶节点的八元码来恢复八元树。

具体的恢复算法如下:

(1)形式一:

①初始化;

②打开数据文件 filename;

③若至文件尾则转⑤,否则转④;

④读入磁盘中一个节点的信息,判断取出的状态信息 status 是 0 还是 1,若 status=0,则开辟 8 个孩子节点空间,且指针指向其孩子节点;若 status=1,则取出的数据信息存入当前所指的节点中,重复前面的步骤;

⑤关闭数据文件,恢复过程结束。

(2)形式二是通过读入文件中的存储信息,根据读出的八元码开辟节点空间,按照树的先根遍历方法和八元码 $O_{L-1}O_{L-2}\cdots O_i\cdots O_0$ 访问八元树中各层的节点,当遇到所访问的节点不存在时创建一个新的节点,直到访问到树的最低层,这个节点必然是叶节点,把该节点的数据域用文件中的数据代替。

文中实验使用头部的 109 幅核磁共振断层图像(尺寸为 256×256),如图 5 所示(其中 6 幅)。取阈值为 25,即图像上灰度值大于 25 的点都被取为物体的体素。总共建立了 266708 个 Octree 型节点(叶节点包括原来的图像亮度信息和密度梯度信息),用形式一所存储的磁盘文件大小为 3 733 912 个字节,用形式二所存储的磁盘文件大小为 4 978 554 个字节。程序用磁盘文件恢复八元树的时间约为第一次建立八元树时间的十分之一,因为第一次建立八元树需要计算八元码和每个叶节点的密度梯度信息,比较花费时间。利用恢复的八元树进行三维显示如图 6 所示,从得到的实验结果可以看出是正确的。

图 5 人头部的 109 幅 MRI 图像的其中 6 幅

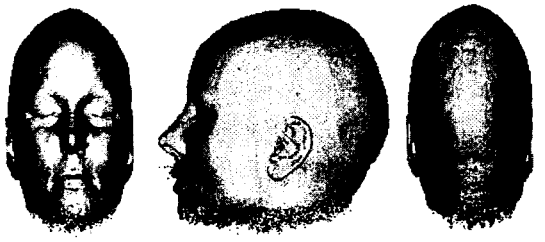

图 6 人头部三维显示的结果

4 结 论

文中根据八元树空间表达的特点,提出一种十分有效的磁盘存储和恢复算法。因此,在基于八元树的三维重建过程中,不再需要从原始数据计算生成八元树,而是快速地从磁盘文件恢复出八元树来进一步处理。文中编程实现了八元树的两种形式的磁盘存储和恢复,获得了正确满意的结果。

参考文献:

[1] Boada I. An octree-based multiresolution hybrid framework [J]. Future Generation Computer Systems, 2004, 11(20): 1275-1284.

[2] Yemez Y, Schmitt F, Multilevel representation and transmission of real objects with progressive octree particles[J]. IEEE Transactions on Visualization and Computer Graphics, 2003, 9(12): 551-569.

[3] Meagher D J. Geometric modeling using octree encoding[J]. computer graphics and image processing, 1982, 19: 129-147.

[4] 赵海峰,罗 斌. 一种改进的八元树三维目标表示方法[J]. 计算机工程与应用,2005,41(29):8-10.

[5] 赵海峰,束学斌. 一种有效的序列断层图像的八元树构造算法[J]. 中国图像学报,1999(5):418-422.

[6] 刘成君,戴汝为. 广义线性八元树表示及物体的广义三维重建[J]. 自动化学报,1997,23(5):694-697.

[7] 罗 斌,汪炳权. 基于 PC 机的断层图象序列 3D 表面重建[J]. 电子科学学刊,1993,15(1):75-78.

(上接第 175 页)

的研究中可以考虑应用数据挖掘的技术来确定权重。

参考文献:

[1] 杨善林,倪志伟. 机器学习与智能决策支持系统[M]. 北京:科学出版社,2004.

[2] Liao T W, Zhang Z, Mount C R. Similarity measures for retrieval in case-based reasoning systems[J]. Applied Artificial Intelligence, 1998, 12: 267-288.

[3] 汤国安. 地理信息系统[M]. 北京:科学出版社,2000.

[4] 叶嘉安. 基于案例的推理和 GIS 相集成的技术在规划申请审批中的应用[J]. 城市规划会刊,2001(3):35-36.

[5] 杜云艳. 地理案例推理及其应用[D]. 北京:中国科学院,2001.

[6] Jone. Model-Based case adaption[A]. In Proc of AAAI-92 [C]. San Jose, CA: [s.n.], 1992. 673-678.

[7] Holt A, Benwell G L. Applying Case-based reasoning technique in GIS[J]. IJGIS, 1999, 13(1): 9-25.