# 一种多策略 GML 应用模式匹配方法

张维蔚 李 超 曾 骁 熊 璋

(北京航空航天大学 计算机学院, 北京 100083)

摘 要: 地理标记语言(GML, Geography Markup Language)应用模式匹配是实现基于GML的地理信息共享的基础问题. 结合现有模式匹配算法和 GML3.0 应用模式的特点, 提出了一种多策略 GML 应用模式匹配方法. 首先, 将输入的 GML 应用模式转化为 GML 模式树, 通过引入基于语言学和基于约束的具体匹配规则, 分别对 GML 模式树进行元素对相似度计算, 并对 2 种计算结果加权合并获得元素水平匹配的相似度值; 然后, 通过基于相似度传播的结构匹配算法对元素的相似度进行修正; 最后, 获得 2 个输入模式的元素匹配映射表. 实验表明, 该算法能够提高 GML 应用模式的元素匹配正确率.

关 键 词: 地理标记语言; 模式匹配; 树状结构; 语言学匹配; 约束匹配; 结构匹配

中图分类号: TP 311

文献标识码: A 文 章 编 号: 1001-5965(2008)05-0520-04

## Multi-strategy GML application schema matching approach

Zhang Weiwei Li Chao Zeng Xiao Xiong Zhang

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

**Abstract** GML(geography markup language) application schema matching is a basic problem in the GML-based share of geography information. Combining the existing pattern matching algorithms and the characteristic of GML3.0 application schema, a multi-strategy GML application schema matching approach was proposed. Firstly, the input GML application schemas were transformed into the tree structure. The specific matching rules including linguistics and constraint were introduced which would help to compute the similarity between two elements, and then the two results were merged into element-level similarity. Secondly, the element-level similarity was modified by a structure-level matching algorithm based on similarity flooding theory. Finally, a mapping table of GML application schema elements was obtained. Experiment results show that the approach can effectively discovery the element similarity, and it can improve the matching result with a high degree of accuracy.

**Key words** GML(geography markup language); pattern matching; trees; linguistics; constraint theory; structure

随着计算机多媒体技术、网络通信技术以及空间信息技术的发展, 建立基于万维网的地理信息系统(GIS, Geography Information System)已成为国内外地理信息系统发展的主要趋势, 几大厂商都竞相发布了各自的产品. 但是, 由于不同开发商之间没有统一开发标准, 他们建立了各自独立的数据结构, 要实现数据共享, 首先必须完成不同数据格式的转换工作. 但因缺乏对空间对象的统一描述, 转换后经常会出现信息丢失的现象.

为了解决多源异构空间数据的共享问题, 开放地理系统联合会(OGC, OpenGIS Consortium)制定了用于地理信息建模、存储和传输的编码标准——地理标记语言(GML, Geography Markup Language)[1]. GML定义了一种通用的数据格式, 不同软件可通过这种数据描述方式进行通信, 从而使地理信息能够在不同领域实现语义共享.

收稿日期: 2007-05-09

作者简介: 张维蔚(1982- ), 女, 辽宁本溪人, 硕士生, vavele@hotmail.com.

GML采用 XML Schema定义各种地理要素, 它提供了一些基本模式作为用户建模的元模式, 用户可选择必要的元模式来构造自己的应用模式. 由于空间数据来源广泛、结构复杂, 并且 GML允许用户灵活建模, 导致不同用户对同一地理要素定义的应用模式可能在命名空间、数据类型、建模结构等方面存在较大差异. 因此, 要真正实现基于 GML的地理信息共享, 必须要解决 GML应用模式的匹配问题[2].

本文以 GML 3.0为基础, 提出一种多策略的 GML应用模式匹配方法, 该方法综合了元素水平和结构水平两种模式匹配方法, 引入基于语言学和基于约束的具体匹配规则, 并充分考虑 GML模式树中相邻节点的相互影响, 采用了基于相似度传播的结构匹配方法对相似度进行修正.

## 1　相关工作及技术

### 1.1　模式匹配

模式匹配过程可以简单概括为: 输入 2个模式, 采用一定的算法进行元素匹配, 输出 2个模式间的元素映射. 现有的匹配方法根据匹配对象不同可分为模式水平匹配与实例水平匹配, 元素水平匹配与结构水平匹配; 根据匹配方法不同可分为基于语言学的匹配与基于约束的匹配[3]. 图 1为模式匹配方法的分类示意图.

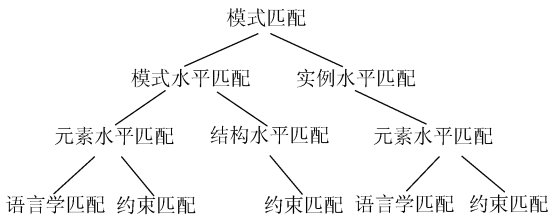



图 1　模式匹配方法分类

模式匹配可以基于一种单独的匹配算法构造匹配器. 但是, 由于每种单独的匹配算法都具有自身的局限性, 因此, 基于多种匹配规则构建混合匹配器, 或对多个匹配器的匹配结果进行加权合并成为目前的研究热点. 此外, 在实际的模式匹配过程中, 还用到大量的辅助信息, 如数据字典、知识库、用户输入以及匹配结果重用等.

### 1.2　GML应用模式匹配

该规范于 2000年 4月正式推出. GML 1.0以文档类型定义 (DTD, Document Type Definition)描述各种地理要素. GML 2.0于 2001年 2月推出, 开始以 XML Schema定义各种地理要素及其属性. GML 2.0只包含 3个元模式, 主要关注简单地理要素. GML 3.0于 2003年 1月正式发布, 元模式增加到 32个, 除了定义简单要素, 还能够描述 2D线型要素以外的地理空间要素, 包括复杂、非线性的 3D要素, 带有拓扑的 2D要素, 时态要素, 动态要素和层; 增加了对复杂的几何实体、拓扑、空间参照系统、元数据、时间特征和动态要素等的支持[4].

目前, 国内外有关 GML应用模式匹配的研究方法大多基于 GML 2.0[2,4-5]. 由于 GML 2.0建模机制较为简单, 对同一地理要素建模的主要差别在于元素命名和数据类型, 因此相应的匹配方法大多局限于元素水平[2,5]. 与以往的版本相比, GML 3.0提供了一套更加丰富的基本标签、公共数据类型以及用户构建应用模式的机制, 这使得基于 GML 3.0对同一要素的建模不但在元素命名和数据类型方面存在差别, 在要素的组织方法上也可能有所不同, 因此仅采用元素水平匹配方法已经无法满足复杂的匹配需求. 文献[4]提出一种结构匹配方法, 该方法首先设定叶节点的相似度为其语言学相似度, 进而通过比较两个节点的子节点相似度来确定节点对的相似度. 该方法虽然在一定程度上挖掘了元素之间相似度的影响, 但是只单方向考虑到子节点对父节点的影响, 而忽略了父节点对子节点的影响.

对于复杂的 GML 3.0应用模式匹配, 有必要全面考虑元素水平和结构水平的影响因素, 并且需要在数据库中预存储大量 GIS领域的辅助信息和 GML 3.0元模式信息.

## 2　GML应用模式匹配算法

本方法提出的 GML应用模式匹配过程如图 2所示, 其核心是构建一种混合匹配器, 采用树状结构作为匹配模型, 分别计算基于语言学和基于约束的元素对相似度, 并对 2种计算结果加权合并获得元素水平匹配的相似度值; 然后, 采用一种基于相似度传播的结构匹配算法进行相似度修正; 最后输出元素匹配映射表.

### 2.1　GML模式树

由于 GML继承了 XML的文档特性, 因此选择树型结构作为 GML应用模式匹配的模型[4], 把这种 GML模式树相应地记为 GSTree. GSTree的根节点表示 GML应用模式的根元素, 叶节点表示 GML应用模式中不包含其它对象的元素, 叶节点的数据类型为 GML基本数据类型. GSTree中父节点与子节点之间的连接关系表示了相应元素在 GML应用模式中的“包含”关系. 采用 GSTree能

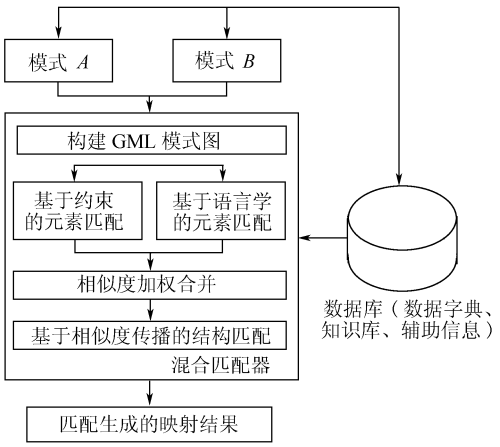



图 2 GML应用模式匹配过程

够消除环的影响, 避免产生相似度无限传播, 并且每个节点至多有一个父节点, 保证了从根节点到某一叶节点有唯一的路径.

一棵典型的 GSTree如图 3所示, 在这里只简单地为“要素集”元素定义一个“要素”子元素. 在实际应用中, 可能存在多个子元素, 按照同样的规则为“要素集”增加子节点就可以构造相应的 GSTree.

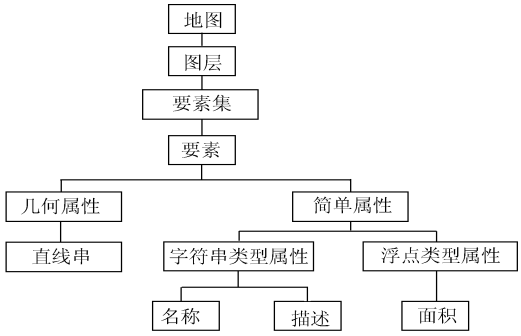



图 3 一棵典型的 GML模式树

## 2.2 基于语言学的元素匹配

基于元素名称, 从语义匹配以及字符串匹配两方面引入具体的匹配规则[4-5].

定义 **1** 二元双向关系符“≌”表示 2个元素之间的匹配关系; 根据规则 R 计算得到的元素 $e_1$ 与 $e_2$ 的相似度记为 $\lambda_R(e_1, e_2)$.

定义 **2** 如果输入数据不满足规则 R 所定义的约束条件, 则称为规则 R 失效.

1) 语义匹配规则

规则 **1** 在相同命名空间内, 如果元素 $e_1$, $e_2$ 具有相同名称, 则 $\lambda_{R1}(e_1, e_2) = 1$; 否则, $\lambda_{R1}(e_1, e_2) = 0$

规则 **2** 在不同命名空间内, 如果元素 $e_1$, $e_2$ 的名称经过专有名称预处理后相同, 则 $\lambda_{R2}(e_1, e_2) = 1$; 否则, 规则 2失效.

规则 **3** 在不同命名空间内, 如果元素 $e_1$, $e_2$ 的名称经过同义词预处理后相同, 则 $\lambda_{R3}(e_1, e_2) = S$; 否则, 规则 3失效.

规则 **4** 在不同命名空间内, 如果元素 $e_1$, $e_2$ 的名称经过近义词预处理后相同, 则 $\lambda_{R4}(e_1, e_2) = H$; 否则, 规则 4失效.

上述规则需要建立地理信息领域的数据字典、专业名称词汇表以及同义词和近义词库作为支持. 其中, $S$和$H$ 分别为用户输入的同义词相似度和近义词相似度, 且 $S > H$.

2) 字符串匹配规则

规则 **5** 设 $s_1$, $s_2$ 分别表示元素 $e_1$, $e_2$ 的名称字符串, 则 $\lambda_{R5}(e_1, e_2) = 1/F(s_1, s_2)$, 其中函数 $F(s_1, s_2)$表示 $s_1$ 与 $s_2$ 的编辑距离, 基于 Levenshtein Distance算法实现[6].

图 4说明了上述 5条规则的优先级关系和组合关系. 当优先级较高的规则生效时, 不再引入优先级较低的规则; 当规则 3或规则 4生效时, 引入规则 5进行组合计算; 对于 2条进行组合的规则, 如果其中之一失效, 则该组合失效, 否则相似度取其中的最大值, 例如: 当规则 4生效时, 相似度取值为 $\max(H, \lambda_{R5}(e_1, e_2))$.

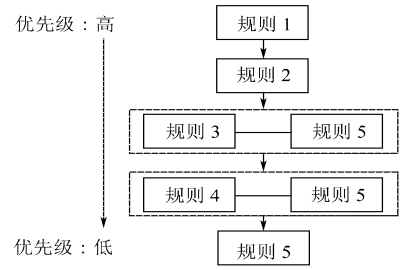



图 4 规则优先级和组合关系

## 2.3 基于约束的元素匹配

模式文档通常包含一些约束来定义元素的关键标识、数据类型、取值范围、唯一性以及可选性等特性. 结合 GML应用模式的特点, 需要构建基于 XML Schema和 GML 3.0的基本标签库和基本数据类型库. 匹配规则如下:

规则 **6** 在相同命名空间内, 如果元素 $e_1$, $e_2$ 的数据类型和标签类型都相同, 则 $\lambda_{R6}(e_1, e_2) = 1$; 否则, $\lambda_{R6}(e_1, e_2) = 0$

规则 **7** 在不同命名空间内, 如果数据类型不为空的 2个元素 $e_1$, $e_2$ 的数据类型相同, 则 $\lambda_{R7}(e_1, e_2) = 1$; 否则, 规则 7失效.

规则 **8** 在不同命名空间内, 如果元素 $e_1$, $e_2$ 的标签类型相同, 则 $\lambda_{R8}(e_1, e_2) = 1$; 否则,

$\lambda_{R8}(e_1, e_2) = 0$

上述 3条规则中, 规则 6具有最高优先级, 而规则 8的优先级最低. 同样, 只有当优先级较高的规则失效时才引入优先级较低的规则.

2 4　相似度加权合并

基于约束的元素匹配方法通常与其它匹配方法结合使用, 有助于限制候选匹配的数量[3].

设元素 $e_1$, $e_2$ 分别为输入模式 $A$ 和模式 $B$ 的 2个节点, 基于语言学匹配方法计算得到的相似度为 $L$, 基于约束匹配方法得到的相似度为 $C$, 用户输入的权值为 $\omega$, 则有

$$\lambda(e_1, e_2) = \omega \cdot L + (1 - \omega) \cdot C \qquad (1)$$

2 5　基于相似度传播的结构匹配

元素水平的模式匹配方法单纯地对 2个输入模式的元素对独立地进行相似度计算, 而忽略了元素之间相似度的相互影响. 在 GSTree中, 元素节点之间具有丰富的"包含"与"被包含"关系, 某特定节点对的相似度变化可能引起其父节点对、子节点对的相似度变化, 因此利用结构匹配算法对元素水平的匹配结果进行相似度修正.

相似度传播的思想来源于 Similarity Flooding 算法[7], 文献 [8]基于该思想提出一种通用的结构匹配算法 (SM, StructureMatching), 包括一种基于有向图的传播机制和相似度修正方法. GSTree与普通有向图的主要区别在于: 根节点到某一叶节点之间具有唯一路径, 在该路径上的各个节点间的相似度相互影响. 本文在 SM 算法上进行改进, 使其适用于 GSTree的树状结构.

定义 **3**　对于 GSTree中的任一节点 $e$, 其父节点表示为 $P_e$, 子节点表示为 $C_e$.

对于 2个 GSTree, 其节点对 $(e_1, e_2)$之间的初始相似度值为 $\lambda(e_1, e_2)_0$, 即基于元素水平匹配方法计算得到的相似度, 则 $(e_1, e_2)$在第 $k(k>0)$次循环中的相似度值可由下面公式计算得到:

$$\lambda(e_1, e_2)_k = \theta\lambda(e_1, e_2)_{k-1} + \theta_P \lambda(P_{e_1}, P_{e_2})_{k-1} + \theta_C \lambda(C_{e_1}, C_{e_2})_{k-1} \qquad (2)$$

其中 $\theta$, $\theta_P$, $\theta_C$ 为用户输入的权值, 并且满足 $\theta + \theta_P + \theta_C = 1$.

可以看出, 节点对 $(e_1, e_2)$的相似度由其前一次循环的相似度和其父节点对、子节点对的相似度共同决定. 当循环计算满足 (3)式时, 终止循环, 其中 $\varepsilon$为用户输入的精度阈值.

$$\lambda(e_1, e_2)_k - \lambda(e_1, e_2)_{k-1} < \varepsilon \qquad (3)$$

通过循环计算, 节点对 $(e_1, e_2)$的相似度传播到其所在的整条路径中.

经过相似度修正后, 可能存在多个候选匹配, 在候选匹配中选择相似度最大者作为输出. 为了提高匹配精度, 可以定义相似度阈值, 只有相似度大于该阈值的匹配才作为最终的正确匹配.

## 3　实验结果及分析

在原型系统中实现了上述算法, 并使用实际数据进行了实验. 选择 6组实验数据, 每组数据包括两个描述城市地理信息 GML应用模式. 采用元素匹配的正确检测率作为评价标准, 即系统检测的正确匹配数与实际正确匹配数的比率. 实验结果如图 5和图 6所示.

由实验结果可知, 当相似度阈值为 0 7时, 仅采用元素水平匹配方法, 正确检测率的平均值为 71 3%, 引入结构水平匹配方法进行相似度修正值后, 平均正确检测率增加到 94 5%; 当相似度阈值提升到 0 8时, 仅采用元素水平匹配方法时, 平均正确检测率为 67 3%, 修正后的平均正确检测率增加到 92 6%. 实验结果表明, 采用多策略 GML应用模式匹配方法能够有效发现 GML应用模式的元素匹配关系.

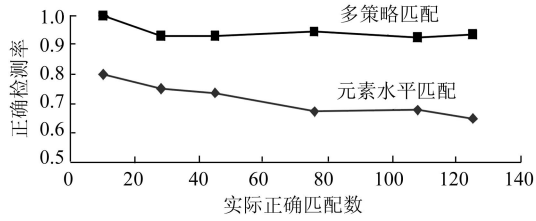



图 5　相似度阈值为 0 7的实验结果

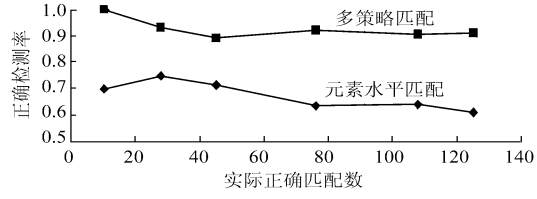



图 6　相似度阈值为 0 8的实验结果

## 4　结束语

本文在分析 GML 3 0应用模式特点基础上, 针对 GML应用模式匹配问题, 提出一种多策略匹配方法. 该方法采用 GSTree作为模式匹配模型, 分别采用基于语言学和基于约束的匹配规则计算元素对的相似度, 对 2种匹配结果进行加权合并. 考虑到相邻节点之间相似度的影响, 采用一种基于相似度传播的结构匹配方法对元素对的相似度进行修正.

(下转第 **528**页)

矩与前两种轮齿车轮接近.

## 4 结束语

利用 DEM,仿真分析了 3种不同轮齿截面的车轮结构与土壤颗粒的细观接触作用,绘制出车轮扰动下土壤颗粒单元的接触力线分布图.通过对离散元细观分析结果的坐标转换和代数求和等数学处理方法,推导出了土壤推力、行驶阻力、挂钩牵引力和驱动转矩等宏观车轮牵引性能参数的函数表达式,并绘制了车轮牵引性能参数-时间步曲线.

仿真分析结果表明,直角梯形截面齿车轮与矩形截面齿车轮和三角形截面齿车轮相比,具有最大的挂钩牵引力,因而其牵引性能最佳.

### 参考文献 (References)


[1] Nakashima H, Fujii H, Oida A. Parametric analysis of lugged wheel performance for a microrover by means of DEM [J]. Journal of Terramechanics, 2007, 44: 153-162

[2] Shibly H, Iagnemma K, Dubowsky S. An equivalent soil mechanics formation for rigid wheels in deformable terrain, with application to planetary exploration rovers[J]. Journal of Terramechanics, 2005, 42: 1-13

[3] Patel N, Scott G P. Application of bekker theory for planetary exploration through wheeled, tracked and legged vehicle locomotion[R]. AIAA-2004-6091, 2004

[4] 郑永春,欧阳自远,王世杰,等. 月壤的物理和机械性质[J]. 矿物岩石, 2004, 24(4): 14-19
Zheng Yongchun, Ouyong Ziyuan, Wang Shijie, et al. Physical and mechanical properties of lunar regolith[J]. Journal of Mineralogy and Petrology, 2004, 24(4): 14-19 (in Chinese)

[5] Zhang D, Whiten W J. A new calculation method for particle motion in tangential direction in discrete element simulations [J]. Power Technology, 1999, 102: 235-243

[6] Liu Sihong, Lu Tinghao. Microscopic shear mechanism of granular materials in simple shear by DEM [J]. Chinese Journal of Geotechnical Engineering, 2000, 22(5): 608-611

[7] Oida A, Momozu M. Simulation of soil behavior and reaction by machine part by means of DEM [J]. Agricultural Engineering International: the CIGR Journal of Scientific Research and Development, 2002, 4: 1-7

[8] Asaf Z, Rubinstein D, shmulevich I. Determination of discrete element model parameters required for soil tillage[J]. Soil & Tillage Research, 2007, 92: 227-242


(责任编辑:张 嵘)

(上接第 523页)

实验结果表明,该匹配方法能够有效地发现不同 GML应用模式间的元素匹配关系,具有较高的正确匹配率,可以广泛应用于基于 GML的地理信息融合应用中.

### 参考文献 (References)


[1] ISO 19136, Geographic information-Geography Markup Language(GML) [S]

[2] Guan Jihong, Zhou Shuigeng, Chen Junpeng. Ontopology based GML schema matching for spatial information integration[C]// Lee J W T. Proceedings of the Second International Conference on Machine Learning and Cybernetics. Washington: IEEE Computer Society, 2003: 2240-2245

[3] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching[J]. The International Journal on Very Large Data Bases, 2001, 10(4): 334-350

[4] 关佶红,虞为,安扬. GML模式匹配算法[J]. 武汉大学学报, 2004, 29(2): 169-174
Guan Jihong, Yu Wei, An Yang. Geography markup language schema matching algorithm [J]. Journal of Wuhan University, 2004, 29(2): 169-174 (in Chinese)

[5] 章勤,孙盛,袁平鹏. 基于模糊集的地理信息模式匹配算法[J]. 华中科技大学学报, 2006, 34(7): 46-48
Zhang Qin, Sun Sheng, Yuan Pingpeng. Fuzzyset-based schema matching algrithm for geographic information[J]. Journal of Huazhong University of Science and Technology, 2006, 34(7): 46-48 (in Chinese)

[6] Zhou Jingtao, Zhang Husheng, Wang Mingwei. Element matching by concatenating linguistic-based matchers and constraint-based matcher[C]// Khoshgoftaar T M. Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence. Washington: IEEE Computer Society, 2005: 265-269

[7] Melnik S, Garcia-Molina H, Rahm Erhard. Similarity flooding a versatile graph matching algorithm and its application to schema matching[C]// Georgakopoulos D. Proceedings of the 18th International Conference on Data Engineering. Washington: IEEE Computer Society, 2002: 117-128

[8] 程伟,周龙骧,林河水,等. 一种多策略通用模式匹配方法[J]. 计算机科学, 2004, 31(11): 121-123
Cheng Wei, Zhou Longxiang, Lin Heshui, et al. A multistrategy generic schema matching approach [J]. Computer Science, 2004, 31(11): 121-123 (in Chinese)


(责任编辑:卢 硕)