第39卷第5期
2011年5月

同济大学学报(自然科学版)
JOURNAL OF TONGJI UNIVERSITY(NATURAL SCIENCE)

Vol.39 No.5
May. 2011

文章编号：0253-374X(2011)05-0758-06

DOI:10.3969/j.issn.0253-374x.2011.05.023

基于统计式机器学习的地理本体融合模型

王晓璇[1,2]，陈　鹏[1]，刘　鹏[2]，刘妙龙[1]

（1.同济大学 测量与国土信息工程系，上海 200092；2.中国人民解放军理工大学 全军网格技术研究中心，江苏 南京 210007）

摘要：针对不同领域对地理事物的认知体系差异造成了地理本体异构的问题，提出了地理本体融合模型，引入统计式机器学习的方法对概念间的关系进行自动处理，并以概念间关系在不同本体出现的频度来产生其可信度，最后形成带有统计信息和领域信息的大型地理概念空间.该模型巧妙规避概念层面繁琐的异构映射过程，融合概念空间将多个地理本体所表达的概念知识融为一体，并保持了领域内的信息，有效实现了不同认知体系之间的共享.

关键词：地理本体融合；统计；样本偏斜；支持向量机

中图分类号：P 208　　**文献标识码**：A

Geography Ontology Fusion Model Based on Statistical Machine Learning

WANG Xiaoxuan[1,2], *CHEN Peng*[1], *LIU Peng*[2], *LIU Miaolong*[1]

(1. Department of Survying and Geoinformatics, Tongji University, Shanghai 200092, China; 2. PLA Research Center for Military Grid Technology, PLA University of Science & Technology, Nanjing 210007, China)

Abstract: The ubiquity of geography ontology heterogeneity is caused by multiple cognitive systems for geographical object. Geography ontology fusion model was proposed by introducing the method of automatic statistical machine learning for processing relationship within the concepts. The credibility was produced according to emergence frequency of relationship between concepts in different ontology and finally a large-scale integrated geographic concept space with statistic and field information was generated. Cumbersome concept mapping process is circumvented through this model, and all the knowledge expressed in ontologies is fused while information within each field is preserved in this concept space which realize sharing between multiple cognitive systems.

Key words: geography ontology integration; statistic; sample skewed; support vector machine

地理本体是地理信息科学中一个新兴和正在发展的研究领域.地理本体的研究对于空间信息的语义共享和互操作有着重要的意义.目前地理本体的数量和规模正在日益增大，然而，地理本体本身存在异构性.一是由于人类的分类认知本身有着相当的复杂性[1]，二是由于空间信息领域的范围具有很大的伸缩性，涉及社会领域的各个部门和组织机构.由于没有统一的标准，导致不同机构不同领域中关于地理事物的认知和分类标准规范种类繁多，相互之间兼容性差，难以实现系统中数据的互操作和共享.Chrisman认为如何保留对现实世界的不同看法，发展可以集成并转换各种分类系统以及解决这些不同分类系统之间语义异质性的科学方法，是一个比较合理的研究方向[2].所以寻找好的地理本体共享和复用集成的途径方法，有利于不同应用领域，不同认知体系之间的有效交流和实质性共享.

目前已有许多关于解决地理本体异构实现本体复用集成的研究.此类研究主要分为2类：一是集成框架模型的研究[3-4]，二是集成方法的研究[5-9].这些研究的主要思路都是寻找地理本体间一对一的映射、建立映射关系，通过改进映射机制和方法发现概念之间的简单的关系，如等价、包含映射等.这种思路在映射方法的效率和性能很大程度上受到本体构建方法的影响，适用性受到一定的限制；并且为保证映射和集成的准确率需要寻找本体元素间精确的相似对应关系，一般都需要在复用本体的前期或后期有不同程度的人工参与决策，在自动化程度方面存在掣肘，因而所利用本体的数量不会太多.

本文依据统计学的思想，建立一种地理本体融合(ontology fusion)模型.该模型从宏观上考虑问

收稿日期：2010-03-08
基金项目：国家自然科学基金(40801060)；国家“八六三”高技术研究发展计划(2009AA12Z214，2008AA01A309)
第一作者：王晓璇(1983—)，女，博士生，主要研究方向为空间信息语义网格及地理信息智能检索等.E-mail:wxx1012@163.com
通讯作者：刘妙龙(1944—)，教授，博士生导师，主要研究方向为城市地理以及GIS理论方法与应用.E-mail:liuml@tongji.edu.cn

题，通过自动统计快速融合符合一定规范的、来自不同领域的地理本体，以概念间关系在不同本体出现的频度来产生其可信度，用支持向量机的思想调整可信度来校正统计失真. 最后形成带有统计信息和领域信息的大型概念空间.

1　地理本体融合模型设计

1.1　本体融合

目前国外关于本体共享和集成的研究较为活跃，相关的概念有本体集成（ontology integration）、本体合并（ontology merging）、本体对齐（ontology alignment）、本体映射（ontology mapping）等. 对这些概念并没有一个公认的定义和界定，但这些概念的核心意义都指向本体复用（ontology reuse）[10-11].

本文所说的本体融合也是本体复用的一种方式，但是与之前提出的概念有所不同，它的内涵包括以下 3 点：①本体融合同样也是针对本体异构问题，是本体复用和共享的过程；②本体融合的目的是将多个不同本体所表达的知识融为一体，为用户提供趋于客观的概念描述，以达成基于这些本体的数据访问的重用和互操作；③本体融合的结果是具有特殊结构的概念空间，该概念空间实现跨领域融合，同时保持不同领域对地理事物的认识.

1.2　模型设计思想

本文依据统计学思想设计融合模型. 模型有两个基本步骤：统计和计算强度因子. 将待融合本体看成子体，将概念空间看成母体，融合过程就像“大鱼吃小鱼”：大鱼（母体，称之为融合概念空间）通过不断“吞食”并“消化”比自己小的鱼（子体，即被融合的本体）而变得越来越大（母体拥有越来越丰富的概念知识）. 融合的基本过程是首先待融合本体和概念空间不断地进行概念的非语义匹配，同时进行领域信息的统计；然后根据母体此次融合前后的状态变化，对统计信息进行机器学习和全局优化，修改所有概念间关系的强度因子，从而达到“消化”子体所携带知识的效果.

2　基于统计式机器学习的地理本体融合

2.1　概念非语义匹配

本模型规避繁琐的语义映射过程，直接对概念做非语义匹配. 本模型的立足点借鉴统计学的思想，将融合过程设计成为一个统计的过程，认为融合是个动态扩充概念空间的过程. 因为本体即是人对客观世界认识的一个形式化的表达，该模型将这些不同本体在概念层面的多样化表达经过统计都呈现在融合后的概念空间中，当越来越多的本体融合在一起的时候，最后形成的概念空间将逐渐趋于客观. 基于这种思想，可以有效规避传统的不断改进相似度算法来提高结果的可信性和客观性的过程.

2.1.1　数据预处理

由于本模型对概念进行非语义匹配，所以必须首先解决语言层面的不一致问题. 诸如不同本体在语法、逻辑关系表达、原语以及语言表达能力上的差异，需要将不同的语言进行归一化处理，转化为同样的语法格式，逻辑表示，再进行统计. 相对而言，语言层面的一致性处理难度要远小于概念层面的一致性处理，因为语言层面的不一致种类和数量基本上是可以确定的.

2.1.2 基于图论的融合方法

将本体看成是概念和概念之间关系的网状图结构，用图论的广度优先算法遍历待融合本体，将待融合本体中的概念节点和融合后的概念空间进行概念非语义匹配，对概念之间的关系及其领域信息进行统计. 流程图如图 1.

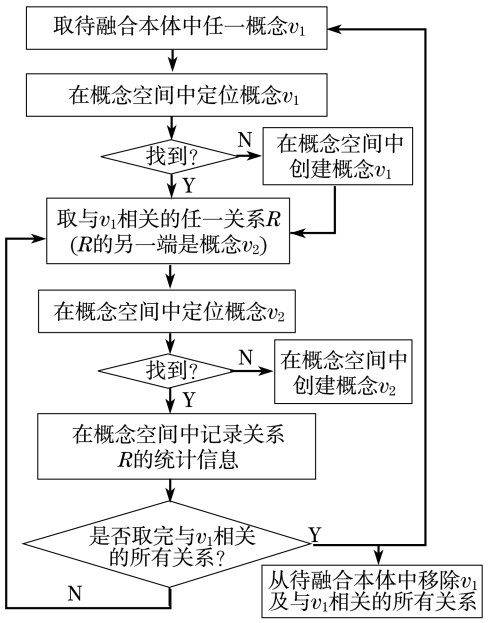


图 1　融合流程图

Fig.1　Flowchart of fusion

初始时，概念空间 S 为空. 下面以概念空间中已有 5 个结点 $v_{11} \sim v_{15}$ 的情况为例，对融合过程进行

简要说明,统计过程是迭代进行的.本例中只对第1次迭代完成后的结果进行表示,在第1次迭代中根据对概念作非语义匹配程度不同,会得到不同的统计结果(图2—4).图中,V_i(V_{ij})表示概念节点,R_i(R_{ij})表示概念间的关系.

①从本体 A 中任取一概念节点 v_1,到概念空间 S 中做非语义匹配.

如有匹配节点:假设 v_1 和 v_{13}匹配成功,v_1 将与 v_{13}融合为一个概念节点.执行步骤②.

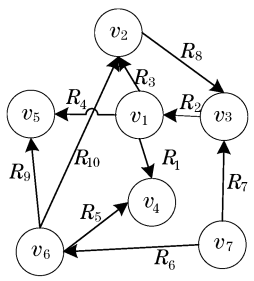


图2 待融合本体 A

Fig.2 Ontology A to fusion

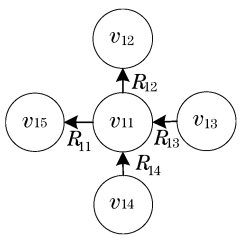


图3 概念空间 S

Fig.3 Concept space S

如没有匹配节点:在概念空间中记录此节点,执行步骤③.

②遍历 v_1 周围节点 v_2,v_3,v_4,v_5,依次与概念空间 S 中的概念节点匹配.

如 v_2,v_3,v_4,v_5 中无节点匹配:记录 v_1 和周围节点的关系和领域信息.执行步骤④,此时概念空间如图4a所示.

如 v_2,v_3,v_4,v_5 中有节点匹配:假设 v_3 和 v_{11}匹配,v_2 和 v_{14}匹配,则 v_3 和 v_{11}融合为一个概念节点,v_2 和 v_{14}融合为一个概念节点,并统计之间的关系和领域信息.执行步骤④.此时概念空间如图4b.

③遍历 v_1 周围节点 v_2,v_3,v_4,v_5,依次与概念空间 S 中的概念节点匹配.

如 v_2,v_3,v_4,v_5 中无节点匹配:记录下这4个节点以及与 v_1 节点的关系,执行步骤④.

如 v_2,v_3,v_4,v_5 中有节点匹配:假设 v_3 和 v_{11}匹配,v_2 和 v_{14}匹配,则 v_3 和 v_{11}融合为一个概念节点,v_2 和 v_{14}融合为一个概念节点,并统计之间的关系和领域信息.执行步骤④.此时概念空间如图4c.

④前面已经以 v_1 为中心,将 v_1 以及与 v_1 有关系的节点 v_2,v_3,v_4,v_5 与概念空间作了匹配,将匹配上的节点进行融合并对相应的关系和领域信息进行统计.此步骤首先在待融和本体中移除节点 v_1 和关系 R_1,R_2,R_3,R_4.再依次以 v_2,v_3,v_4,v_5 为中心,重复①～④,直到待融和本体移除为空时融合完成.

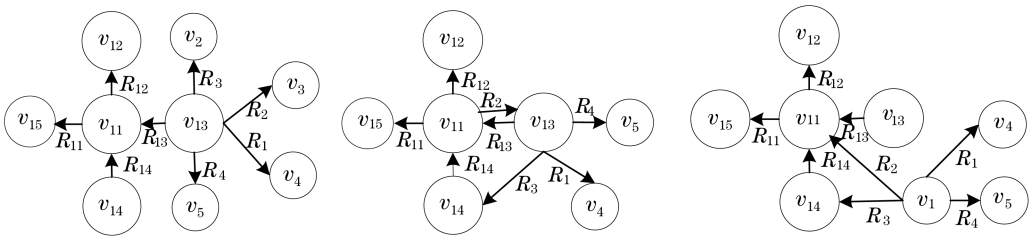


图4 统计结果

Fig.4 Statistical results

2.2 基于支持向量机思想的强度因子计算方法

2.2.1 支持向量机解决样本偏斜问题

由于目前地理本体构建的分布性、异构性和自治性等特点,所融合的本体的分布往往是偏斜(skewed)或称不均衡的,即不同领域的样本的数量可能存在数量级的差距.在数据偏斜的情况下,样本无法准确反映整个空间的数据分布,即存在概念关系和属性统计可信度失真现象.应对措施分为两方面,一方面,随着融合本体的领域范围越广,数量越多,偏斜的程度就会降低,但是事实上地理本体的研究正在逐步引起重视和普及阶段,信息化程度高的领域和部门对地理本体的重视会多于信息化程度低的领域和部门,所以仍然存在热门领域的地理本体的数量会远远超过非热门领域的本体数量的问题.另一方面可以全局优化调整,通过机器学习寻找合适的数学模型参数对统计信息进行调整,设计合理的强度因子计算方法.

本文引入支持向量机(support vecton machine,SVM)的思想,样本数目不对称时,SVM模型的关键在于如何确定惩罚参数 C 的值,通过对每类样本分别设置不同的惩罚参数 C,控制对不同类别中错分样本的惩罚程度来解决[12].合理的惩罚因子优化特征选择框架或改进特征选择方法获得对小类别特征的重视.

2.2.2　基于惩罚参数的强度因子计算

设 n 个领域 $F_1, F_2, \cdots, F_n$，样本数目分别是 $N_1, N_2, \cdots, N_n$，S_{ij} 表示关系 i 在领域 j 中的统计次数，V_{ij} 表示关系 i 在领域 j 中的强度因子. 设定样本数目最小的领域的惩罚参数为 1，假设样本最小的领域为 F_j，那么合理的 C 值应使下式成立：

$$C_i = \begin{cases} C_1, X_1 \in F_1 \\ C_2, X_2 \in F_2 \\ \vdots \\ C_n, X_n \in F_n \end{cases}, \text{同时满足 } C_i = \frac{N_j}{N_i} \quad (1)$$

式(1)表明，对不同领域的惩罚与样本数目成反比，则强度因子为

$$V_{ij} = C_i S_{ij} \quad (2)$$

2.3　融合结果

融合结果为一个融合概念空间. 这个概念空间保留了对地理现实世界的不同的看法. 它是一张巨大的概念关系网，它要记录所有子体(原本体)中曾经出现过的概念以及概念间的关系. 融合概念空间可不断扩展，它初始为空，随着所融合的概念和关系的增加而不断扩展. 与本体相同的地方是，它们都可以图示为若干顶点(即概念)和若干条连接顶点的边(即概念间的关系)；不同的地方在于，融合概念空间里的每一条边，要记录概念关系强度的统计信息. 每一条边的统计信息具体包括：该关系在所有子体中出现的次数及换算出来的强度，以及在某些特定领域里该关系出现的次数及强度. 每一条边的统计信息可以表示为：$\{(S, V), [(s_i, v_i)]_{i=1,\cdots,n}\}$. S 代表该关系在所有子本体中出现的次数，V 代表使用机器学习方法根据 S、每一个 (s_i, v_i) 及全局情况计算出来的关系强度因子. s_i 和 v_i 分别代表关系在第 i 个领域的出现次数和计算出来的强度. $[(s_i, v_i)]_{i=1,\cdots,n}$ 记录从领域 1 到 n 每一个 (s_i, v_i) 的情况. 因此，可以形象地理解，融合概念空间的关系网里的关系有的粗、有的细，还可以“进到任意一条关系里面”观察该关系在不同领域里的粗细程度.

3　应用实例

应用实例来说明本文提出的融合模型如何有效地解决不同应用领域、不同认知体系之间的异构问题. 图 5 上半部分表示了已经存在的概念空间统计信息，里面已经融合了一些地理本体，集成了多个领域对地理现象的认识，表示为一些地理概念和它们之间的关系，并用二元组记录了关系的统计信息.

3.1　非语义匹配

将新的本体融合进去的时候，会存在各种各样的异构现象. 本实例实现 2 个本体融合到概念空间的过程. 农业部门在进行土地利用分类时，参照《中华人民共和国草原法》，建立了农业本体(图 5 下左). 数字海湾建设过程中，针对海湾研究中各类地理概念及其语义关系，也需要对海湾地理本体进行建模，图 5(下右)中的海湾本体取自文献[13]对海湾地理本体的建模.

可以看出，相比已存在的概念空间中的概念和它们之间的关系，海湾本体和农业本体有不同的分类系统，农业本体的对于用地的分类粒度更为细致，并且对“草地”这个概念，存在同名异义的现象. 另外海湾本体中“陆域”属于“海湾”的一部分(本体中表示为“Part-of”)，然而在概念空间中“海湾”和“陆域”的关系是不相交的两个概念(本体中表示为“Disjoint”)，出现了概念之间关系的多样性.

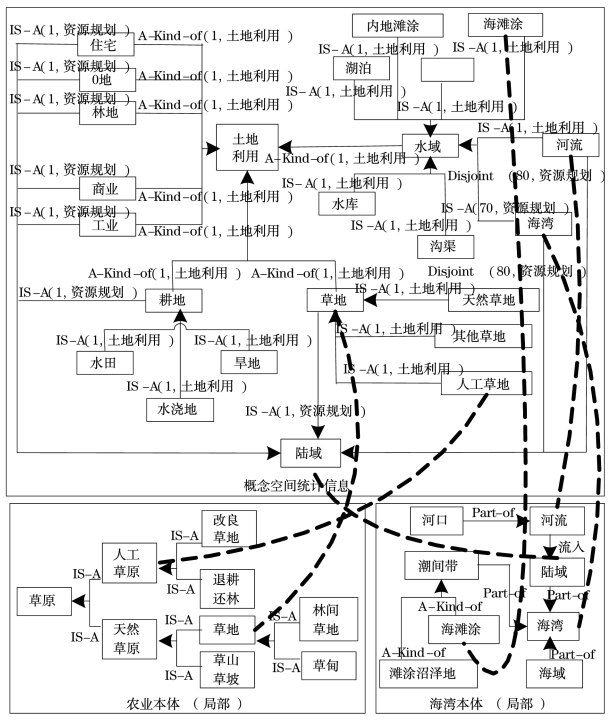
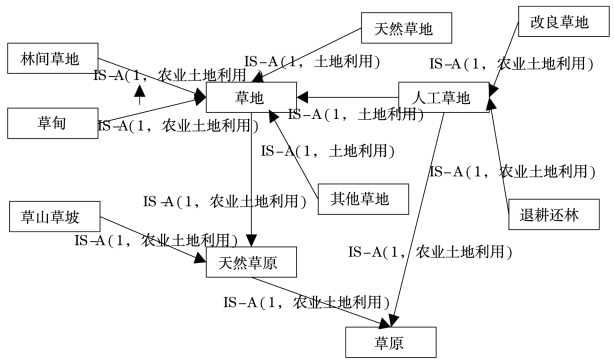


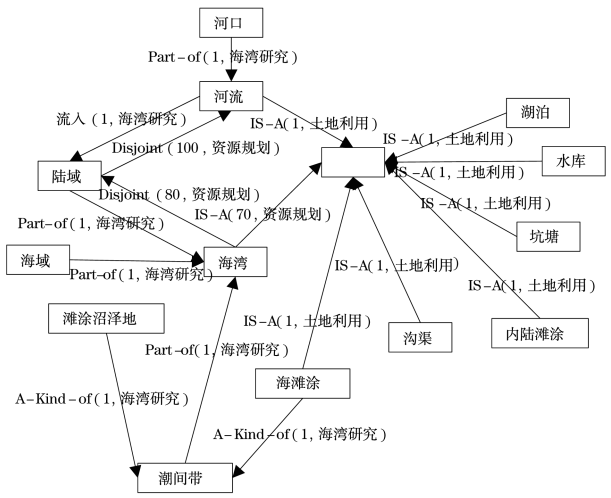
图 5　待融合本体和概念空间中概念的匹配关系

Fig.5　Matching relationship of ontology and concept space

根据本文提出的融合模型，用图的广度优先算法遍历城镇土地利用本体和海湾本体，匹配相同的概念，并记录概念之间的关系和该关系所存在的领域. 图 5 表示待融合本体和概念空间中概念的匹配关系. 图 6a 和图 6b 分别表示农业利用本体和海湾地理本体融合到概念空间后的结果，由于空间有限，只显示了概念空间中变动的部分.



a 与农业利用本体融合后的概念空间(局部)



b 与海湾地理本体融合后的概念空间(局部)

图6 地理本体统计结果

Fig.6 Statistical results of geography ontology

通过实例可以看到，概念层面的不一致与数量极多、形态各异的概念本身直接相关，非语义匹配的方法比较巧妙地规避了传统本体复用技术必须要解决的不同本体在概念层面表示的不一致问题. 对于因为领域不同导致相同概念有不同的含义，统计后在概念空间中表现为概念有不同的关系；对于不同本体用不同的词表达相同的含义的情况，经过统计学习后会在这些词之间形成强度比较大的等价关系；对于不同本体建模采用了不同的概念描述粒度的情况，经过统计学习后会综合各种粒度的信息. 随着本体样本的增多，融合空间会逐步复杂并越来越趋于客观.

3.2 强度因子计算

从统计结果可以看到，“陆域”和“海湾”的关系呈现多样性，城市规划部门大多认为“陆地”和“海湾”的关系是不相交的两个概念，而海洋部门则认为“陆域”属于“海湾”的一部分，造成多样性的原因，是因为对“陆域”这个概念的所指是不一样的，海湾本体中的“陆域”，实际上是指“海湾”中的“陆域”，而非一般的“陆域”. 这些认知在本领域内都是正确的，可以采纳的. 但是城市规划的本体数量较大，根据统计信息可以得到两个概念之间的关系“Disjoint”的强度为80，而关系“part-of”的强度只为1，可信度很低，但是现在海洋部门的本体只有1个，由于样本数量的差距导致可信度的差距会影响概念空间的客观性，需要用本文中提出的强度因子计算方法对其进行调整.

现在假设两个概念之间的关系涉及4个领域，呈现4种关系. 如表1，有 $i=1,2,3,4$ 表示4个领域，已知4个领域的本体数量分别为：150，2，80，20，R_1—R_4 表示两个概念间4种不同的关系.

表1 两概念之间的统计信息

Tab.1 Statistical information between two concepts

关系	$i=1$	$i=2$	$i=3$	$i=4$
R_1	8	0	0	0
R_2	70	0	80	0
R_3	0	2	0	17
R_4	3	0	0	2

分析表中的统计信息，领域1和领域3中大多数的本体用 R_2 描述两概念之间的关系，领域2和领域4中大多数的本体用 R_3 描述两概念之间的关系，但是关系 R_2 的强度为70，领域2的关系 R_3 的强度仅为2，典型由于样本悬殊造成的统计失真. 这种情况是非常普遍的，如在海洋部门对海湾的认知体系中，“陆域”(海湾中的陆域)属于“海湾”的一部分，然而在土地利用认知体系中，“海湾”和“陆域”(非海湾中的陆域)的关系是不相交的，两种语义信息在各自领域内都是高可信的. 但是由于现在城市规划方面的本体样本比较多，而造成了在统计的时候，“Disjoint”这种关系的可信度高，这样的统计数据是不符合客观事实的. 按照本文提出的强度因子计算方法，对统计信息进行调整，根据公式(1)、(2)，调整后的强度信息如表2.

表2 调整后的强度信息

Tab.2 Intensity information after adjustment

关系	$i=1$	$i=2$	$i=3$	$i=4$
R_1	0.107	0	0	0
R_2	0.933	0	2	0
R_3	0	2	0	1.7
R_4	3	0	0	0.2

关系 R_1—R_4 在概念空间中的可信度调整为

$$R_1(s_i, v_i)_{i=1} = 0.107$$

$$R_2(s_i, v_i) = \begin{cases} 0.933, & i = 1 \\ 2, & i = 3 \end{cases}$$

$$R_3(s_i, v_i) = \begin{cases} 2, & i = 2 \\ 1.7, & i = 4 \end{cases}$$

$$R_4(s_i, v_i) = \begin{cases} 3, & i = 1 \\ 0.2, & i = 4 \end{cases}$$

从结果数据可以看出，调整后领域 2 的统计可信度明显增加，领域 1 的统计可信度明显减小，其他领域的可信度也有相应的调整. 试验证明，该方法对解决本体数量不均衡所带来的可信度差异是有效的.

4　结语

本文提出的基于统计式机器学习的地理本体融合方法具有如下优点：①从宏观上考虑问题，它规避了概念层面繁琐的异构映射过程，利用统计方法巧妙解决异构问题；②自动化程度高，不需要繁复的人工参与；③所融和的本体数量越多，概念空间的可信度越高，既实现跨领域融合，又保持领域内信息. 利用本文提出的模型，不同领域不同部门的人只需要查询融合后的概念空间，就可以看到各个领域对同一概念的描述方式，并通过可信度的值判断每种描述方式的可信程度，有效实现了不同认知体系、不同分类系统之间的共享. 本模型虽然从地理信息领域存在的实际问题出发进行设计，但是地理信息领域存在的这些问题在其他领域同样存在，所以本文提出的本体融合模型具有很好的通用性，同样适合于其他领域.

但模型还有待完善：首先，惩罚因子的确定是根据样本数量线性确定的，现实中的影响因素复杂程度往往较高，需要对影响因素进行更加全面的分析，进一步完善强度因子计算方法；另外，由于融合的结果是一个大型的概念空间，里面记录了大量的概念关系强度信息和领域信息，难以用标准的 OWL 语言来表示，不能直接被 OWL 应用所使用，所以对融合结果的应用也是需要进一步研究的问题.

参考文献：

[1] Mark D M. Human spatial cognition [M]// Human Factors in Geographical Information Systems. London: Belhaven Press, 1993:51.

[2] Chrisman N. Building GIS without foundations: ontology from a social practice perspective [C]//Proceedings of GIScience, Savannah: [s. n.], 2000.

[3] Kavouras M. A unified ontological framework for semantic integration [M]//Next Generation Geospatial Information: From Digital Image Analysis to Spatiotemporal Databases. London: Taylor and Francis, 2005:147.

[4] 谭喜成，边馥苓. 基于本体协同的空间信息互操作方法[J]. 武汉大学学报：信息科学版，2005，30 (2)：178.
TAN Xicheng, BIAN Fuling. Heterogeneous spatial information interoperability based on cooperative ontologies[J]. Geomatics and Information Science of Wuhan University, 2005, 30 (2):178.

[5] Kavouras M, Kokla M. Fusion of top-level and geographical domain ontologies based on context formation and complementarity [J]. International Journal of Geographical Information Science 2001, 15(7):679.

[6] Sunna W, Cruz I F. Structure-based methods to enhance geospatial ontology alignment [C]//GeoSpatial Semantics, Lecture Notes in Computer Science, Berlin: Springer - Verlag, 2007:82 - 97.

[7] Tomai E, Prastacos P, Kavouras M. A framework for intentional and extensional integration of geographic ontologies [J]. Transactions in GIS, 2007, 11(6):873.

[8] Duckham M, Worboys M F. An algebraic approach to automated information fusion [J]. International Journal of Geographic Information Science, 2005, 19(5):537.

[9] 李德仁，崔巍. 空间信息语义网格[J]. 武汉大学学报：信息科学版，2004，29(10)：847.
LI Deren, CUI Wei. Semantic grid of spatial information[J]. Geomatics and Information Science of Wuhan University, 2004, 29 (10):847.

[10] Kalfoglou Y, Schorelmmer M. Ontology mapping: the state of the art [J]. The Knowledge Engineering Review, 2003, 18 (1):1.

[11] Pinto H S, Martins J P. A methodology for ontology integration [C]//Proceedings of the International Conference on Knowledge Capture, Victoria: ACM Press, 2001:131 - 138.

[12] Veropoulos K, Cambelu C, Cristianini N. Controlling the sensitivity of support vector machines[C] //Proceedings of the International Joint Conference on AI, Stockholm: [s. n.], 1999: 55 - 60.

[13] 杜云艳，张丹丹，苏奋振，等. 基于地理本体的海湾空间数据组织方法——以辽东湾为例[J]. 地球信息科学，2008，10(1)：7.
DU Yunyan, ZHANG Dandan, SU Fenzhen, et al. Geospatial data organization of the bay based on the geo-ontology—a case study of the Liaodong Bay [J]. Geo-information Science, 2008, 10 (1):7.