引用格式：孙凯,诸云强,潘鹏,等.形态本体及其在地理空间数据发现中的应用研究[J].地球信息科学学报,2016,18(8):1011-1021. [Sun K, Zhu Y Q, Pan P, *et al.* 2016. Research on morphology-ontology and its application in geospatial data discovery. Journal of Geo-information Science, 18(8):1011-1021.] **DOI:**10.3724/SP.J.1047.2016.01011

形态本体及其在地理空间数据发现中的应用研究

孙　凯[1,2],诸云强[1,3*],潘　鹏[1],罗　侃[1,2],王东旭[1,2],侯志伟[1,2]

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101;
2. 中国科学院大学,北京 100049; 3. 江苏省地理信息资源开发与利用协同创新中心,南京 210023

Research on Morphology-Ontology and Its Application in Geospatial Data Discovery

SUN Kai[1,2], ZHU Yunqiang[1,3*], PAN Peng[1], LUO Kan[1,2], WANG Dongxu[1,2] and HOU Zhiwei[1,2]

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

Abstract: The semantic heterogeneity of geospatial data is the main bottleneck for the realization of data association, the intelligent recommendation and the accurate discovery of data. Geospatial data ontology is known as an effective approach to solve the semantic heterogeneity of geospatial data. The morphological characteristic is an important feature of geospatial data besides its temporal, spatial and attribute characteristics, and it is the primary research content of geospatial data ontology. Based on the systematic analysis on the morphological characteristics of geospatial data, this paper studies and puts forward a concept system. Furthermore, this paper creates the morphology-ontology model of geospatial data, defines the ontology representation method of morphological information, and ultimately constructs the morphology-ontology. In the last part, a prototype system for the semantic retrieval of geospatial data has been programmed by taking the morphology-ontology product as the ontology library and using the Jena Java framework. The constructed morphology-ontology have been applied to the retrieval of metadata from the National Earth System Science Data Sharing Infrastructure. Verification test shows that the morphology-ontology of geospatial data can effectively solve the semantic heterogeneity existing in the morphological characteristics of geospatial data and improve the precision and recall rate of data discovery result. The research methods and results of this paper have great reference values in solving the semantic heterogeneity problems occurred in other research fields.

Key words: geospatial data; morphological characteristic; semantic heterogeneity; data discovery; ontology

***Corresponding author:** ZHU Yunqiang, E-mail: zhuyq@igsnrr.ac.cn

摘要：地理空间数据语义异构是实现数据关联、数据智能推荐和精确发现的主要瓶颈。地理空间数据本体被认为是解决地理空间数据语义异构的有效方法。形态特征是地理空间数据(除时空、要素内容外)的重要特征,是地理空间数据本体的重要研

收稿日期：2015-04-30;**修回日期：**2015-05-11.

基金项目：国家自然科学基金项目“基于元数据语义的地理空间数据关联方法研究”(41371381);科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”(2013FY110900);国家留学基金项目(201504910358);国家科技基础条件平台-地球系统科学数据共享平台(2005DKA32300)。

作者简介：孙　凯(1990-),男,硕士生,山西长治人,研究方向为地学数据共享和地理信息技术与应用。
E-mail: sunk.14s@igsnrr.ac.cn

***通讯作者：**诸云强(1977-),男,博士,研究员,江西广丰人,研究方向为地学数据共享、资源环境信息系统。
E-mail: zhuyq@igsnrr.ac.cn

究内容。本文首先在系统分析地理空间数据形态特征的基础上,提出地理空间数据形态特征的概念体系。然后,建立地理空间数据形态本体模型,提出形态信息的本体表示方法,并构建地理空间数据形态本体。最后,基于形态本体的本体库,利用Jena本体推理技术,开发地理空间数据语义检索原型系统,并将形态本体应用于国家地球系统科学数据共享平台的元数据检索中。实验结果表明,地理空间数据形态本体可以有效地解决数据形态特征的语义异构,提高数据发现的查全率和查准率。本文的研究方法和成果对解决其他领域数据的语义异构,有重要的参考意义。

关键词:地理空间数据;形态特征;语义异构;数据发现;本体

1 引言

地学研究工作的快速发展,以及对地观测手段的极大丰富,使地理空间数据的存量呈爆炸性增长。海量的数据资源,给科学研究和应用带来了便利,但同时对数据发现的查全率、查准率提出了更高的要求。地理空间数据资源的生成往往伴随着复杂的地理科学过程,所以地理空间数据常以多种形式存在且带有复杂的语义异构现象,这是实现地理空间领域数据精确发现和共享的主要障碍[1]。传统信息检索方式普遍基于关键字匹配或全文检索技术,借助于目录、索引和关键词匹配等方式实现,忽略了数据本身丰富的语义特征,无法有效地解决由语义异构带来的数据检索问题[2]。

本体是共享概念模型的形式化规范说明[3],可以用来描述数据的语义信息[4],对多源异构数据的领域概念和关系进行显式的解释说明,使隐性知识显性化,不同数据集间的各种联系能够被应用系统识别[5]。在地理空间领域,作为共享地理概念模型或地理认知模型的形式化说明,地理本体为地理数据提供了形式化语义说明,对于解决地理数据的语义异质性和实现语义层次上的互操作具有很大的潜在优势,通过地理本体可以在集成地理数据时将数据的语义也集成在一起[6]。

地理空间数据的空间、时间和要素等基本特征,是数据内容的决定因素,是区分不同数据的本质属性,称为本质特征。地理空间数据的形态特征是数据内在结构特征和外在形状特征的描述,包含了数据基准、格式、类型、比例尺等内容。国内外针对本质特征的地理时空本体已经开展了大量的研究,并得到了广泛的应用[7-13]。形态特征作为除本质特征外地理空间数据的重要特征,对于解决该领域数据语义异构有非常重要的意义,同时形态特征也是用户发现目标数据的必要条件,所以研究地理空间数据形态特征,并结合本体理论,构建形态本体是非常必要的。近年来,相关学者已经就数据形态特征及其与本体理论的结合开展了研究。Cristian等[14]提出了一种基于OWL(Web Ontology Language)知识的数据表达方法,实质上是关系模式以及其相关数据项的描述技术和一个用于描述数据库相关元数据的OWL本体。RDA(Research Data Alliance)数据类型注册工作组[15],设计了数据类型注册的可用规范说明,并提出了多重数据类型注册策略。李庭波等[16]以森林资源经营管理为例,探索了森林资源数据结构本体学习技术。郝亚楠等[17]提出了一种基于语义的数据格式转换方法,并以Word文档为例,采用基于学习的策略,自动地将Word文档转换为具有语义信息的XML文档。苏里等[18]研究地图的形式化语义表达问题,将语言学的理论和方法应用于地图领域,把地图符号系统作为一种二维图形语言,采用扩展的描述逻辑,建立了能完备表达地图语言的语义模型。杨小忠等[19]研究了基于应用本体的多卫星遥感数据检索,总结了遥感卫星传感器的属性,例如波段数、空间分辨率等。

目前,地理空间数据形态特征以及形态本体的研究存在以下问题:文献检索中相关内容极少;现有文献虽涉及地理空间数据形态特征的不同部分,但仍没有提出完整的概念体系;以形态本体为基础开展的数据检索实践很少。基于上述情况,本文以地理空间数据形态特征为研究对象,明确地理空间数据形态特征的概念,提出了完善的地理空间数据形态特征概念体系;基于UML(Unified Modeling Language)建立地理空间数据形态本体模型,并构建本体;最后,以国家地球系统科学数据共享平台中的元数据为数据源,以地理空间数据形态本体为本体库,提取并标注元数据的形态特征,进行了该领域本体的数据发现应用。

2 地理空间数据形态特征分析

2.1 地理空间数据形态特征

形态(Morphology)一词起源于希腊,强调“对

形状的研究”，它最早应用在生物学中。生物形态学是研究有机体的形成、结构和结构特征的科学，包含对生物体外观（形状、结构、颜色、纹理等）和内部形态（骨骼、器官等）的研究，生物形态学与研究生物功能的生物生理学是相对而言的，二者紧密相关，不能分离。

借鉴生物形态学，本文提出地理空间数据的形态特征。图1为生物形态学的研究对象、研究内容和互补内容3方面与数据形态特征的映射关系。地理空间数据形态特征的含义亦可从这3方面阐述：（1）形态特征的研究对象为“数据的形状”，“数据形状”的含义决定了形态特征的研究边界和内容；（2）根据数据形状是否能够直接清晰地判别，可分为外部形态（能够清晰判别和感官的数据形状）和内部形态（无法直接判别，但对数据形状有直接或者间接的影响），外部形态表征内部形态，内部形态决定外部形态；（3）本质特征是用户数据发现过程中最为重要的检索词，形态特征是本质特征的重要补充，二者密切关联，但同时二者有明确的界限，没有重叠。本质特征是指标识空间数据唯一性的特征，决定数据内容，具有严格的不可变换性，一旦本质特征变了，空间数据就变为另一个空间数据。形态特征是对数据内容的内在结构特征和外在形状特征的描述，是可以变换的，但它的变换不会导致空间数据本质的改变，这是界定本质特征和形态特征的主要依据。例如，北京市2010年土地覆被数据，如果空间范围、时间范围和要素变化（如变成河北省或2015年或社会经济），该数据就会成为另外一个数据，而变换该数据的数据格式、存储介质和投影等，数据本身不会改变。

形态特征表现为数据的外在形式和附加特征，描述地理空间数据的结构、格式、存储、基准等内容。本文根据数据资源的生产流程总结了地理空间数据形态特征具体的研究内容：数据采集过程需确定比例尺、精度和尺度等；数据表达及可视化过程会涉及地图符号系统、数据基准、数据单位、数据语言和字符编码等；数据组织过程需确定数据类型和数据结构；数据存储过程需确定数据格式和存储介质。根据上述内容是否能够直接清晰的判别，将其分为外部形态和内部形态2类（表1）。内部形态和外部形态有密不可分的关系，例如，比例尺决定地图符号系统，同样的地物在不同比例尺的地图上，其地图符号是不同的；数据格式表征数据结构，若数据的格式为shapefile，则表明其数据结构应是矢量数据结构。

元数据是关于数据的数据，是对空间数据标识、内容、时空范围、质量、分发方式的描述。其中，关于数据的基准、类型、格式等内容与形态本体有较强的相关性，二者都是对数据的内在结构和外在形状进行了描述。但二者有本质的区别，且形态本体有明显的优势：（1）元数据中有对形态概念的描述，但只停留在描述层面上，而形态本体的概念表达需要符合本体严格的层次关系，是对形态概念及实例，概念、实例间关系进行的形式化表达，由此可以确定概念的描述规则和值域范围，在更细的粒度上、更规范的描述数据，有利于数据共享；（2）本体的推理机制是元数据不具备的，形态本体作为收集了常用的词以及词和词之间关系的语义字典库，可以为关联数据发现和推荐提供支持。

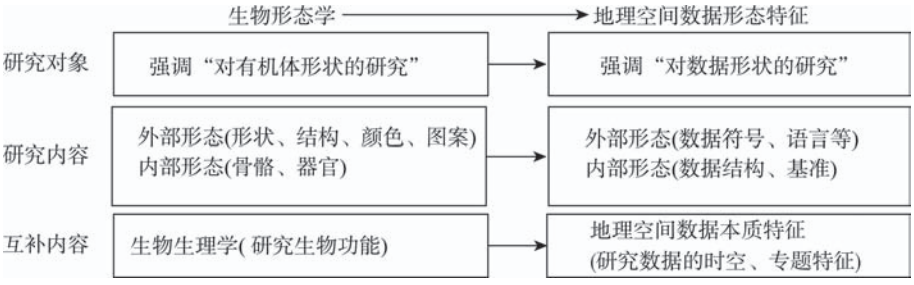

图1　地理空间数据形态特征与生物形态学的映射关系

Fig.1　The mapping relationship between the morphological characteristics of geospatial data and the biological morphology

表1　地理空间数据形态特征研究内容分类表

Tab.1　The classified research contents for the morphological characteristics of geospatial data

类别	研究内容
外部形态	数据语言、数据格式、存储介质、数据单位、地图符号系统、数据量
内部形态	比例尺、精度、尺度、数据基准、数据类型、数据结构、字符编码

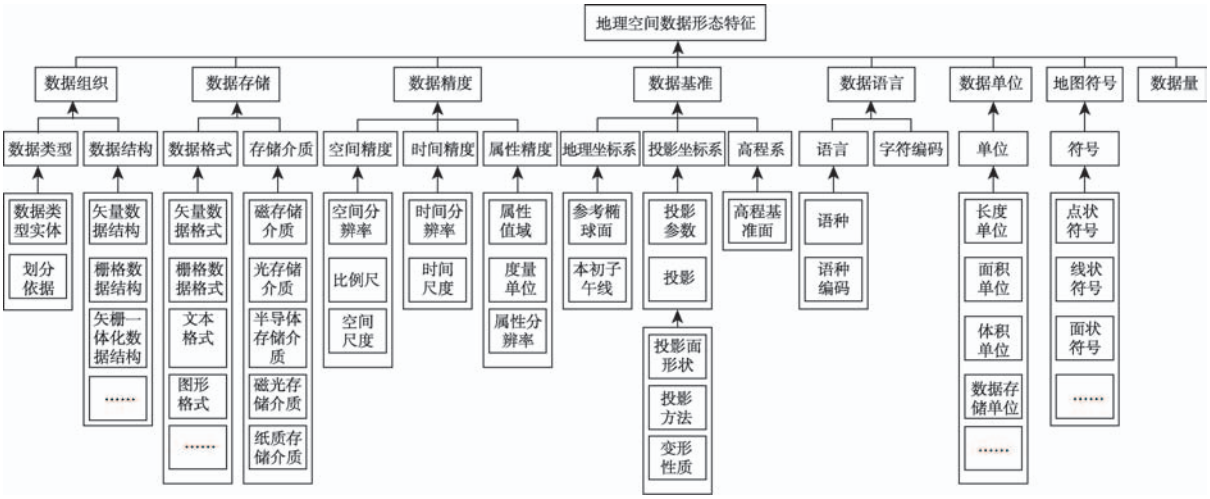

图2 地理空间数据形态特征概念体系图

Fig.2 The concept system for the morphological characteristics of geospatial data

2.2 地理空间数据形态特征概念及其关系

地理空间数据形态本体是对地理空间领域共享的形态特征概念体系的形式化规范说明，形态特征概念体系是构建本体的基础。图2为地理空间数据形态特征概念体系图，由于概念较多，在图中只列出了概念层次的一部分。数据组织和数据存储是形态特征的主要内容，也是本体构建的重点，其子特征数据类型、数据结构、数据格式和存储介质之间存在着密切的关系。数据采集完成后，首先需要完成数据的组织，明晰数据类型。区别于程序设计中变量的数据类型，这里的数据类型指矢量数据或栅格数据，文本数据、图形数据或图像数据，不同的分类标准可以得到不同的数据类型分类体系。数据类型决定了数据结构。数据结构指用于在计算机中表示地理信息的数据组织方式。矢量数据类型使用矢量数据结构（如双重独立编码结构），栅格数据采用四叉树等栅格数据结构。由于不同的公司、组织的标准和软件工具的处理方式不同，使同样的数据结构会由于几何或属性信息组织方式、文件头和扩展名以及是否含有拓扑关系等不同，生成不同的数据格式。至此，原始数据已经以一个文件形式存在，将其保存在移动硬盘、U盘等存储介质中即可。

数据基准和数据精度、尺度特征对于空间数据非常重要，因而也是本文研究的重点。数据基准决定数据所采用的空间参考系统。地理坐标系指通过经纬度确定地球表面点位位置；投影坐标系指按照一定数学法则（投影）将地球表面上的经纬线网表示到平面上，即平面参考系，二者都有各自的参数，参数以属性的形式存在。数据精度包含空间精度、时间精度和属性精度。通常，空间精度是数据生产者和使用者最为关注的，包含空间分辨率、比例尺和空间尺度（分县、分省等）等含义。时间精度包括时间分辨率、时间尺度（按年、月、日）。属性精度包含属性值域（数据集中属性值的取值范围）、度量单位和属性分辨率（数据集中能区分的最小属性值数量）。此外，数据集所采用的语种和字符编码、数据单位、地图数据的符号系统以及数据量都是数据形态特征的研究内容。

3 地理空间数据形态本体建模

3.1 地理空间数据形态本体建模原语

国内外学者从各自的专业领域出发，提出了不同的本体建模原语，例如金芝[20]等提出的三元组、王洪伟[21]等提出的四元组、Naing[22]博士等提出的六元组等，此外，还有Perez等提出的目前应用最广泛的五元组结构。其用分类法组织本体，总结出本体应包含5个基本的建模元语，即本体可以表示为5元组O=<C, R, F, A, I>[23]。C代表了概念（或者类），是具有相同属性的对象的集合。R代表了关系，是指领域中概念之间的交互作用。F代表函数，表示一种特殊的关系。这种关系中，前 $n-1$ 个元素可以唯一决定第 n 个元素，形式化的定义为F：$C1\times C2\times\cdots\times Cn-1\rightarrow Cn$。A代表公理，对概念和关系进行约束。I代表实例，即对象，是概念的具体化表现。

结合不同建模原语的含义和构建本体的实践，本文将地理空间数据形态本体的逻辑结构定义为

包含概念、属性、关系、约束和实例的五元组模型（M_{GDO}），如式（1）所示。

$$M_{GDO}=<MC, MR, MP, MC_{on}, MI> \quad (1)$$

式中：MC（Morphological Concepts）表示地理空间数据形态本体的概念，指一系列对象的集合；MR（Morphological Relations）表示地理空间数据形态本体的概念与概念、概念与实例、实例与实例之间的关系集合，包含本体所共有的语义关系以及形态本体所特有的形态关系集合；MP（Morphological Properties）表示地理空间数据形态本体的概念的属性，包括对象属性和数据属性，对象属性表示实例与实例之间的关系（如投影与投影面形状的“投影面形状是”关系），数据属性表示实例与数值的关系（如TM全色波段影像的“空间分辨率是”30 m）；MC_{on}（Morphological Constraint）表示地理空间数据形态本体的规则集合，包括对属性的取值类型、取值范围和取值基数等的约束（如数据量的数值不能为负数）和函数（例如不同投影之间的转换公式）；MI（Morphological Individuals）表示地理空间数据形态本体的实例，即对象。

3.2 地理空间数据形态本体建模方法

3.2.1 基于UML的本体建模方法

近年来，随着本体理论及其应用研究的逐渐深入，产生了多种本体建模及其形式化表达的方法。例如，基于一阶谓词逻辑、描述逻辑等逻辑语言的方法、基于集合理论或元组理论的方法、基于Web的OWL、RDF（Resource Description Framework）等方法，均基于复杂语法和逻辑。此外，基于UML的本体建模方法也得到了很大发展。UML面向对象的思想，可以体现本体概念的属性、方法的继承。UML类图描述领域概念及其之间的关系，由类名、属性、操作3部分组成，而本体中的概念、属性、方法则可以映射为类图的类、类的属性、类的方法等。UML作为本体表示语言，具有直观的优点，但同时由于图形表达能力有限，无法表达本体约束，为此引入OCL（Object Constraint Language）[24]。OCL是一种形式化的语言，用基于数学的、精确的语言表达式说明一些在图形化的模型中不能充分表达的模型规则和约束。

基于上述描述，本文采用UML和OCL结合作为本体建模方法。本体的概念、关系、属性、方法用UML类图中相应的映射元素表示，本体中的规则约束用OCL约束语言表示。地理空间数据形态本体的元素对应关系如表2所示。

表2　本体元素与UML模型元素映射关系

Tab.2　The mapping relationship between the ontology elements and the UML model elements

形态本体的元素	对应UML的映射元素
概念	类名
数据属性	类名+属性名
对象属性	关联关系
kind-of关系	泛化关系（类名1、类名2）
part-of关系	聚合关系（类名1、类名2）
基本语义关系	语义关系名（类名1、类名2）
形态关系	形态关系名（类名1、类名2）
方法	类名+方法（）
规则集	OCL约束条件
本体模型图	UML类图
实例	UML对象图

3.2.2 基于UML的地理空间数据形态本体建模

地理空间数据形态本体建模是对该领域本体的概念集、属性集、关系集、方法集和规则集等的形式化规范说明，有利于保证本体构建的完整性和规范性。本文采用本体元素与UML类图元素映射的方式建模，并以数据基准子本体为例进行阐述。

地理空间数据形态本体概念集映射为类集，概念MC映射为类Class，形态本体属性MP映射为类的属性，用类名加属性名表示，具体描述如表3所示。

地理坐标系（gcs.n）表示类地理坐标系（gcs）的name属性，其他属性表示具有相似的含义。

表3　概念及属性的UML表示

Tab.3　The UML representation of concepts and properties

概念	类	概念属性	类属性
地理坐标系	gcs	地理坐标系名	gcs.n
		参考椭球面是	gcs.sre
		参考椭球面名	sre.n
参考椭球面	sre	椭球长半轴是	sre.sae
		椭球扁率是	sre.if
		投影坐标系名	pcs.n
		地理坐标系是	pcs.gcs
		投影方式是	pcs.pr
		中央经线是	pcs.cm
投影坐标系	pcs	原点纬线是	pcs.lo
		水平偏移量是	pcs.fe
		垂直偏移量是	pcs.fn
		比例因子是	pcs.sf
		EPSG代码是	pcs.epsg
高程系	es	高程基准面是	es.vd

形态本体关系MR表示本体所共有的语义关系以及形态本体所特有的形态关系集合，前者共有4种可映射为类的关系，描述如下：

（1）kindOf关系：表示为generalization（子类，父类），如generalization（地心坐标系，地理坐标系）；

（2）partOf关系：表示为composition或aggregation（部分类，整体类），如composition（投影，投影坐标系）；

（3）instanceOf关系：表示为instanceOf（对象，类），如instanceOf（高斯克吕格投影，投影）；

（4）disjointOf关系：表示为disjointOf（互斥类，类名），如disjointOf（地心坐标系，参心坐标系）。

而形态本体所特有的形态关系集合的主要内容为概念的变换关系例如数据格式变换、投影变换等，变换关系的特性（如是否可逆）以及变换信息损失度对变换关系的度量。此外，概念实例的关系还包含同义、近义、反义等。形态特征的变换关系可映射为对象的set( )方法，描述如下：

（1）投影转换方法：pr.setPr():Data；

（2）高程系转换方法：es.setEs():Data；

（3）地理坐标系转换方法：gcs.setGcs():Data；

（4）参考椭球转换方法：sre.setSre():Data。

形态本体的实例MI映射为类的对象，用UML对象图表示，对象图表示类图的一个实例，二者的基本概念相似，故本文不做详细讨论。

综上分析，形态本体模型的UML类图如图3所示。

OCL约束语言为UML类图提供规则约束MCon，示例描述如下：

（1）context sre inv

self.n→includes（gcs.sre）//约束地理坐标系的参考椭球面必须是该类的实例；

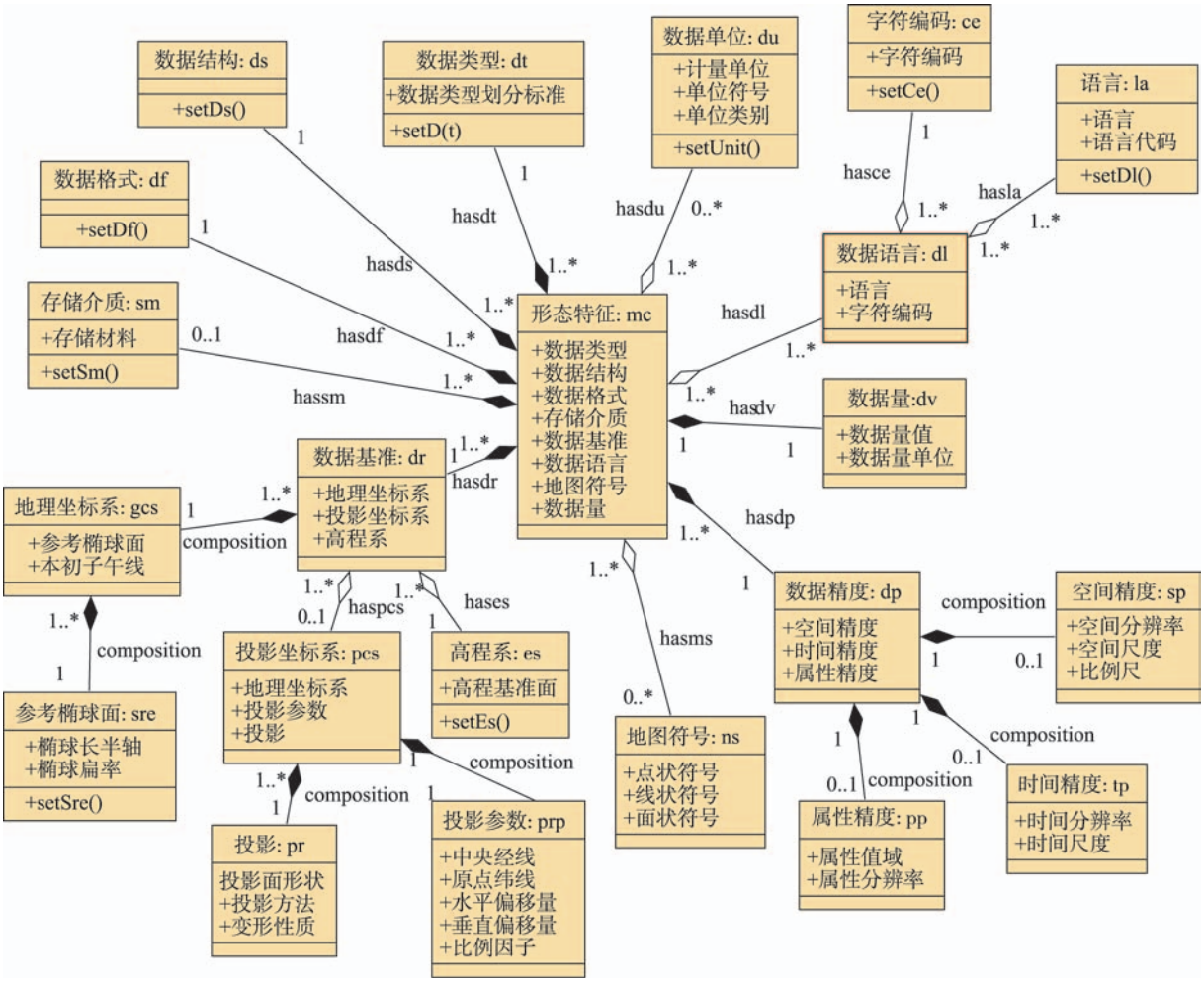

图3 地理空间数据形态本体模型的UML类图表示

Fig.3 The UML class diagram of the morphology-ontology model for geospatial data

（2）context pcs inv

self.allinstances()→isUnique(pcs.epsg) //约束投影坐标系类的所有实例必须具有唯一的EPSG代码；

（3）context pcs inv

pcs.sf<=1 //约束投影坐标系的比例因子必须小于等于1；

（4）context pr inv

self.n→includes(pcs.pr) //约束投影坐标系的投影必须是投影类的实例。

4 地理空间数据形态本体构建与应用实践

4.1 地理空间数据形态本体构建实践

本体构建是将本体概念、属性、关系、实例和约束等，实现在本体构建工具中的一个系统化工程。目前，国内外学者已经提出了很多本体构建的方法论，例如骨架法[25]，TOVE法[26]，METHONTOLOGY法[27]等。但现有的本体构建方法没有成熟的理论指导，且过于抽象，指导性、实用性不强。本文结合骨架法理论，确定本体构建的范围和边界，同时提出“自上而下”的本体概念设计和“自下而上”的本体构建方法的思想，以保证本体集成依赖程度最小以及优良的可扩展性。

地理空间数据形态本体的构建是指利用某种本体描述语言，以形态本体模型为理论指导，以上述本体构建方法为方法论，将本体的概念体系、属性、关系、约束和实例等进行表达的过程。形态本体的构建充分考虑了对现有语料库的继承，例如，投影坐标系本体采用了“欧洲石油调查组织”（European Petroleum Survey Group，EPSG）的投影坐标系代码对照字典，语言本体采用了ISO 639国际标准化组织语言编码标准。

本文根据形态本体概念体系的层次结构，“自上而下”逐级延伸，将形态本体拆分为依赖程度较小的子本体。在本体构建工具protégé中构建子本体的基本步骤：在Classes模块中添加形态特征概念以及概念之间的语义关系；在ObjectProperties和DataProperties中添加概念的对象属性和数据属性；在Individuals中添加概念的实例及其关系、属性和规则约束。从最底层子本体开始，“自下而上”逐级完善，最后通过protégé的imports功能集成为总本体。本文已经构建的子本体如表4所示。

4.2 地理空间数据形态本体在数据发现中的应用

当前，多源异构的地理空间数据发现大多以元数据为核心，图4为基于传统关键词匹配技术的数据发现策略。地理空间数据集以及其元数据存储在数据库中，用户输入检索语句后，用分词技术对检索语句进行处理并获取关键词列表，然后将关键词逐个与元数据内容进行字符匹配，并返回匹配成功的数据记录。

基于本体理论的数据发现称为语义检索，它是根据资源对象的语义关系、概念匹配以及有关推理机制来引导用户的查询和检索反馈[28]。语义检索的优势在于在信息检索过程中的检索匹配不是基于字面的机械匹配，也不是基于字段的匹配，而是基于知识单元的、面向语义的匹配，从而大大提高信息检索反馈的相关性和准确度[29]。图5为基于形

表4　已构建地理空间数据形态特征子本体

Tab.4　The constructed sub-ontologies in the morphological characteristics of geospatial data

本体名称	主要概念	实例（示例）	关系（示例）
数据类型本体	数据类型、数据类型分类标准	图像文件、文件类型	类型包含
数据结构本体	栅格数据结构、矢量数据结构	四叉树、双重独立编码结构	
数据格式本体	数据格式	.png, .tif, .shp	格式相似、格式互逆
存储介质本体	存储介质、存储材料	硬盘、磁存储材料	存储材料为
比例尺本体	比例尺、比例尺等级	1:1 000 000, 小比例尺	比例尺大于、小于
地理坐标系本体	地理坐标系、参考椭球面	WGS84坐标系、WGS 84椭球	坐标系可变换
投影坐标系本体	投影、投影面形状	高斯克吕格投影、圆柱面	投影为、投影参数为
高程系本体	高程系	黄海高程系	
语言本体	语言、语言代码	中文、英文，chi、eng	语言代码为
计量单位本体	计量单位、计量单位符号	米、平方米，m、m^2	符号为、单位可换算

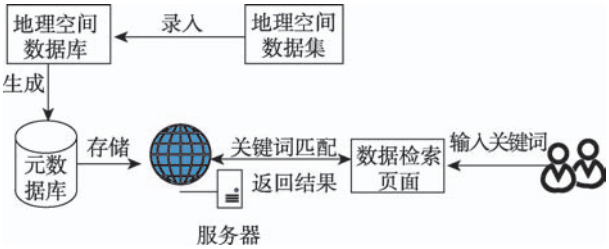

图4 基于关键词匹配的数据发现策略

Fig.4 Data discovery technology based on keywords matching

态本体的语义数据发现策略流程。对比图4,可以清晰地看到后者的改进。后者在检索过程中加入了元数据语义处理中心和形态语义转换中心,分别从元数据和关键词,即数据发现的首尾2个层面进行改进。元数据语义处理中心对元数据进行形态语义抽取和语义标注,处理完成后,输出包含有语义信息的元数据,并提交至服务器。形态语义转换中心则利用地理空间数据形态本体和形态语义推理规则库对用户输入的关键词进行解析和扩展,输出更加丰富的关键词列表组合,并提交至服务器。最后,由服务器返回数据检索结果。

语义检索的关键步骤是对形态语义推理规则库的定义。本文采用Jena推理机,实现形态本体库的推理。Jena推理机内部的推理引擎需要规则库来定义其行为,Jena自身包含了一系列通用规则,如不同类之间的关系,属性的传递、互逆等[30]。另外,用户可以定制自己的推理规则,实现个性化需求。本文按照Jena推理规则的语法,基于形态本体中构建的语义关系,自定义推理规则库,并通过引入自定义规则库文件的方式实现本体推理。形态本体自定义推理规则如下。

(1)数据格式子本体推理规则

Rule1: [格式相似: (?x GeoDataOnt:格式相似 ?y) -> (?y GeoDataOnt:格式相似 ?x)]

Rule2: [格式相同: (?x GeoDataOnt:格式相同 ?y) -> (?y GeoDataOnt:格式相同?x)]

Rule3: [格式互逆: (?x GeoDataOnt:格式可变换 ?y)(?y GeoDataOnt:格式可变换 ?x) -> (?x GeoDataOnt:格式互逆 ?y)]

(2)数据类型子本体推理规则

Rule4: [类型包含: (?x GeoDataOnt:类型包含?y)(?y GeoDataOnt:类型包含?z) -> (?x GeoDataOnt:类型包含 ?z)]

(3)数据单位子本体推理规则

Rule5: [单位可换算: (?x GeoDataOnt:单位可换算 ?y)(?y GeoDataOnt:单位可换算 ?z) -> (?x GeoDataOnt:单位可换算 ?z)]

(4)比例尺子本体推理规则

Rule6: [比例尺大于: (?x GeoDataOnt:比例尺大于?y)(?y GeoDataOnt:比例尺大于?z) -> (?x GeoDataOnt:比例尺大于?z)]

Rule7: [比例尺小于: (?x GeoDataOnt:比例尺小于?y)(?y GeoDataOnt:比例尺小于?z) -> (?x GeoDataOnt:比例尺小于?z)]

Rule8: [比例尺大小: (?x GeoDataOnt:比例尺大于 ?y) -> (?y GeoDataOnt:比例尺小于 ?x)]

Rule9: [比例尺大小: (?x GeoDataOnt:比例尺小于 ?y) -> (?y GeoDataOnt:比例尺大于 ?x)]

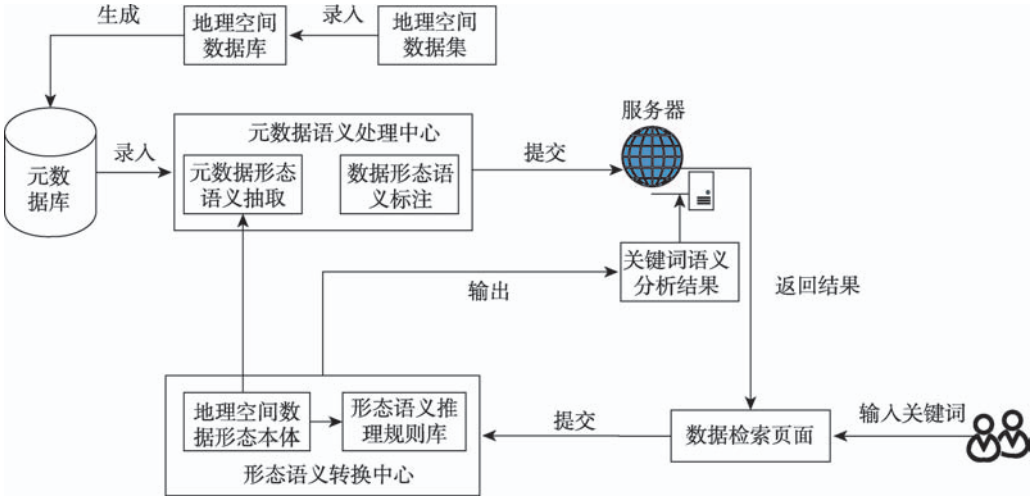

图5 基于形态本体的数据发现策略

Fig.5 Data discovery technology based on morphology-ontology

（5）坐标系子本体推理规则

Rule10: [坐标系相同: (?x GeoDataOnt:坐标系相同 ?y) -> (?y GeoDataOnt:坐标系相同?x)]

Rule11: [坐标系可变换: (?x GeoDataOnt:坐标系可变换?y)(?y GeoDataOnt:坐标系可变换?z) -> (?x GeoDataOnt:坐标系可变换?z)]

（6）投影子本体推理规则

Rule12: [投影相同: (?x GeoDataOnt:投影相同 ?y) -> (?y GeoDataOnt:投影相同?x)]

4.3 地球系统科学数据共享平台中的应用分析

国家地球系统科学数据共享平台是23个国家认定的科技平台之一，是唯一以整合共享分散的地球系统科学研究数据为重点的国家科技基础条件平台，旨在整合集成分散的、历史的、现状的和未来的地球系统科学研究产生的数据资源，为全球变化创新研究和区域可持续发展提供数据服务[31]，是一个非盈利的地球系统科学前沿研究与全球变化研究的数据支撑平台[32]。截止2015年底，平台已经整合集成了1864个元数据集，共约138 TB数据。平台数据内容复杂，来源多样，导致数据存在较为复杂的语义异构现象，且平台仍使用基于关键字匹配的数据发现策略，给平台的数据发现和推荐关联数据等服务带来了障碍。

本文应用分析的数据源是国家地球系统科学数据共享平台导出的1403条元数据，数据格式为Excel，内容覆盖全国层面的地表过程与人地关系、典型区域地表过程与人地关系、全球变化与区域响应、日地系统与空间环境等，包含了元数据标题、关键词、摘要、空间和时间范围等信息。本文首先对数据进行人工的形态语义特征提取和标注，具体过程是对应数据集，以地理空间数据形态特征概念体系为依据，在Excel表中添加其元数据的规范化形态信息；其次，以地理空间数据形态本体为本体库，以关联数据开发框架Jena为基础，开发了地理空间数据语义检索原型系统，系统的应用流程如图5所示。通过输入检索词，得到检索结果后，人工确定数据源和检索结果的真值（即数据源和检索结果中与检索词相符的数据条目），并计算结果的查全率和查准率，计算公式如式（2）–（3）所示。

$$\text{查全率}=\frac{\text{检索结果中相关数据总数}}{\text{数据源中相关数据总数}}\times 100\% \quad (2)$$

$$\text{查准率}=\frac{\text{检索结果中相关数据总数}}{\text{检索结果数据总数}}\times 100\% \quad (3)$$

本文以“矢量数据”（数据类型子本体实例）、“图片数据”（数据类型子本体实例）、“栅格数据结构”（数据结构子本体实例）和“txt”（数据格式子本体实例）为例，分别进行基于关键词和形态本体的数据检索，检索结果的统计如表5所示。以“矢量数据”为检索条件的检索结果列表如图6所示，本文以此结果为例对关键词匹配方法和语义检索的检索结果作如下分析：前者结果列表呈无序排列，难以快速定位目标数据，而后者按关联权重大小呈有序排列，关联权重越大越靠前显示，同时与用户的目标数据也越接近，方便用户迅速定位；在本体中shp、shapefile、coverage等实例与“矢量数据”存在“数据类型为”的关系，因而后者通过智能关联得到了更多的相关数据，查全率得到了显著提高；原型系统没有对检索词作分词处理，以及数据源中形态特征标注过于简单，使2种方法的查准率都较高，未能显著体现基于形态本体的语义检索方法对于提高查准率的有效性。

5　结语

地理空间数据在空间、时间、内容、形态和来源等层次的语义异构是实现数据精确发现和智能推荐的主要障碍，也是目前的研究热点。本文以地理空间数据形态特征为研究对象，深入分析并总结了形态特征的内涵，建立了包含数据组织、数据存储、数据精度、数据基准和数据语言等特征的概念体系。基于此，结合本体理论，提出了形态语义信息

表5　关键词匹配和语义检索的检索结果对比

Tab.5　Comparative results of methods based on the keywords matching and the semantic retrieval

检索词	数据资源真值	基于关键词检索				基于形态本体检索			
		检索结果数	检索结果相关数	查全率/(%)	查准率/(%)	检索结果数	检索结果相关数	查全率/(%)	查准率/(%)
矢量数据	256	53	40	15.63	75.47	218	218	85.16	100
图片数据	81	30	24	29.63	80.00	79	79	97.53	100
栅格数据结构	208	84	76	36.54	90.48	169	169	81.25	100
txt	285	254	237	83.16	93.31	285	285	100.00	100

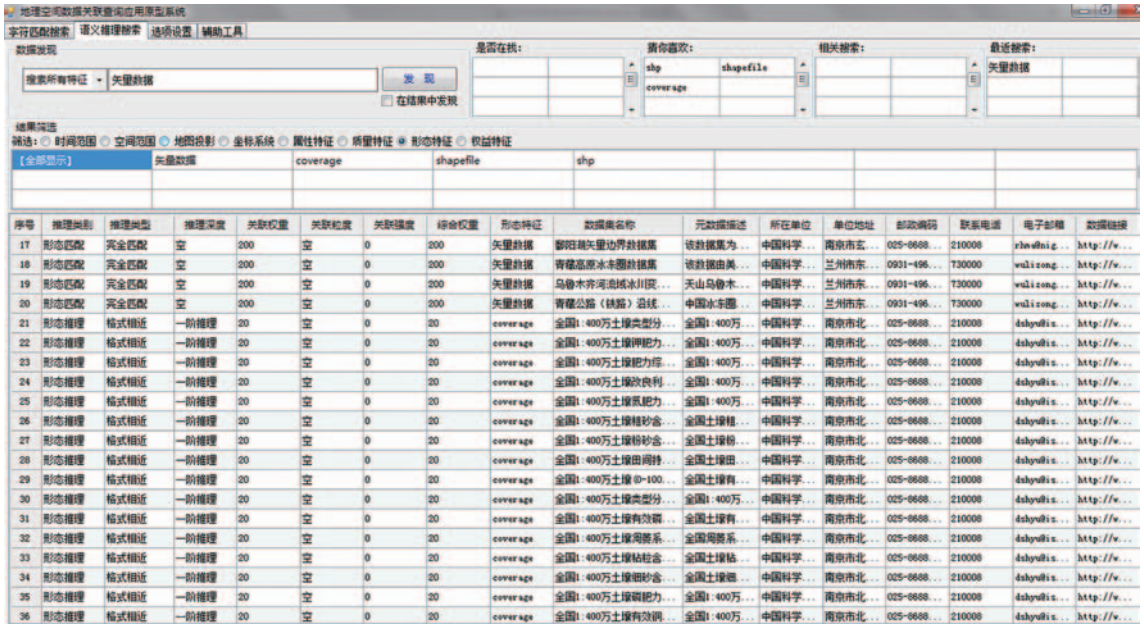

图6 基于形态本体的语义检索结果

Fig.6 The result of semantic retrieval based on the morphology-ontology of geospatial data

的建模和本体表示方法,并构建了地理空间数据形态本体。最后,将形态本体在国家地球系统科学数据共享平台中进行了模拟应用分析,应用结果表明,地理空间数据形态本体可以有效地解决地理空间数据形态特征的语义异构,提高数据发现的查全率和查准率,提升数据发现服务。

基于形态本体的语义数据发现策略仍有很多亟待解决的难题:形态本体构建、数据语义信息抽取和标注等技术仍然依赖人工处理,自动化远未实现;应对数据形态特征的变更和扩展的形态本体自动实时更新技术也尚未实现;囊括地理空间数据空间、时间、内容、形态和来源等完整数据描述的关联指标体系仍没有重大进展。尤其对于数据形态语义的关联指标体系(即数据形态语义关联权重的计算研究),目前在文献检索中还很少,今后需要开展进一步的研究。

致谢:感谢国家地球系统科学数据共享平台提供的实验数据。

参考文献(References):

[1] Yuan J, Yue P, Gong J Y, *et al.* A linked data approach for geospatial data provenance[J]. IEEE Transactions on Geoscience and Remote Sensing, 2013,51(11):5105-5112.

[2] 董少春,尹宏伟,许刚.地质时间本体在异构数据检索中的应用[J].地球信息科学学报,2010,12(2):194-199. [Dong S C, Yin H W, Xu G. Heterogeneous data searching based on geologic time ontology[J]. Journal of Geoinformation Science, 2010,12(2):194-199.]

[3] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods[J]. Data & Knowledge Engineering, 1998,25(1):161-197.

[4] 瞿裕忠,胡伟,郑东栋,等.关系数据库模式和本体间映射的研究综述[J].计算机研究与发展,2015,45(2):300-309. [Qu Y Z, Hu W, Zheng D D, *et al.* Mapping between relational database schemas and ontologies: the state of the art [J]. Journal of Computer Research and Development, 2015,45(2):300-309.]

[5] Wache H, Voegele T, Visser U, *et al.* Ontology-based integration of information- a survey of existing approaches [C]. IJCAI- 01 Workshop: Ontologies and Information Sharing, 2001:108-117.

[6] 陈建军,周成虎,王敬贵.地理本体的研究进展与分析[J].地学前缘,2006,13(3):81-90. [ChenJ J, Zhou C H, Wang J G. Advances in the study of the geo-ontology[J]. Earth Science Frontiers, 2006,13(3):81-90.]

[7] Arpinar I B, Sheth A P, Ramakrishnan C, *et al.* Geospatial ontology development and semantic analytics[J]. Transactions in GIS, 2006,10(4):551-575.

[8] Grüninger M, Ong D. Verification of time ontologies with points and intervals[C]. 2011 Eighteenth International Symposium on Temporal Representation and Reasoning (TIME), 2011:31-38.

[9] Guo M. The application of ontology in semantic discovery for GeoData web service[J]. Communications and Network, 2013,5(3):678-680.

[10] Kuhn W. Modeling the semantics of geographic catego-

ries through conceptual integration[A]. In: Geographic Information Science[M]. Berlin: Springer Berlin Heidelberg, 2002:108-118.

[11] Tomai E, Kavouras M. From "onto-geonoesis" to "onto-genesis": the design of geographic ontologies[J]. Geoinformatica, 2004,8(3):285-302.

[12] Vandecasteele A, Napoli A. Spatial ontologies for detecting abnormal maritime behaviour[C]. Oceans 2012 MTS/IEEE Yeosu Conference, 2012:1-7.

[13] Liu W, Gu H, Peng C, *et al.* Ontology-based retrieval of geographic information[C]. IEEE 18th International Conference on Geoinformatics, 2010:1-6.

[14] de Laborda C P, Conrad S. Relational.OWL: a data and schema representation format based on OWL[C]. Proceedings of the 2nd Asia-Pacific Conference on Conceptual Modelling, 2005,43:89-96.

[15] Research Data Alliance. Data type registries[EB/OL]. https://rd- alliance.org/groups/data- type- registries- wg.html, 2015-09-09.

[16] 李庭波,陈平留,郑德祥.基于森林资源数据结构的本体学习探索[J].西南林学院学报,2009,29(2):57-61. [Li T B, Chen P L, Zheng D X. Ontology learning based on data structure of forest resource[J]. Journal of Southwest Forestry University, 2009,29(2):57-61.]

[17] 郝亚南,陈少飞,李天柱,等.基于语义的数据格式转换[J].计算机系统应用,2005(11):40-43. [Hao Y N, Chen S F, Li T Z, *et al.* Data transformation with semantics[J]. Computer Systems & Applications, 2005,11:40-43.]

[18] 苏里.基于描述逻辑的地图语义模型初步研究[J].测绘科学,2009,34(3):92-93. [Su L. Study of cartographic semantics model based on description logic[J]. Science of Surveying and Mapping, 2009,34(3):92-93.]

[19] 杨小忠,贾占军,刘士彬,等.基于应用本体的多卫星遥感数据检索[J].遥感信息,2007(1):30-36. [Yang X Z, Jia Z J, Liu S B, *et al.* Multi-satellites remote sensing data retrieval based on application ontology[J]. Remote Sensing Information, 2007,1:30-36.]

[20] 金芝.基于本体的需求自动获取[J].计算机学报,2000,23(5):486-492. [Jin Z. Ontology-based requirements elicitation[J]. Chinese Journal of Computers, 2000,23(5):486-492.]

[21] 王洪伟,吴家春,蒋馥.基于描述逻辑的本体模型研究[J].系统工程,2003,21(3):101-106. [Wang H W, Wu J C, Jiang F. A study on ontology model based on description logics[J]. Systems Engineering, 2003,21(3):101-106.]

[22] Naing M M, Lim E P, Goh D H L. Ontology-based web annotation framework for hyperlink structures[C]. Proceedings of Third International Conference on Web Information Systems Engineering, 2002.

[23] Gómez-Pérez A, Benjamins R. Overview of knowledge sharing and reuse components: ontologies and problem-solving methods[C]. IJCAI and the Scandinavian AI Societies. CEUR Workshop Proceedings, 1999.

[24] 李嘉丽,王念滨,孙玮鸿.基于 UML 的本体表示方法研究[J].计算机工程,2009,35(12):41-43. [Li J L, Wang N B, Sun W H. Research on UML-based ontology representation method[J]. Computer Engineering, 2009,35(12):41-43.]

[25] Uschold M, Gruninger M. Ontologies: principles, methods and applications[J]. The Knowledge Engineering Review, 1996,11(2):93-136.

[26] Grüninger M, Fox M S. Methodology for the design and evaluation of ontologies[C]. Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[27] Gómez-Pérez A. Knowledge sharing and reuse[J]. Handbook of Applied Expert Systems, 1998,10:1-36.

[28] 宋佳,王卷乐,诸云强,等.基于地理空间本体的语义检索相关度研究[J].计算机工程与应用, 2011,47(5):114-117. [Song J, Wang J L, Zhu Y Q, *et al.* Research on relevancy for semantic retrieval based on geographic spatio-ontology[J]. Computer Engineering and Applications, 2011, 47(5):114-117.]

[29] 颜端武,丁晟春,李岳蒙,等.基于语义 Web 和 Jena 插件的语义检索系统实验研究[J].情报理论与实践,2006,29(3):349-352. [Yan D W, Ding S C, Li Y M, *et al.* Research on experimental system of semantic retrieval based on semantic Web and Jena plugin[J]. Information Studies: Theory & Application, 2006,29(3):349-352.]

[30] 田宏,马朋云.基于 Jena 的城市交通领域本体推理和查询方法[J].计算机应用与软件,2011,28(8):57-59. [Tian H, Ma P Y. A reasoning and query method for urban transportation domain ontology based on Jena[J]. Computer Applications and Software, 2011,28(8):57-59.]

[31] 诸云强,宋佳,冯敏,等.地球系统科学数据共享软件研究与发展[J].中国科技资源导刊,2012(6):11-16. [Zhu Y Q, Song J, Feng M, *et al.* Research and development of software of earth system science data sharing[J]. China Science & Technology Resources Review, 2012,6:11-16.]

[32] 诸云强,孙九林,廖顺宝,等.地球系统科学数据共享研究与实践[J].地球信息科学学报,2010,12(1):1-8. [Zhu Y Q, Sun J L, Liao S B, *et al.* Earth system scientific data sharing research and practice[J]. Journal of Geo-information Science, 2010,12(1):1-8.]