**DBpedia知识库本体分析**

**[摘要]** 在现有的语义网项目架构中，基于关联数据形式的知识库项目往往处于整个语义网络的核心，如何对于这些知识库的知识内容进行组织、储存和查找就成为了决定整个语义网络运行效率的关键因素。在目前的关联数据知识库项目中，DBpedia是较为典型且成熟的一个，DBpedia网站使用本体的方法来对其条目内容进行组织和存储，本文旨在通过对DBpedia现有本体结构的分析来说明知识库的本体结构对于知识库的组织、存储和查找有着怎样的影响，并试图从该例中分析归纳得出类似网站知识库内容的本体构建的一般要点。

**[关键词]** DBpedia 关联数据 本体 本体构建

1. **概述及相关简介**

1998年，WWW网络的发明者Berners-Lee提出了语义网的概念。这一概念的核心在于致力提高万维网络及其互联的资源的可用性和有效性，使得下一代的互联网更加智能和高效，能够有效处理目前网络中的大量信息内容。这一概念和其具体的技术实现几经波折，从一开始的基于本体的构想到2006年Berners-Lee提出的关联数据概念，在目前的语义网构想中，关联数据成为了其技术实现的核心概念。

关联数据是一种推荐的最佳实践，用来在与以往中使用URI和RDF发布、分享、连接各类数据、信息和只是，发布和部署实例数据和类数据，从而通过HTTP协议解释并获取这些数据同时强调数据的相互关联、相互联系以及有益于人际理解的语境信息。在目前的具体实践中，数据往往以RDF文件的形式发布到互联网络上，存储在关联数据知识库中。而大多数需要使用这些关联数据的网站可以直接从在线关联数据知识库的数据接口获取RDF文件并提取其中的相关信息反馈给用户，从而实现信息和数据的跨网站共享。从上面不难看出，在线关联数据知识库在当前的关联数据语义网构想中占据着核心位置。

DBpedia就是这样一个在线关联数据知识库项目。它从维基百科的词条中抽取结构化数据，以提供更准确和直接的维基百科搜索，并在其他数据集和维基百科之间创建连接，并进一步将这些数据以关联数据的形式发布到互联网上，提供给需要这些关联数据的在线网络应用、社交网站或者其他在线关联数据知识库。同时，与一些同时期的关联数据知识库项目不同，DBpedia提供的关联数据知识内容并不受限于某一较小的专业或学科领域，到2012年为止，DBpedia知识库项目已有超过3640000个条目，这些条目涉及人名、地名、音乐专辑、电影、组织、种族等等多个类目。在大数量和宽范围的前提下，DBpedia知识库的条目组织就成为决定其运行效率的关键因素。DBpedia知识库以构建本体的形式对条目进行组织，起到了一定的效果。

本体一词是指对客观世界存在的事物的系统描述，在信息科学的领域中，1991年Neches等人最早给出的定义是：给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成，规定这些词汇外延规则。1993年Gruber将其定义表述为“概念模型的明确的规范说明”。1997年Borst进一步完善了该本体的定义，表述为“共享概念模型的形式化规范说明”，增加了“共享”和“形式化”两个限定语。1998年Studer等人在对Gruber和Borst的定义进行了深入研究后，将本体进一步表述为“共享概念模型的明确的形式化规范说明”并从此形成了目前信息学界对本体概念比较一致的看法。

本文以DBpedia的本体结构为实例，从类与条目属性以及本体结构两个方面，分析DBpedia本体的构建思路。

1. **DBpedia本体的类和条目属性分析**

如前所述，本体是对共享概念模型的明确的形式化规范说明。根据斯坦福大学医学院发本体的构建七步法，构建本体有如下七个步骤：

1. 确定本体的专业领域与范畴
2. 考察复用现有本体的可能性
3. 列出本体中的重要术语
4. 定义类和类的等级关系
5. 定义类的属性
6. 定义属性的分面
7. 创建实例

但以上所述的七步法仅适用于某一专业或学科领域本体的构建。在DBpedia知识库一例中，由于所需要构建本体的是维基百科条目，因此其专业领域与范畴无法简单确定，因此在DBpedia的本体构建中，采用了类似于以叙词表为基础的自动化本体转换的方式。

但同样的，维基百科的词条表并非规范的叙词表，其类目划分也无法简单归于某一专业分类之下，词语的专指度不可避免的出现不足，因此，在DBpedia的本体构建中，类目划分和定义基本是人工完成，之后的条目转换才使用了自动转换的方式。

**2.1 DBpedia类共有属性分析**

本体的体系结构主要包括3个要素：核心元素集、元素间的交互作用以及这些元素到规范语义间的映射关系。对于本体的一般构造过程，可以用以下公式进行表示：

*本体 = 概念（Concept） + 属性（Property） + 公理（Axiom）+ 取值（Value）+ 名义（Nominal）*

从以上的一般公式中，我们还可以归纳出本体当中的基本关系类型，共有四类：

1. part-of关系，表达概念之间部分和整体的关系
2. kind-of关系，表达概念之间的集成关系
3. instance-of关系，表达概念的实例和概念之间的关系
4. attribute-of关系，表达某个概念是另一个概念的属性

由于DBpedia知识库的本体从分类上属于知识本体，因此在类和条目的属性设置上需要体现以上四种关系。

如图1所示，DBpedia的类除了名称这一唯一标示之外，还有如下共有属性：



图 1 DBpedia类属性实例

1. Type属性，类型是类与条目的根本性区分，以图中‘Film’这个类为例，它的类型属性值是‘Class’，而该类下属的具体电影条目的属性则是‘Film’。
2. Istypeof属性，这一属性列出了该类下属的所有条目
3. SubclassOf属性，这一属性表明该类的上级类
4. IsSubclassOf属性，这一属性表明该类的下级类
5. Sameas属性，这一属性值指向一个URI，其中说明了所有跟该类同义的词语

以上是DBpedia本体类的五个共有属性，从该属性定义中可以看出，由于条目数量多，类目归属复杂，DBpedia类的属性定义较为简单。但仍能体现概念之间的关系。此外，由于名称的复用可能存在，因此DBpedia类和条目的属性命名匹配规则都较为严苛。仍以图中Film类为例，当以film为索引词查找该类时，无法直接得到反馈结果。由于DBpedia的使用用户生成内容较为随意，在创建本体时为类和条目添加了Sameas属性，以反映不同名称的同一事物。

**2.2 DBpedia条目属性分析**

在DBpedia本体体系当中，条目与类是两种不同属性的概念。简单的区分而言，DBpedia本体体系中的条目相当于一般本体体系中的概念、属性和取值的集合，三者共同描述某种客观事物，而类则相当于一般意义上的分类概念，并不指代某种具体事物，而是一类事物的共有概念。

因此，DBpedia本体的条目属性除了表明其与类和其他概念的关系和联系之外，还根据其所属的类拥有各自独特的属性。

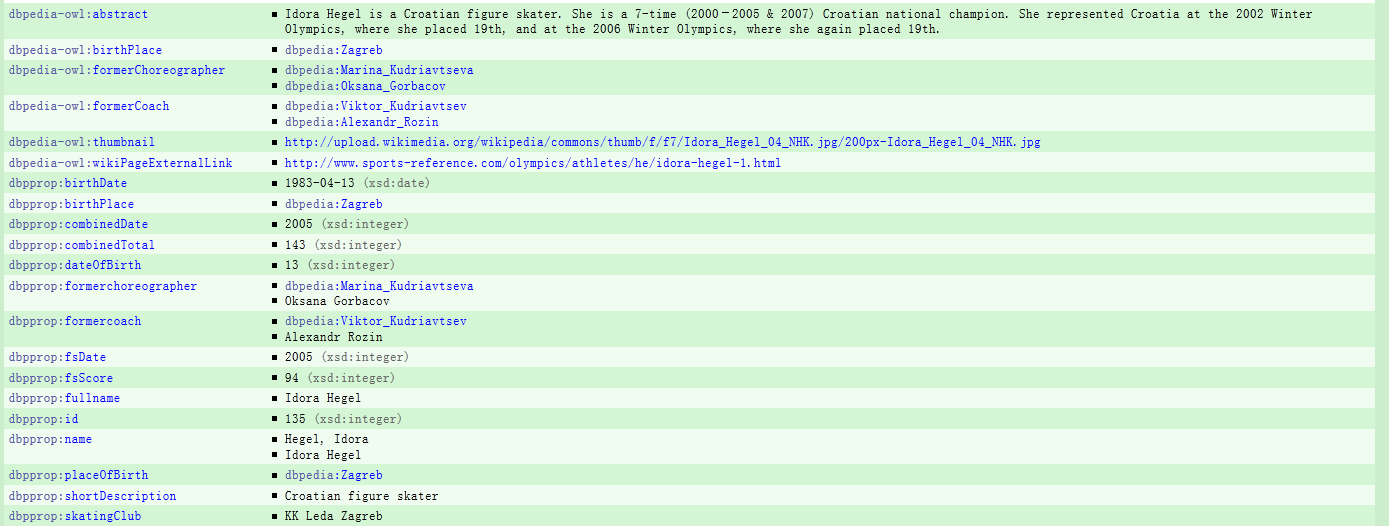


图2 DBpedia条目属性实例

从实例条目中，DBpedia条目所共有的关系属性仅有一条，即：

1. Type，如类属性，在不同的本体语言下其Type定义不同。在DBpedia本体语言下，条目的Type属性是其所有上级类。如例所示，该条目所描述的运动员从属于“运动员”、“花样滑冰运动员”和“人”三个类，因此其Type属性在DBpedia本体语言下有三个值。

而根据其从属类的不同，DBpedia本体体系中的条目拥有独特的用以说明其概念的属性，如上例的运动员有“出生年月”、“性别”等属性。

**2.3 DBpedia本体属性设置的评析**

从实例分析中，不难看出，DBpedia的本体设计和构建符合以上本体构建原则。尽管由于原维基百科的条目存在重名，以及单个概念在不同领域的不同含义等问题，但在DBpedia本体的构建过程中，以名称为区分条目和类的最根本区别。使用了添加后缀等命名规则来区分在原维基百科条目中重名条目。同时，针对不同类或条目所涉及的不同概念，合理的添加了能够反映其概念特质的属性。

同时，针对不同本体语言，DBpedia还对同一属性做了不同取值，这样在使用和查询的过程中，能够用多种本体语言对DBpedia知识库的数据内容进行查找。

**3. DBpedia本体结构分析**

基于DBpedia类和条目的属性，DBpedia本体的结构较为简明，规则大致如下：

1. 以“事物（Thing）”这一概念来总括描述DBpedia本体中所涉及到的所有客观事物。
2. 事物（Thing）类下分有不同的子类，每一个子类都可以有它自己的子类
3. 条目用以描述具体的事物，从属于某个类和它的全部上级类

**3.1 DBpedia本体结构表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 事物 | 活动 | 游戏 |  |  |  |
| 运动 |  |  |  |
| 解剖结构 | 动脉 |  |  |  |
| 骨骼 |  |  |  |
| 脑 |  |  |  |
| 胚胎 |  |  |  |
| 淋巴 |  |  |  |
| 肌肉 |  |  |  |
| 神经 |  |  |  |
| 静脉 |  |  |  |
| 奖杯 |  |  |  |  |
| 化学物质 | 化合物 |  |  |  |
| 颜色 |  |  |  |  |
| 货币 |  |  |  |  |
| 数据库 | 生物数据库 |  |  |  |
| 装置 | 汽车引擎 |  |  |  |
| 武器 |  |  |  |
| 疾病 |  |  |  |  |
| 药物 |  |  |  |  |
| 民族 |  |  |  |  |
| 事件 | 协约 |  |  |  |
| 选举 |  |  |  |
| 电影节 |  |  |  |
| 武装冲突 |  |  |  |
| 音乐节 |  |  |  |
| 太空任务 |  |  |  |
| 体育项目 | 足球赛 |  |  |
| 国际长途汽车大奖赛 |  |  |
| 武术混合项目 |  |  |
| 奥林匹克 |  |  |
| 赛跑 |  |  |
| 女子网球联合会赛事 |  |  |
| 摔跤 |  |  |
| 太空航行年份 |  |  |  |
| 食物 | 饮料 |  |  |  |
| 假日 |  |  |  |  |
| 语言 |  |  |  |  |
| 案件 | 美国最高法院案件 |  |  |  |
| 交通方式 | 飞机 |  |  |  |
| 汽车 |  |  |  |
| 火车 |  |  |  |
| 火箭 |  |  |  |
| 轮船 |  |  |  |
| 太空舱 |  |  |  |
| 空间站 |  |  |  |
| 航天飞行器 |  |  |  |
| 音乐流派 |  |  |  |  |
| 姓名 | 名 |  |  |  |
| 奥林匹克成绩 |  |  |  |  |
| 组织 | 乐团 |  |  |  |
| 广播电视台 | 广播网络公司 |  |  |
| 广播台 |  |  |
| 电视台 |  |  |
| 公司 | 航空公司 |  |  |
| 唱片公司 |  |  |
| 教育组织 | 学院 |  |  |
| 学校 |  |  |
| 大学 |  |  |
| 政府机构 |  |  |  |
| 立法机关 |  |  |  |
| 军事单位 |  |  |  |
| 非营利性组织 |  |  |  |
| 政党 |  |  |  |
| 运动联赛 | 美式橄榄球联赛 |  |  |
| 澳式足球联赛 |  |  |
| 高速赛车联赛 |  |  |
| 篮球联赛 |  |  |
| 棒球联赛 |  |  |
| 加拿大式橄榄球联赛 |  |  |
| 板球联赛 |  |  |
| 曲棍球联赛 |  |  |
| 高尔夫球联赛 |  |  |
| 冰球联赛 |  |  |
| 手球联赛 |  |  |
| 轮滑曲棍球联赛 |  |  |
| 长曲棍球联赛 |  |  |
| 武术混合联赛 |  |  |
| 摩托赛车联赛 |  |  |
| 马球联赛 |  |  |
| 英式橄榄球联赛 |  |  |
| 足球联赛 |  |  |
| 垒球联赛 |  |  |
| 速滑联赛 |  |  |
| 机车联赛 |  |  |
| 网球联赛 |  |  |
| 电子游戏联赛 |  |  |
| 排球联赛 |  |  |
| 运动队伍 | 美式橄榄球队 |  |  |
| 篮球队 |  |  |
| 加拿大式橄榄球队 |  |  |
| 曲棍球队 |  |  |
| 足球俱乐部 |  |  |
| 摩托车队 |  |  |
| 公会 |  |  |  |
| 人物 | 大使 |  |  |  |
| 建筑师 |  |  |  |
| 艺术家 | 演员 | 成人演员 |  |
| 喜剧演员 |  |  |
| 动漫画家 |  |  |
| 音乐人 |  |  |
| 作家 |  |  |
| 宇航员 |  |  |  |
| 运动员 | 澳式足球运动员 |  |  |
| 羽毛球运动员 |  |  |
| 棒球运动员 |  |  |
| 篮球运动员 |  |  |
| 拳击手 |  |  |
| 板球手 |  |  |
| 自行车手 |  |  |
| 花样滑冰运动员 |  |  |
| 一级方程式赛车手 |  |  |
| 盖尔运动员 |  |  |
| 高尔夫球手 |  |  |
| 美式橄榄球运动员 | 美国橄榄球运动员 |  |
| 冰球运动员 |  |  |
| 武术家 |  |  |
| 全美汽车比赛选手 |  |  |
| 大学生运动会运动员 |  |  |
| 英式橄榄球运动员 |  |  |
| 斯诺克运动员 | 斯诺克冠军 |  |
| 足球运动员 |  |  |
| 网球运动员 |  |  |
| 摔跤手 |  |  |
| 英国皇室 |  |  |  |
| 牧师 | 红衣主教 |  |  |
| 基督主教 |  |  |
| 圣徒 |  |  |
| 足球教练 |  |  |  |
| 罪犯 |  |  |  |
| 虚构人物 | 动漫人物 |  |  |
| 记者 |  |  |  |
| 法官 |  |  |  |
| 军事人员 |  |  |  |
| 模特 |  |  |  |
| 君主 |  |  |  |
| 官员 |  |  |  |
| 哲学家 |  |  |  |
| 《花花公子》玩伴女郎 |  |  |  |
| 牌手 |  |  |  |
| 政客 | 大法官 |  |  |
| 国会议员 |  |  |
| 州长 |  |  |
| 市长 |  |  |
| 议会议员 |  |  |
| 总统 |  |  |
| 首相 |  |  |
| 参议员 |  |  |
| 科学家 |  |  |  |
| 足球经理 |  |  |  |
| 地点 | 建筑结构 | 建筑 | 历史建筑 |  |
| 医院 |  |
| 旅店 |  |
| 灯塔 |  |
| 博物馆 |  |
| 餐厅 |  |
| 超市 |  |
| 体育场 |  |
| 剧院 |  |
| 基础设施 | 机场 |  |
| 火箭发射台 |  |
| 发电厂 |  |
| 道路交通 | 桥梁 |
| 公交中转系统 |
| 铁路 |
| 公路 |
| 三岔路口 |
| 隧道 |
| 车站 |  |
| 公园 |  |  |
| 历史名胜 |  |  |
| 纪念碑 |  |  |
| 自然景点 | 水体 | 湖泊 |  |
| 流动水体 | 运河 |
| 河流 |
| 山洞 |  |  |
| 环形山 |  |  |
| 山 |  |  |
| 山脉 |  |  |
| 峡谷 |  |  |
| 人类居住地 | 行政区划 |  |  |
| 大陆 |  |  |
| 国家 |  |  |
| 岛屿 |  |  |
| 定居地 | 城市 |  |
| 城镇 |  |
| 村庄 |  |
| 保护区 |  |  |  |
| 滑雪区 |  |  |  |
| 产酒区 |  |  |  |
| 世界文化遗产保护地 |  |  |  |
| 星球 |  |  |  |  |
| 编程语言 |  |  |  |  |
| 计划 | 研究计划 |  |  |  |
| 蛋白质 |  |  |  |  |
| 单曲 |  |  |  |  |
| 斯诺克世界排名 | |  |  |  |
| 种群 | 古细菌 |  |  |  |
| 细菌 |  |  |  |
| 真核生物 | 动物 | 两栖类 |  |
| 蛛型类 |  |
| 鸟类 |  |
| 甲克类 |  |
| 鱼类 |  |
| 昆虫 |  |
| 哺乳类 |  |
| 软体类 |  |
| 爬行类 |  |
| 真菌 |  |  |
| 植物 | 石松类 |  |
| 针叶类 |  |
| 苏铁类 |  |
| 蕨类 |  |
| 有花植物 | 葡萄 |
| 银杏类 |  |
| 苞果类 |  |
| 绿藻类 |  |
| 苔类 |  |
| 劳动成果 | 电影 |  |  |  |
| 音乐 |  |  |  |
| 乐曲 | 专辑 |  |  |
| 单曲 | 欧洲歌唱大赛参赛曲 |  |
| 软件 | 电子游戏 |  |  |
| 电视剧 |  |  |  |
| 真人秀 |  |  |  |
| 网站 |  |  |  |
| 书面成果 | 书 |  |  |
| 期刊文献 | 学术期刊 |  |
| 杂志 |  |
| 报纸 |  |
| 戏剧 |  |  |
| 其他类 |  |  |  |  |

**3.2 DBpedia本体结构的优点和存在的问题**

从上表可以看出，DBpedia的本体结构从整体上仍借鉴和使用了维基百科的类目分类方法。由于其本体结构较为简单，类之间只有不同层之间的从属关系而没有同层之间的其他关系，因此在维护和添加数据时较为方便，可扩展性良好。对于数据扩充极为迅速的DBpedia项目而言，这种结构有利于它的数据扩充和维护。

但DBpedia的本体结构也存在问题，主要是本体类的设置较为随意，同层次之间的类概念语义联系极弱。在某些领域的类划分过于详细，其他则比较粗糙。这就造成在目前的阶段，DBpedia在线知识库的搜索和查询效率不高，在线查询的稳定性差等问题。

**4. 结论**

DBpedia本体结构和其类目及条目属性的设置既有劣势，也有优点。其劣势主要在于同层次概念之间缺乏关系所导致的类概念之间的联系较为松散，类定义较为随意在特定领域难以准确体现从属关系，其优势在于对于概念的明确表达、本体结构的优秀可扩展性、对于多种本体语言的支持。

因此，我们在构建知识本体时需要注意借鉴其优势，避免其不足，构建类似知识本体的原则总结归纳如下：

（1）明确性和客观性。本体建立的基本原则就是要使概念的表达准确明晰,符合实际,因此对本体的描述要尽可能使用标准术语。

（2）连贯性和可扩展性。本体中通过推断与学习获得的概念知识和已定义的概念知识必须保持一致;同时,本体中要预留对已建本体作进一步补充说明的空间,以方便人们对特殊应用定义新的概念项,换句话说,就是在向 Ontology中添加通用或专用的术语时,可以不再修改原有的内容。

（3）层次性和交互性。构建本体时使用概念层次关系和多重继承机制可以增加本体的表达能力但同层次概念间应保持最小的语义距离; 本体原型建立之后,还要保持本体的交互性,以便于领域专家的评估、检验和修正。

（4）最小编码偏差。本体的概念知识不应依赖于编码符号,如果不是知识内容的需要,而是为了标注或实现上的方便而做出某种表示选择, 就可能导致编码偏差。这种偏差应保持最小, 以不影响本体的知识表示。

（5）最小本体承诺。当某个知识系统使用知识的行为与某个本体中的定义相一致时, 就可以称这个知识系统承诺了该本体。这种承诺是基于知识层级的,在实际应用中,本体中定义了用于在知识系统间进行交换的词表,而本体承诺则是以连贯和一致的方式使用这个共享词表的协议。在本体设计时,应该在支持知识共享活动要求的基础上尽可能地减少本体承诺,即允许对本体进行承诺的知识系统根据自身需要自由地对本体进行专门化和实例化。

**参考文献**

1. Mike Uschold, Michael Gruninger. Ontologies: Principles, Methods and Applications. Knowledge Engi neering Review , 1996 , 11( 2) : 5- 33.
2. Bizer, Christian; Lehmann, Jens; Kobilarov, Georgi; Auer, Soren; Becker, Christian; Cyganiak, Richard; Hellmann, Sebastian (September 2009). "DBpedia - A crystallization point for the Web of Data". Web Semantics: Science, Services and Agents on the World Wide Web 7 (3): 154–165.
3. 邓志鸿, 唐世渭, 张铭等. Ontology研究综述. 北京大学学报(自然科学版). 2002，38(5):730-738.