Journal of Computer Applications
计算机应用,2014,34(8):2192 - 2196,2201
ISSN 1001-9081
CODEN JYIIDU
2014-08-10
http://www.joca.cn

文章编号:1001-9081(2014)08-2192-05
doi:10.11772/j.issn.1001-9081.2014.08.2192

电影本体的构建与一致性分析

高小龙[1],朱信忠[1*],赵建民[1],曹存根[2],徐慧英[1],吴　德[3]
(1.浙江师范大学 数理与信息工程学院,浙江 金华 321004;　2.中国科学院 计算技术研究所,北京 100080;
3.金华市高新技术产业园区管委会 高新技术产业局,浙江 金华 321017)
(*通信作者电子邮箱 zxz@zjnu.cn)

摘　要:针对移动网络对电影服务系统提出的更高要求和电影领域知识描述的不足,阐述构建电影领域本体(MO)的必要性和可行性:首先,概述电影本体的对象和组成,提出构建电影领域本体模型的原则和方法,利用 OWL 和 Protege 4.1 构建模型;其次,详细说明电影本体的类、属性、实例、公理和推理规则的具体表述;最后,分析电影本体的一致性,包括类间关系的一致性分析和基于公理的一致性分析。

关键词:电影本体;OWL;推理规则;一致性分析
中图分类号:TP302;TP391　　**文献标志码**:A

Building and consistency analysis of movie ontology

GAO Xiaolong[1], ZHU Xinzhong[1*], ZHAO Jianmin[1], CAO Cungen[2], XU Huiying[1], WU De[3]
(1. *College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua Zhejiang* 321004, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*;
3. *High-Tech Industry Bureau, Administrative Committee of Jinhua High and New Technology Industrial Park, Jinhua Zhejiang* 321017, *China*)

Abstract: To tackle the higher requirement of mobile network for movie service system and the lack of description of movie domain knowledge, the necessity and feasibility of establishing the Movie Ontology (MO) were illustrated. Firstly, the objects and components of MO were summarized, and the principle and method for building the MO model were also put forward, with using the Web Ontology Language (OWL) and Protege 4.1 to build the model. After that, the concrete representation of the class, property, individual, axioms and inference rules in the MO were explained. Finally, the consistency of MO was analyzed, including the consistency analysis of relationship between classes and the consistency analysis based on axioms.

Key words: Movie Ontology (MO); Web Ontology Language (OWL); inference rule; consistency analysis

0　引言

科学技术的发展让人类由信息化时代逐步迈入知识化时代,因此计算机应用领域迫切需要学科专业知识的支撑。作为一门社会学科,电影在影视检索系统和影视教学系统等方面具有重要的作用。近年来网络中电影信息海量出现,产生的资源无结构化、非标准化问题日益严重。利用传统组织文本和图片的方式管理电影资源,已经无法满足信息系统的要求和网民消费的需求。

本体(Ontology)[1]是近些年人工智能领域快速发展的一项技术,它在知识获取、分析、表示和处理等方面具有优势。本体研究的主要目标是实现某种程度的知识共享和重用。本体能够辅助用户实现这一目标的原因包括:首先,本体分析澄清领域知识的结构,识别相关领域概念的本质,为知识共享打好基础;其次,本体对领域知识的分析结果可以重用,避免重复分析;最后,本体采用语义明确、定义统一的术语和概念,使知识共享成为可能。目前本体在知识工程、自然语言处理和基于知识的信息系统等方面得到广泛的应用[2];因此构建电影本体,利用本体思想处理电影领域的知识,是一项具有重要意义的工作。

领域本体的研究主要集中在医学、地理、农业和生物等领域[3],在电影领域本体的建立和电影知识获取方面,国内外的工作相对较少。具有代表性的电影领域本体是 W3C 利用 OWL(Web Ontology Language)[4]开发的 MO(Movie Ontology)[5]。但是 MO 存在两方面的缺陷:1)MO 中缺乏能够反映电影本质信息的知识,比如电影的主题、场景和票房等;2)MO为英文版本的电影本体,不利于描述国内电影知识。针对以上不足,本文提出一种"三阶段"电影本体构建方法,建立一个涵盖电影本质知识、可共享的中文版本电影本体,并对电影本体模型进行一致性分析。

1　电影本体

1.1　电影对象

首先确定电影本体的内容和范围。电影与传统的文本、

收稿日期:2014-04-15;**修回日期**:2014-05-15。
作者简介:高小龙(1988 -),男,山西临汾人,硕士研究生,主要研究方向:本体理论、云计算;　朱信忠(1975 -),男,江苏沛县人,副教授,主要研究方向:云计算、模式识别与数字工程、多媒体图像检索、智能信息处理;　赵建民(1951 -),男,上海人,教授,博士生导师,主要研究方向:制造业信息化、人工智能、模式识别;　曹存根(1964 -),男,江苏镇江人,研究员,博士生导师,主要研究方向:知识获取与共享、文本挖掘;徐慧英(1977 -),女,浙江金华人,副教授,主要研究方向:智慧物流、物联网通信、模式识别与图像处理、智能信息处理;　吴德(1981 -),男,浙江金华人,硕士,主要研究方向:电子信息与通信工程、信息服务业与电子商务及信息管理。

《计算机应用》Journal of Computer Applications

音乐和图片等信息媒介相比,囊括的信息更丰富,它是一门容纳文学、摄影、绘画、音乐和科技等多种门类,同时又具有独特性的综合艺术。因此,电影对象的划分会根据分类标准的不同而不同。电影作品是电影领域最基本的对象,其他电影对象都围绕电影作品进行刻画和描述。

综合各种电影对象的分类方法,电影对象可分为人、地域、时间、影视公司、民族、电台频道、语言、视频格式、画面色彩、画面质量、语言风格、主题、事件、制作技术、场景、电影奖项、电影节、类型、题材、角色、电影分类方法共 21 个方面。为方便用户理解和检索,电影本体中增加 rdfs:label 和 rdfs:comment 对知识进行形式化表示。电影本体包含与电影对象相对应的 21 个子本体。这 21 个子本体均自成体系,同时又存在关联性。这些子本体有机地组合在一起,形成完整的电影本体。

1.2　电影本体的组成

电影本体的知识结构包括类(Class)、属性(Property)、实例(Individual)、公理(Axioms)和推理规则(Rules)5 部分。该结构通过"类-属性-实例"的三元组将电影和它的外延知识进行关联,利用公理和推理规则对知识进行约束,完成电影知识网络的架构。

1)类是具有相似性质的所有个体的抽象。本体类的缺省父类为 owl:Thing,电影本体中设置电影以及与电影相关联的同级类共 21 个,每个类包含相应的子类。特别地,本文将"电影分类方法"作为电影本体的一个类。这样的设置不仅更好地体现本体的继承关系(Is-a)和整体-部分(Is-Part-Of)关系,而且对电影的分类体系结构起到优化作用。

2)属性是对类概念的刻画和描述,包括对象属性(Object Property)和数据属性(Datatype Property)两种。对象属性约束两个类的实例之间的关系,定义域为类,值域为某个类的实例。数据类型属性约束类的实例与 RDF 文字或 XML Schema 数据类型间的关系,定义域为某个类的实例,值域为 any、Boolean、string、float 等。

3)实例即个体,是本体类所包含的基本元素,也是本体中最基本的对象。构建本体时需要根据本体的描述粒度来决定哪些对象可以作为类的实例。

4)本体的公理表示一些永真式,是 OWL 类公理的核心部分,用来描述类概念之间存在的各种约束条件。

5)推理规则是通过形式化地描述领域专家知识而形成的系统规则。这些规则表示该领域的一些问题以及与这些问题相应的答案,可以利用它们来模仿专家在求解中的关联推理能力。

1.3　电影本体的构建原则

本体的构建原则对本体建模具有指导意义。本文根据电影本体的特点和作用,制定出电影本体的构建原则,包括以下几方面:

1)尽最大可能使用标准术语。

2)定义清晰。对电影本体中标准术语的含义给出准确的定义和详细的说明。

3)保持一致。本体的前后保持一致,即支持与本体定义相一致的推理。

4)可扩展性。电影本体提供一个可共享的词汇表,为预期的任务提供概念基础。

5)编码偏好程度最小。电影本体位于知识的表达层次,与特定的符号级编码无关。

6)同层次概念保持最小的语义距离。

7)本体约定最小。只要能够满足当前特定的知识共享需求即可。

1.4　电影本体的构建方法

根据电影本体的组织结构和构建原则,本文提出一种电影本体构建方法。该方法包括 3 个阶段:电影知识收集与分析、电影本体的模型构建和电影本体的一致性分析。与文献[6]、TOVE[7] 和七步法[8] 等本体构建方法相比,该方法具有两点优势:

1)增加知识的需求分析。参考《中国国家图书馆分类法》(第 4 版)[9] 中关于电影艺术的概念描述、《辞海》[10]、百度百科、维基百科、《中国大百科全书:电影》(第 2 版)[11]、电影类书籍、豆瓣电影[12] 和 IMDb[13] 等主流电影网站等权威信息媒介,保证电影本体中知识的标准化和精确性;

2)模型构建阶段明确给出电影本体的词汇表,合理复用已有的本体,最终形成电影本体的文档结构,保证电影本体的可扩展性和易维护性。

本文以 OWL DL 作为本体描述语言,Protege 4.1[14] 作为本体编辑工具[15]。电影本体构建方法 3 个阶段的具体步骤如下。

第 1 阶段　电影对象的收集与分析。该阶段包括电影知识的获取、分析和文档化,对电影对象进行初始化的管理和保存。

第 2 阶段　电影本体的模型构建。该阶段主要包括以下几步:

1)电影本体内容和范围的合理界定。电影与音乐和文学等学科之间存在密切联系,很难确定所表示的知识是否被完全覆盖。电影本体完全涵盖电影作品的基本内容,并增加电影节、电影改编、场景等电影的外延知识,形成一个电影作品库。

2)刻画电影本体的领域词汇表。领域词汇表标识并收集所有有用和潜在有用的领域概念及语义、属性和实例等。构建本体之前,本体建立者应当写入词汇表的各个项:词条名称、类型、语义描述、词汇所属类别(类/属性/实例)等。电影本体的词汇表包含类词汇表、属性词汇表等。

3)确定电影本体的类和实例。一要确定电影的对象哪些可以作为类;二要根据本体的描述粒度决定对象作为类还是实例,同时确定本体类的实例集和每个实例所属的类。

4)划分电影本体的类间结构和关系。电影对象的分类结构对本体建立起奠基作用。具体的表示方法主要有自顶向下、自底向上和综合法 3 种。电影本体的构建采用综合法,形成以"电影"为父类的多级树状结构。电影本体的分类结构如图 1 所示。

5)考察和复用已有本体。本体复用是本体知识共享、与相关领域知识增加应用可行性的重要步骤。本体复用需要结合系统自身的需求,适当地借鉴和使用。电影本体中复用"人、地域、时间、民族、语言"等公共信息本体。

6)描述电影本体类的属性。根据最实用、最符合观众心理欣赏特点的原则,电影本体设置 25 个对象属性和 9 个数据

Journal of Computer Applications 计算机应用学报

属性,从不同的侧面对电影本体的类进行刻画。

7)描述电影本体的约束公理和推理规则。深入分析电影本体的对象,明确对象之间的关联约束,利用 OWL 描述公理和规则,为电影本体的一致性分析作铺垫。

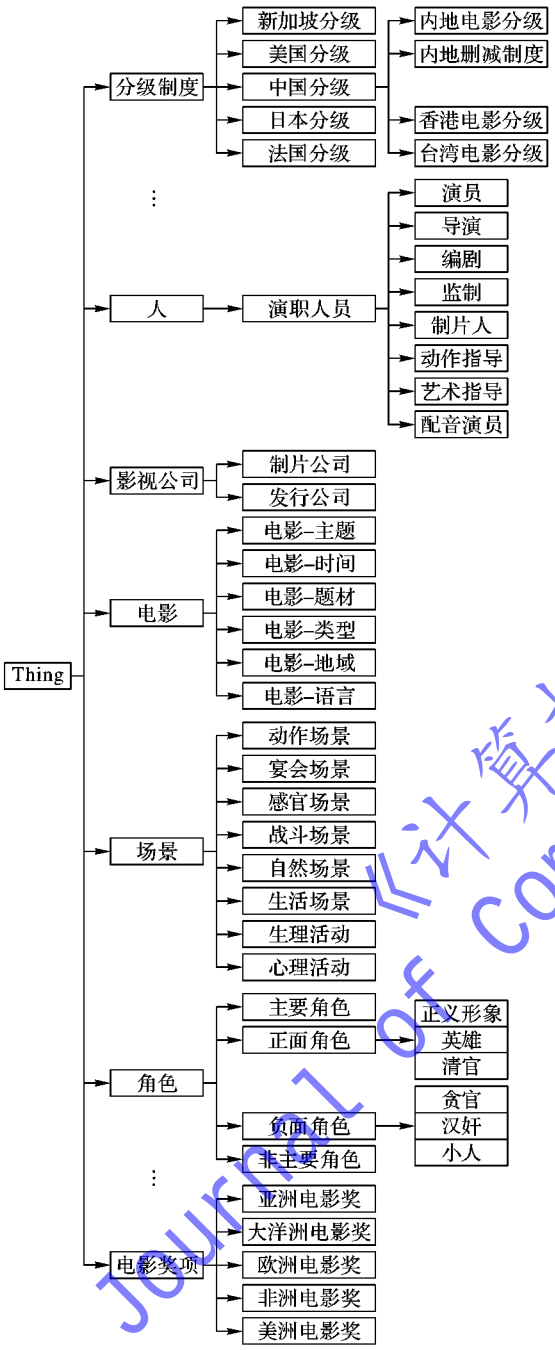

图 1 电影本体的分类结构

8)添加标签和注释。为本体的概念增加 rdfs:label 和 rdfs:comment,便于计算机处理,同时方便用户理解概念和使用本体,进一步完善电影本体。

9)形成电影本体文档。领域本体的建立是不断修改和扩展的过程,文档化有利于本体的编辑和维护,通过不断的改进使本体逐渐接近人类对客观世界的认知。

第 3 阶段 电影本体的一致性分析。主要包括类间关系的一致性分析和基于公理的一致性分析,具体参见本文第 3 章。

2 电影本体的详细描述

2.1 类

本体的类有两方面作用:1)描述类本身的含义和类所包含的知识;2)对类的子类和实例作出限定。电影本体的类词汇包括:电影、类型、题材、主题、时间、语言、地域、人、角色、场景、事件、情节等。其中"人"划分为"职员""演员","职员"包括"导演""编剧"等,"演员"包括"电影演员""配音演员"等。电影本体的类词汇表如表 1 所示。

表 1 电影本体的类词汇(部分)

词条名称	类型	语义描述	所属类别
电影	类	所有电影的父类	Thing
人	类	所有参与电影的人员的父类	Thing
导演	类	电影的导演,人的子类	人
演员	类	电影中参与演出的人,人的子类	人
配音演员	类	为电影配上对白的人	演员
时间	类	与电影相关的时间的父类	Thing
时间点	类	电影中具体时间的父类,时间的子类	时间
场景	类	电影中所有场景的父类	Thing
动作场景	类	场景的子类	场景
类型	类	电影所属类型的父类	Thing
语言	类	电影中对白、字幕语言的父类	Thing

2.2 属性

属性描述本体中类或实例的一个侧面。子类和父类的属性之间存在继承和发展关系。电影本体的属性包括上映时间、主题、事件、场景、拍摄地区等 25 个对象属性,剧情介绍、电影票房、电影经费、电影评分等 9 个数据类型属性。电影本体的属性词汇表标明各个属性的类型、定义域、值域和语义描述,具体描述如表 2 所示。

表 2 电影本体的属性词汇(部分)

词条名称	类型	语义描述	所属类别	定义域	值域
上映时间	属性	电影在某地区的首映的时间	对象属性	电影	时间
主题	属性	电影有主题	对象属性	电影	主题
分类方法	属性	电影按照某种方式属于哪一类	对象属性	电影	电影分类方法
拍摄地区	属性	电影在哪里拍摄	对象属性	电影	地理区域
角色	属性	电影中包含的人物形象	对象属性	电影	角色
电影票房	属性	电影的商业销售情况	数据属性	电影	String
电影经费	属性	制作电影所付出的费用	数据属性	电影	String
电影评分	属性	用户对电影的评价	数据属性	电影	Float
剧情介绍	属性	对电影内容的简单介绍	数据属性	电影	String
点击率	属性	统计电影被点击的比率	数据属性	电影	Float

电影本体中关于电影的主题、角色和剧情介绍等信息源依据豆瓣电影,电影评分的数值来源依据 IMDb 中文网。豆瓣电影提供每一部电影的 IMDb 链接,二者之间的关联性保证电影本体中描述的每一部电影信息都具有一致性和准确性。

本体中另一个描述概念之间联系的重要方面是关系。它通常表现为动词,表示一个命题或断言,描述类、实例之间的各种联系。关系将一个概念与其他概念或实例联系起来,起到知识连通的作用。电影本体的类关系如表 3 所示。

Journal of Computer Applications 《计算机应用》

表 3　电影本体中的关系(部分)

关系	注释
属整体	某一地区属于某地域
有部分	某地域包含某地区
改编自	表示电影的取材来源
改编为	文学或电影资源通过艺术加工改编成电影
属于电影	表示事件、场景等在某电影中出现
限制	表示电影在某地区属于某种分级限制
早于	演员或角色出生时间早于电影的拍摄时间
晚于	“早于”的反义词

2.3　实例描述

为了直观描述电影本体的类和实例,表 4 给出电影作品《白鹿原》的实例。

表 4　电影本体中的实例(部分)

实例名称	所属类别	说明
白鹿原	电影	一部电影
王全安	导演/编剧	中国导演,导演电影作品《白鹿原》等
俄语	语言	目前世界语系中流行的一种语言
亲情	主题	电影反映的主题思想是关于家庭骨肉亲情的
大众电影百花奖	亚洲电影奖	中国内地的一类电影奖项

电影《白鹿原》的 OWL DL 表示如下:

```
<!--http://www.ict.ac.cn/Ontology/电影#白鹿原-->
<owl:NamedIndividual rdf:about="http://www.ict.ac.cn/Ontology/电影#白鹿原">
  <rdf:type rdf:resource="http://www.ict.ac.cn/Ontology/电影#电影"/>
  <电影票房 rdf:datatype="&xsd;string">1.3 亿</电影票房>
  <播放时长 rdf:datatype="&xsd;string">156 min</播放时长>
  <电影经费 rdf:datatype="&xsd;string">1 亿</电影经费>
  <电影评分 rdf:datatype="&xsd;float">7.1</电影评分>
  <剧情介绍 rdf:datatype="&xsd;string">影片以黑娃、小娥的爱情故事为引子讲述白鹿原上的故事,从小人物的卑微爱情,到轰轰烈烈的家族情仇,再到翻天覆地的历史剧变。</剧情介绍>
  <导演 rdf:resource="&人;王全安"/>
  <编剧 rdf:resource="&人;陈忠实"/>
  <电影发行公司 rdf:resource="&公司;华夏电影发行有限责任公司"/>
  <拍摄地区 rdf:resource="&地域;中国"/>
  <上映时间 rdf:resource="&时间;2012 年 9 月 15 日"/>
  <电影节 rdf:resource="&电影节;柏林国际电影节"/>
  <类型 rdf:resource="&类型;历史"/>
  <角色 rdf:resource="&角色;田小娥"/>
  <改编自 rdf:resource="&电影改编;白鹿原"/>
</owl:NamedIndividual>
```

2.4　公理和推理规则

领域本体的公理集包括领域知识的一些永真断言,例如本体概念之间的不相交关系、等价关系和继承关系等。领域本体的公理集如表 5 所示。为了便于描述,这里设定 $C_i(i=1,2,\cdots,n)$ 表示领域本体中的一个概念。

根据电影本体的属性和关系,列出电影本体中的公理,如表 6 所示。公理中使用的符号定义如下:小写字母 a,b,c 分别表示类名或实例名,p 表示它们之间的属性,mo 表示 MovieOntology。@prefix mo:http://www.ict.ac.cn/Ontology/电影.owl#表示以 mo 为前缀的词条所代表的含义。

表 5　领域本体的公理集(部分)

公理	描述
公理 1	$\mathrm{SubClassOf}(C_1,C_2)\cap\mathrm{SubClassOf}(C_2,C_3)\Rightarrow\mathrm{SubClassOf}(C_1,C_3)$
公理 2	$\mathrm{SubClassOf}(C_1,C_2)\cap\cdots\cap\mathrm{SubClassOf}(C_i,C_{i+1})\cdots\cap\mathrm{SubClassOf}(C_n,C_1)=\mathrm{False}$
公理 3	$\mathrm{UnionOf}(C_1,C_2)\cap\mathrm{UnionOf}(C_2,C_3)\Rightarrow\mathrm{UnionOf}(C_1,C_3)$
公理 4	$\mathrm{UnionOf}(C_1,C_2)\cap\cdots\cap\mathrm{UnionOf}(C_i,C_{i+1})\cdots\cap\mathrm{UnionOf}(C_n,C_1)=\mathrm{False}$
公理 5	$\mathrm{EquivalentWith}(C_1,C_2)\cap\mathrm{EquivalentWith}(C_2,C_3)\Rightarrow\mathrm{EquivalentWith}(C_1,C_3)$
公理 6	$\mathrm{SubClassOf}(C_1,C_2)\cap\mathrm{EquivalentWith}(C_2,C_3)\Rightarrow\mathrm{SubClassOf}(C_1,C_3)$
公理 7	$\mathrm{SubClassOf}(C_1,C_3)\cap\mathrm{EquivalentWith}(C_1,C_2)\Rightarrow\mathrm{SubClassOf}(C_2,C_3)$
公理 8	$\mathrm{DisjointWith}(C_1,C_2)\cap(\mathrm{SubClassOf}(C_3,C_1)\cup\mathrm{EquivalentWith}(C_3,C_1))\cap(\mathrm{SubClassOf}(C_4,C_2)\cup\mathrm{EquivalentWith}(C_4,C_2))\Rightarrow\mathrm{DisjointWith}(C_3,C_4)$

表 6　电影本体的公理集

公理	描述
函数属性	(?p rdf:type ?FunctionalProperty),(?a mo:p ?b),(?a mo:p ?c)⇒(?b sameAs ?c)
反函数属性	(?p rdf:type ?inverseFunctionalProperty),(?a mo:p ?c),(?b mo:p ?c)⇒(?a sameAs ?b)
类间传递	(?a rdf:subClassOf ?b),(?b rdf:subClassOf ?c)⇒(?a rdf:subClassOf ?c)
属性传递 1	(?a rdf:subPropertyOf ?b),(?b rdf:subPropertyOf ?c)⇒(?a rdf:subPropertyOf ?c)
属性传递 2	(?p rdf:type ?TransitiveProperty),(?a mo:p ?b),(?b mo:p ?c)⇒(?a mo:p ?c)
对称属性	(?p rdf:type ?SymmetricProperty),(?a mo:p ?b)⇒(?b mo:p ?a)
自反属性	(?p rdf:inversrOf ?q),(?a mo:p ?b)⇒(?b mo:p ?a)

根据知识和公理得出的推理规则是扩展和完善电影本体的重要依据,主要有两方面的作用:一是完善电影知识的外延;二是检测和维护电影知识,保证本体的一致性和完整性。推理规则会通过知识的更新,不断得到补充和完善。电影本体中的推理规则如表 7 所示。

3　电影本体的一致性分析

本体的一致性是指本体包含的概念知识不存在矛盾,各个部分构成一个合理的整体。电影本体的一致性分析包括类间关系的一致性和基于公理的一致性[16]两方面。首先给出电影本体的类间关系定义。

定义 1　对于本体 O 中的任意类 C_1、C_2 和任意个体 i,如果存在 i 是类 C_1 的实例,同时也是类 C_2 的实例,则称作类 C_1 和 C_2 之间存在继承关系。记作

$$\left.\begin{array}{l}\forall C_1\in O,C_2\in O,i\in O\\ \exists[i\in\mathrm{instanceOf}(C_1)]\cap[i\in\mathrm{instanceOf}(C_2)]\end{array}\right\}\Rightarrow \mathrm{is\text{-}a}(C_1,C_2)$$

定义 2　对于本体 O 中的任意类 C_1、C_2 和任意个体 i,如

《计算机应用与软件 Journal of Computer Applications

果存在 i 既不是类 C_1 的实例,同时也不是类 C_2 的实例,则称作类 C_1 和类 C_2 是不相交的。记作

$$\left.\begin{array}{l}\forall C_1 \in O, C_2 \in O, i \in O \\ \exists i \notin [\text{instanceOf}(C_1) \cap \text{instanceOf}(C_2)]\end{array}\right\} \Rightarrow \text{DisjointWith}(C_1, C_2)$$

表 7 电影本体的推理规则(部分)

推理规则	描述
规则 1	(?电影 mo:题材 ?红色经典)⇒ (?电影 mo:电影反映时间 ?抗日战争时期)
规则 2	(?电影 mo:事件 ?事件),(?电影 mo:反映地区 ?地理区域),(?事件 mo:发生地点 ?地理区域)⇒ (?事件 mo:发生地点)包含于(?电影 mo:反映地区)
规则 3	(?电影 mo:上映时间 ?时间)⇒ (?演职人员 mo:出生日期 ?早于"上映时间")
规则 4	(?电影 mo:拍摄时间 ?时间),(?电影 mo:上映时间 ?时间)⇒(电影.拍摄时间 早于 电影.上映时间)
规则 5.1	(?电影 mo:人 ?职员)⇒(?职员 mo:出生日期 xsd:Datetime)≤(?电影 mo:拍摄时间 ?时间)
规则 5.2	(?电影 mo:类型 ?纪录片),(?电影 mo:人 ?演员)⇒ (?演员 mo:出生日期 xsd:Datetime) 早于 (?电影 mo:拍摄时间 ?时间)
规则 5.3	(?电影 mo:类型 ?非"纪录片"), (?电影 mo:人 ?演员)⇒(?演员 mo:出生日期 xsd:Datetime) 晚于 (?电影 mo:拍摄时间 ?时间)
规则 6	(?电影 mo:演员 ?演员),(?演员 mo:饰演 ?角色)⇒ (?角色 mo:属于电影 ?电影)
规则 7	(?电影 mo:角色 ?角色),(?角色 mo:饰演者 ?演员) ⇒(?演员 mo:主演 ?电影)
规则 8	(?电影 mo:人 ?人),(?人 mo:国籍 ?国家)⇒ (?人 mo:母语 ?语言)
规则 9	(?电影 mo:分级制度 ?分级制度), (?电影 mo:拍摄地区 ?地理区域)⇒(?分级制度 mo:限制 ?(?电影 mo:拍摄地区 ?地理区域))
规则 10	(?影视公司 mo:主要成就 ?电影)⇒ (?电影 ?制片公司 or 出品公司 ?影视公司)

3.1 类间关系的一致性分析

电影本体的类间关系不一致主要由不相交关系 DisjointWith 引起。电影本体只定义不相交关系,因此本节只讨论由不相交关系引起的不一致性[17]。

假设在本体 O 中,类 C 的所有父类和子类的集合为 S,所有与类 C 不相交的类的集合为 S',通过比较 S 和 S' 是否有交集来判定类是否一致。同理类的属性也可以利用相同的方法进行判定。判定算法具体描述如下。

算法 1 领域本体不一致性判定算法。

输入:领域本体知识。

输出:领域本体不一致性判断结果。

步骤如下:

1) 假设 $C_i(i = 1,2,\cdots,n,n \geqslant 1)$ 表示本体中所有不相同的类;

2) $i = 1$;

3) $S \neq \varnothing, S' \neq \varnothing$;

4) 对于本体 O 中所有的类 C',如果存在 is-a(C',C) 或者 is-a(C,C'),则 $S = S \cup \{C'\}$;如果存在 DisjointWith(C,C'),则 $S' = S' \cup \{C'\}$;

5) 如果 $S \cap S' \neq \varnothing$,算法终止。输出:本体不一致;

6) $i = i + 1$;

7) 如果 $i \leqslant n$,转到步骤 3);

8) 算法结束。输出本体一致。

下面分析算法的可靠性和完备性。

如果本体中的类存在不相交关系或者出现关系冲突,那么由算法 1 一定能够得到类间关系是不相交的,这是算法 1 的完备性;如果算法 1 得到类间关系是不相交的,那么本体的类之间一定存在不相交关系或者出现关系冲突,这是算法 1 的可靠性。下面给出相关的定理。

定理 1 算法 1 一定会终止。

定理 2 算法 1 是完备的。

定理 3 算法 1 是可靠的。

3.2 基于公理的一致性分析

基于电影公理的一致性分析是指电影知识和电影公理之间不存在矛盾。由于电影知识来源的多样性和人工整理知识的疏漏,本体工程师所获取的知识和概念不一定完全正确,这些原因都可能造成本体的不一致。基于表 6 和表 7 列出的电影公理和规则,给出以下分析。

例 1:

规则 4 [(?电影 mo:拍摄时间 ?时间),(?电影 mo:上映时间 ?时间)⇒(电影.拍摄时间 早于 电影.上映时间)]。

该规则说明对于任意一部电影,它的拍摄时间一定早于上映时间。如果电影本体中出现拍摄时间大于上映时间,则报告本体出现不一致。依据此类规则可以推断电影本体知识的一致性与否。

需要说明的是,本体的一致性分析涉及本体的类、属性和实例等各个方面,是一项非常复杂的工作。本体工程师需要随着知识的更新,逐步找出有效的分析方法,循序渐进地完善本体。本文主要针对电影领域本体的初始工作,在完成电影本体局部分析的基础上,进一步分析整个电影本体的一致性。

4 结语

本文结合当前电影领域的形势,对电影领域知识本体建立的必要性和可行性进行说明,提出一种"三阶段"电影领域本体构建方法,并阐述电影本体构建的步骤。该方法涵盖本体构建和维护的整个生命周期,相比于其他本体构建方法具有明显优势。本文所建立的电影本体模型,在国内具有超前性和实用性,对于下一代互联网业务的扩展具有重要意义。随着知识的发展和演变,本体模型的建设和完善工作将继续进行。

参考文献:

[1] LIU Z. Ontology and semantic Web [J]. Chongqing Library and Information Science Research, 2006, 7(3): 1 - 4. (刘植惠. 本体(Ontology)与语义网(Semantic Web) [J]. 重庆图情研究, 2006, 7(3): 1 - 4.)

[2] LU R. Knowledge engineering and knowledge science at the turn of the century [M]. Beijing: Tsinghua University Press, 2001. (陆汝钤. 世纪之交的知识工程与知识科学[M]. 北京: 清华大学出版社, 2001.)

[3] WANG L, CAO C. Acquiring and analyzing knowledge of nationalities: an ontology-based approach [J]. Computer Science, 2003, 30(15): 47 - 54. (王丽丽, 曹存根. 基于本体论的民族知识获取与分析[J]. 计算机科学, 2003, 30(15): 47 - 54.)

(下转第 2201 页)

Journal of Computer Applications 计算机应用

准确率不高,本文将传统的统计模型结合统计规则对兼类词进行词性标注,并介绍了规则的形式化描述及基于规则的兼类词识别算法,实验结果表明了改进的互信息算法在规则的获取上的有效性。下一步的工作是针对更多的兼类词尝试用规则的方法进行识别,完善规则库,并且尝试用聚类的方法对兼类词的识别进行研究。

参考文献:

[1] BRILL E. A corpus-based approach to language learning [D]. Philadelphia: University of Pennsylvania, 1993.

[2] HAMMERTON J, OSBORNE M, ARMSTRONG S, *et al*. Introduction to special issue on machine learning approaches to shallow parsing [J]. Journal of Machine Learning Research, 2002, 13(2): 551 – 558.

[3] BRILL E. Unsupervised learning of disambiguation rules for part-of-speech [C]// Proceedings of the Third Workshop on Very Large Corpora. Piscataway: IEEE Press, 1995: 1 – 13.

[4] SCHMID H. Probabilistic part-of-speech using decision tree [C]// Proceedings of the 1994 International Conference on New Methods in Language Processing. Piscataway: IEEE Press, 1994: 44 – 49.

[5] YUAN C. Improved hidden Markov model for speech recognition and POS tagging [J]. Journal of Central South University, 2012, 19(2): 511 – 516.

[6] RATNAPARKHI A. A maximum entropy model for part-of-speech tagging [C]// Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1996: 132 – 142.

[7] LIU T, LEI L, CHEN L. A parallel training research of Chinese part-of-speech tagging CRF model based on MapReduce [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 147 – 152. (刘滔, 雷霖, 陈荦, 等. 基于 MapReduce 的中文词性标注 CRF 模型并行化训练研究[J]. 北京大学学报: 自然科学版, 2013, 49(1): 147 – 152.)

[8] EMBAL A, SAHA S. Simulated annealing based classifier ensemble techniques: application to part of speech tagging [J]. Information Fusion, 2013, 14(3): 288 – 300.

[9] ZHAO Y, WANG X, LIU B, *et al*. Fusion of clustering trigger-pair features for POS tagging based on maximum entropy model [J]. Journal of Computer Research and Development, 2006, 43(2): 268 – 274. (赵岩, 王晓龙, 刘秉权, 等. 融合聚类触发对特征的最大熵词性标注模型[J]. 计算机研究与发展, 2006, 43(2): 268 – 274).

[10] BRANTS T. TnT: a statistical part-of-speech tagger [C]// Proceedings of the 6th Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2000: 224 – 231.

[11] COHEN W, CARVALHO V. Stacked sequential learning [C]// IJCAI'05: Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2005: 671 – 676.

[12] ZHAN Y, CLARK S. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model [C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2010: 843 – 852.

[13] AULI M, LOPEZ A. Training a log-linear parser with loss functions via softmax-margin [C]// EMNLP'11: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 333 – 343.

(上接第 2196 页)

[4] W3CSchool. OWL introduction [EB/OL]. [2013-09-18]. http://www.w3school.com.cn/rdf/rdf_owl.asp. (W3CSchool. OWL 简介[EB/OL]. [2013-09-18]. http://www.w3school.com.cn/rdf/rdf_owl.asp.)

[5] Movieontology.org. MO — the movie ontology [EB/OL]. [2013-09-18]. http://www.movieontology.org/.

[6] USCHOLD M, KING M, MORALEE S, *et al*. The enterprise ontology [J]. The Knowledge Engineering Review, 1998, 13(1): 31 – 89.

[7] University of Toronto. Faculty of Applied Science and Engineering. TOVE ontologies [EB/OL]. [2013-09-18]. http://www.ie.utoronto.ca/EIL/tove/toveont.html.

[8] NOY N F, McGUINNESS D L. Ontology development 101: a guide to creating your first ontology: knowledge systems laboratory, SMI-2001-0880 [R]. Stanford: Stanford University, 2001.

[9] Editorial Committee of China Library Classification. China library classification [M]. 4th ed. Beijing: Beijing Library Press, 1999: 197 – 199. (中国图书馆分类法编辑委员会. 中国图书馆分类法[M]. 4 版. 北京: 北京图书馆出版社, 1999: 197 – 199.)

[10] XIA Z, CHEN Z. Unabridged dictionary: art [M]. 6th ed. Shanghai: Shanghai Lexicographical Publishing House, 2010. (夏征农, 陈至立. 辞海: 艺术分册[M]. 6 版. 上海: 上海辞书出版社, 2010.)

[11] Editorial Department in Encyclopedia of China Publishing House. Encyclopedia of China: movie [M]. 2nd ed. Beijing: Encyclopedia of China Publishing House, 2004. (中国大百科全书出版社编辑部. 中国大百科全书: 电影[M]. 2 版. 北京: 中国大百科全书出版社, 2004.)

[12] Douban. Douban movie [EB/OL]. [2014-04-09]. http://movie.douban.com/. (豆瓣. 豆瓣电影[EB/OL]. [2014-04-09]. http://movie.douban.com/.)

[13] IMDb Chinese Website. Movie database [EB/OL]. [2014-04-09]. http://www.imdb.cn/. (IMDb 中文网. 电影资料库[EB/OL]. [2014-04-09]. http://www.imdb.cn/.)

[14] Stanford Center for Biomedical Informatics Research. Protégé [EB/OL]. [2014-04-09]. http://protege.stanford.edu/.

[15] USCHOLD M, GRUNINGER M. Ontologies: principles, methods and applications [J]. The Knowledge Engineering Review, 1996, 11(2): 93 – 136.

[16] GAO Y, CAO C, SUI Y. Musical domain-specific ontology building and analysis [J]. Computer Science, 2004, 31(1): 103 – 107. (高颖, 曹存根, 眭跃飞. 音乐领域本体的建立和分析[J]. 计算机科学, 2004, 31(1): 103 – 107.)

[17] XIE N, WANG W. Ontology and acquiring of agriculture knowledge [J]. Agriculture Network Information, 2007(8): 12 – 16. (谢能付, 王文生. 农业知识本体构建方法[J]. 农业网络信息, 2007(8): 12 – 16.)

《计算机应用》 Journal of Computer Applications