

文章编号:1001-9081(2008)05-1170-03

基于贝叶斯网络的本体不确定性推理

杨喜权, 曹雪亚, 国頤娜, 周建园

(东北师范大学 计算机学院, 长春 130117)

(caoxy290@nenu.edu.cn)

摘要:运用 OWL 语言扩展了本体对领域知识的不确定性表示,并基于贝叶斯网络实现了本体领域知识的不确定性推理。实验表明将贝叶斯网络与本体结合起来,能够充分发挥本体在知识描述方面的优势和贝叶斯网络的推理能力,实现依据部分信息的概率描述获取知识,指导实践。

关键词:本体; 贝叶斯网络; 不确定性

中图分类号: TP311 **文献标志码:**A

Research of ontology uncertainty reasoning based on Bayesian network

YANG Xi-quan, CAO Xue-ya, GUO Di-na, ZHOU Jian-yuan

(School of Computer Science, Northeast Normal University, Changchun Jilin 130117, China)

Abstract: This paper adopted OWL to extend the uncertainty representation in domain ontology and support uncertainty reasoning by Bayesian network. The experiment shows that Bayesian network together with ontology can preserve the advantages of both, for ontology has power of knowledge presenting and Bayesian network provides ability of reasoning. It can obtain knowledge from partial probabilistic-described information and instruct practice.

Key words: ontology; Bayesian network; uncertainty

随着计算机科学和信息技术的飞速发展,人类所面临的知识和信息成倍增长,在丰富的信息中形成知识日益重要。不同领域的人们都在期待能够从这些堆积如山的信息中找到自己想要的知识。但是,由于教育背景和研究侧重点的差异,同一领域的研究人员对同一个研究对象也可能有不同的理解和表述,导致难于知识的共享和重用,使得基于知识的推理更加困难。知识是对有用信息按其内在联系进行的组织与分类,要使得知识能够在一定范围内共享、使用,就需要使用一种概括性强又能较为具体表示出知识之间关系的表示模型。知识模型系统的研究一直是知识工程领域的一个研究重点。

本体(Ontology)作为一种能在语义和知识层次上描述知识模型的建模工具,提供了概念的规范化描述,为知识的共享奠定了基础^[1]。但本体不能表示概念之间的重叠或相交程度,也不能支持只知道概念或个体的部分信息的推理。Bayesian 决策理论为处理不确定的事件或推理提供了理论基础。在不确定知识的表示和推理方面,贝叶斯网络被证明是获得非确定性知识的置信度的最有效方法之一^[2,3]。本文结合贝叶斯网络(Bayesian Network)作为不确定性推理依据,对 Web 本体语言(Web Ontology Language,OWL)进行概率扩展,使它能支持不确定知识和不完备或不精确信息的推理。

本文利用斯坦福大学的本体开发工具 Protégé3.3^[4]构建了一个简单的鱼本体,并将表示不确定信息的概率信息同时附加在它上面,把概率概念转化成贝叶斯网络中对应的节点,通过贝叶斯网络进行推理。

1 基于本体的知识模型

本体最早是一个哲学的范畴,后来随着人工智能的发展,被人工智能界赋予了新的定义。在人工智能界,本体的目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定

该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。

1.1 本体的形式化定义及建模原语

本体是共享概念模型的明确的形式化规范说明。文献[5]用分类法组织了 Ontology, 归纳出五个基本的建模原语或者五个基本元素。

1)类(Class)或概念(Concept):指任何事物的抽象,如工作描述、功能、行为、策略和推理过程等。从语义上讲,它表示的是对象的集合。

2)关系(relations):在领域中概念之间的交互作用。关系对应于对象元组的集合,形式上定义为 n 维笛卡尔积的子集。 $R: C_1 \times C_2 \times \dots \times C_n$, 如子类关系(subClassOf)。

3)函数(functions):一类特殊的关系。该关系的前 $n-1$ 个元素可以唯一决定第 n 个元素。形式化的定义为映射 $F: C_1 \times C_2 \times \dots \times C_{n-1} \rightarrow C_n$ 。如:函数 $father-of(x,y)$ 表示 x 是 y 的父节点。

4)公理(axioms):代表永真断言,如概念乙属于概念甲的范围。

5)实例(instances):代表元素,从语义上讲实例表示的就是对象。

1.2 鱼本体的知识模型系统

在本文中我们依据鱼的分类、产地、捕鱼的季节等信息,以及它们对应的先验概率和条件概率,构建鱼分类的本体。在鱼本体的开发过程中,采用斯坦福大学提出的 7 步法^[6],抽取出实际鱼分类领域中的概念、关系、实例等,统称为术语。具体步骤如下:

1)创建类。与表示概率相关的类有 ProbObj、States 和 Variable,其中 ProbObj 类下面的子类 CondProb 和 PriorProb 分别表示节点的条件概率和先验概率。把与鱼本身相关的属性

收稿日期:2007-11-13;修回日期:2008-01-04。 基金项目:国家自然科学基金资助项目(60473042)。

作者简介:杨喜权(1963-),男,吉林四平人,副教授,主要研究方向:Web 信息处理、信息安全; 曹雪亚(1981-),女,浙江嘉兴人,硕士研究生,主要研究方向:语义网、信息挖掘; 国頤娜(1982-),女,吉林长春人,硕士研究生,主要研究方向:数据挖掘; 周建园(1981-),男,吉林松原人,硕士研究生,主要研究方向:信息挖掘、遗传算法。

也建成相应的类,分别是:Season类,表示捕鱼的季节;Location类,表示捕鱼的地点;ShiningDegree类,表示鱼的光泽度;WidthDegree类,表示鱼身的宽窄。

2)构建关系。从语义上讲,基本的关系共有4种:part-of, kind-of, instance-of, attribute-of。在本文中较多地用到了attribute-of关系,如hasClass, hasCondition, hasProbValue等,表示某个概念是另一个概念的属性。

3)实例化。选择恰当的类,以OWL/RDF格式表示,为它创建个体并赋值。例如鱼的个体,其属性包括捕鱼的季节、地点、光泽度等,那么可以创建这样一个实例,即给相应的属性赋值,如捕鱼季节:冬季;地点:北大西洋;光泽度:亮。定义好本体并进行实例化,把二者结合起来构成了一个知识库^[7]。将构建的鱼本体作为特定领域内的知识库,充分发挥本体在概念模型上明确的形式化描述特性。

在Protégé3.3下构建鱼本体的分类,类图如图1所示,利用本体可直观表达知识之间的确定关系,但是却不能直接为人类活动提供有价值的信息。例如在北大西洋,什么季节可以捕到鲑鱼等。因此我们引入概率表示,扩展本体对不确定性知识的表述,进而利用贝叶斯网络推理出有价值的知识。

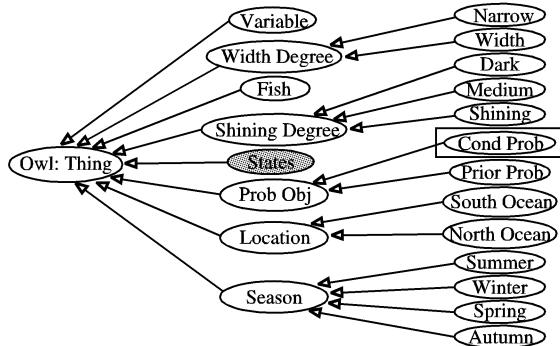


图1 鱼本体的类图

1.3 OWL语言概率描述

在领域本体中引入概率可以较好地表示概念间的不确定信息和潜在的因果关系^[8]。下面给出条件概率 $P(\text{fish_gui} \mid \text{spring}) = 0.3$ 的一般表述:

```
<!--probability for being fish_gui-->
<Variable rdf:ID = "fish_gui">
    <hasClass> Fish </hasClass>
    <hasState> True </hasState>
</Variable>
<Variable rdf:ID = "spring">
    <hasClass> Season </hasClass>
    <hasState> True </hasState>
</Variable>
<CondProb rdf:ID = "P(fish_gui | spring)">
    <hasCondition> spring </hasCondition>
    <hasVariable> fish_gui </hasVariable>
    <hasProbValue> 0.3 </hasProbValue>
</CondProb>
```

这样可以构建基于概率约束的鱼本体,在该本体中,蕴含了两个图形化的模型,OWL图形模型和贝叶斯网络模型。OWL图形模型是基于RDF的有向图,该图形模型能可视化一个特定本体的所有相关信息,例如类层次关系等;而贝叶斯网络模型,通过概率概念来表示,它可以更好地表示概念之间的不确定信息和依赖关系,可以通过Jena^[9]对OWL本体读取的支持,从本体中提取出概率表示,进而实现推理。

2 Bayesian推理

2.1 不确定性推理

在传统的经典逻辑中,OWL语句所表述的事实、领域知

识和推理的结果,只能为真或为假,而不是一个介于两者之间的值。但是,现实世界中却充满了不确定知识或不精确的信息,它只能部分为真。

不确定性推理是建立在非经典逻辑基础上的一种推理,是对不确定性知识的应用和处理。严格地说,不确定性推理就是从不确定性的初始证据出发,通过运用不确定的知识,最终推出具有一定程度的不确定性,但却是合理或者近乎合理的结论的思维过程^[10]。

2.2 贝叶斯网络

贝叶斯网络又称为置信网络。它是表示变量间概率依赖关系的有向无环图(Directed Acyclic Graph, DAG),它提供了一种自然的表示因果信息的方法,用来发现数据间的潜在关系。在这个网络中,节点表示变量,有向边表示变量间的依赖关系,同时每个节点都对应着一个条件概率分布表(Conditional Probability Table, CPT),指明了该变量与父节点之间概率依赖的数量关系^[11]。

在贝叶斯网络中,有两类节点, X 节点之前的节点集合称为父节点, X 节点之后的节点集合称为子节点。节点 X 上的一系列命题 $x = (x_1, x_2, \dots)$ 的“置信度”描述了在给定网络所有其余部分的证据 e 的前提下这些变量之间的相关概率,即 $P(x \mid e)$ 。可以将依赖于父节点的置信度同子节点分离为: $P(x \mid e) \propto P(e^c \mid x)P(x \mid e^p)$ 。其中 e^c 表示子节点上的证据, e^p 表示父节点上的证据。因式 $P(e^c \mid x) = P(e_{c_1}, e_{c_2}, \dots, e_{c_{|C|}} \mid x) = \prod_{j=1}^{|C|} P(e_{c_j} \mid x)$, 其中 c_j 表示第 j 个子节点, e_{c_j} 表示其状态的概率值, $|C|$ 表示集合中的元素个数。因式 $P(x \mid e^p) = \sum_{\text{all } P_{mn}} P(x \mid P_{mn}) \prod_{i=1}^{|P|} P(P_i \mid e_{p_i})$, 表示父节点上的值的先验概率在所有可能的状态组合上的总和,以及给定那些父节点值时的 x 变量的条件概率的总和。最后将公式 $p(x \mid e)$ 的值归一化,即得所需的后验概率值。

对图1的鱼本体应用Bayesian得到的鱼的贝叶斯网络^[12],如图2所示。

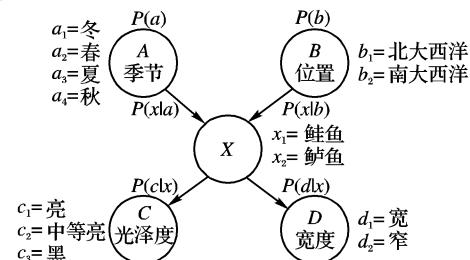


图2 鱼分类贝叶斯网络

表1 鱼的类别与各属性对应的CPTs

类	属性										
	季节			区域		光泽度		宽度			
别	冬	春	夏	秋	北大	南大	亮	中等	暗	宽	窄
鲑鱼	0.9	0.3	0.4	0.8	0.65	0.25	0.33	0.33	0.34	0.40	0.60
鲈鱼	0.1	0.7	0.6	0.2	0.35	0.75	0.80	0.10	0.10	0.95	0.05

本文所用鱼分类例子是一个简单贝叶斯网络,捕鱼季节同捕鱼地区统计独立,但所捕获的鱼类确实依赖于这些因素,并且,鱼的宽度和光泽度又依赖于鱼的类别本身。通过专家经验和相关研究,给出每个节点的CPTs,如表1。本文对于节点CPTs的计算分两种情况,如果它没有父节点,则利用它的先验概率作为它的CPTs,如果它的父节点存在且不为空,则通过计算条件概率 $P(C \mid O_e)$ 作为它的CPTs,其中 O_e 为 C 的不为空的父节点。

3 实验与分析

通过 Norsys 公司开发的贝叶斯网络推理平台 Netica 辅助我们的可视化推理。实验方法如下：

1) 贝叶斯网络的初始化

读取本体中的概率表示,如果将它表示成如图 2 中的节点。对节点有两种情况处理。对于父节点存在,可以将它的先验概率作为它的 CPT 表输入。在本例中,由于不知道先

验分布,因此可以给定一个足够小的非零实数 λ ,赋予概率概念主观概率值 λ 符合贝叶斯理论基本思想,可以简化计算,提高贝叶斯网络的推理效率。在图形化界面中,我们可以看到它们的后验概率是均匀分布的。对于父节点存在,如果只有一个父节点,可以直接将它的条件概率作为对应的 CPTs。如果有多个父节点,则输入相对应的 CPTs,但需要依据统计独立性假设和边缘条件概率,计算它的联合条件概率。经过编译形成贝叶斯网络推理图,如图 3 所示。

表 2 不同变量推理的鱼的分布

分布 情况	鱼类	概率	宽度		光泽度			季节				区域	
			宽	窄	亮	中	暗	冬	春	夏	秋	北大西洋	南大西洋
情况 1	鲈鱼	0.566	0.711	0.289	0.800	0.100	0.100	0.214	0.284	0.273	0.229	0.463	0.537
情况 2	鲑鱼	0.762	0.531	0.469	0.100	0.100	0.800	0.311	0.192	0.211	0.286	0.562	0.438
情况 3	鲈鱼	0.536	0.800	0.200	0.582	0.207	0.211	0.223	0.275	0.268	0.234	0.472	0.528
情况 4	鲑鱼	0.819	0.200	0.800	0.415	0.288	0.297	0.327	0.178	0.200	0.295	0.579	0.421

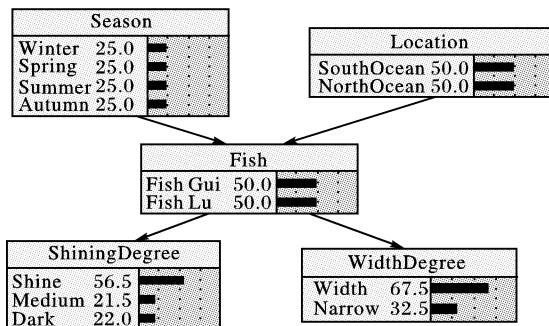


图 3 贝叶斯网络推理图

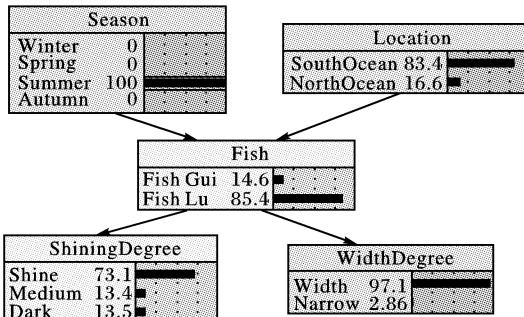


图 4 贝叶斯推理结果

2) 利用贝叶斯网络进行不确定性推理

在本实例中,收集每一节点上的证据,假定只知道鱼相关信息的部分信息,来确定鱼的分布。已知现在是夏季,即 $P(a_i | e_A) = 1$,对于 $i = 1, 2, 4$,则 $P(a_i | e_A) = 0$,假设并不知道渔船来自哪个地区,但是我们发现渔民喜欢在南大西洋捕鱼,那么可设 $P(b_1 | e_B) = 0.8, P(b_2 | e_B) = 0.2$ 。通过对鱼的观察,发现其形状较宽,于是我们手工设置成 $P(e_D | d_1) = 0.8, P(e_D | d_2) = 0.1$ 。假设由于光线不好,而不能测出鱼的光泽度,因而设置 $P(e_C | c_1) = P(e_C | c_2) = P(e_C | c_3)$ 。将上述数据作为证据输入到贝叶斯网络中,可以得到如图 4 的贝叶斯推理结果。由此,可以知道,鲈鱼的后验概率较大,因此这种鱼是鲈鱼的可能性较大。

同样可以依据实际需要给定证据,用来指导渔民的捕鱼。表 2 反映了以变量光泽度或宽度为依据进一步推理鱼的分布情况(表中加下划线的数据为已知的推理变量)。

4 结语

本文构建了一个简单的鱼本体,可以在领域范围内达到知识的共享。同时将贝叶斯网络作为底层的推理机制,在只

知道实例的部分信息的情况下具有较好的推理准确性。同时推理的粒度可以依赖于构建的本体的粒度,实现了在变量级上推理,较好地挖掘出信息中的知识。在研究的过程中我们也发现了一些问题:

1) 本体的可重用性和面向特定领域

在本文中所运用的鱼本体是针对鱼分类的情况开发的,它可以指导渔民捕鱼和鱼类加工厂对鱼进行分类。但是如果针对鱼类研究专家,这样的本体显然在知识的表述和概念的共享上具有局限性。

2) 本体和贝叶斯网络的无缝连接

将本体转化为贝叶斯网络需要借助手工来实现,我们将进一步利用 Jena 对本体读取的支持和 Netica 提供的 Java API 接口,开发直接将本体转化为贝叶斯网络的组件。

参考文献:

- [1] HENDLER J. Ontologies on the Semantic Web[J]. IEEE Intelligent Systems, 2002, 17(2): 73 - 74.
- [2] HECKERMAN D, MEEK C, COOPER G. A Bayesian Approach to Causal Discovery, MSR-TR-97-05[R]. Microsoft Research, 1997.
- [3] HECKERMAN D. Bayesian networks for data mining[J]. Data Mining and Knowledge Discovery, 1997, 1(1): 79 - 119.
- [4] Stanford Center for Biomedical Informatics Research. PROTEGÉ3.3 [EB/OL]. [2007-08-30]. <http://protege.stanford.edu/>.
- [5] PEREZ A G, BENJAMINS V R. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods[C]// Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5). San Francisco, CA, USA: Morgan Kaufmann, 1999: 1 - 15.
- [6] 李景, 孟连生. 构建知识本体方法体系的比较研究[J]. 现代图书情报技术, 2004(7): 17 - 22.
- [7] 陈宏. 基于本体的知识表示研究[D]. 长沙: 长沙理工大学, 2006.
- [8] DING ZHONG-LI, PENG YUN. A probabilistic extension to ontology language OWL[C]// Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04). Washington, DC: IEEE Computer Society, 2004, 4: 111 - 120.
- [9] Jena - A semantic Web framework for java[EB/OL]. [2007-11-10]. <http://jena.sourceforge.net/>.
- [10] 李德毅, 杜鵑. 不确定性人工智能[M]. 北京: 国防工业出版社, 2005.
- [11] 李国仿. 基于贝叶斯网络的本体映射模型的研究[D]. 武汉: 武汉科技大学, 2006.
- [12] (美) DUDA R O, HART P E, STORK D G. 模式分类[M]. 2 版. 李宏东, 姚天翔, 等译. 北京: 机械工业出版社, 2006.