机器学习在本体映射中的应用

李 琼,李宝敏
(西安工业大学 计算机科学与工程学院,陕西 西安 710032)

摘 要:文中针对语义网中同领域内的本体异构现象,以及无法实现领域内本体库共享的问题,提出利用人工智能研究中的机器学习算法来解决。通过概念匹配映射使异构本体的语义更好地得到映射,并在果品领域经过实例的应用,其效果还是客观的。在本体映射匹配研究中,机器学习算法发挥了很大的作用,对语义重叠的概念进行高效率的推理匹配映射,为语义网本体异构的环境下实现信息在语义上的共享互操作提供了一种解决的途径。

关键词:语义 Web;机器学习算法;决策树算法;分类树;概念映射
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2009)12-0081-04

The Application of Machine Learning in Ontology Mapping

LI Qiong, LI Bao-min
(College of Computer Science and Engineering, Xi'an Technology University, Xi'an 710032, China)

Abstract: *In order to solve the problem of ontology heterogeneous in semantic web, it is proposed the algorithm of machine learning of artificial intelligence. Get the better match of semantic meaning through the mapping of concept, and use in the domain of fruit with living example and objective effect. In the research of ontology mapping and match, machine learning plays a very significant role for the efficient matching of reasoning mapping to the semantic overlap concept. Besides, machine learning provides an approach to solve the ontology heterogeneous to share the information in semantic web.*

Key words: semantic web; machine learning algorithm; decision tree algorithm; classification tree; concept mapping

0 引 言

语义网(Semantic Web)是在本体(Ontology)理论基础之上对现有 Web 所进行的扩展,其目标是使 Web 上的信息具有计算机可以理解的语义,在本体的支持下实现信息在语义上的互操作,以及对 Web 资源所进行的智能访问和检索。语义网服务是以语义网和本体为基础的一个重要的基础研究领域,其目标是通过将语义网技术和 Web 服务技术相结合,提供下一代网络的集成平台[1]。

语义网服务中大量的本体由于应用的领域不同,组织、设计者不同,其结构和概念的表达形式也存在着一定的差异,因此造成了本体的异构性。本体映射指的是在多个本体之间找到语义相同或相似的对应元素,从而在多个本体之间建立语义联系,消除不同本体或本体不同版本之间知识表达时的不一致现象,进而达到真正意义上的知识共享。本体映射是解决不同本体间的知识共享和重用问题的有效方法[2]。目前本体映射大多是由人工手动来完成的,不仅过程繁杂,而且很容易出错。这极大地影响了本体映射自动化程度和准确性。机器学习是解决这类问题的有效方法之一[3]。在此将对目前机器学习研究的主要趋势、理论与技术以及存在的问题,在本体映射中的应用做一详细的介绍。

收稿日期:2009-04-03;**修回日期**:2009-07-18
基金项目:国家"星火计划"项目(2004EA850069)
作者简介:李 琼(1984-),女,河南三门峡人,硕士研究生,研究方向为计算机网络与语义网;李宝敏,教授,硕士生导师,研究方向为计算机系统结构、计算机网络与语义网。

1 机器学习

机器学习是研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习的研究是根据生理学、认知科学等对人类学习机理的了解,建立人类学习过程的计算模型或认识模型,发展各种学习理论和学习方法,研究通用的学习算法并进行理论上的分析,建立面向任务的具有特定应用的学习系统。这些研究目标相互影响相互促进。

机器学习在人工智能的研究中具有十分重要的地位。机器学习系统是一个具有学习能力的智能系统,

它可以推理演绎加以归纳知识,证明已存在的事实、定理,并能归纳总结新的规则、定理和定律,遇到错误能自我校正,通过经验对自身性能加以改进,可以不断地自动获取和发现所需要的知识。

随着人工智能的深入发展,机器学习逐渐成为人工智能研究的核心之一。它的应用已遍及人工智能的各个分支,如专家系统、自动推理、自然语言理解、模式识别、计算机视觉、智能机器人等领域。

从机器学习的执行部分所反映的任务类型上看,目前大部分的应用研究领域基本上集中于以下两个范畴:分类和问题求解。

2 机器学习算法

机器学习的中心问题是从特殊的训练样例中归纳出一般函数。发展至今,机器学习算法已经有很多,并有了广泛的应用。常见的算法有:概念学习、决策树学习、支持向量机(SVM)、人工神经网络、贝叶斯(Bayes)学习、遗传算法、基于实例的学习、规则学习等等[4]。

2.1 决策树算法描述

决策树算法是应用最广泛的归纳推理算法之一,在数据挖掘的分类、生成规则方面都有很多应用。大多数已开发的决策树学习是 ID3 算法一种核心算法的变体。

决策树是通过把实例从根节点排列到某个叶子节点的方法来分类实例,叶子节点即为实例所属的分类。树上每个节点说明了对实例的某个属性的测试,并且该节点的每个后继分支对应于该属性的一个可能值。决策树分类实例的方法是从这棵树的根结点开始,测试这个结点指定的属性,然后按照给定实例的该属性值对应的树枝向下移动。然后这个过程在以新结点为根的子树上重复。

2.2 ID3 算法描述

ID3 算法采用自顶向下的贪婪搜索遍历可能的决策树空间,这种方法也是后继的 C4.5 算法的基础。基本的 ID3 算法通过自顶向下构造决策树进行学习。构造过程是从"哪个属性将在树的根结点被测试"这个问题开始的[5]。在算法中使用统计测试来确定每一个实例属性单独分类训练样例的能力。分类能力最好的属性被选作树的根结点的测试,然后为根结点属性的每一个可能值产生一个分支,并把训练样例排列到适当的分支,也就是,样例的该属性值对应的分支之下。然后重复整个过程,用每一个分支结点关联的训练样例来选取在该结点被测试的最佳属性,这形成了对决策树的贪婪搜索,也就是算法从不回溯重新考虑以前的选择。

3 领域本体映射

文中是依据国家"星火计划"的项目而建立的果品领域本体,果品领域本体库就是刻画果品领域类、属性、关系和实例的一种模型,目的是让果品知识能够被计算机理解和处理。基于本体思想建立果品体系,基本就是对果品知识按照果品领域的要求进行分类;但在果品领域内的本体由于语法结构各不相同,在未来的语义搜索时需要涉及到各个果品领域的本体库。如何把果品领域中的异构本体整合起来?映射技术就是为解决本体整合问题而产生的。

在本体映射中核心问题是机器学习算法,也就是依据定义和规则来进行推理、归纳、综合,来解决映射问题,并将映射结果存储,作为下次映射的依据。在不同本体分类树中,对其中的各个节点一一进行分析比较,根据语义匹配映射规则,找出语义重叠和相关的概念节点,再根据映射各项关系结果做出标记,进行映射匹配。

映射的过程可以看作是 Ontology 匹配(或者映射)的过程。典型的映射过程通过分析、比较 Ontology 来判断概念之间的对应关系,最终建立本体之间的映射关系。例如:给出两个本体,其本体分类树分别为 Ta 和 Tb,要寻找 Ta 到 Tb 的映射,即为找到分类树 Ta 中概念 A 在另一个分类树 Tb 中的匹配位置,需要找到 Tb 中三个关键的概念节点:

- Tb 中最相似的候选概念节点 B;
- Tb 最接近 A 的父节点 C;
- Tb 最接近 A 的子节点 D。

其中映射关系在这里可以分为两种:同义关系映射和实例映射。同义关系映射表示相似概念之间对称的等价关系,即不同本体树中的两个描述语言表示相同样的语义。实例映射,指的是一个概念是另一个概念的实例。

同义关系映射,对于同一领域的两个 Ontology ,它们的元素在很大程度上是存在重叠的,因此,可以在一个 Ontology 中为另一个 Ontology 的一个概念结点寻找一个最相似的结点。实例关系映射,对于结构不同的两个果品本体树,实例概念在各个本体树中的节点位置不相同,根据属性为实例概念节点找到在另个果品树中的对应节点,如图 1~图 3 所示。

图 1~图 3 为果品领域中的三个本体树,在图 1 本体树 3 中,苹果在农业生物学领域属于仁果类,在本体树 1 中,苹果相对于仁果类就是实例关系,苹果是仁果类的实例;在图 2 本体树 2 中,山楂又称为山里红,两个概念表示同一个事物和同一个概念,在图 1 本体树 1 中它和山楂就是同义词的关系。

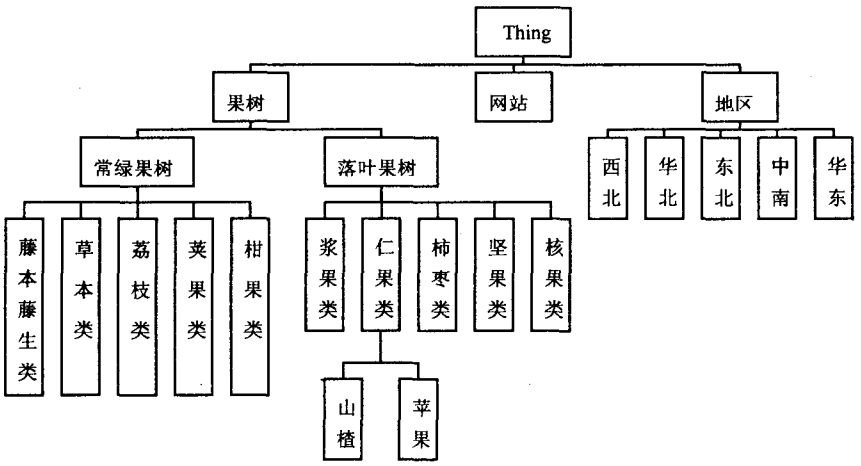

图 1　本体树 1

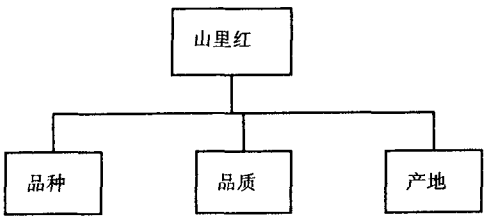

图 2　本体树 2

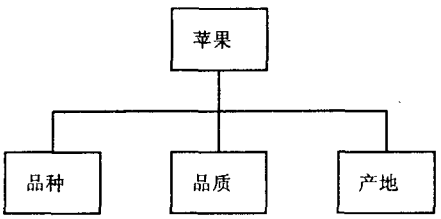

图 3　本体树 3

4　基于机器规则映射算法

本体是一组概念、属性和关系对一个领域进行规范化的描述。概念是领域中实体的抽象,可用一棵本体分类树来描述。

定义本体 T 为一个领域本体分类树,用一个 5 元组描述:

T=(N,R,H,S,X)

其中:N 表示概念的集合,是树中的节点;R 是一个关系集合,是树中的边,如:用 is－apart－of,ame－of,attribute－of,Instance－of 等来说明这些关系。H 描述树中概念间以及概念与实例间的层次分类关系;S 是实例的集合;X 是描述概念节点及实例的元数据属性集合,如:DublinCore。

映射 F:对于领域中的两个领域本体 Ontology 概念树 T=(N,R,H,S,X)、T′=(N′,R′,H′,S′,X′),F:N×N′→ref。其中,ref 是一个可以接受的解释语义关系的名词集合。

在文中果品领域概念映射的关系为两种,$f_{instance}()$和 $f_{same}()$,即实例映射关系和同义映射关系。在图 1～图 3 所示三棵本体分类树 T1,T2,T3 中,其中 T1=(N1,R1,H1,S1,X1),N1={Thing,果树,网站,地区,常绿果树,落叶果树,西北,华北,东北,中南,华东,藤本藤生类,草本类,荔枝类,荚果类,柑果类,浆果类,仁果类,柿枣类,坚果类,核果类},R1 为关系集合,R1={is－a-part－of,same－of,attribute－of,Instance－of};T2=(N2,R2,H2,S2,X2),N2={山楂,品种,品质,产地},R2={is－apart－of,same－of,attribute－of,Instance－of};T3=(N3,R3,H3,S3,X3),N3={苹果,品种,品质,产地},R3={is－apart－of,same－of,attribute－of,Instance－of}。

果树本体中,仁果类的实例为苹果、梨、枇杷、木瓜和山楂;核果类的实例为桃、杏、李、樱桃和榔;坚果类的实例为核桃、栗、银杏、阿月浑子和榛子;浆果类的实例为葡萄、草莓、醋栗、猕猴桃和树莓;柿枣类的实例为柿、君迁子(黑枣)、枣和酸枣;柑果类的实例为橘、柑、柚子、橙、柠檬、枳、黄皮和葡萄柚;荔枝类的实例为荔枝、龙眼和韶子;荚果类的实例为酸豆、角豆树、四棱豆和苹婆等;草本类的实例为香蕉和菠萝;藤本蔓生类的实例为西番莲和南胡颓子。

地区本体中,华北的实例为北京市、天津市、河北省、山西省和内蒙古自治区;东北的实例为辽宁省、吉林省和黑龙江省;华东的实例为上海市、江苏省、浙江省、安徽省、福建省、江西省和山东省;中南的实例为河南省、湖北省、湖南省、广东省、广西壮族自治区和海南省;西南的实例为重庆市、四川省、贵州省、云南省和西藏自治区;西北的实例为陕西省、甘肃省、青海省、宁夏回族自治区和新疆维吾尔自治区;港澳台特区的实例为香港特别行政区、澳门特别行政区和台湾省。

分类树的节点是概念,每个概念都用同一组实例来支撑说明。概念是客观世界中任何事物的抽象描述,形式上,概念定义为一个四元组:C={Id, L, P, IC},其中 Id 为概念的唯一标识符,用 URI 表示;L 为概念的语言词汇;P 为概念属性的集合;IC 为属于该概念的实例的集合。每个概念可用一组属性描述:Dublin Core 属性组(Title:资源的名字;Author or Creator:资源内容的创建者或组织者;Subject and Keywords:资源的主题及关键字;Description:资源内容的原文描述)。概念之间、概念和实例之间的关系可以用

一组关系来描述，如：用 is－a，is－part－of，ame－of，attribute－of，Instance－of 等用来说明这些的关系。语义匹配(映射)关系，有 1 对 1、1 对多、多对多匹配[6]。

例如：在本体分类树 T2 中，山里红这个概念节点，C1＝{Id，L，P，IC}，Id＝'蔷薇科植物山里红或山楂的干燥成熟果实'，L＝'蔷薇科落叶小乔木，树皮暗灰色，有浅黄色皮孔，小枝紫褐色，单叶互生或于短枝上簇生，叶片宽卵形，伞房花序，花白色，后期变粉红色，果实球形，熟后深红色，表面具有淡色小斑点'。

在本体分类树 T1 中，山楂这个概念节点，C2＝{Id，L，P，IC}，Id＝'蔷薇科植物山里红或山楂的干燥成熟果实'，L＝'果实较小，类球形，直径 0.8cm～1.4cm，有的压成饼状。表面棕色至棕红色，并有细密皱纹，顶端凹陷，有花萼残迹，基部有果梗或已脱落。质硬，果肉薄，味微酸涩'。

在本体分类树 T3 中，概念节点苹果，C3＝{Id，L，P，IC}，Id＝'落叶乔木，叶椭圆形，有锯齿，花白微红，果实圆形，味甜，是普通的水果'，L＝'果实类球形，直径 10cm～30cm，从外到里依次是果皮、果肉、果核。成熟果实的果皮成红色或黄色，有果梗。果皮薄可食，果肉多，汁多'。

在本体分类树 T1 中，概念节点仁果类，C4＝{Id，L，P，IC}，Id＝'多汁的果肉包着几个种子的果核'，IC＝{苹果，梨，山楂，木瓜，枇杷，沙果，海棠果}。

在进行映射推理时，采用机器学习算法，依据机器映射算法中的规则、定义，比较本体分类树中概念节点的属性，对两个本体之间的概念节点进行映射。

首先在本体分类树 TI 与 T2 之间进行映射，T2 中山里红 C1 与本体树 T1 中的各概念节点进行比较，当与山楂节点 C2 比较时，采用机器学习算法推理出 C1_Id＝C2_Id，C2_L≌C1_L，得出 T2 中 C1 与 T1 中的 C2 表示的是同一个概念，可以在 C1 与 C2 之间建立同义映射关系。又根据机器学习算法，C2 的父节点是 C1 的父节点，C2 的兄弟节点与 C1 相似。

在本体分类树 T1 和 T3 之间进行映射，T3 中的苹果节点 C3 依次与分类树 T1 中的概念节点进行比较，机器算法推理出 C3 与本体分类树 T1 中的 C4 有关系 C3_Id⊂ C4_Id，C3_IC⊂ C4_IC，得出 C3⊂C4，C3 是 C4 的实例，则在 C3 和 C4 之间建立实例映射关系[7]。

从而得出：

f_{same}(C1)＝ C2，C1 与 C2 之间是同义映射关系，即 C1 与 C2 表达是同一种概念。

$f_{instance}$(C3)＝ C4，C3 与 C4 之间是实例映射关系，即 C3 是 C4 的实例。

这就完成了一个可行的本体映射。

下面是基于机器学习映射方法的一些相关规则[8]：

● s 是概念 C 的实例，若 C 是 C′的子孙(表示 C< C′)，那么 s 也是概念 C′的实例。

● 在不同本体中，如果两个概念属于两个本体树中同一个父概念，那么这两个概念是相似的，即兄弟概念是相似的。本体树 T 与 T′，有两个同概念节点 P，则在两个本体树中 P 节点的子节点是相似的。

● 如果在不同本体中两个概念的父概念相似，那么这两个概念也可能相似，并且这两个概念的部分子概念也可能相似。

● 如果某个概念的兄弟概念结点 L 与某一概念 X 相似，那么该概念 L 与概念 X 也可能相似。

● 如果两个概念相似，那么它们的子概念在一定程度上也相似。

● 如果所有子概念都与概念 Y 相似，那么它们的父概念也与概念 Y 相似。

● 如果两个概念具有相同的兄弟，则这两个概念可能是相似的。

● 如果两个概念具有相同的实例，则这两个概念可能是相似的。

● 如果两个概念具有相同的属性，则这两个概念可能是相似的。

5 结束语

文中依据果品领域本体，结合机器学习算法详细描述了概念映射算法。但在进行映射比较的同时，很多属性要进行参考，如层次结构描述、关系、约束等，还有机器算法规则需要更详细的定义。

异构的领域本体匹配是语义网发展面临的最富有挑战性的问题之一，本体描述没有同一标准，映射算法也不相同，所以实现完全程度上本体匹配尚不可能，但该论文的概念映射匹配的研究为同领域的语义互取提供了可能，从这重意义上讲，对今后语义网领域研究工作具有一定参考价值。

参考文献：

[1] 张 娜，李宝敏. 语义检索及其关键技术研究[J]. 计算机技术与发展，2006，16(11)：22－25.

[2] 万 彬，王卫疆，汪秉文. 语义 Web 服务及其在 WWW 上的应用研究[J]. 微机发展(现更名：计算机技术与发展)，2005，15(7)：106－108.

[3] Quinlan J R. Induction of decision trees[J]. Machine Learn-

(下转第 88 页)

缩,以最好点为中心,将所有点向最好点收缩一半,最后计算收缩后各顶点的函数值,转到开始,循环迭代。

图 1 为程序流程图。

5 实验结果及分析

5.1 多态蚁群算法仿真实验

以 TSP 问题为例,试验中所用的 TSP 问题数据来源于 oliver30 城市问题。对应的参数得出相应的计算结果如表 1 所示。

表 1 几组参数下多态蚁群算法仿真实验结果

α	β	ρ	Q	代数	最优结果
1	2	0.40	200.00	132	423.9117
1	3	0.70	100.00	294	423.9117
1	3	0.50	50.00	222	423.9117
1	4	0.40	100.00	372	423.9117
1	4	0.30	200.00	245	423.9117

5.2 用单纯形算法确定多态蚁群算法的组合参数匹配实验

根据多态蚁群算法得出的结果(见表 1),将其放在 a 数组里作为初始值,算法中的 α,β,ρ,Q 四个参数映射单纯形空间的一个点,这一组参数通过蚁群算法求出一个结果,也就是单纯形算法中顶点的函数值,在单纯形算法中,先给定空间几个点,再通过上述单纯形的加速算法求出最好的解及其相应的参数配置,结果如表 2 所示。

表 2 最优的解及其相应的参数组合匹配实验结果

α	β	ρ	Q	代数	最优结果
1	3	0.30	100.00	70	423.7406

5.3 实验分析

(1)在单纯形算法中调用多态蚁群算法,蚁群算法的形参是浮点型的 α、β、ρ、Q 四个参数,返回的是这一组参数以 TSP 问题为例所得出的结果,也就是单纯形算法中顶点的函数值。根据单纯形算法的加速原理,比较函数值的大小,丢掉最坏的点,代之以新的点,再重新计算单纯形各顶点及其函数值,然后再循环迭代,直至找到最优的解及其相应的参数组合匹配。

(2)在表 1 中五组参数,分别在 132、294、222、372 和 245 代,取得最好结果仅 423.9117。且在 400 代内也未找到更好的结果。而表 2 中参数组合在第 70 代,就取得最好结果 423.7406。

(3)从国内外研究结果看,针对 oliver30 城市问题,423.7406 是目前已公布的最好解。

(4)由于多态蚁群算法是一种随机搜索算法,种子选择的不一样,使得每次实验得出最优结果的代数稍有偏差。使用表 2 中参数组合,经过多次实验,都能较快找到最优结果 423.7406。实验证明用单纯形算法确定蚁群算法中参数的组合配置是可行的。

6 结束语

在对多态蚁群算法模型及参数进行理论分析的基础上,讨论了组合参数的选择对寻优结果的影响,提出了用单纯形算法来确定多态蚁群算法中参数的最优组合方法,设计了用单纯形算法确定多态蚁群算法中参数的最优组合模型及流程。并进行了仿真实验,实验结果表明这种方法确定的最优参数组合较快搜索到目前已公布的最好解。

参考文献:

[1] Dorigo M, Maniezzo V, Colorni A. The Ant System: Optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 1996, 26(1):29-41.

[2] 常晓磊,闫仁武. 一种基于蚁群算法的分类规则挖掘算法[J]. 计算机技术与发展,2007,17(7):114-116.

[3] 徐精明,曹先彬,王煦法. 多态蚁群算法[J]. 中国科学技术大学学报,2005,35(1):59-65.

[4] 徐精明,曹先彬,王煦法. 蚁群算法求解问题时易产生的误区及对策[J]. 计算机工程,2004,30(16):25-27.

[5] 詹士昌,徐 婕,吴 俊. 蚁群算法中有关算法参数的最优选择[J]. 科技通报,2003,19(5):381-386.

[6] 叶志伟,郑肇葆. 蚁群算法参数 α、β、ρ 设置的研究——以 TSP 为例[J]. 武汉大学学报,2004,29(7):597-601.

[7] 段海滨. 蚁群算法原理及其应用[M]. 北京:科学出版社,2005:100-160.

[8] 孔锐睿,仇汝臣,周田惠. 单纯形的加速算法[J]. 南京理工大学学报,2003,27(2):38-41.

(上接第 84 页)

ing,1986(1):81-106.

[4] Noy N F. Semantic Integration: A Survey of Ontology-based Approaches[J]. SIGMOD Record,2004,33(4):65-70.

[5] 班 瑞. 基于语义 web 的机器学习算法研究与应用[D]. 南京:南京理工大学,2006.

[6] Ding Y, Foo S. Ontology Research and Development, Part 2-A Review of ontology Mapping and Evolving[J]. Journal of Information Science,2002,28(5):375-388.

[7] 欧 灵,张玉芳,吴中福,等. 基于机器学习的本体概念相似性研究[J]. 计算机科学,2006(11):188-191.

[8] 李选如,何洁月. 语义集成:本体映射方法研究[J]. 计算机技术与发展,2007,17(2):56-58.