文章编号：1001-8360(2016)10-0052-06

基于领域本体模型的概念语义相似度计算研究

朱雪

(大连海事大学 信息科学技术学院，大连 116026)
(通信作者电子邮箱 1874110235@qq. com)

摘 要： 随着本体在信息检索、人工智能等领域的广泛应用,面向本体的概念相似度计算成为本体研究的一大热点。目前领域本体中概念相似度的研究主要是利用概念的上下位关系进行计算,但这并没有完整反映出概念的语义信息。本文首先在本体模型的基础上提出领域本体模型的八元组表示方法和领域本体概念的九元组表示方法,给出领域本体模型的 DCG 图。然后以提出领域本体模型为基础构建概念语义相似度计算的 MD4 模型,该模型全面考虑概念的属性、上下位语义结构关系、自定义语义关系和实例特征对相似度的影响,通过综合计算,得到领域本体中概念的实际相似度。本文最后以动车组专业本体作为实验对象,对 MD4 模型进验证。实验结果表明,该计算模型充分利用概念的语义信息,得到的结果也比较合理。

关键词： 本体;领域本体;语义相似度;MD4 计算模型

中图分类号： P391.4 **文献标志码：** A

Research on Concept Semantic Similarity Computation Based on Domain Ontology Model

ZHU Xue
(Institute of Computer Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract : Along with the widespread application of ontology in the fields of information retrieval and artificial intelligence etc, the concept similarity computation of domain ontology has become the focus research field. Currently, most research on concept similarity computation is based on“is a” relation between concepts, however, it does not utilize the concept semantic information completely. Firstly, on the basis of ontology model, this paper proposes an Eight-tuples model for domain ontology and a Nine-tuples model for its concept. We can use the Directed Cyclic Graph(DCG) to formalize the domain ontology. Then this paper builds the fourfold Matching-Distance Model(MD4)to compute concept semantic similarity based the on domain ontology model. This MD4 model has comprehensive consideration about the property,“is-a” relation,“user-defined” relation and instance of concept. And through integration computation, we can get the true concept similarity. Finally, this paper makes an experiment on the EMU ontology that was built in the high-speed railway domain. The ex-periment shows that the method utilizes the concept semantic information fully and the computation result is reasonable.

Key words: Formal Concept Analysis(FCA)； intent weight value；bag-of-visterm；category visual vocabulary

1 绪论

本体在软件工程、人工智能、信息检索、Web 服务等领域中扮演着越来越重要的角色[1]。依照领域依赖程度,可将本体分为通用本体(Top Ontology)、领域本体(Domain Ontology)、任务本体(Task Ontology)和应用本体(Application Ontology)[2]。领域本体可以有效地组织领域中的知识,使知识更好地共享、重用。有一些领域,如农业领域、铁路领域、高速铁路领域、航空领域等,这些领域都由不同的专业领域构成,如高速铁路领域由工务工程、牵引供电、动车组、运营管理等专业领域构成。这些领域本体的构建可以先基于叙词表或范畴与主题词表构建各专业本体,然后集成为一个统一的领域本体。在利用领域本体的同时,如何提高概念相似度计算是进行概念语义扩展的重要步骤，如何提高概念相似度精度计算就成为提高检索质量的关键技术之一。目前基于概念距离计算概念相似度的算法大多是针对上下位关系，而忽略其他关系，这就导致算法不

收稿日期：2016-10—30：修回日期：2016—11—12。

基金项目：国家自然科学基金资助项目(61373099)；国家青年科学基金资助项目(61402316)。

作者简介：朱雪(1994 一)，女，山东滨州人，硕士，主要研究方向：概念格、数据挖掘。

能完整反映概念的语义,从而影响概念相似度计算的准确性。

本文在本体模型的基础上提出领域本体的模型,并以此模型为基础构建领域本体概念语义相似度计算的 MD4 模型,最后以高速铁路领域中动车组专业本体作为实验对象,对本文提出的 MD4 模型进行验证。实验结果表明:该计算模型能够比较准确地反映概念之间的语义关系,为由多专业构成的领域本体概念之间的语义关系提供一种有效的量化方法。

2 概念语义相似度研究综述

传统基于领域本体的概念之间相似度计算模型主要有基于距离的语义相似度计算模型、基内容的语义相似度计算模型和基于属性的语义相似度计算模型 3 种。

基于距离的语义相似度计算模型基本思想是把概念之间的语义距离用两个概念在层次网络中的几何距离来量化。最简单的一种计算方法就是把网络中的所有有向边的距离都看成同等重要,都看成 1。这样,两个概念间的距离就等于这两个概念对应的节点在层次网络中构成最短距离的有向边数量。因此可得出一种简单的语义相似度计算模型[3]。

$$\mathrm{im1}(w1,w2)=\frac{2\ (H-1)\ -L}{2\ (H-1)}$$

式中,H 为网络结构的最大深度;L 为概念节点 w1 和 w2 之间有向边的数量。

该计算模型可以简单地反映出:如果两个概念之间的距离越远,它们之间的语义相似度就越小;反之越近,则越大。但上述计算模型在计算概念之间的语义相似度是很粗糙的,没有考虑网络结构中有向边的差异，于是有人提出基于距离的改进语义相似度计算模型,Leacock[4]在其基础上提出的计算模型。

基于内容的语义相似度计算模型基本原理是[5]:如果两个概念共享的信息越多,它们之间的语义相似度也就越大;反之越少,则越小。在层次网络中,每一个概念都可以认为是对它祖先节点的细化,因此可以近似理解为每一个子节点包含它所有祖先节点的信息内容。这样,两个概念的语义相似度就可以用其最近共同祖先节点的信息内容来衡量,可以得到层次网络中任意两个概念之间的语义相似度计算模型[6]

$$\mathrm{sim2}(w1,w2)=\frac{2IC[Anc(w1,w2)]}{IC(w1)+IC(w2)}$$

式中,Anc(w1,w2)表示概念节点 w1 和 w2 在层次网络中的最近共同祖先节点。IC(w)表示概念 w 所拥有的信息量。

在现实世界中,如果两个事物有很多属性相同,则说明这两个事物很相似;反之,则相反。因此,基于属性的语义相似度计算模型的基本原理也就是通过判断两个概念对应的属性集的相似程度。Tversky 提出一种基于属性的计算概念语义相似度的方法[7]

$$\mathrm{sim3}(w1,w2)=\theta f(w1\cap w2)-\alpha f(w1-w2)-\beta f(w2-w1)$$

式中,w1∩ w2 表示概念 w1 和 w2 所共同拥有的属性集;w1- w2 表示概念 w1 拥有而概念 w2 没有的属性集;w2- w1 表示概念 w2 拥有而概念 w1 没有的属性集。

其他改进的语义相似度算法实质都是基于传统的语义相似度计算模型,即基于距离、基于内容和基于属性进行相似度的计算。文献[8]通过实验比较分析现有的方法,结果显示结构相似往往比词汇相似更可靠。本体概念的结构相似大多利用本体的树结构来进行计算。在领域本体概念语义相似度计算方面,黄果等[9]针对 3 种传统语义相似度计算模型的优缺点和领域本体所特有的性质,提出一种改进的基于领域本体语义相似度计算模型。在该模型中,对影响本体层次网络有向边权重的 4 种因素(类型、密度、深度、强度）中的密度和深度进行新的量化,并且把概念的属性因素考虑到其中,从而更加全面地量化本体网络中概念节点之间的语义相似度，提高了概念之间语义相似度量化的准确性。

3　领域本体模型

3.1　本体模型

关于 Ontology 的定义有许多,目前获得较多认同的是 R. Studer 的解释[10]:“Ontology 是对概念体系的明确的、形式化的、可共享的规范说明”。在最简单的情况下,本体只描述概念的分类层次结构;在复杂的情况下,本体可以在概念分类层次的基础上,加入一组合适的关系、公理、规则来表示概念间的其它关系,约束概念的内涵解释。

定义 1　一个完整的本体应由概念、关系、函数、公理和实例等 5 类基本元素构成。本体可以表示为如下形式:O={C, R, F, A, I},其中,C 为概念(或称为类)。概念是指客观世界中任何事物的抽象描述,在本体中通常按照一定的关系形成一个层次结构;R 2C× C:概念之间的关系,如“subclass-of”关系、“part-of”关系等;F　Rn 是一种特殊的关系,其中第 n 个元素 cn 相对于前面 n- 1 个元素是惟一确定的,函数 F 可表示为 c1× c2× … × cn- 1※cn;A 为概念或者概念之间的关系所满足的公理,是一些永真式;I 为领域内概念实例的集合。

3.2　领域本体模型

(1)领域本体模型根据本体的定义和描述,领域本体反映一个对给定领域的通用观点,其通过定义概念与概念之间的关系来描述概念的语义信息。在实际的领域本体中,由于概念之间不仅仅存在着上下位关系,概念之间通过其他各种关系可以连接,尤其在多专业构成的领域本体中还有许多自定义的关系,这使得概念的组织形式并不完全是一个树型结构,而是一个网状结构。因此,根据领域本体的特点,我们在本体模型的基础上重新构建领域本体模型。

定义 2　领域本体模型是一个 8 元组:DO={C,P, Hc, Rs, Rud, I, F, A},其中,DO 表示领域本体;P 表示领域本体中 Datatype 类型属性;Hc 表示类间的上下位(subclass-of)二元关系;Rs 表示类间的同义(synonymy)关系;Rud 表示类间的用户自定义(user-defined)关系(包括 part-of 关系也用自定义关系来描述),也就是类的 Object Property。

定义 3　概念 C 的模型是一个 9 元组:C={P,Csc, Cuc, Cs, Cr, Hc, Rs, Rud, Ic},其中,P 表示概念 CDatatype 类型属性;Csc 表示概念 C 的子概念(subclass);Cuc 表示概念 C 的父概念(upperclass);Cs 表示概念 C 的同义概念(equivalentclass);Cr 表示与概念 C 有关系的概念,这里主要指通过用户自定义关系联系起来的概念;Hc 表示概念 C 的上下位关系;Rs 表示概念 C 的同义(synonymy)关系;Rud 表示概念 C 的用户自定义(user-defined)关系;Ic 描述概念 C 的实例。

概念之间的关系主要分为 3 类:①上下位关系,用 Csc,Cuc 和 Hc 表示;②同义关系,用 Cs 和 Rs 表示;③用户自定义关系,用 Rud 表示。

(2)领域本体形式化描述

本体是将某个应用领域抽象概括成一组概念及概念之间的关系,它主要是要描述概念间的层次结构:上层概念的语义相对于底层概念更为抽象,共享的程度高;而底层概念较为具体,更贴近具体的应用,概念之间是泛化和特例的关系。

由于语义网络表示法在描述概念层次结构时具有天然的优越性,因此本文用基于语义网络的表示方法来表示本体的概念模型。在领域本体中,只考虑上下位关系时的本体模型为树型结构,可以使用语义网络的有向非循环图(DAG)[11]的方法表示。领域本体概念间的自定义关系包括多种形式,用户可以根据实际情况自己定义。考虑用户自定义关系的领域本体模型可以用有向循环图(DCG)来表示,如图 1 所示。其中,C14 和 C15 是用户自定义关系。下面是本体概念模型的 DCG 图。

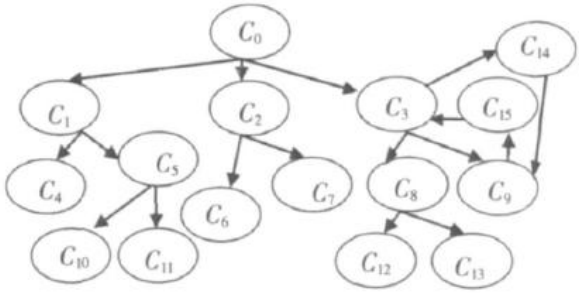

图 1 本体概念模型的 DCG 图

4 领域本体概念语义相似度计算 MD4 模型

首先定义领域本体概念语义相似度:

定义 4 Sim:C1× C2※[0, 1],相似值在 0 和 1 之间。C1 和 C2 是基于定义 3 的两个概念的术语集合,Sim(A, B)表示 A 和 B 之间的相似度。Sim(e, f)=1:表示概念 e 和概念 f 是相同的两个概念;Sim(e, f)= 0:表示概念 e 和概念 f 是两个完全不同的概念。

4.1 MD4 模型组成描述

传统本体概念间相似度计算的不足在于其语义邻居只考虑层次语义关系,没有考虑语义关系中非层次关系的影响,同时对象实例对于概念的影响也没有考虑。本文在定义 3 的基础上,提出计算概念之间相似度的 MD4 模型,如图 2 所示。该模型全面考虑本体概念模型中各种元素对相似度的影响。

图 2 MD4 模型组成描述

4.2 MD4 模型概念语义相似度算法

在同一本体中,概念相似度计算首先需要检查两个概念是否同义。如果两个概念同义,那么两个概念是完全相似的,其相似度为 1。

定义 5 如果领域本体中概念 C1 和 C2 成同义关系,那么 Sim(C1, C2)= 1。

(1)上下位关系语义相似度计算

在领域本体中,只考虑上下位关系时的本体模型为树型结构。计算上下位关系语义相似度时采用基于距离的概念相似度计算方法。之所以选择这种方法,是因为在领域本体中,概念主要以树状结构排列,而利用距离计算概念的初始相似度可以合理地利用概念的这种组织形式,从而使算法比较直观、易于理解。本文参考陈杰、蒋祖华的算法,综合考虑概念距离和层次对概念相似度的影响[12],算法式为

$$Sim_h(C_i, C_j) = [\alpha \times (dl(C_1) + dl(C_2))] / [(Dist(C_1, C_2) + \alpha) \times 2 \times Maxdl \times \max(|dl(C_1) - dl(C_2)|, 1)] \quad (1)$$

式中,dl(C1)和 dl(C2)分别是 C1 和 C2 所处的层次;Dist(C1, C2)是概念 C1 和 C2 之间的本体树中的最短路径;Maxdl 是指本体树的最大深度,在这里除以该参数是便于计算结果的归一化处理;α 是一个可调节参数,一般不小于零。

(2)概念 Datatype 类型属性相似度计算

当两个 Datatype 型的属性进行比较时,如果两个属性是相同的,那么相似度为 1,否则相似度为 0。首先确定 Ci 和 Cj 的属性集 Pi 和 Pj,概念 Ci 和 Cj 分别对应 m 和 n 个 Datatype 类型的属性(Datatype Proper-ty),然后对属性集合 Pi 和 Pj 进行笛卡尔乘积 Pi×Pj,得到配对集,再计算 Ci 和 Cj 的属性相似度 Simp,得到 Ci 和 Cj 的属性相似度计算公式为

$$Sim_p(C_i, C_j) = \frac{\sum_{i=1}^{m_2}\sum_{j=1}^{n_2} Sim(P_i, P_j)}{\max(m_2, n_2)}$$

式中,m2 和 n2 分别是念 Ci 和 Cj 的 Datatype 类型属性的个数。

5 结束语

本文首先对语义相似度计算模型进行综述。针对

领域本体的实际情况,提出领域本体模型的 8 元组表示方法和领域本体概念的 9 元组表示方法。在领域本体概念模型表示的基础上,构建计算领域本体概念之间语义相似度的 MD4 模型,并给出该模型的详细算法。最后以动车组专业本体作为实验对象,采用本文提出的 MD4 计算模型、传统的基于距离的语义相似度计算模型分别计算每对概念的语义相似度,并和专家根据经验给出的语义相似度进行比较。结果表明,该计算模型能够比较准确地反映概念之间的语义关系,为领域本体概念之间的语义关系提供一种有效的量化方法。

参考文献:

[1] 丁轶.基于 LDA 的图像区域标注模型的研究[D].南京：南京大学，2012.

[2] LI F，PERONA P. A Bayesian hierarchical model for learning natural scene categories [C]Piscataway：IEEE Press，2005，2：524 -531.

[3] 王敏.基于 LDA 主题模型的图像场景分类[D].西安：西安电子科技大学，2013.

[4] 李晓旭. 基于概率主题模型的图像分类和标注的研究[D].北京：北京邮电大学，2012.

[5] BLEI，NG A，JORDAN M . Latent Dirchlet allocation[J]Journal of Machine Learning Research，2003，2(3)：993—1022.