《智能信息处理》课程考试

基于本体的文本相似度计算分析

刘程

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 8 日

基于本体的语义相似度计算分析

刘程

（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　语义相似度计算在知识获取、信息检索、文本聚类等方面有着广泛的应用，而本体能够准确的描述概念之间的内在联系，帮助人们准确快速获取的人们所需要的内容， 应用在语义相似度的计算上是非常合适的。本文主要对本体、语义相似度以及二者之间的相互作用进行梳理，归纳了基于本体的语义相似度的几种计算模型，探讨了目前此计算模型的应用场景。有关讨论结果将为未来基于本体的语义相似度的计算分析研究提供借鉴。

关键词　本体；语义相似度；语义相似度计算；

Analysis of semantic similarity calculation based on ontology

Liu Cheng

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract　Semantic similarity calculation has a wide range of applications in knowledge acquisition, information retrieval, text clustering, etc. Ontology can accurately describe the internal connections between concepts, help people accurately and quickly obtain the content that people need, and apply to semantic similarity The calculation of the degree is very suitable. This paper mainly sorts out the ontology, semantic similarity and the interaction between the two, summarizes several calculation models based on ontology, and discusses the application scenarios of the current calculation model. The results of the discussion will provide reference for the calculation and analysis of semantic similarity based on ontology in the future.

Key words　Ontology; Semantic similarity; Semantic similarity calculation;

1　引言

随着互联网技术的迅速发展和在线资源容量的迅速膨胀, 互联网已成为一个庞大而杂乱无章的信息库。因此，如何在节约时间的同时又能更好地检索到自己所需类别的信息已成为人们迫切需要解决的问题。一方面, 用户对信息获取的准确性、系统性越来越难；另一方面，由于语义的异构特征，利用关键词实现的简单匹配，缺失了有针对性的语义信息和数据关联，缺少对用户意图的精准推测，因而导致了“信息孤岛”现象。诸如此类的“差异化表达”，无法满足用户从信息到知识层面探索的深层次体验。因此，能够对文本进行自动化的处理与检索是人们一直关注的话题。语义相似度作为机器学习、自然语言处理领域的底层框架，在过去的20 年里发展迅速，成果丰富。其中利用本体解决语义层面的信息共享问题，成为该领域的一个核心研究方向。将本体的相关特征应用到语义相似度的建模型中，改进语义相似度的计算模型是非常有必要的。通过总结经典方法、梳理汇报最新研究成果，对于完善基于本体的语义相似度研究进展具有重要应用价值。

2　相关知识描述

2.1　本体定义

本体一词最开始是哲学上的一个概念，并且一

直以来存在着许多不同的用法，也有很多学者对此进行研究和定义。在计算机科学领域，本体的核心意思是指一种模型，用于描述由一套对象类型（概念或者说类）、属性以及关系类型所构成的世界，本体之中模型的那些特征应当非常类似于相应的现实世界。

本体这一概念的确定也经历了一个过程。首先由 Neches 等人进行本体的定义，其定义本体为“组成主题领域的词汇表的基本术语和关系，以及组合这些术语和关系来定义词汇表外延和规则”。随后，在 1993 年 Gruber 给出的本体定义被现在大家广泛接受“本体是概念模型明确规范说明”。之后 Studer 等人认为本体是共享概念模型的明确的形式化规范说明。[1]虽然对于本体的定义个不相同，但是，其本质含义具有以下四层：概念化、形式化、明确、共享。“概念”指通过抽象出客观世界中一些现象的相关概念而得到的概念模型；“明确”指所使用的概念及使用这些概念的约束都有明确的定义；“形式化”指本体是计算机可读的；“共享”指本体中体现的是共同认可的知识，反映的是相关领域中公认的概念集 [2]。

2.2 语义相似度

语义相似度是通过语义距离来衡量的概念，也就是说，如果两个语义之间的距离是无穷大的，那么这两个语义之间的相似度很低，反之，如果两个语义之间的距离无限逼近0，那么可以认为这两个语义非常相似。Dekang L.认为任何两个对象的相似度取决于它们的共性(Commonality)和差异性(Differences)。即两个对象的共性越多, 相似度越大；两个对象的差异性越大, 相似度越小。用Com(i,j)来表示对象i和j的相似度, 用Dif(i,j)表示两者的相异度（即语义距离）。如果两个对象i和j不相似，则Com(i,j)=0；如果相似性值越高，那么对象之间的相似性越大，当相似值达到1时，则表示两个对象完全相似，即对象是等同的[3]。

2.3 基于本体的语义相似度分析

想要对本体技术语义相似度进行分析，就需要有意识的建立以距离为基础的语义相似度计算模型。在此模型建立的过程中，需要结合以下因素进行：

(1) 语义重合度，即本体内部概念中上位关系概念相同概念的数量，此数据可以在一定程度上反映出本体概念的相同内容, 在计算中直接将共同具有的内容以公共节点的形式表示；

(2) 语义深度，即本体内部概念所具有的层次深度，其通常与语义的相似度之间具有较显著的正相关性；

(3) 语义距离，即本体中两个节点连接通路中最短路径所要经过的边数，通常情况下， 其具体的大小与语义相似度之间具有较显著的负相关性；

(4) 语义密度，即与概念具有兄弟关系的阶段的数量，通常情况下，其具体大小与语义相似度之间具有较显著的正相关性。

另外，需要注意的是，除通过语义距离对语义相似度进行表示外，基于语义的属性、领域本体也可以构建出反应语义相似度的模型[4]。

3 本体结构对语义相似度计算的影响

本体借助树形结构来表征概念之间的语义关系，树的节点表示概念，边表示概念节点之间的关系。一般来说，较为抽象宽泛的概念位于数中较高的位置，周围节点相对来说比较稀疏，反之，较为具体详细的节点处于树中较低的位置或者是末端的叶子节点，周围节点相对来说比较稠密。本体结构主要由概念节点深度、概念节点密度、连接概念节点的中间路径类型、有向边关联强度以及概念节点属性五大类。本章我们就不同本体结构对语义相似度的影响进行分析[5]。

(1) 概念节点深度

概念节点在树中的位置侧面反映了不同层面的概念关系。概念节点层次越高，说明概念越抽象，则两个概念间的相似度越低；概念节点层次越低、表达的概念越具体，则两个概念之间的相似度越高。

(2) 概念节点密度

对于局部本体结构，概念节点的密度越大，语义相似度越小。

(3) 连接概念节点的中间路径类型

概念节点所代表的概念词对应到现实世界就会体现多种关系类型，因而节点的语义相似度也不同。

(4) 有向边关联强度

在本体结构中，一个子节点可以有一个或多个父节点。当一个父节点相连了很多子节点时，如果在同一层次中，某一子节点相对其他的子节点更重要，那么该子节点和父节点相应构成的有向边的权重也就更大。因此在本体结构树中，不同节点之间的重要程度存在差异。

(5) 概念节点属性

对于领域本体来说，概念节点之间使用的相同属性越多，对其相应的语义相似度影响越大。

4 基于本体的语义相似度计算模型

基于本体的语义相似度计算分为基于距离的方法、基于信息内容的方法、基于概念属性的方法和混合式方法 4 种代表性方法[6]。

4.1 基于距离的语义相似度计算

基于距离的语义相似度计算方法的核心是测量两个概念词，通过本体概念结构树中的路径长度以计算其语义相似度，以路径长度的方式来体现语义之间的差异。基本思想是：两个概念词在本体层次树中的路径长度越大，相似度越小。Rada 利用本体的层次结构中两个概念词的距离来表征相互之间的语义距离。计算公式如 (1)所示。通过这种方法来计算相对比较简单，但是不足之处是大多数都没有考虑边的类型影响因素，这一算法成立是以“假设本体分类体系中所有边的距离权重相等”为前提得到的结果。另外，所表征的关联强度以及位置信息都会影响边的重要程度。

$$Com(i,j)=\frac{1}{Dif(i,j)} \tag{1}$$

4.2 基于信息内容的语义相似度计算

基于内容的方法是将信息熵的计算与本体关系相结合。其基本思想主要就是通过比较概念对之间的共享信息量来计算概念之间的语义相似度，概念之间共享信息越多，熵越小，差异信息越小，相似度越大。共享的信息内容通过共有的父节点信息量计算表示，差异信息量由各概念和共有的父节点之间的信息量差来表示。因为在本体中子节点往往是其上一层父节点的细化，所以在整个树形结构中任意一个子节点的信息内容能够反映其所有的祖先节点的信息内容，这也说明了可以通过比较概念词的共有的父节点概念词所包含的信息内容来衡量它们之间的相似度。单个概念节点信息量的计算公式如(2)所示。在本体中计算任意两个节点之间的相似度由公式(3)所示。

$$Info=\sum_{i=1}^{m}(C_i)\log_2(p(C_i)) \tag{2}$$

$$Com(c_i,c_j)=\frac{2IC(Lnncan(c_i,c_j))}{IC(c_i)+IC(C_j)} \tag{3}$$

4.3 基于概念属性的语义相似度计算

基于概念属性的语义相似度计算模型是借助于两个概念对应的属性集来进行计算的，它的基本原理是衡量两个概念对应的属性集的相似程度。概念属性分为数据类型属性和对象类型属性两种。这种计算方法的优劣是通过本体属性集的完备性来衡量的。也就是说，如果两个概念对应的属性集中所共同拥有的属性越多，那么说明它们的相似度越高，反之，相似度越低。基于概念属性的语义相似度方法是为了尝试去解决基于距离的语义相似度计算方法所反映的问题，不再是通过路径长度来衡量语义相似度，而是通过利用本体属性的重叠程度来衡量。从这一点我们也能清楚地看出基于概念属性的语义相似度计算方法弥补了基于距离的语义相似度计算方法不能解决跨本体的语义相似度的问题，使得在语义相似度的计算上应用更加广泛。从两个概念的属性集角度来计算语义相似度的计算公式如等式(4)-(7)所描述。

$$Com(c_1,c_2)=Xf(x)-Yf(y)-Zf(z) \tag{4}$$

$$x=c_1\cap c_2 \tag{5}$$

$$y=c_1-c_2 \tag{6}$$

$$z=c_2-c_1 \tag{7}$$

其中 x 是两个概念之间的共同属性，y 是两个概念之间的不同属性。通过这一计算方法就将共同属性和不同属性进行了综合比较，利用相同属性增加概念间的语义相似度，不同属性减少语义相似度这一性质来得到所期望的效果，属性的选择上比较灵活。当然，该算法也存在需要进一步改进的地方。其不足之处在于缺乏对数据类型的区分，一些细节问题还没有被完全考虑进去，例如被比较概念词的位置信息、祖先节点和所包含的信息内容等方面都有待完善。

4.4 基于混合方法的语义相似度计算

基于混合方法的语义相似度计算，顾名思义就是对以上所用方法的一个整合，将路径长度、概念深度、局部密度、概念属性和信息内容等因素同时

考虑在内的语义相似度计算方法。由于考虑的因素比较多，所以计算效果比较好。

4.5 基于本体的语义相似度的应用

语义相似度计算在自然语言处理、智能检索、文本聚类、文本分类、自动应答、词义排歧和机器翻译等领域都有广泛的应用，它是一个基础研究课题，正在为越来越多的研究人员所关注。根据现有的研究进行归纳整理，可以发现目前语义相似度的计算主要在以下几个方面有着广泛的应用[7]：

(1) 在多文档文摘系统中，语义相似度可以反映出局部主题信息的拟合程度。

(2) 在信息检索中，相似度更多的是反映文本与用户查询在意义上的符合程度。

(3) 在基于实例的机器翻译中，语义相似度主要用于衡量文本中词语的可替换程度。

(4) 在自动应答系统领域，语义相似度的计算主要体现在计算用户问句和领域文本内容的相似度上。

(5) 在文本分类研究中，语义相似度可以反映文本与给定的分类体系中某类别的相关程度。

(6) 在文本聚类的研究中，语义相似度计算是研究的基础，通过相似度计算，把文档集合按照文档间的相似度大小分成更小的文本簇。

专家学者还在不断的进行研究，以期基于本体的语义相似度能够在更多层面为人类提供方便。

5 总结与展望

从目前的研究现状来看，基于本体的语义相似度计算研究已经取得不错的进展，语义相似度的计算模型的不断优化也在很大程度上使得计算质量和计算效率得到提升，从而使得信息检索的效率可以得到提升。但是任何技术都是会随着人类社会的进步而不断优化的，基于本体的语义相似度的计算还有很大的提升空间，未来我们的研究可以从以下几个方向予以深入：

(1)利用本体进行语义相似度计算的前提是要将被比较词语转换成本体中的概念词，因此，准确有效实现被比较词语向本体概念词的映射很重要。

(2)网络自身的分布性使得各个领域，甚至是同一个领域， 都必然使用自己的本体来描述数据，这就带来了本体异构问题。相似性的度量可以在同一本体内进行，也可以在不同本体内进行。因此，应充分利用网络资源，加强跨本体，尤其是异构本体的语义相似度相关研究。

(3)除了本体的结构信息和被比较概念词在本体中的位置信息，应加强基于本体实例的混合式算法研究，充分利用本体库的统计特性，将两类语义相似度计算方法的特性融合起来。

(4)任何一个本体语义相似度算法都不可能解决所有问题，因此，要加强相似度融合技术研究。

(5) 基于本体的语义相似度研究决不是某个领域技术或专家能够解决的问题，因此要加强领域之间的合作。

参 考 文 献

[1] 张云中.基于形式概念分析的领域本体构建方法研究[D].吉林大学,2009.

[2] 王晋,孙涌,王璁玮.基于领域本体的文本相似度算法[J].苏州大学学报(工科版),2011,03:13-17+25.

[3]Jiawei Han,Michelle Kamber,Jian Pei.Data Mining Concepts and echniques Third Edition[M].Beijing: China Machine Press,2012:66-68.

[4]李晓红.基于本体技术的语义检索及其语义相似度分析[J].电子技术与软件工程,2017(01):187.

[5]裴培,丁雪晶.基于本体的语义相似度计算综述[J].合肥学院学报(综合版),2020,37(05):68-74.

[6]李国钊.基于领域本体的语义相似度计算方法解析[J].科技经济导刊,2016(16):160+165.

[7]史俊冰.一种基于《知网》的词语相似度计算方法[J].太原学院学报(自然科学版),2017,35(01):69-72.