《智能信息处理》课程考试

基于本体的语义相似度计算方法研究

陈佳慧

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

2020 年 12 月 5 日

基于本体的语义相似度计算方法研究

陈佳慧 [1)]

[1)]（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　随着科技的发展，数据已经成为人们生活中不可或缺的物品，科技的发展，使得人们所面临待解决的数据信息问题原来越多，想要满足人们对信息数据的实际需求，减轻人们的负担，提供有效的检索信息的方式，已经成为现在科技信息发展的重要课题和趋势，而本体能够准确描述概念含义以及概念之间的内在联系，所以基于本体的语义相似度计算方法，可以有效地解决网络信息检索时间问题，提高搜索信息数据的准确率。本文就基于本体的语义相似度计算方法展开了研究。

关键词　本体；语义相似度；计算方法

中图法分类号　TP36　　DOI 号

Research on semantic similarity calculation method based on Ontology

CHEN Jia Hui[1)]

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

Abstract　With the development of science and technology, data has become an indispensable item in people's life. With the development of science and technology, people are faced with more and more data information problems to be solved. It has become an important topic and trend for the development of science and technology information to meet the actual needs of people for information data, reduce people's burden, and provide an effective way to retrieve information The ontology based semantic similarity calculation method can effectively solve the problem of network information retrieval time and improve the accuracy of search information data. This paper studies the semantic similarity calculation method based on ontology.

Key words　Ontology；semantic similarity；　calculation method

1　相关概念

1.1　语义相似度

相似性最早出现在心理学领域，是人们进行感知体验后进行定性的比较，而非定量的表示。而两个对象之间的相似度或者相关度计算早已经成为数据挖掘和信息提取领域中的基本问题，具体地来说，它已经是文本处理的核心问题。例如，语义相似度已经被应用于词义消歧，信息提取，语音自动摘要等方面。一般来说，语义相关度涵盖语义相似度，语义相似度是指两个概念本身的相关程度，但是着两个概念之间可能不存在相似关系，但是可以通过其他关系相关联而形成一定的关系。语义相似度是语义相关度的一种特例。在语义相似度计算中，本体是一个通用的载体。

两个对象之间的语义相似度取决于它们的共性和差异性。通俗的理解来说就是两个对象之间的共性越多，相似度就越大；两个对象之间的差异性越打，相似度越小。

1.2　本体

本体论，最先起源于哲学领域，有的西放学家认为本体仅仅是理念[1]，又有人认为本体是“自在之物”，而本体在计算机学科领域内，是指一种“形

式化的，对于共享概念体系的明确而又详细的说明”。简单的来说，本体就是一种概念，比如人这个概念集合，它是一种抽象集合用来表达世界上的具体的实际的物体，而在人工智能领域，我们主要将本体论的观念用在知识表达上，即借助本体论中的基本元素，概念及概念之间的关联，作为描述真实世界的知识模型，如下图 1 所示：

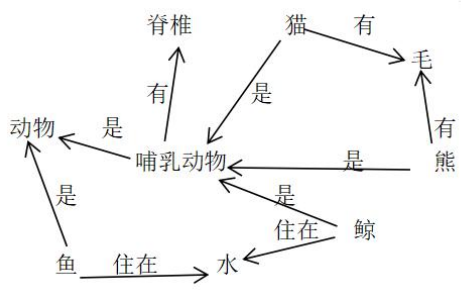

图 1　真实世界的知识模型

本体提供的是一种共享词表，也就是特定领域之中那些存在着的对象类型或其属性和相互关系，或者说，本体就是一种特殊类型的术语集，具有结构化的特点，且更加适合于在计算机系统之中使用。其中本体的一种特例为领域本体，可以称之为具有专业性的本体，通常情况下的领域本体可以对一个学科进行涵盖，包括学科中的概念，概念属性之间的关系特征，而且，领域本体可以更为合理的对知识进行表述，此外，由于领域本体具有显著的邻域特性，使得邻域本体可以更为合理有效的对知识进行表示。

2　本体结构

2.1　基于树状本体结构

所谓基于树状本体结构的相似度算法是指在相似度计算过程中主要基于上下位相连的关系，例如 WordNet 中的“is a”关系。而以树为主体的图结构是指上下位关系作为主要关 系连接概念节点，同时除了上下位关系，还有少量其它类型的 关系编织于概念之间。这些算法在进行语义相关度计算时，不仅考虑上下位关系，还考虑了其它类型的关系。例如，WordNet 中除考虑“is a”关系外，还要考虑“part of”关系等。基于 WordNet 名词概念的以树为主体的图结构示例，如下图 2 所示：

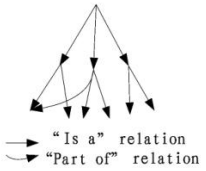

图 2　基于 WordNet 名词概念的以树为主体的图结构示例

2.2　基于树状本体结构中影响相关度的因素

被比较概念词在本体层次树中所处的深度。层次越高，越抽象；层次越低，越具体。高层次的概念词间的语义相似度一般小于低层次概念词间的语义相似度。

被比较概念词在本体层次中所处区域的密度。局部区域密度越大，该区域对节点概念的细化程度就越大。区域内概念间语义相似程度较大。

被比较概念词连通路径上各个边的类型。在本体中，不同的概念关系所表征的语义相似度和相关度是不同的。例如，“同义关系”所表征的语义相似度和相关度应大于上下位关系与整体和部分关系所表征的语义相似度和相关度。

被比较概念词相隔路径长度。由于不同的概念关系所表征的语义相似度和相关度是不同的，另外相同类型的边处于不同深度，不同密度区域所代表的语义距离也有所不同，因此可以得出结论：当一对概念的路径包含在另一对概念之中时，这对概念间的相似度大些。

3　语义相似度计算方法

基于本体的语义相似度计算方法主要分为基于距离的方法、基于信息量的算法、基于属性的方法。

3.1　基于距离的方法

基于语义距离的计算模型，主要是对邻域本体中的两个概念作为主要研究对象，通过对两个节点之间的最短路径来衡量概念之间的语义相似度，该类模型的计算，需要对两个节点的基本情况进行分析，明确的各个向边的差异情况，并按照相关有效的方式，实现对语义相似度的计算。一般情况下，语义相似度的区间为[0,1]，如果出现语义距离为 0 的时候，就可以断定，语义间的相似度为 1，那么就可以认定本体中的两个概念是相同的，相反，如果语义距离特别大，那么证明两个语义的相似度相差特别大。适当的对转化模型进行应用，实现对近似度的计算。

但是这种方法应用大规模的本体时，就会忽视掉多种其他的继承关系，该方法没有考虑其他影响语义相似度的因素，比如公布祖先节点的分布和数量等，因此只选取最短路径的方法而没有考虑本体中的很多概念结构知识。

3.2 基于信息量的算法

基于信息量的语义相似度计算算法的基本原理是：如果两个概念词共享的信息越多，它们之间的语义相似度也就越大；反之，共享的信息越少，相似度也就越少。在本体分类体系树中，每个概念的子节点都是对其祖先节点概念的一次细化和具体化，因此，可以通过被比较概念词的公共父节点概念词所包括的信息内容来衡量它们之间的相似度。

根据信息论可知，概念词所包含的信息内容可通过其在给定的文献集中出现的频率来衡量，频率越高，信息内容就越贫乏；反之，所含的信息内容也就越丰富。在一个树状结构中，概念节点的频率是指次概念实例所出现的频率，任何一个非叶子节点所对应的概念出现的频率是它所对应的所有子孙节点出现的频 率之和，显然根节点实例出现的频率为 1 ，其所含的信息含量最低；随着节点的下移，其所对应的概念就越具体，对应实例出现的频率就越低，所含的信息含量就越高，叶子节点信息含 量最高；尽管树状结构中的理论是如此，但依靠大规模语料库统计的频率并不一定完全符合以上规律。所有基于信息内容的语义相似度计算算法都建立在被比较概念词对共享父节点所含信息内容基础上。

Zhang[2]等人提出一种新的信息量计算模型，采用上位词、下位词、相对深度和最大节点等因素，计算 WordNet 中单父节点或多父节点的节点信息量．该算法补充了多父节点信息量的计算方法，可以敏锐感知上位词、下位词、相对深度和最大节点等因素的不同引起的信息量差别，且能有效解决单父节点和多父节点相似度计算问题．Lord[3]等人接着提出使用共享父节点所包含的信息内容来计算概念词间的语义相似度，也就是直接使用最近公共父节点概念词的信息量来计算被比较概念词对间的相似度。

3.3 基于属性的算法

基于属性的方法针对两个概念对应的属性集进行相似度计算。该方法的计算效果依赖于本体属性集的完备性。两个概念间共有更多的相同属性，则相似度更高，反之概念间不同属性越多，相似度降低。在实际的领域本体中，属性可以有效的对概念进行详细的补充和说明，是概念中不可或缺的重要部分，基于概念属性计算模型，主要是借由不同概念之间的属性概念实现对语义相似度的判断。

3.4 三种算法的比较

综合以上影响语义相似度和相关度的因素和相似度算法 的评价标准，具体分析如下： 基于树或以树为主体图的算法中，基于结构的方法分为基于简单结构和基于复杂结构的算法。基于简单结构的算法 也就是基于路径的方法简单，易于实施，不依赖附加信息。但是不能体现概念对所在位置深度，使得该方法计算出的树中位置深度较深的概念间的相似度偏小。也不能体现概念对所在位置局部密度，使得方法计算出的本体中局部密度较大的概念对的相似度偏小。基于复杂结构的方法充分地利用了隐藏在本体内固有的、丰富的结构信息，提高了语义相关度计算效果，同时相对简单，且不依赖附加信息，普遍具有较好的关 联度。基于内容的方法相对比较客观，能综合反映概念在句法、 语义、语用等方面的相似性和差异，但也存在一些问题：比较依赖于训练所用的语料库，受数据稀疏和数据噪声的干扰较大，有时会出现明显的错误。另外，当建立一个新的应用时，尤其是应用到某些领域本体时，针对领域本体的语料库不全或者尚未建立，使得此类算法很难实施。

基于属性的方法因各个方法差异，优缺点也各有不同，其共同局限性是：此类办法必须依赖于概念具备完备的属性集，对于不存在针对概念完备属性集的情况，此类办法则无法实施。

4 总结

本文主要总结了几种代表性的语义相似度计算方法的发展脉络，分析了内容方法，并且比较了这几种计算方法的优缺点和实验效果，随着知识图谱的发展，相信未来基于本体的语义相似度计算研究以后会面向更多的领域，比如在医学领域，使用本体的语义相似度进行基因之间的相似度检验，可以提高对比效率，加强精准度，也会是一个值得期待的事情。

参 考 文 献

[1] Kitamura Y.An Ontology-based Human Friendly Message Generation in a Multiagent Human Media System for Oil Refinery Plant Operation[C] InternationalConference on Systems．Tokyo: IEEE，1999: 648-653．

[2] ZHANG X,SUN S.An information content-based approach for measuring concept semantic similarity in WordNet [J/OL].Wireless PersonalCommunications,2018:1-16[2018-03-12].https://doi.org/10.1007/s11277-018-5249-7.

[3] Lord P W，Stevens R D,Brass A ,et al.Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship Between Swquence and Annotation[J].Bioinformatics,2003,19(10):1275-1283.