《智能信息处理》课程考试

基于本体的语义相似度分析

靳少宁

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分析

靳少宁 [1)]

[1)]（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　在计算机领域，本体可以在语义层次上描述知识，可以看成描述某个学科领域知识的一个通用概念模型。本体是从客观世界抽象出来的一个概念模型，这个模型包含了某个学科领域内的基本术语和术语之间的关系（或者称为概念和概念之间的关系）。本体不等同于个体，它是团体的共识，是相应领域内公认的概念集合。而数以百万计的文本数据正在渗透到我们的日常生活中。文本对象之间的语义相似度分析是文本挖掘中的基本问题之一，包括文档分类、聚类、推荐、查询扩展、信息检索、相关性反馈、词义消歧等。虽然常识和领域知识的结合可以让人快速判断两个物体是否相似，但计算机对人类的思维却知之甚少。本体等知识资源可以很好地捕获文本对象的语义，从而实现领域知识和上下文信息的数值表示[1]。

关键词　本体；语义推理；语义相似度；领域本体

Semantic similarity analysis based on ontology

Shaoning Jin[1)]

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

Abstract　In computer domain,ontology can describe knowledge at semantic level,and can be regarded as a general conceptual model to describe domain knowledge of a subject.Ontology is a conceptual model abstracted from the objective world, which includes the relationship between basic terms and terms in a subject area (or the relationship between concepts and concepts) . Ontology is not equal to the individual, it is the consensus of the group, and it is the accepted concept set in the corresponding field. And millions of text data are permeating our daily lives. Semantic similarity analysis between text objects is one of the basic problems in text mining, including document classification, clustering, recommendation, query expansion, information retrieval, relevance feedback, word sense disambiguation and so on. While the combination of common sense and domain knowledge makes it possible to quickly determine whether two objects are similar, computers know very little about the human mind. Ontology and other knowledge resources can capture the semantics of text objects well, and thus realize the numerical representation of domain knowledge and context information.

Key words　ontology; semantic reasoning; semantic similarity; domain ontology

0　引言

随着网络上信息的爆炸式增长，人们对于 Web 信息语义层次上的需求越来越迫切。基于本体的语义网为信息的语义理解及共享提供了重要贡献。本体在信息检索、人工智能、语义网等领域越来越得到重视。基于

本体语义信息检索利用本体知识模型中的概念来表达用户的查询需求，从而需要分析概念之间的相似度来判断概念与用户查询之间的差异程度。因此概念之间的语义相似度计算是语义信息检索的重要部分，决定了检索结果的准确与效率，如何提高语义相似度计算的精度成为语义信息检索的关键。

语义相似度算法的研究近年来在国内外得到广泛的研究与发展，这些研究分为两类：一是基于语料库的统计法，虽然基于大规模语料库的计算方法比较客观，但是这种方法比较依赖于训练所用的语料库，计算量大，计算方法复杂，另外，受数据稀疏和实际噪声的干扰较大。二是基于语义距离的计算法。该方法利用近年来出现的一些大规模、可计算的本体来进行研究，以本体层次结构模型中的集合距离来量化概念之间的距离，简单、直观、便于实现。但有些研究仅仅考虑了层次网络中的概念点之间的距离差异，并没有考虑本体图形结构中概念节点的共同祖先节点之间的关系，这些研究都没有考虑影响语义距离的其他因素[2]。

1 本体简介

1.1 概念定义

本体不等同于个体，它是团体的共识，是相应领域内公认的概念集合。数据挖掘被认为是从（大规模）数据中发现有趣的（非琐碎的、以前未知的、有洞察力的和潜在有用的）信息或模式，以及描述性的、可理解的和预测性的模型的自动化过程。尽管结构化字段中的大量数据也可能包含数字、日期和事实，但非结构化信息通常是文本（文章、网站文本、博客文章等）。本体可以在语义层次上描述知识，可以看成描述某个学科领域知识的一个通用概念模型。人们用许多不同的方式交流，通过说和听，做手势，或各种形式的文本。非结构化文本信息的存在使得有效地执行知识管理活动变得更加困难。计算机能够理解普通语言并与人类进行对话的想法是几个世纪以来的一个梦想。自 21 世纪以来，这一愿景开始显得更加可信。许多网站(如谷歌翻译)现在提供自动翻译;移动应用可以理解语音指令;搜索引擎可以自动完成或纠正您的查询，并找到与查询条件紧密匹配的相关结果。然而，对于自然语言，我们离成熟的机器理解还很远。例如，只有在遇到短语或短句时，自动翻译才能很好地执行;语音指令对背景噪音很敏感;在搜索引擎中，由于缺乏对用户意图的真正理解，仍然存在一些无关的搜索结果。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。

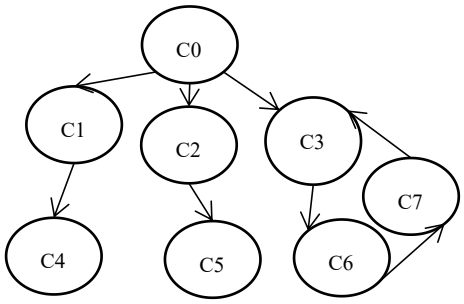

图 1 本体概念模型的 DCG 图

1.2 本体组成

Perez等人将本体按分类法来组织并归纳出本体的5个基本构成元素,分别为:许多专业领域的认可。概念、关系、函数、公理和实例。概念反应出实物的基本属性,包括实物本身所具有的功能、事物之间的关联及时间的行为推理关系等;关系是对概念之间的相互作用的一种界定和表达;函数是事物之间关系的一种特例,反应出事物关系的通用性;公理是判断实物之间关系的概念、鉴定规则是否正确的依据;实例表示概念的具体对象和表示。

2 语义网的概念及层次结构

2.1 概念

语义相似性是指两个概念相似(或不相似)的程度。虽然人类对概念之间的相似性没有正式的定义，但他们可以很容易地确定

两个概念在某些方面是否相似。例如，大多数人会同意“pencil”和“pen”比“car”和“phone”更相似。在人类对相似性和亲缘关系的感知背后，应该有一个深刻的心理学解释。知识必须以一种有效和经济的方式储存在人脑中。为了确定“金丝雀会飞”这句话的真实性，我们需要考虑人类记忆组织的两种可能方式。首先，人们可能储存了每种鸟都能飞的事实。然后他们可以直接检索这个事实来决定这个句子是正确的。另一种组织方式是只存储鸟会飞的概括，并从预先存储的金丝雀是鸟而鸟会飞的信息中推断出“金丝雀会飞”。根据心理学研究[Collins 和 Quillian, 1969, Olivera，， Quillian, 1969]，人们倾向于以后者的方式存储和访问语义相似的知识，因为它在存储空间上更经济，但在需要进行推断时需要更长的检索时间。虽然人类如何获取知识的确切本质仍然是一个有趣的问题，在这篇论文中，我们从一个更实际的观点来考虑语义相似性。我们试图观察人类在日常生活中是如何使用语义相似性概念的。人类常识的一部分可能包括知道哪些概念是相似的(或不相似的)。例如，一个小孩可以很容易地看出“apple”和“orange”比“apple”和“desk”更相似。考虑一下下面的句子，“the child is eating an apple”和“the child is eating an orange”比“the child is eating a desk”更有说服力。因为“橘子”和“苹果”放在“吃”的语境中比放在“桌子”更合理。食物制造的常识和领域知识的结合。很明显，“橘子”和“苹果”都是食物，所以可以吃，而“桌子”不是。从这个角度来看，我们可以说“orange”和“apple”比“desk”更相似。领域知识和常识的结合也表明，“橙色”指的是一种水果，与众所周知的颜色没有关联。这些是人类可以快速解决的问题，不需要很多有意识的思考，基于现实世界的知识和常识的结合。语义相似性度量的研究一直是信息检索和自然语言处理的重要组成部分。实体之间的语义相似性会随着时间和领域的不同而变化。

2.2 语义网的层次结构

语义网的层次结构如图 2 所示。共分为六层：Unicode 和 URI、XML、RDF 和 RDF Schema、本体、逻辑、证明和信任。

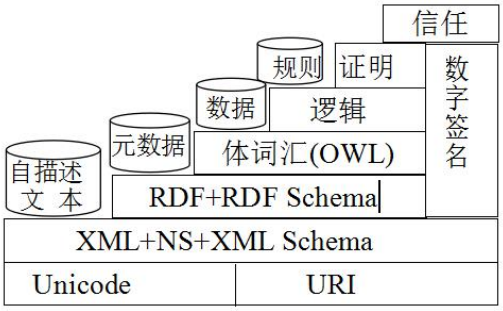

图 2 语义网的层次模型

3 语义相似度分析过程

在 NLP 领域，语义相似度的计算一直是个难题：搜索场景下 query 和 Doc 的语义相似度、feeds 场景下 Doc 和 Doc 的语义相似度、机器翻译场景下 A 句子和 B 句子的语义相似度等等。关于概念的知识是巨大的，而在传统的字典中查找可能是冗长和费时的，它必须以一种高效和经济的方式存储在人脑中。根据心理学研究，人类倾向于在语义上相似地存储和获取知识。关于概念的知识是通过对一般概念的访问“即时”计算出来的。例如，我们知道“金丝雀会飞”是因为“鸟会飞”和“金丝雀是鸟的一种”。这种联想过程可能表明了动物→鸟→金丝雀的层级结构。一般来说，知识被储存在尽可能高的位置，并被较低（更具体）的概念继承，而不是随机储存。基于本体的语义相似度度量可以大致分为边缘计数、基于信息内容、基于特征、基于注释和混合度量。其中，边数度量试图基于语义链接的数量和给定本

体中两个概念之间的最小路径长度来度量相似性；基于特征的方法是根据共同特征和非共同特征数量的加权和来估计相似度；基于注释的方法利用本体论提供的定义，以量化两个概念的注释与其语义邻居之间的重叠。

语义相似度是以语义距离为衡量的概念，换言之如果两个语义之间的距离是无穷大的，则这两个语义之间的相似度是非常低的，而如果两个语义之间的距离无限接近于0，则可以视为两个语义是高度相似的，所以要对本体技术语义相似度进行分析，要有意识的建立以距离为基础的语义相似度计算模型，在此模型构建的过程中，又要结合以下因素进行。首先是语义重合度，即本体内部概念中上位关系概念相同概念的数量；其次，是语义深度，即本体内部概念所具有的层次深度；最后，语义距离，即本体中两个节点连接通路中最短路径要通过的边数。在语义相似度的计算模型确定之后，在信息检索的过程中，要利用语义相似度进行信息检索，可以在概念初始化后，对相似度阈值进行确定，然后利用相似度计算模型进行语义相似度计算，并按照序列输出，为用户提供检索的结果，这在概率方面可以提升用户获得预期检索结果的概率。

4 总结

在本论文中，我们发展了一系列的技术来衡量物体在多个领域中的语义相似性。通过利用已经建立的结构化知识，我们从现有的词汇资源中挖掘领域知识，并将其整合到不同领域的具体应用中。

混合单词嵌入的特性对于执行 nlp 任务非常有用。例如，除了在文本分类中表示文档向量之外。这也有利于文件的总结。通过查找文档的最重要的嵌入点，可以从单词嵌入中提取关键字，这些关键字可以被视为文档的标签。大多数 nlp 系统失败的主题之一是短文档，如 tweets，microblog 等，这些文档中包含的信息有限，阻碍了大多数 nlp 模型的培训。将领域知识与词嵌入中的上下文信息相结合，有望缓解语义和多义词严重降低语言系统性能的局面。

参 考 文 献

[1] Xuebo Song,Ontology-based Domain-Specific Semantic Similarity Analysis and Applications[J]. 2020,04

[2] Lemaire B, Denhiere G. Effect of High-Order Co-occurrences on Word Semantic Similarities[C]. IEEE Workshop on Mobile Computing System and Application, 2012: 85-90

[3] 唐小波, 金钟鸣. 基于本体与规则的语义推理研究[J]. 情报学报, 2011, 30(7): 695~703

[4] 宋炜, 张铭. 语义网简明教程[M]. 北京：高等教育出版社,2020