基于本体的语义相似度计算综述

杨帆

作业	分数
得分	

2020 年 12 月 10 日

基于本体的语义相似度计算综述

杨帆

（大连海事大学 软件工程 辽宁省大连市 中国 116026）

摘 要 大数据时代,如何从海量的信息中精确的检索到满足需求的信息成为当前信息检索领域的首要任务。语义信息检索是语义网技术和信息检索结合紧密的研究领域,旨在提供一种高效率、智能化的信息检索系统。由于语义异构问题的存在，传统的语义查询扩展方式已无法满足用户深层次需求。本体论的提出为语义信息检索提供了一种全新的途径，本体能够准确描述概念含义以及概念之间的内在联系，对于帮助人们快速准确获取与需求相关的内容具有非常实际的意义。目前,基于 Web 语义的信息检索广泛应用于自然语言处理、知识提取以及智能推理等领域。基于本体的语义相似度计算方法也积累了丰富的成果。本文基于距离、基于内容、基于属性和混合式方法四个方面进行阐述,梳理总结国内外算法进展。

关键词 本体；语义相似度；语义相似度计算

中图法分类号 TP311.20 **DOI 号** 10.3969/j.issn.1001-3695.2014.01.030

Ontology Based Semantic Similarity Review

Yang Fan

Software engineering, Dalian Maritime University, City Dalian

Abstract In the era of big data, how to accurately retrieve the required information from the massive information has become the primary task in the field of information retrieval. Semantic information retrieval is a research field combining semantic Web technology and information retrieval, which aims to provide an efficient and intelligent information retrieval system. Due to the problem of semantic heterogeneity, the traditional semantic query extension can no longer meet the deep needs of users. Ontology provides a new way for semantic information retrieval. Ontology can accurately describe the meaning of concepts and the internal relations between concepts, and has very practical significance for helping people quickly and accurately obtain the content related to requirements. At present, information retrieval based on Web semantics is widely used in the fields of natural language processing, knowledge extraction and intelligent reasoning. Ontology-based semantic similarity calculation method has also accumulated a wealth of achievements. Based on distance, content-based, attribute-based and hybrid methods, this paper summarizes the progress of algorithms at home and abroad.

Key words ontology; semantic similarity; semantic similarity measuring

0 引言

随着 Internet 的飞速发展、网络资源的爆炸式增长，人们对于 Web 信息的获取提出了更高的要求。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。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。

1 相关概念

1.1 本体

本体的概念最早源自于古希腊哲学家亚里士多德对于客观事物存在本源的研究，近年来作为信息抽象和知识表示的主要手段被计算机领域所采用。Studer 等对本体做了深入的研究，定义本体“共享概念模型的明确的形式化规范说明”被广泛接受，本体的形式化定义为: O ={C,I,R,F,A}，其中，C 代表概念集合，即抽取出来用来描述事物对象的集合；I 代表概念的实例集合；R 表示定义在概念集合上的关系集合；F 为定义在概念集合上的函数集合；A 表示公理集合，用于约束概念、关系、函数的一阶逻辑谓词集合。

一个本体主要由概念的集合以及概念间的语义关系的集合组成，可以用层次树状结构表示，如图 1 所示。其中，树中节点表示概念节点，边表示概念之间的关系。

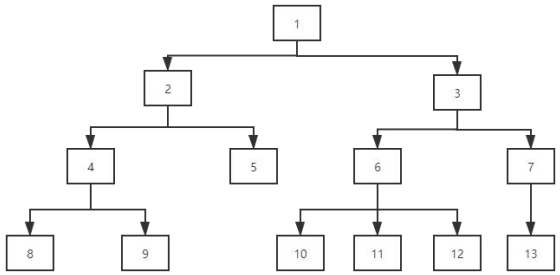

图 1 本体结构图

1.2 语义相似度

两个对象之间的相似度或相关度计算早已成为数据挖掘和信息提取领域中的基本问题，具体地说，它是文本处理的核心问题。例如，语义相似度或相关度算法已经被应用于词义消歧、音频识别错误的检测、信息提取、语音自动摘要、人的姓名解析、文本相似度计算、文本分类和聚类等。

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。借助数据挖掘中二维数据特征向量的表达形式，建立一个对象－属性结构数据矩阵(n 个对象被 p 个属性刻画，其中 xij 是对象 xi 的第 j 个属性的值，如公式(1) 所示) ，可以从中得到任意两个数据对象之间的相似性矩阵(公式(2)) 。

$$\begin{pmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{pmatrix} \tag{2}$$

信息检索系统中概念语义相似度指的是两个概念之间词义的符合程度。当两个概念之间具有某些相同属性时，称它们是相似的，记概念 X、Y 之间的相似度为 Sim(X，Y) ，相似度的计算满足以下条件:

$$\mathrm{Sim}(X,Y)=\begin{cases} 0 & \text{两个概念之间没有相同的属性} \\ P & P\in(0,1) \\ 1 & \text{两个概念之间所有属性都相同} \end{cases}$$

计算概念之间的语义相似度应遵循以下规则：相似度是一个位于 0 和 1 的闭区间内的数值;算法应简洁、准确;要将影响相似度的因素考虑全面。

2 基于本体的语义相似度计算研究进展

国内外学者在语义相似度领域已形成了较为成熟划分体系和研究成果， 本文参 Hlianoutakis、Batet[1]等的研究，结合陈二静[2]等文本相似度综述划分体系，将基于本体的语义相似度计算为 4 种: 基于距离的语义相似度计算、基于内容的语义相似度计算、基于属性语义计算相似度和混合方法。

2.1 基于距离的语义相似度计算

其基本思想是通过两个概念词在本体树状分类体系中的路径长度来量化它们之间的语义距离。语义相似度和语义距离之间存在着密切的关系：两个词语的语义距离越大，其相似越低；反之，两个词语的语义距离越小，其相似度越大。

基于距离的语义相似度计算方法是通过测量两个概念节点在本体层次树中的位置，以路径长度的方式体现差异。路径越短，相似度越大；路径越长，相似度越小。后期称之为“利用最短路径(Shortest Path)计算文本相似度的模型”。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[3]。计算公式如下(3) :

$$\mathrm{sim(i,j)}=\frac{1}{\mathrm{dis(i,j)}} \tag{3}$$

该算法的计算复杂度相对较小，缺陷是大部分均未考虑边的类型影响因素，算法成立的前提是假设“本体分类体系中所有边的距离权重相等”。此外，结合前文所提，边的重要性还会受到位置信息、所表征的关联强度等因素的影响。Hao 等提出从概念距离和概念深度两方面计算相似度，但缺少对同一层次的最小公共父节点相似度结果的对比计算，算法准确性有待进一步探索。

Hirst-Stronge 认为，如果概念对间存在一条 较短的路径，且在遍历过程中改变路径方向的次数较少，那么这两个概念词语义相关。概念间路径涉 及多种类型的关系。Hirst-Stronge 法开辟了一个新的视角，但由于其在很大程度上取决于“方向”问题而不是概念关系，因此表现似乎不是很好。

Yang 和 Power 法提出基于本体结构中有向边关系的语义相关度计算方法，有向边关系包括: “is a”“equivalence”和“part of”关系。设计了 BDLS 和 UBFS 两种搜索算法和两种语义相关度计算方法，由于该算法的实现需要涉及到 7 个可以自由参数，因而不稳定。

2.2 基于内容的语义相似度计算

基于信息内容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，相似度也就越小。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分和具体化，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(4) 如下:

$$Info(C_i)=\sum_{i=1}^{m}(C_i)\log_2(p(C_i)) \tag{4}$$

在本体中计算任意两个节点之间的相似度公式(5) 如下:

$$Sim(c_i,c_j)=\frac{2IC(Lunncan(c_i,c_j))}{IC(c_i)+IC(c_J)} \tag{5}$$

Lord 等人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。他们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。Resnik 使用共享父节点信息内容来计算概念词间的语义相似度，该算法与 Lord 等人的算法不同之处在于它并不是基于最近公共父节点的信息内容，而是基于公共父节点概念词中信息量最大的父节点的信息内容。上述两种算法都只考虑了被比较概念词对的共享信息内容，Lin 认为还应该考虑被比较概念词对各自所包含的信息内容。当被比基于属性的语义相似度计算较概念词集属于同一个本体 时，使用 Lin 法可以获得比 Resnik 法更好的基于相似度的排列结果。Jiang 和 Conrath 法与上述 3 种算法不同的是：直接通过对语义距离的计算来表征被比较概念词间的相似度。和 Lin 法一样，该算法也同时使用了共享父节点和被比较概念词所包含的信息内容。荣河江[4]等利用基因本体携带的语义关系、基因产物属性，改进了基于信息量的计算方法，将信息量均值纳入考虑，在 Li 方法的基础上做了拓展，实验结果进一步提升。

2.3 基于属性的语义相似度计算

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。Tversky 算法从属性角度研究两个概念之间的语义相似度，计算模型如下公式(6) :

$$Sim(c_1,c_2)=Xf(c_1\cap c_2)-Yf(c_1-c_2)-Zf(c_2-c_1) \tag{6}$$

该模型从属性的角度出发，综合比较了两个概念之间的共同属性 (f(c1∩ c2) 的返回值) 和不同属性((c1 －c2) 和 (c2 －c1)) 。利用相同的属性增加概念间的相似度，不同的属性减少概念间相似度进行计算。该算法的特点在于属性的选择，但缺乏对数据类型属性的区分，没有考虑被比较概念词的位置信息、以及祖先节点和所包含信息内容。

2.4 混合方法

混合式语义相似度和相关度计算算法实际上

是对上述 3 种算法的综合考虑，即同时考虑了概念词的位置信息、边的类型、概念词的属性信息等。代表算法有 Li 等人法，该算法同时考虑了被比较的词语队间的最短路径 L 和最近公共父节点在分类体系树中所处的深度，以及词语所处位置的局部密度信息，是个非线性函数。Marco 等人提出的基于图模型的语义相似度和相关度计算方法（简称 SSA）是基于以树为主体的本体结构和释词方法的混合，也取得了不错的关联度和值连续性。

另外，近年来国内外出现的大量相关研究成果大都属于基于以上计算方法的改进方法、混合方法和具体应用。例如 Shibin 等[5]人提出了 OHICC 基于局部密度、信息量和概念深度的混合算法，只不过算法中的局部密度扩展成由当前节点与其儿子节点连接边的个数决定。

3 算法分析比较

基于距离的方法直观、易于理解、具有较低的时间复杂度，对于小规模的本体结构具有一定的实用价值。当面对复杂结构大型本体时，因其较少对本体特征的关注，导致忽略结构中存在的多种继承性以及其他相似度的影响因素(公共父节点的分布与数量等) ，算法效果不是很突出。此外，该方法较多的依赖于本体结构的完备性和覆盖力，适用于 WordNet 这种大型专业的通用本体库。[5]

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：例如，Resnik 单纯依据信息内容的方法不能体现概念间的距离，而且忽略了密度深度等结构信息，即使有的办法考虑了概念间的部分结构因素，得出较粗略的结果，例如一个节点与同一子树中任意节点的相似度值都相同。另外这类算法还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。

混合方法不能减少对附加信息的依赖，没有从根本上克服它基于方法的局限性。

4 结束语

基于本体的语义相似度计算方法不依赖于语料库，其计算结果与领域专家经验值较为一致，在轻量数据集中表现稳定，是一个较好的选择。但是随着大数据的兴起，研究者应将基于大规模语料库的计算方法和基于本体的研究者应将基于大规模语料库的计算方法和基于本体的计算方法融合起来，根据数据集的规模选择性使用上述方法。同时可以借鉴关系抽取中的弱监督学习方法，加强基于释词的研究，将本体中的概念与可靠的网络知识库(如 Wikipedia、FreeBase 等) 中的词条进行匹配，充分挖掘知识库中的语义信息。

任何一种相似度计算算法都不能解决所有问题，算法也没有绝对的好与坏，可能因为应用场合不同而表现各异。Wikipedia 的覆盖范围更加广泛，知识描述更加全面，信息内容更新速度更加迅速。另外，以 WordNet 为基础的算法得到较充分的研究，结果已经取得了较好的关联度，而以 Wikipedia 为基础的算法还比较少，关联度还有较大的提升空间，所以不失为未来语义相似度和相关度计算研究的趋势。

参 考 文 献

[1] Batet M , David Sánchez, Valls A . An ontology-based measure to compute semantic similarity in biomedicine[J]. Journal of Biomedical Informatics, 2011, 44(1):118-125. 陈二静, 姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017, 1(6):1-11.

[2] 陈二静, 姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017, 1(6):1-11.

[3] Rada R , Mili H . Development and application of a metric on semantic nets[J]. IEEE Transaction on System Man & Cybernetics, 1989, 19(1):17-30.

[4] 荣河江, 王亚东. 基于基因本体的相似度计算方法[J]. 智能计算机与应用, 2019, 009(001):P.108-113.

[5] Yan S , Luan L . A Concept Similarity Method in Structural and Semantic Levels[C]// Second International Symposium on Information Science & Engineering. IEEE, 2010.

[6] 刘宏哲, 须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学, 2012(02):14-19.