《智能信息处理》课程考试

基于本体的文本相似度计算分析

韩炜

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

2021 年 12 月 15 日

基于本体的文本相似度计算分析

韩炜 [1)]

[1)]（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　本文主要讨论了对于文本相似度计算的三种理论。这三种理论包括：词语相似度分析、句子相似度分析、文本相似度分析。然后对目前国内外关于文本信息相似度处理的方法进行查询、归纳和总结。最后提出四点关于文本相似度处理的研究方向。

关键词　本体；相似度分析；文本相似度；本体异构；距离相似度；内容相似度；属性相似度；混合式相似度；

中图法分类号　TP36　　DOI 号

Base on the Ontology of Text Similarity Calculation Analysis

HAN Wei[1)]

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

Abstract　This paper mainly discusses three kinds of theories about text similarity calculation. The three theories include: word similarity analysis, sentence similarity analysis, text similarity analysis. Then we have inquired, induced and summarized the some methods of text information similarity processing which are come from domestic and abroad . Finally, four research directions on text similarity processing are proposed.

Key words　Ontology; Similarity Analysis; Text Similarity; Ontology Isomerism; Edge Counting Measures; Information Content Measures; Feature-based Measures; Hybrid Measures;

1　问题描述

1.1　本体定义

本体概念最开始在哲学领域开始广泛使用起来，在哲学领域本体的含义为“根本的、实质的”。当本体一字引入到计算机界、人工智能领域时候，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，很多人开始研究本体并给出不同的含义，其中，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”。[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。通俗的讲，本体就是：某个领域的概念的集合、概念和概念关系的集合、以上两中集合的集合。这样的本体在共享范围内有着明确的唯一的定义，达成一种共识，便于人和及其进行交流。本体的形式化表示：

$$O = (V, C, P, H, R) \qquad (1)$$

其中，V表示词汇集合，C表示各个词汇之间的关系与约束，P表示各个属性之间的关系和性质，H表示对同意词集和单词的实例声明，R表示对实例的具体描述。

另外，词汇之间的关系可分为三类：上下位关系、同义关系、除去上述两种的其他关系。[2]

1.2　本体异构

导致本体异构出现的原因主要有三个方面：

a）本体的创建具有主观性，不同的创建者都有可以自己掌握的领域知识来构建和维护本体；

b）本体的创建具有分布性，同领域的本体可以由不同的用户创建，创建过程不具有协同性；

c）本体的创建具有自治性，不同的创建者都

可以根据自己特定的应用需求来构建和维护本体。

本体异构主要包括模式层异构和数据层异构（也称实例异构），模式层异构又包括语言性异构和结构性异构。图 1 给出了两个文献领域的异构本体局部，图中：

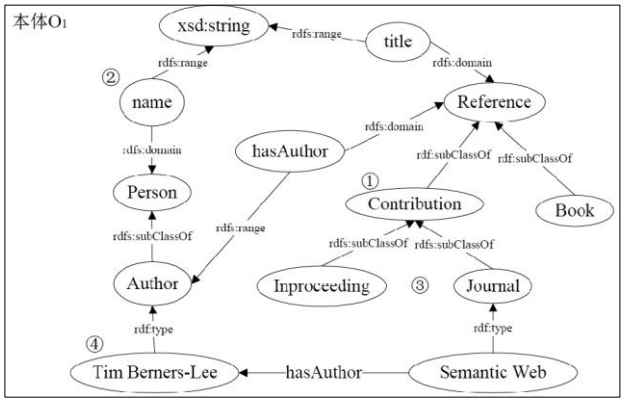
（1）本体 O_1 中概念“Contribution”与本体 O_2 中概念“paper”具有不同的名称，却有相同的含义，都表示“论文”类（见图 1 中①）；

（2）本体 O_1 中属性“name”与本体 O_2 中属性“name”具有相同的名称，但表示完全不同的含义，分别表示“人名”和“参考文献的名称”（见图 1 中②）；

（3）本体 O_1 中概念“Contribution” 划分为子类“Inproceeding”和“Journal”，而本体 O_2 中对应概念“paper”划分为子类“unpublished”和“published”（见图 1 中③）；

（4）本体 O_1 中实例“Tim Berners-Lee”与本体 O_2 中实例“Tim B Lee”表示同一个人（见图 1 中④）。

其中，①和②为语言性异构，③为结构性异构；④为实例异构。[5]



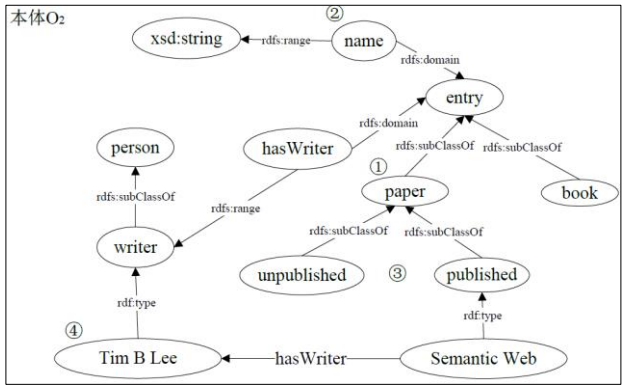


图 1 两个异构本体

对于上述问题的出现，就需要我们进行词语的相似度分析。很明显，但当我们输入某个查询条件的时候总是希望可以找到比较准确的结果，一般说来，我们用查准率和查全率来表示。所以，现在研究的主要目标和衡量标准就是：如果提高查准率和提升查全率。

1.3 应用场景

当然，词语相似度分子，不仅仅限制在刚才所述例子中。

词语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。

自然语言处理、智能检索、文本聚类、文本分类、数据挖掘、信息提取、自动应答、词义排歧和机器翻译等领域有更多的应用。

根据文献查询、总结和归纳。对词语相似度计算的应用背景、研究成果进行了汇总，词语相似度计算的应用主要有以下几点：

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

(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) 在多文档文摘系统中，相似度可以反映出局部主题信息的拟合程度。

(4) 在自动应答系统领域，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，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) 相似度计算是文本聚类的基础，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。[3]

2 文本相似度分析理论

对于文本相似度问题，可以分为一下三类：词语相似度、句子相似度、文本相似度。

2.1 词语相似度

词语相似度是指词语之间自身语义的相似程度，在领域本体中词语表示为概念，词语相似度又可以归结为概念相似度的计算。所有的概念根据上下位关系构成了一个树状的层次体系，可以通过计算语义距离的方法来计算词语相似度。[1]

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，其相似度越低；反之，两个词语的语义距离越小，其相似度越大[3]。

对于两个词语 w_1和 w_2，记$Sim(w_1,w_2)$为其相似度，$Dis(w_1,w_2)$为词语语义距离，则$Sim(w_1,w_2)$和 $Dis(w_1,w_2)$存在下列对应关系：$Dis(w_1,w_2)$ 和 $Sim(w_1,w_2)$ 成反向关系，即$Dis(w_1,w_2)$越大，则 $Sim(w_1,w_2)$越小:

①当 $Dis(w_1,w_2)$为 0 时 ，$Sim(w_1,w_2)$为 1，表示两个词语完全相似;

②当 $Dis(w_1,w_2)$为无穷大时 ，$Sim(w_1,w_2)$为 0，表示两个词语完全不相似或不相关。

两者之间的对应关系可通过下列公式来揭示：

$$\mathrm{Sim}(w_1,w_2)=\frac{\alpha}{Dis(w_1,w_2)+\alpha} \qquad (2)$$

其中,α为调节因子。

2.2 句子相似度

在计算句子相似度方面，主要关注文本中的实词，因为实词具有实在的意义。所以在本文中，使用实词之间词语相似度来统筹计算句子相似度值，流程如下。

设两个句子为s_1,s_2，而句子是由多个词语组成，有：

$$s_1=\{w_{11},w_{12},w_{13},\dots,w_{1m}\} \qquad (3)$$

$$s_2=\{w_{21},w_{22},w_{23},\dots,w_{2n}\} \qquad (4)$$

对两句子建立一个语义相关矩阵$R[m,n]$，其中 m 为s_1的长度值，n 为s_2的长度值，$R[i,j]$表示s_1中位置 i 的词和s_2中位置 j 的词最相似语义的语义相关度。因此 $R[i,j]$也是从 i 到 j 的弧上的权重。此后使用匈牙利算法来求二分图的最大加权匹配即可得两个句子的相似度数值。[1]

2.3 文本相似度

在得到词语及句子相似度算法的基础上，就可以进行文本相似度的计算，其计算过程如下。

设两文本为t_1、t_2，而文本是由多个句子组成，有：

$$t_1=\{s_{11},s_{12},s_{13},\dots,s_{1m}\} \qquad (5)$$

$$t_2=\{s_{21},s_{22},s_{23},\dots,s_{2n}\} \qquad (6)$$

则可得两个文本的句子相似度特征矩阵：

$$\boldsymbol{T}_{12}=\boldsymbol{T}_1\times\boldsymbol{T}_2^T=\begin{bmatrix} s_{11}s_{21} & \cdots & s_{1m}s_{21} \\ \vdots & & \vdots \\ s_{11}s_{2n} & \cdots & s_{1m}s_{2n} \end{bmatrix} \qquad (7)$$

其中，$s_{1i}s_{2j}=simS(s_{1i},s_{2j})$

计算时首先遍历相似度特征矩阵，取出相似度最大的句子组合，再将其所属行列从相似度矩阵中删除，继续选取余下矩阵中相似度最大组合，直到矩阵中元素为零，此时可得到句子最大组合序列[1]。

$$MAXsim=\{simS_{\max1},simS_{\max2},\dots,simS_{\max}\} \quad (8)$$

其中，$simS_{\max}$是两文本中句子的最大相似度组合，则可得文本相似度计算公式为：

$$sim(t_1,t_2)=\frac{1}{k}\sum_{i=1}^{k}simS_{max} \qquad (9)$$

3 主要算法

3.1 国外

学者们一般将基于本体的语义相似度计算方法划分为 4 类：基于距离的语义相似度计算(Edge Counting Measures)、基于内容的语义相似度计算(Information Content Measures)、基于属性的语义相似度计算(Feature-based Measures)和混合式语义相似度计算(Hybrid Measures)。在不作具体说明情况下，本文介绍的 4 类算法都是建立在“IS-A”关系树状分类体系基础上的。[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的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量化它们之间的语义距离。代表算法有：Shortest Path 法、Weighted Links 法、Wuand Palmer 法、Lietal 法、Leacockand Chodorow 法等。

(2) 基于信息内容的语义相似度计算

基于信息内容的语义相似度计算方法的基本原理是：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。

(3) 基于属性的语义相似度计算

事物由其属性特征反映其本身，人们用以辨识或区分该事物的标志就是属性特征。事物之间的关联程度与其所具有的公共属性数相关。基于属性的语义相似度计算正是基于此提出的。对于两个被比较概念词而言，公共属性项越多，相似度越大。

(4) 混合式语义相似度计算

混合式语义相似度计算方法实际上是对上述三种方法的综合考虑，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 Rodriguez 等人和 Knappe 提出的算法模型。其中 Rodriguez 等人提出的算法模型既可以用来计算单个本体中概念词间的相似度，也可以用来计算多个本体中概念词间的相似度。

3.2 国内

基于 CNKI 和万方数据库中检索到的文献来看，国内在这方面的研究相对较晚，且目前主要成果集中在三个方面：[3]

(1)在对国外上述经典模型进行介绍的基础上进行理论比较或相关改进，如黄果、周竹荣等人提出了一种以 Leacock 模型为基础,将概念的信息内容和概念的属性作为决策因子的基于领域本体的语义相似度计算模型。

(2) WordNet 是一个联机英语词汇检索系统，由 Prince-ton 大学研制。它作为语言学本体库同时又是一部语义词典，在自然语言处理研究方面应用很广。它采用语义网络作为其词汇本体的基本表示形式。在 WordNet 中，网络节点由字形(Wordform)标识，分为名词、动词、形容词、副词和功能词等 5 种。节点之间的关系分为同义关系(Synonymy)、反义关系(Antonymy) 、继承关系(Hypony-my) 、部分/整体关系(Meronymy)、形态关系(Morpholog-icalrelation)等。WordNet 提供了很好的概念层次结构。

(3) 知网是一个以汉语和英语词语所代表的概念为描述对象、以揭示概念与概念之间以及概念所具有属性之间的关系为基本内容的常识库和知识库。其中包含丰富的词汇语义知识和本体知识，这些关系都隐含在知网的知识词典和义原的特征文件中。知网中有以下两个主要的概念：

a） 义项。它是对词汇语义的一种描述，每一个词可以表达为几个义项。义项是用一种知识表示语言来描述的，这种知识表示语言所用的词汇叫做义原。

b）义原。它是用于描述一个概念的最小意义单位，从所有词汇中提炼出的可以用来描述其他词汇的不可再分的基本元素。

与一般的语义词典(如同义词、词林或 WordNet)不同，知网并不是简单地将所有的概念归结到一个树状的概念层次体系中，而是试图用一系列的义原来对每一个概念进行描述。

知网的汉语知识库中每个词汇由一个四元组表示：[5]

< W_X = 词语E_X = 词语例子G_X = 词语词性DEF = 概念定义 >

DEF 部分是表示词与义原的关系，也是词汇描述中最重要的部分，可以简单地认为词是由义原通过某种关系构成的。

4 总结

目前，基于本体的相似度计算研究已经取得很多成果，本文仅从算法角度，对国外 4 类代表算法进行了系统的阐述和比较，并简单综述了国内代表性的研究。基于当前研究成果，笔者认为，今后基于本体的语义相似度研究还需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，如:如何根据具体任务选择调用相关算法和确定相关参数。

此外，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] 崔韬世,麦范金.词语相似度计算方法分析[J].网络安全技术与应用,2012,05:55-56+72.

[4] 孙海霞,钱庆,成颖.基于本体的语义相似度计算方法研究综述[J].现代图书情报技术,2010,01:51-56.

[5] 马良荔,孙煜飞,柳青.语义 Web 中的本体匹配研究[J].计算机应用研究,2017,05:1-3.