本体语义相似度自适应综合加权算法研究*

白津宁

(大连海事大学 信息科学技术学院 大连 116026)

摘 要：本体语义相似度计算是解决语义网中语义异构的关键环节。通过对传统语义相似度计算方法的分析研究，引入本体层次结构，给出基于信息内容、距离、属性的语义相似度改进计算方法，并采用主成分分析法，提出一种自适应相似度综合加权计算方法(ACWA)，以解决传统综合加权计算时人工赋权的不足。实验结采表明，提出的 ACWA 算法的计算结果与参照标准之间的皮尔森系数较传统算法平均高出了 8.1% ，有效才是升了本体语义相似度计算的准确性。

关键词：本体；语义相似度；主成分分析法

中图法分类号：TP391 **文献标志码**： A **文章编号**：1001-3695(2016)05-1403-04

Adaptive Ontology Semantic Similarity Comprehensive Weighted Algorithm

Bai Jinning

(College of Information Science & Technology, Dalian Maritime University, Dalian Liaoning 116026, China)

Abstract：Ontology semantic similarity computation is the key to solve the semantic heterogeneity in semantic Web. Through analysis and study of the traditional ontology semantic similarity computation, this article introduced the ontolσ gy hierarchy, proposed an improved semantic similarity method, which is based on information content, distance , and at tribute , and put forward a ACWA using the principal component analysis, to address the deficiencies of artificial weight in traditional comprehensive weighted calculation. The experimental results show that the pearson cofficient of the pro posed ACWA algorithm results compared with the reference value is 8. 1 % higher than that of the traditional method, and the accuracy of ontology semantic similarity calculation is effectively increased.

Keywords：Ontology；Semantic similarity； Principal component analysis

0 引言

当前语义相似度计算已广泛应用于本体学习与合并、语义标注、知识管理中的信息抽取及自然语言理解等相关领域。随着本体日益广泛的应用，各领域都出现了大量的本体，然而本体的建立却没有统一的规范约束，因此产生了语义异构等许多问题。为了解决语义异构问题并把这些领域本体进行合并，本体映射(Ontology Mapping)的研究应用而生。本体映射是本体集成(Ontology Integration)、本体串联(Ontology Alignment)、本体合并(Ontology Merging)等的技术基础，是解决知识共享与重用的有效途径[1] ，而本体概念间的语义相似度计算是本体映射的关键技术之一，已成为当今信息技术研究的一个热点[2]。目前本体概念的语义相似度计算方法主要有基于信息内容的计算、基于距离的计算、基于属性的计算及混合 3 种方法的综合加权计算。本文在对传统概念语义相似度计算方法进行分析研究的基础上，提出一种以自适应综合加权为核心的语义相似度计算方法。该方法综合考虑概念在本体树中的节点密度信息、深度信息、公共父节点信息，采用改进的基于信息内容的概念语义相似度算法，结合基于距离和属性的改进方法计算，利用主成分分析法对 3 个因素计算出的相似度进行动态加权线性求和。实验表明，与传统算法相比，本文提出的 ACWA 算法的计算结果与参照标准值之间 的皮尔森系数比传统算法平均高出 8.1%。

收稿日期：2015-02-01；**修回日期**：2015-06-06 **基金项目**： “863”高技术研究发展技术项目（2013AA01A605），国际自然科学基金项目(61373099)资助

作者简介：白津宁(1992 一)，女，硕士生，主要研究方向为概念格、数据挖掘、图像处理（2624997157@qq.com）。

1 相关工作

目前国内外已有大量研究者对本体概念语义相似度的计算进行了研究，并形成了许多成熟的相似度计算方法。主要包括 3 种计算方法。

1.1 基于信息内容的计算方法

基于信息内容的计算方法通过一个给定本体中两个概念之间的共享信息内容来确定概念之间的相似度[3] 。本体结构中的子节点表示的概念通常都是对其祖先节点表示概念的细化，所以一个节点的内容能够代表其所有祖先节点的信息内容，最近邻公共祖先节点代表的是在本体树结构中距离两个概念最近的共同祖先节点，可以采用两个概念最近公共祖先节点的信息内容对其语义相似度进行量化，得到本体树中任意两个概念之间的语义相似度计算[4]，如式(1)所示:

$$Sim(a,b)_{ic} = \frac{2IC(Lcan(a,b))}{IC(a)+IC(b)} \quad (1)$$

其中 Lcan(a,b)表示 a，b 在本体树中的最近邻公共祖先节点 IC(α)，IC(b)的表示 a 和 b 所拥有的信息内容，也称信息量，这种信息的量化表示提供了一种测量语义相似性的方法，信息量的计算公式如式(2)所示:

$$IC(c) = -\log\frac{N(c)}{N} \quad (2)$$

其中 N(c)为概念 c 在训练样本中出现的次数，N 为训练样本的总数。

1.2 基于距离的计算方法

基于距离的计算方法通过本体树中两个概念词的几何距离来计算它们之间的语义距离，计算模型如式(3) 所示:

$$Sim(a,b) = \frac{1}{Dis(a,b)} \quad (3)$$

其中 Dis(a,b)表示概念之间的最短距离。WuAndPalmer 算法与 LeacockAndChodorow 算法是基于语义距离的两种 经典算法。 WuAndPalmer 算法通过与概念词最近的公共父结点概念词的位置关系来计算其相似度; LeacockAndChodorow 算法则是将两概念间的路径长度转化为信息量来进行相似度计算。

1.3 基于属性的计算方法

本体概念通过属性来表明概念特征，基于属性的计算方法通过统计概念所具有的公共属性的个数来获得概念的相似度。概念的相似度与概念拥有的公共属性个数成正比。Tversky 提出基于属性最经典的语义相似度计算方法，计算模型如式 (4) 所示：

$$Sim(a,b) = \alpha \times \mathrm{Pr}\,operties(a \cap b) - \beta \times \mathrm{Pr}\,operties(a-b) - \gamma \times properties(b-a) \quad (4)$$

其中 Properties(a∩b)的表示概念 a 和 b 所具备的公共属性集合； Properties(a-b) 表示概念 a 具备而概念 b 不具备的属性集合；Properties(b-a) 则表示概念 b 具备而 a 不具备的属性集合。基于属性的计算方法可以对人类认识和辨别现实生活中各种事物的过程进行模拟，但需要给出事物各个属性的详细信息。

2 自适应加权算法

本体概念语义相似度综合计算分别考虑多个因素计算概念间的语义相似度，并线性加权求和计算出最终的语义相似度。综合计算过程中权值的确定是关键的一环，权值确定的准确与否直接影响相似度计算的准确性，为了使权值确定方法适用于不同的本体，本文基于经济学领域中经典的主成分分析法，提出一种自适应的综合加权算法 (Adaptive Comprehensive Weighted Algorithm, ACWA) 。首先考虑本体概念的信息内容、概念距离和概念属性，分别计算出概念的相似度， 然后用主成分分析法对计算的个结果进行加权求和，计算出最终的相似度。

2.1 多因素相似度计算

为了计算结果更加全面准确，从多个因素出发，分别计算概念间的相似度。本文从本体概念信息量、距离和属性因素分析，给出相应的语义相似度计算公式。

(1)概念信息量

提出一种利用本体概念自身的信息量进行求解的方法。如式（5）所示：

$$IC(c) = \left(\frac{mNode(c)}{mNode(T)} + \frac{\log(D(c))}{\log(mD(T))}\right) * \left(1 - \frac{\log(h(c)+1)}{\log(mNode(T))}\right) \quad (5)$$

其中 mNode(c) 是以概念 c 节点所在的以根节点为直接父结点的子树所拥有的概念节点的总个数，mNode(T) 是概念 c 在本体树所拥有的概念节点的总个数，h(c) 是概念 c 的所有子节点个数,D(c) 是概念的深度 mD(T) 是本体树 T 最大深度。求出概念的信息量后再使用式(1)计算两个概念的相似度。

(2)概念间距离

本文将边的关系类型作为权重加入到计算过程中，针对本体 3 种概念关系，定义对应边的权重如式(6) 所示：

$$Wedge(a,b)=\begin{cases}0.9, & \text{is } A\\ 0.5, & \text{partof}\\ 0.1, & \text{other}\end{cases} \quad (6)$$

并提出改进算法如式(7) 所示：

$$Sim(a,b)_{dis}=\frac{2*sPath(c,r)+sPath(a,b)}{sPath(a,b)+sPath(a,c)+sPath(b,c)+2*sPath(c,r)} \quad (7)$$

$$sPath(a_1,a_n)=\sum_{i=1}^{n} wedge_i * path(a_i,a_{i+1}) \quad (8)$$

其中 wedge 为相邻概念间边的权重，sPath 为相邻概念间直接距离，例 sPath(a,b) 表示从概念 a 到概念 b 的加权最短路径。

(3)概念属性

结合属性结构信息，提出一种改进的计算方法，如式 (9) 所示：

$$Sim(a,b)_{pro}=\frac{Properties(a\cap b)}{Properties(a\cap b)+\alpha*Properties(a-b)+\beta*Properties(b-a)} \quad (9)$$

$$\alpha=\begin{cases}\dfrac{d(a)}{d(a)+d(b)}, & a\leqslant b\\ 1-\dfrac{d(a)}{d(a)+d(b)}, & a>b\end{cases},\alpha+\beta=1 \quad (10)$$

d(α) ，d(b) 分别为概念在本体层次结构中的深度。

2.2 基于主成分分析法动态权值的计算

PCA 的思想是在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。通常转化生成的综合指标称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，这就使得主成分比原始变量具有某些更优越的性能。PCA 根据主成分的贡献率来分配各主成分的权重，而不是人为确定的，从而克服了多因素分析中人为确定权值的缺陷，保证结果客观、合理、准确[4]。

ACWA 中基于 PCA 的动态权值计算方法的

主要思想如下：

(1)将 3 个因素计算出的 3 个相似度作为 3 个维度，通过多个样本的计算得到相似度矩阵作为输入样本矩阵；

(2) 将样本矩阵矩阵标准化变换为标准矩阵 Z，并求出相关系数矩阵 R；

(3) 解样本相关矩阵 R 的特征方程得 3 个特征根，确定主成分；

(4) 解方程组 R；

(5) 将标准化后的指标变量转换为主成分；

(6) 对 3 个主成分进行加权求和，即得最终相似度值，权值为每个主成分的贡献率。

2.3 综合加权计算

本文考虑概念信息量、距离和属性 3 个因素分别计算出相似度后，对 3 个相似度用改进的主成分分析法进行动态加权线性求和。假设被比较概念对集合中有 m 对概念词，设 $X=(Sim_{ic(i)}\ \ Sim_{dis(i)},\ \ Sim_{pro(i)}$为主成分输入样本集合中的一个向量，其中每一维变量分别代表综合相似度计算模块中各部分语义相似度计算的结果，则概念相似度矩阵表示为 $X_{sim}=(x_{i1},\ x_{i2},x_{i3})$ (i= 1, 2, .., m) 。对构建出来的相似度矩阵 X_{sim} 进行主成分分析，提取出主成分为 $Y=(y_{sim1},\ y_{sim2},\ y_{sim3}$)，各主成分的贡献率为$(r_1,\ r_2,\ r_3)$，则最终的概念语义相似度计算公式为：

$$Sim_{total}=r_1*y_{sim1}+r_2*y_{sim2}+r_3*y_{sim3} \quad (11)$$

ACWA 算法:

输入：节点概念 ab，公共节点 c，根节点概念 r

输出：综合加权语义相似度 Sim_{total}

Begin

1.计算节点概念信息量 IC(a)、IC(b)、 IC(c)

2.基于信息量的语义相似度计算：

$$Sim(a,b)_{ic}=\frac{2IC(Lcan(a,b))}{IC(a)+IC(b)}$$

3.确定节点概念之间边上权重 ：wedge

4.计算节点概念之间最短加权距离：sPath(a，b)

5.基于距离的语义相似度计算 ：$Sim(a,b)_{dis}$=

$$\frac{2*sPath(c,r)+sPath(a,b)}{sPath(a,b)+sPath(a,c)+sPath(b,c)+2*sPath(c,r)}$$

6.确定影响因子

7.基于属性的语义相似度计算：$Sim(a,b)_{pro}$=

$$\frac{Properties(a\cap b)}{Properties(a\cap b)+\alpha*Properties(a-b)+\beta*Properties(b-a)}$$

8.PCA 计算 3 个因素各自的动态权值

9.综合加权计算 :Sim_{total}

End

3 实验与结果分析

验证自适应综合加权算法优越性：

3.1 实验过程

采用部分旅游本体作为实验数据，以验证算法的准确性和有效性数据的统计信息如表 1 所列：

表 1　数据统计信息表

本体	概念个数	属性个数	边类型数
旅游本体	41	33	2

实验步骤如下：

(1)选取数据集中的概念对；

(2)从信息内容、距离和属性计算本体概念对的语义相似度并构造相似度矩阵；

(3)用相似度矩阵作为样本矩阵，实现改进的 PCA 算法，输入样本矩阵计算出各个因素的贡献率，并将其作为权值；

(4 对应相似度值加权并线性相加计算出最终的概念语义相似度；

(5)计算结果与经典的算法计算结果比较。

3.2 实验结果分析

ACWA 与传统算法相比：将 3 个因素的贡献率作为权值进行加权计算得到最终相似度结果，与传统算法计算的结果比较如表 2 所示：

表 2 相似度计算结果比较

本体概念对的相似性	相似度算法			人工判定
	所提算法(ACWA)	基于信息内容(Resnik)	基于距离(Wu & Palmer)	
Sim(人文资源，自然资源)	0.456	0.500	0.540	0.430
Sim(人文资源，中心站)	0.215	0.300	0.360	0.192
Sim(景点，交通站点)	0.313	0.486	0.400	0.380
Sim(人文资源，少林寺)	0.495	0.668	0.600	0.550
Sim(人民公园，中原区)	0.193	0.293	0.300	0.190
Sim(人民公园，二七区)	0.193	0.293	0.300	0.190
Sim(少林寺，龙门石窟)	0.646	0.568	0.400	0.645
Sim(少林寺，塔林)	0.597	0.601	0.600	0.600
Sim(火车站，郑州站)	0.609	0.650	0.600	0.620
Sim(住宿，二七区)	0.187	0.340	0.300	0.200

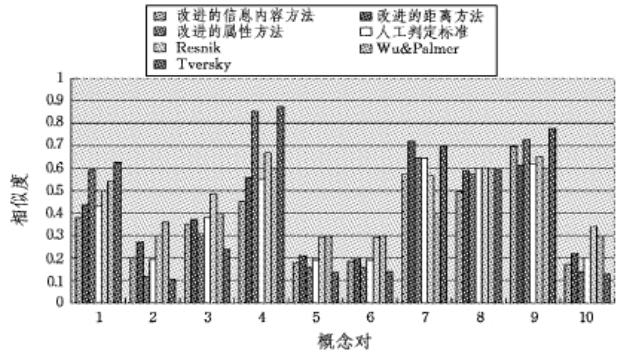

图 1　各算法计算结果与专家人工判断比较

(1)在信息内容值相差较小的两个弱概念计算中，ACWA 算法较基于信息内容的 Resnik 方法更加准确。

(2)ACWA 计算结果与人工判断匹配更多。

(3)ACWA 算法使权值的确定可以根据计算出的样本动态确定，具有自适应性，更加客观。

总的来讲，基于本体　　层次结构和主成分分析法提出的 ACWA 算法的计算结果更 接近人类专家的判断结果，相似度计算准确性较高。

3 结语

语义网的快速发展以及本体的广泛应用产生了语义异构问题，本体语义相似度计算是解决语义网中语义异构的关键环节。针对传统本体概念语义相似度计算的不足，充分考虑相似度计算的影响因素和目前综合加权算法中专家人工赋权的严重不足，提出了改进的相似度计算算法 ACWA。ACWA 考虑影响本体概念的 3 个主要因素并对相应的已有算法进行改进，结合改进的主成分分析法对各个因素计算出的样本结果动态加权并线性相加计算出最终相似度。实验结果证明，该算法与人工判断的相似度值的相关度优于传统算法，尤其是主成分分析法自适应动态赋权，降低了人工赋权的误差，有效确保了计算的全面性、准确性。ACWA 算法使权值的确定可以根据计算出的样本动态确定，具有自适应性，更加客观。

参考文献：

[1] 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 ，39(2):8-13

[2] Bae M，Kang S， Oh S. Semantic similarity method for keyword query system on RDF[J]. . Neurocomputing， 2014 ，146 (C)： 264 275

[3] Zhang C，Yang Y，Guo X，et al. The Improved Algorithm of Semantic Similarity Based on the Multi-dictionary [J]. Journal of Software ，2014 ，9(2)：324-328

[4] 王桐，王磊，吴吉义. Word Net 中的综合概念语义相似度计算方法[J]. 北京邮电大学学报，2013，36(2)：98-106

[5] Sun HX，Qian J，Cheng Y. Review of Ontology-based Semantic Similarity Measuring[J]. New Technology of Library and Infor rnation Service， 2010 (1) ： 51-56