

基于模糊语义的本体概念相似度计算算法

罗俊丽¹, 王 亚¹, 路 凯²

(1 许昌学院 计算机科学与技术学院, 河南 许昌 461000; 2 许昌电气职业学院, 河南 许昌 461000)

摘 要: 针对现有本体概念相似度计算方法无法处理概念特征属性的模糊性问题, 基于本体的语义结构特点提出了通过语义扩展集来对概念特征进行描述的方法, 并在此基础上定义概念的模糊特征集合. 在特征的量化过程中, 充分考虑了本体概念相似度的不对称性、概念的层次深度及区域密度对相似度的影响, 最后, 通过模糊集合之间的相似度来衡量本体概念的相似度. 在通用数据集上的实验表明, 该方法的相似度计算性能要优于以往的计算方法.

关键词: 模糊集合; 语义相似度; 边权重; 本体

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2013)07-0128-05

An Algorithm of Ontology Concepts Similarity Computing Based on Fuzzy Semantics

LUO Jun-li¹, WANG Ya¹, LU Kai²

(1 College of Computer Science and Technology, Xuchang University, Xuchang 461000, China;

(2 Xuchang Electrical Vocational College, Xuchang 461000, China)

Abstract: For the existing ontology concepts similarity measuring methods cannot describe the fuzzy features of concept, based on the semantic structure information of ontology, a method which describes the features of the concept by semantically expanded set is proposed, then a fuzzy set is defined based on the semantically expanded set. As quantifying the features, the algorithm considers influence of asymmetry, depth and local density of ontology concepts on similarity computation. Finally, the semantic similarity between two fuzzy sets is used to measure the concepts similarity of ontology. The result of experiment on a commonly used dataset indicates that our method outperforms traditional similarity measures.

Key words: fuzzy set; semantic similarity; edge weight; ontology

1 引言

随着本体技术和语义 Web 的发展, 基于本体的概念相似度计算被广泛用于解决网络信息资源的语义冲突和资源共享问题. 为了有效地计算概念之间的相似度, 需要考虑三个基本因素: 概念特征的描述、特征权重的量化、特征集合之间的相似度计算.

目前, 针对本体概念相似度的方法有很多, 概括起来都是基于网络距离模型和基于信息论模型的计算方法^[1]. 相关研究表明, 虽然网络距离模型和信息理论模型分别从不同的角度来反映语义相似度. 但

这两种方法都可以认为是 Tversky 相似度模型的不同变化形式. 但 Tversky 模型的二值特征无法描述本体概念间语义关系特征的模糊性, 此外, 已有的概念相似度方法很多都是对称的, 这并不适合描述本体概念结构的性质. 通常认为在本体中每个子概念节点都是对其父概念节点的一次细分和具体化, 因而子概念与父概念节点间的相似程度要高于父概念与其之间的相似程度.

针对上述问题, 本文提出了一种基于模糊特征集合的本体概念相似性度量方法 FSSC(Fuzzy Semantic Similarity Compute).

收稿日期: 2012-10-15; **修回日期:** 2012-11-08

基金项目: 国家自然科学基金项目(61003035)

2 基本概念

2.1 Tversky 模型

Tversky 模型是基于概念特征进行相似性度量方法的典型代表^[2],它通过特征集合描述概念对象.该模型基于二值特征,通过 0 或 1 表示特征的有或无,相似度被定义为关于特征共同性和差异性的函数.其计算公式为:

$$Sim(C_1, C_2) = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A - B) + \beta \cdot f(B - A)} \quad (1)$$

其中 A、B 分别是对象 C_1 和 C_2 的特征集合. Tversky 模型具有不对称性, α 和 β 的取值是造成不对称的重要原因.

2.2 模糊集合

与经典集合不同,模糊集合 A 中的元素不是确定地属于 A,而是某种程度地隶属于 A^[3]. 属于集合 A 的隶属程度可用模糊隶属度值表示为: $\mu_A(x)$. 即定义在论域 U 上的模糊集合 A,对于任何的 $x \in U$ 都有一个数值 $\mu_A(x) \in [0, 1]$ 与之对应,可以把模糊集合 A 表示为 $A = \{x, \mu_A(x) \mid x \in U\}$. Zadeh 给出了定义在模糊集合上的基本操作算子如下:

$$\mu_{\neg A}(x) = 1 - \mu_A(x) \quad (2)$$

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) \quad (3)$$

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) \quad (4)$$

2.3 本体概念信息量

根据信息论中信息量的定义,一个概念 C 的信息量 IC 可以采用如下的公式计算:

$$IC(c) = -\log p(c) \quad (5)$$

其中, $P(C)$ 为概念 C 的实例在文本集中出现的概率. 由于文本集统计方法的数据稀疏及概念词汇的多义性问题会影响概念信息量度量的准确性,因此在文中把本体本身作为统计源,根据其良好的概念层次结构信息来计算各概念的出现概率值. 根概念节点的出现概率被认为是 1,其信息量为 0,随着概念层次的不断深入,概念出现的概率越来越小,其信息量越来越大. 由于任何一个属于父概念的实例,必定属于其某个直接子概念,因此我们把本体中子概念的出现概率看作是对其父概念的分享. 在文中,概念的信息量定义如下:

$$P(C) = \begin{cases} 1, C \text{ 是根节点} \\ \left(\frac{1}{n}\right) * P(P), C \text{ 为非根节点} \end{cases} \quad (6)$$

其中, P 为 C 的父节点, n 为 P 的子节点个数.

3 基于模糊语义的概念相似度计算

该算法的具体步骤为:

输入:本体图 $G = (V, E)$

输出:本体中概念 C_i 和 C_j 的相似度

步骤 1:考虑本体概念相似度的不对称性、概念层次深度及区域密度的影响,对本体图中的单位连接边赋权重,得到单位层次语义关系相似度的矩阵 T_0 .

步骤 2:计算 T_0 传递闭包 T^+ 来层次语义关系的传递性,得到各概念间的语义相似度.

步骤 3:由矩阵 T^+ 分别定义概念 C_i 和 C_j 的语义扩展集,并在此基础上定义概念的模糊集合.

步骤 4:计算模糊集合之间的相似度作为概念 C_i 和 C_j 之间的语义相似度.

3.1 本体语义关系矩阵 T_0

(1)边权重的计算

边权重的计算是衡量本体概念间语义相似度的基础, Rada 等^[4]认为对语义网络中相似性的评估实际上应该只限于 IS-A 型的层次分类结构,而排除其它的连接类型. 因此我们主要研究的语义相似性度量方法也是基于 IS-A 类型的层次结构.

在本体层次网络中边权重主要与三个因素有关:一是语义关系边所处的层次深度,层次越高的边,其连接的父子概念越抽象,则边权重的值越小,而层次越低的边,其连接的父子概念越具体,则语义越相似,边权重值越大;二是语义关系边所处的节点密度,节点密度越大,则父子概念间的语义相似性越大,边权重越大,反之则越小;三是语义关系边的类型,一个概念跟它的父概念的相似程度要高于其父概念与它的相似程度. 语义距离权重一般被认为是边权重的相反度量,因此这里先给出语义距离权重的计算公式:

$$Dist(c_i, c_j) = \begin{cases} \frac{1}{n} \cdot \frac{IC(c_i) - IC(c_j)}{IC(c_j) + IC(c_i)} \cdot \frac{Dep(c_j)}{Dep(c_i) + Dep(c_j)}, & c_i \text{ 为 } c_j \text{ 的子概念} \\ \frac{1}{n} \cdot \frac{IC(c_j) - IC(c_i)}{IC(c_j) + IC(c_i)} \cdot \frac{Dep(c_i)}{Dep(c_i) + Dep(c_j)}, & c_j \text{ 为 } c_i \text{ 的子概念} \end{cases} \quad (7)$$

其中, n 表示父概念 P 的子概念数目,而 $IC(c_i) - IC(c_j) / IC(c_i) + IC(c_j)$ 则是表示父子概念间的连接边强度,该公式内置了层次深度及区域密度的影响; $Dep(c_j) / Dep(c_j) + Dep(c_i)$ 则反映了语义关

系边对概念相似度的不对称性影响。

而边权重一般被认为是语义距离权重的相反度量,因此,给出边权重的计算公式为:

$$wt(c_i, c_j) = \frac{1}{e^{Dist(c_i, c_j)}} \quad (8)$$

为了能够清晰的反映边权重的各影响因素,按照文中的语义距离权重计算方法对如图 1 所示的本体片段计算其语义距离权重。

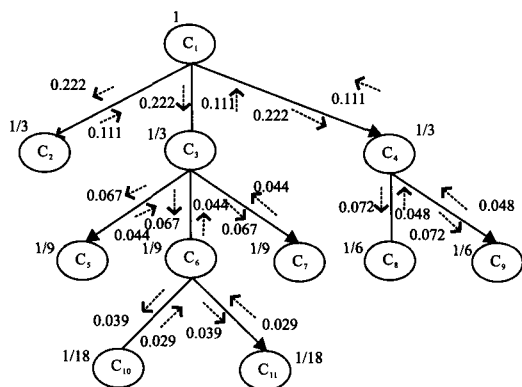


图 1 本体片段语义距离权重示例

图中各概念旁边的分数是按照公式 $P(C) = (1/n) * P(P)$ 为各概念分配的概率;概念连接边旁的数字为按照公式(7) 计算出的连接边语义距离权重,数字旁边的虚线箭头表示语义关系边的连接类型,指向父节点的箭头表示的是子 $\rightarrow$ 父概念边语义距离权重,指向子节点的箭头表示父 $\rightarrow$ 子概念边的语义距离权重。从图中可以看出,随着层次结构树深度的增加,对应连接边的语义距离权重逐渐减少,而概念边权重加大;而随着概念区域密度的增加,语义距离权重逐渐减小,则概念边权重逐渐增大;而子 $\rightarrow$ 父概念边语义距离权重要大于父 $\rightarrow$ 子概念边的语义距离权重,这符合我们对于边权重影响因素的分析。

(2) 本体语义关系矩阵 T_0

得到本体中各个单位有向边的边权重之后,可以将其表示为语义关系矩阵 T_0 。

$$T_0 = \begin{matrix} & \begin{matrix} c_1 & c_2 & \cdots & c_n \end{matrix} \\ \begin{matrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{matrix} & \begin{bmatrix} 1 & c_{12} & \cdots & c_{1n} \\ c_{21} & 1 & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & \vdots & \cdots & 1 \end{bmatrix} \end{matrix}$$

其中, C_{ij} 的值表示概念 C_i 和 C_j 之间的单位有向边权重,则矩阵 T_0 体现了语义关系边的不对称性。

3.2 本体各概念语义相似度矩阵 T^+

层次语义关系是具有传递性的,为了体现这种

传递性,通过定义矩阵合成运算 $\otimes$ 来描述。若概念 i 和概念 j 通过节点 k 相连,则定义概念 i 和概念 j 之间的语义相似度为:

$$[T \otimes T]_{ij} = \max(T_{ik} \cdot T_{kj}) \quad (9)$$

令 $T_{r+1} = T_r \otimes T_0$,对语义关系矩阵 T_0 定义传递闭包运算:

$$T^+ = \lim_{r \rightarrow \infty} T^r \quad (10)$$

对于本体来说,传递闭包将在经过本体层次最大深度次数的计算后达到收敛状态。矩阵 T^+ 较全面地反映了本体中概念间的语义非对称性、层次深度、区域密度、语义传递性等问题。

3.3 概念的模糊特征集合

对象特征的描述是相似度计算的基础,针对本体概念的相似度研究,目前的方法没有充分考虑本体的语义结构特点,认为本体概念是较小的信息描述,无法用更小的特征信息对本体概念进行描述,为此我们通过选择概念的语义扩展集来描述概念特征。

(1) 概念的语义扩展集

虽然对于本体中的任何两个概念来说,理论上上都存在着某种程度的语义关系,但是对于本体概念 C ,不可能将所有的相关概念都作为其特征定义的描述。文中认为,只需要找出那些在本体各概念语义相似度矩阵 T^+ 中与概念 C 的相关度大于某个阈值的所有概念来作为它的特征描述即可。即本体概念 C 的特征描述集为其语义扩展集 $SE_C = \{C, C_1, C_2, \dots\}$,其中 C_i 为矩阵 T^+ 中与 C 的相似度大于阈值 θ 的本体概念。

(2) 概念的模糊特征集合

得到概念的语义扩展集合之后,就可以进一步定义概念的模糊特征集合。对于本体中给定的概念 C ,本体中的所有概念作为论域 U , C 的语义特征扩展集 $SE_C = \{C, C_1, C_2, \dots\}$,则定义 C 的模糊集合为论域 U 上的一个模糊子集 $F_C = \{C/1, C_1/w_{c,c_1}, C_2/w_{c,c_2}, \dots\}$ 。其中概念 C_i 对 C 的隶属度 w_{c,c_i} 是从本体各概念语义相似度矩阵 T^+ 中得到的。

3.4 模糊相似度计算

对于模糊集合之间的相似度计算,仍采用经典的 Tversky 计算公式及 Zadeh 模糊操作算子来计算概念之间的相似度。

4 实验

4.1 实验数据

鉴于 WordNet 本体的名词词库中大约 80% 是

层次语义关系(即上下位语义关系),因此采用 WordNet 作为实验本体^[5]. 文中的实验是基于 Rubenstein 数据集的子集 M&C 数据集来评估本文 FSSC 算法的性能,M&C 用了 30 对名词作为实验数据,并且通过 38 个试验者来对数据进行评估,在文中使用了其中的 28 组概念对作为本文的实验数据集.

4.2 评价标准

文中通过采用皮尔森相关系数来对本文的概念相似度计算方法进行评价,相关系数计算公式为:

$$Correl(X,Y)=\frac{\sum_{i=1}^n(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n(X_i-\bar{X})^2(Y_i-\bar{Y})^2}}$$

(11)

其中, X_i 和 Y_i 分别表示由概念相似度计算公式和专家判断的第*i*对概念的相似度值, $\bar{X}$ 和 $\bar{Y}$ 分别表示样本概念对相似度的平均值.

4.3 实验结果及分析

实验过程分为两个阶段:边权重的计算和本体概念相似度的计算. 为了对比 FSSC 算法的有效性,在计算边权重时,分别采取了基于信息论的算法、基于语义距离模型算法和本文算法. 这里基于信息论的方法采用的相似度计算公式为 Lin 提出的方法,概念的信息量采用本文的计算方法得出. 基于语义距离模型的相似度公式采用的是 Wu-Palmer 等提出的方法^[6]. 由于目前基于 WordNet 的概念相似度方法一般都是对称的,因此实验中本文 FSSC 算法中使用的 Tversky 相似度计算公式中的 α 和 β 的取值相同,经过多次尝试,当 $\alpha=0.3$ 时,本文算法与专家判断值有较好的相关度,各算法的相似度计算结果如表 1 所示. 表 2 列出了各算法与人工判断的相关系数.

从实验结果可以看出,该文提出的方法与专家判断结果较为相近,取得了较好的相似度计算结果. 这是因为本文算法在计算边权重时充分考虑了概念层次深度、区域密度及连接边类型对概念相似度的影响,同时通过层次语义关系矩阵的闭包运算实现了层次语义关系的传递性,从而保证了概念的语义扩展集合能够较好的表达概念特征,克服了传统概念相似度计算方法无法对概念进行细致描述的不足,提高了相似度计算的准确性.

表 1 各相似度计算方法结果比较

概念 1	概念 2	基于信息论的方法	基于语义距离的方法	FSSC 方法
cord	smile	0.002	0.016	0.000
rooster	voyage	0.000	0.004	0.000
noon	string	0.000	0.021	0.000
glass	magician	0.322	0.255	0.078
monk	slave	0.598	0.610	0.311
coast	forest	0.700	0.511	0.050
monk	oracle	0.623	0.566	0.120
lad	wizard	0.320	0.433	0.244
forest	graveyard	0.822	0.750	0.003
food	rooster	0.743	0.825	0.000
coast	hill	0.655	0.739	0.252
journey	car	0.452	0.467	0.000
crane	implement	0.820	0.720	0.601
lad	brother	0.211	0.420	0.310
bird	crane	0.420	0.300	0.557
bird	cock	0.890	0.765	0.810
food	fruit	0.486	0.358	0.129
brother	monk	0.711	0.623	0.750
asylum	madhouse	0.632	0.533	0.762
furnace	stove	0.798	0.821	0.914
magician	wizard	0.410	0.653	0.762
journey	voyage	0.687	0.691	0.708
coast	shore	0.862	0.701	0.758
tool	implement	0.710	0.790	0.826
boy	lad	0.814	0.820	0.774
car	automobile	0.910	0.891	0.983
midday	noon	0.924	0.914	1.000
gem	jewel	0.893	0.889	1.000

表 2 各方法与人工判断的相关系数

计算方法	相关系数
基于信息论的方法	0.744
基于距离模型的方法	0.732
FSSC 方法	0.865

5 结束语

本文提出了一种基于模糊语义的本体概念相似度计算方法. 借助于本体良好的层次语义关系,选择本体中与概念 C 具有较大语义关联的概念作为其特征描述,克服了传统相似度方法无法对概念进行细粒度特征描述的不足,同时将概念之间的相关度作为其特征描述的模糊隶属度,在此基础上定义概念的模糊特征集合,弥补了 Tversky 模型二值特征的缺陷,通过计算模糊特征集合之间的相似度得到

概念之间的相似度值,提高了概念相似度计算的准确率. 下一步我们将在本体映射、语义信息检索等领域中应用该方法,进一步验证该方法的有效性.

参考文献:

- [1] 宋玲. 语义相似度计算及其应用研究[D]. 济南: 山东大学, 2009.
- [2] 李文清, 孙新, 张常有, 等. 一种本体概念的语义相似度计算方法[J]. 自动化学报, 2012, 38(2): 229-235.
- [3] Cross V. Fuzzy semantic distance measures between ontological concepts[J]. IEEE Annual Meeting of the Fuzzy Information. 2004(2): 635-640.
- [4] Roy Rada, Hafedh Mili, Ellen Bicknell, et al. Development and application of a metric on semantic nets [J]. IEEE Transaction on Systems, Man, and Cyber-

netics, 1989; 19(1): 17-30.

- [5] 罗俊丽, 李慧娜, 路凯. 基于词义消歧的语义查询扩展研究[J]. 微电子学与计算机, 2012, 29(1): 71-75.
- [6] 曾诚, 韩光辉, 李兵, 等. 一种改进的基于路径的语义相似度计算算法[J]. 微电子学与计算机, 2011, 28(8): 129-135.

作者简介:

罗俊丽 女, (1986—), 硕士, 助教. 研究方向为语义信息处理、数据挖掘.

王亚 女, (1986—), 硕士, 助教. 研究方向为复杂信息处理、数据挖掘.

路凯 男, (1981—), 硕士, 助教. 研究方向为信息处理和信息安全.

(上接第 127 页)

- [6] Po DDY, Do M N. Directional multiscale modeling of images using the contourlet transform[J]. Image Processing, IEEE Transactions on, 2006, 15(6): 1610-1620.
- [7] 梅雪, 李久贤, 夏良正. 基于 Contourlet 变换的加权小波特征抽取算法[J]. 数据采集与处理, 2008(1): 23-26.
- [8] Sing J K. Self-adaptive RBF neural network-based segmentation of medical images of the brain[J]. Intelli-

gent Sensing and Information Processing, 2005: 447-452.

作者简介:

岳丽娟 女, (1988—), 硕士研究生. 研究方向为人工智能与模式识别.

吴锡生 男, (1959—) 博士, 教授, 硕士生导师. 研究方向为人工智能与模式识别、图像处理等.