基于本体的概念相似度计算及其应用*

冉 婕[1,2],孙 瑜[1],漆丽娟[2]

(1.云南师范大学 计算机科学与信息技术学院,云南 昆明 650092;
2.云南昭通师范高等专科学校 计算机科学系,云南 昭通 657000)

摘 要:提出了基于语义相似度和相关度的综合概念相似度计算方法。语义相似度考虑了语义距离和本体库特征,加入概念的信息量、概念的深度、概念的密度和不对称因子的辅助影响;语义相关度从直接相关、间接相关、直接继承和间接继承几个方面考虑。通过实验和两种传统的语义相似度计算方法进行对比,本方法能更好地区分本体树中不同关系的概念对,验证了该方法的有效性。

关键词:本体;相似度;语义距离;信息量

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2010)11-0014-03

Concept similarity computation based on ontology and its application

RAN Jie[1,2], SUN Yu[1], QI Li Juan[2]

(1.Institute of Computer Science and Information Technology, Yunnan Normal University, Kunming 650092, China;
2.Department of Computer Science, Zhaotong Teacher's College, Zhaotong 657000, China)

Abstract: This paper puts forward an integrated method based on semantic similarity and semantic relevancy. In semantic similarity, we think about the semantic distance and the characteristics of ontology, the concept's amount of information, the depth of concept, the density of concept and symmetry factor; in semantic relevancy, we think about the directly relation and indirectly relation, direct inheritance and indirect inheritance. Compated with two traditonal semantic similarity computation methods, this method can better distinguish the different concept in ontology tree and this method is effective.

Key words: ontology; similarity; semantic distance; amount of information

目前,信息检索大多基于关键字进行,查准率及查全率均不高,而本体能描述数据的语义,基于本体进行信息检索,检索效率显然要高。参考文献[1]指出,本体在信息检索中的应用能够显著地提高检索的精确率和返回率。在信息检索领域中,概念的语义相似度计算起着重要的作用,因此可以利用本体计算概念间的语义相似度。语义相似度在不同的应用领域中可能会有不同的含义。在信息整合领域中,相似度一般指的是文本与文本能够匹配的程度;而在信息检索领域中,相似度则反映与用户查询在语义上的匹配程度。相似度越高,表明该文本与用户的请求越接近[2]。本文的研究背景为基于本体的信息检索。

利用本体计算概念间相似度的基础是:2 个概念间具有一定的语义相关性,它们在概念间的结构层次网络图中存在一条路径[3]。Resnik[4]根据 2 个词的公共祖先节点的最大信息量来衡量 2 个词的语义相似度;Agirre[5]在利用 WordNet 计算词语的相似度时,考虑了语义距离、概念层次树的深度和概念层次树的区域密度;参考文献[6]提出基于距离的语义相似度计算模型,这种模型简单直观,但它依赖于预先建立好的本体层次网络;参考文献[7]引入计算语言学中的语义距离思想来计算概念相似度,但其考虑概念间的相似度影响因素较少。针对上述研究情况,本文提出了一种基于语义相似度及相关度的综合概念相似度计算方法。

1 概念相似度计算

当 2 个概念具有某些共同特征时,则定义它们是相似的,用 sim(x,y)表示概念 x、y 之间的相似度。形式上,相似度计算满足[7]:(1)相似度的值为[0,1]区间中的一

*基金项目:国家自然科学基金项目(60903131);云南省社会发展科技计划应用基础研究项目(2009ZC052M);云南省教育厅重点项目(07Z10661)

个实数，即 $sim(x,y)\in[0,1]$；(2)如果 2 个概念是完全相同的，则相似度为 1，即 $sim(x,y)=1$ 当且仅当 $x=y$；(3)如果2 个概念没有任何共同特征，则相似度为 0，即 $sim(x,y)=0$。

本文基于骨架法构建了一个成语典故本体 ILQO (Idiom Literary quotation Ontology)，为了减小本体的规模，将本体的范围确定在楚汉相争时期，并在该本体上实现语义检索，试图通过 ILQO 查询出相关的历史知识。ILQO 树中最大深度为 4，共分为 11 个大类 79 个小类，图 1 中的本体片断具有一定的代表性，节点旁的数字表示概念的信息量。在 ILQO 上实现语义检索，提出了一种基于语义相似度和语义相关度的综合概念相似度计算方法。

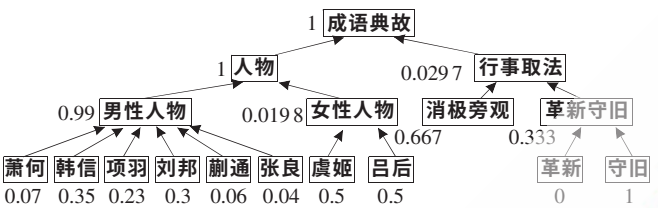

图 1 ILQO 片断

1.1 语义相似度

为了准确计算概念之间的语义相似度，本文充分考虑概念的语义距离、概念的信息量、概念深度、密度及不对称等因素，并在现有技术的基础上，多方面、多角度地给出概念语义相似度的综合计算。

1.1.1 概念的语义距离

2 个概念间的语义距离，是指在本体树中连接这 2 个节点的最短路径所跨的边数。本文用 $Dist(C_i,C_j)$来表示概念 C_i 与 C_j 之间的语义距离。1 个概念与其本身的距离为 0。语义距离是决定相似度的一个重要因素。一般而言，2 个概念的语义距离越大，其相似度越低；反之，2 个概念的语义距离越小，其相似度越大。两者之间可以建立一种简单的对应关系[3]。这种对应关系需要满足以下条件：(1)2 个概念距离为 0 时，其相似度为 1；(2)2 个概念距离为无穷大时，其相似度为 0；(3)2 个词语的距离越大，其相似度越小(单调下降)。C_i，C_j 是本体层次树中的任意 2 个节点，2 个概念 C_i 和 C_j 的语义距离 $Dist(C_i,C_j)$对相似度的影响可由以下公式决定[8]：

$$P1=\frac{\alpha}{Dist(C_i,C_j)+\alpha} \tag{1}$$

α 是可调节参数，表示当语义相似度为 0.5 时的语义距离值。

1.1.2 概念的信息量

在计算概念的相似度时，不同的概念拥有的实例不同，概念拥有的实例越多，说明概念在本体树中的重要性越大，2 个概念拥有的实例数越多，其相似的可能性越大。在本体树中，给概念分配不同的信息量，信息量的设定思想是根据概念下的实例数来设定各个概念的权重系数。具体公式为：

$$Info(C_i)=\frac{C_i\text{ 下的实例个数}}{C_i\text{ 的父节点的实例个数}} \tag{2}$$

鉴于 ILQO 的特征，从分类来讲，可看成 2 个大类，即基于人物的分类(成语典故中涉及的主要人物)和基于其他(除人物分类外的其他 10 个类)的分类。人物的 2 个子类中，男性人物拥有绝大多数实例，女性人物的实例相对偏少，从所查阅的资料显示涉及女性人物的目前只收集了 2 个实例。对于其他类，也存在类似的情况，分类下的实例数目也偏少，甚至有为 0 的情况(某些分类下暂无实例，主要为了便于以后本体库的扩充)。这样，在计算概念的信息量时，会出现不平衡的情况，如图 1 中行事取法类的信息量是 0.029 7，女性人物的信息量为 0.019 8。为了解决同一层中分类间的不均衡问题，故对信息量的处理分 2 种情况考虑：(1)男性人物和其父节点的信息量计算仍采用公式(2)；(2)女性人物及其他分类同父节点的信息量的公式调整为：

$$Info(C_i)=\frac{C_i\text{ 下的实例个数}}{C_i\text{ 的父节点的实例个数}}\times 10 \tag{3}$$

通过实验证明，信息量的调整有助于相似度的提高。通过以上分析可知，2 个概念拥有的实例越多，其相似度越大。本文提出 2 个概念 C_i 和 C_j 的信息量对相似度的影响，可由以下公式决定：

如果 $Info(C_i)$、$Info(C_j)$不同时为 0，则 $P2=1-|Info(C_i)-Info(C_j)|$，否则 $P2=0$。 (4)

从前面的分析可知，若概念 C_i 和 C_j 下皆无实例，无相似度可言，故 P2 为 0。

1.1.3 概念的深度

在本体的层次树中，概念的组织自顶向下，分类由大到小、由粗到细，处在离根较远的概念间的相似度要比离根近的概念间的相似度要大[3]。节点的深度是指概念与树根的最短路径所包括的边数。在本体树中，每一层都是对上一层概念的细化，由此可见，在语义距离相同的前提下，2 个节点的深度和越大，概念之间的相似度越大；2 个节点的深度差越小，概念之间的相似度越大。同样距离的 2 个概念，其相似度随着它们所处层次的总和的增加而增加，随着它们之间层次差的增加而减小[9]。根节点的深度为：$Dep(Root)=1$，其他节点的深度为：$Dep(c)=Dep(Parent(c))+1$，2 个概念 C_i 和 C_j 的深度对相似度的影响可由以下公式决定：

$$P3=1-\frac{|Dep(C_i)-Dep(C_j)|}{Dep(C_i)+Dep(C_j)} \tag{5}$$

1.1.4 概念的密度

节点的密度是指概念的直接子节点的数目。本体中不同概念节点的密度是不同的，有的节点可能有上百个子节点，而有的节点可能只有几个子节点。一般来说，某个节点的子节点密度越大，说明细化的概念越具体，这

些子节点间的语义相似度也就越小，反之越大。

用 $Width(C_i)$ 来表示概念 C_i 的密度，概念 C_i 和 C_j 的密度对相似度的影响可由公式(6)决定：

$$P4=\begin{cases}\dfrac{Width(C_i)}{Width(C_j)}, & Width(C_i)\leqslant Width(C_j)\\ \dfrac{Width(C_j)}{Width(C_i)}, & 其他\end{cases} \tag{6}$$

从前面的分析可知，概念的相似度值是在[0，1]区间内的一个实数，超出该范围即认为是不合法的，故在考虑其密度时，有上述 2 种可能。

1.1.5 语义相似度的不对称性分析

参考文献[10]对语义相似度的不对称性进行了分析，即概念间的语义相似度具有不对称性。一般来说，一个概念跟它的祖先相比的相似程度高于其祖先与它相比的相似程度，树中一个处于较大深度的概念跟一个深度较浅的概念相比的相似程度要大于反过来相比的相似程度即：

$$\begin{cases}Sim(C_i,C_j)>Sim(C_j,C_i) & Depth(C_i)>Depth(C_j)\\ Sim(C_i,C_j)\leqslant Sim(C_j,C_i) & Depth(C_i)\leqslant Depth(C_j)\end{cases}$$

在 ILQO 中，如图 1 所示，对于概念“男性人物”和“刘邦”，根据主观判断，“男性人物”和“刘邦”相比的相似程度要低于“刘邦”和“男性人物”的相似程度，也即概念间的语义相似度具有不对称性。因此，引入不对称因子：

$$P5=\frac{Depth(C_j)}{Depth(C_i)+Depth(C_j)} \tag{7}$$

1.1.6 概念的语义相似度

综合本体树中概念的语义距离、信息量、深度、密度和不对称因素几方面的影响，提出概念语义相似度计算的公式：

$$Sim(C_i,C_j)=P1\times P2\times P3\times P4\times P5 \tag{8}$$

但在具体的实验中发现，相似度的 5 个方面其值均在[0，1]区间中。简单地将各部分乘积起来，测试了本体树中 5 对不同关系的概念对，相似度最大值为 0.232 5，最小值为 0.008 33，这显然不符合日常经验，故对上述公式进行调整。调整后的公式为：

$$Sim(C_i,C_j)=(P1+P2+P3+P4+P5)/5 \tag{9}$$

公式(9)不仅考虑了语义距离、概念的信息量、概念的深度和宽度，还考虑了语义相似度的不对称性，能较为合理地体现概念间的语义信息。

1.2 语义相关度

概念间的相关性包括一些能够体现概念之间客观存在的联系内涵。相关度作为相关性的量化指标，用来衡量概念间的相关程度。一般地，相关度的取值区间为[0，1]。若 2 个概念间没有联系，则这 2 个概念的相关度为 0；若 2 个概念之间有直接联系，则相关度为 1。由相关概念间的联系，相关性可分为直接相关、间接相关、直接继承相关和间接继承相关 4 种[10]。

(1)若 C_i，C_j 2 个概念直接相关(兄弟关系)：则 $Rel(C_i,C_j)=1$，如图 1 中萧何与刘邦的关系；

(2)若 C_i，C_j 2 个概念通过 n 个概念间接相关：则 $Rel(C_i,C_j)=\dfrac{1}{n+1}$，如刘邦和吕后，通过男性人物、人物和女性人物相关；

(3)若 C_i，C_j 2 个概念直接继承相关(父子关系)：则 $Rel(C_i,C_j)=0.5$，如男性人物和刘邦的关系；

(4)若 C_i，C_j 2 个概念通过 n 个概念间接继承相关：则 $Rel(C_i,C_j)=\dfrac{1}{2(n+1)}$，如人物和刘邦，通过男性人物间接继承相关。

由以上分析可知，相关度在很大程度上依赖于本体树中各概念间的关系，故对概念的相似度计算有较大的影响。

1.3 概念相似度

综合上面的分析，在基于本体的信息检索领域用概念关联度来衡量概念间的联系，概念的相似度主要考虑语义相似度和语义相关度 2 方面的因素，提出如下的概念语义相似度计算公式：

$$Sim\_Rel(C_i,C_j)=\lambda_1\times Sim(C_i,C_j)+\lambda_2\times Rel(C_i,C_j) \tag{10}$$

概念相似度是一个主观性相当强的概念，对于不同的应用，概念的相似度也不同。调节参数正是根据系统应用的不同来设计的，这里用 λ_1 和 λ_2 表示调节参数。在计算概念间的相似度时，可以调整参数值来确定系统需要的相似度。

2 实验分析

公式(1)中的 α 参数，参考文献[12]对其测试发现 α 取值为 2 时能获得和人们的日常经验相符的相似度值，本文实验结果也得到相同的结论，故取参数 α 为 2。在公式(10)中，考虑语义相似度和相关度在相似度计算中具有相同的重要程度，故令 λ_1 和 λ_2 为 0.5。实验证明，该取值的结果和人们的经验值一致。表 1 是本体树中不同概念对的相似度值。

表 1 信息量调整前后对照

概念 1	概念 2	说明	Sim	Rel	Sim_Rel
刘邦	行事取法	信息量调整前	0.529 28	0.250 0	0.389 64
刘邦	行事取法	信息量调整后	0.583 28	0.250 0	0.416 64
刘邦	女性人物	信息量调整前	0.562 98	0.333 3	0.448 14
刘邦	女性人物	信息量调整后	0.598 62	0.333 3	0.465 96
女性人物	吕后	信息量调整前	0.761 10	0.500 0	0.630 55
女性人物	吕后	信息量调整后	0.796 74	0.500 0	0.648 37
女性人物	人物	信息量调整前	0.683 96	0.500 0	0.591 98
女性人物	人物	信息量调整后	0.719 60	0.500 0	0.609 80
行事取法	守旧	信息量调整前	0.505 94	0.250 0	0.377 97
行事取法	守旧	信息量调整后	0.559 40	0.250 0	0.404 70

从表 1 中的数据可以看出，信息量调整后相同概念

对的相似度值有一定程度的提高，说明调整是合理的。

徐德智[11]等提出了概念间语义相似度计算方法，陈沈焰[12]等提出了概念语义相似度计算公式，利用其计算可以获得不同概念间的相似度值。

本文基于本体的概念层次树所提供的丰富语义信息，提出了一种基于语义相似度和语义相关度的综合概念相似度计算方法。语义相似度分别从语义距离、概念的信息量、概念深度、概念密度和不对称因子几方面考虑；语义相关度从直接相关、间接相关、直接继承和间接继承几个方面考虑。实验证明，与两种传统的相似度计算方法比较，本方法所获取的相似度值能更好地体现本体树中不同概念对的重要程度。该计算方法在 ILQO 中能获得较好的相似度计算，如何使其应用于其他领域本体的信息检索，是下一步的研究方向。另外，对影响语义相似度的几个因素，将其视为相同的重要程度，如何区分不同的因素对相似度的影响，也是下一步要改进的地方。

参考文献

[1] GUARINO N, GVERTER M C. OntoSeek: content-based access to the Web [J]. IEEE Intelligent Systems, 1999,14(3):70-80.

[2] 王家琴，李仁发，李仲生，等.一种基于本体的概念语义相似度方法的研究[J].计算机工程，2007，33(11):201-203.

[3] 张忠平，赵海亮，张志惠.基于本体的概念相似度计算[J].计算机工程，2009，35(7):17-19.

[4] PESNIK P. Using information content to evaluate semantic similarity [C]//Proc. of the 14th IJCAI. Montereal, Canada: [s.n.], 1995:448-453.

[5] AGIRRE E, RIGAU G. A proposal for word sense disambiguation using conceptual distance [C]//Proc. of the 1st International Conference on Recent Advances in NLP. Tzigov Chark, Bulgaria. [s.n.], 1995.

[6] LEACOCK C, CHODOROW M. Combining local context and wordNet similarity for word sense identification [J]. Computational Liguistics, 1998,24(1):147-165.

[7] 朱礼军，陶兰，刘慧.领域本体中的概念相似度计算[J].华南理工大学学报(自然科学版)，2004，32(11):147-150.

[8] 刘群，李素建.基于《知网》的词汇语义相似度计算[J].中文计算语言学，2002，7(2):59-76.

[9] 吴健.基于本体论和词汇语义相似度的 Web 服务发现[J].计算机学报，2005，28(4):595-602.

[10] 姜华.基于本体的语义检索技术研究与实现[J].现代图书情报技术，2008(4):39-43.

[11] 徐德智，郑春卉，K.Passi.基于 SUMO 的概念语义相似度研究[J].计算机应用，2006，26(1):180-183.

[12] 陈沈焰，吴军华.基于本体的概念语义相似度计算及其应用[J].微电子学与计算机，2008，25(12):96-99.

(收稿日期：2010-01-14)

作者简介：

冉婕，女，1975 年生，讲师，硕士研究生，主要研究方向：本体构建及语义检索。

孙瑜，女，1974 年生，博士，教授，硕士生导师，主要研究方向：智能信息处理。

漆丽娟，女，1977 年生，讲师，硕士研究生，主要研究方向：人工智能。