第35卷 第19期 计 算 机 工 程 2009年10月
Vol.35 No.19 Computer Engineering October 2009
·软件技术与数据库· 文章编号：1000—3428(2009)19—0083—03 文献标识码：A 中图分类号：TP311

基于本体的概念语义相似度度量

史 斌，闫健卓，王 普，方丽英
(北京工业大学电子信息与控制工程学院，北京 100124)

摘 要：针对概念语义相似度度量问题，提出结合基于图理论和信息量2种方法的语义相似度度量算法。计算2个概念在概念图中连接的路径长度、局部密度以及在连接2个概念之间的路径上连接关系的连接力度，结合连结路径权重和信息量来度量概念之间的语义相似度。实验结果表明，该算法能取得较好的度量效果。

关键词：本体；语义相似度；信息量

Ontology-based Measure of Semantic Similarity Between Concepts

SHI Bin, YAN Jian-zhuo, WANG Pu, FANG Li-ying
(College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124)

【Abstract】 This paper proposes a new measure which combines the graph-based measure and information content-based measure. The measure computes the path length of the two concepts in the concept graph, local density and the connect power of the edge, and integrates them with edge weight and information content. Experimental result indicates the measure performs well.

【Key words】 ontology; semantic similarity; information content

1 概述

随着网络上信息量的迅速膨胀以及被利用率的逐渐提高，Web信息的获取和理解成为一个重要课题。人们对Web信息的利用不再满足于网页元素的重用和获取，对Web信息语义层面的理解需求越来越迫切。虽然目前网络形式仍然以万维网为主，但是语义网已成为其发展的必然趋势，基于本体的语义网为Web信息的机器理解以及信息的共享提供了便利。

本体是知识共享的工具和载体，由概念、属性、关系、层次结构和实例五元组组成。本体的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇(术语)和词汇间相互关系的明确定义。本体的特点是能够层次化地管理大量的概念属性以及实例。因此，对于基于本体的语义网来说，概念之间的语义相似度度量成为本体应用到各个方面的基础工作。例如，网页的语义标注、语义搜索引擎、自然语言问答以及语义冲突的消解等方面的应用，都是以语义距离的度量为基础的。

2 概念间的语义相似度

通常人们比较的并不是2个概念之间的语义距，而是2个具体词汇之间的语义距离。例如，人们在讨论狗与猫的相似度时一定是指某种狗和某种猫的相似度，而不是狗的集合和猫的集合的相似度。词汇之间的语义相似度与概念之间的语义相似度原理相同。

传统的基于本体的概念之间的语义相似度度量方法可分为2类：

(1)基于图形化原理，以概念之间路径的长短作为衡量语义距离的标准；

(2)基于信息量的方法，利用2个概念共享的信息量多少来衡量它们之间的语义距离。

2.1 定义

定义1(本体模型) 定义本体模型为一个五元组：$O(C, R, H, P, A)$，其中，C代表概念集合；R为关系集合；H表示层次结构，记录了$C\times C$的概念之间的层次信息，$H(c_1, c_2)$表示c_1与c_2之间是子类的关系，c_1是c_2的祖先概念；P为属性集合；A是所有实例的集合。

定义2(最短路径) 在层次关系中连接c_1和c_2的边的权重和最小的路径即为c_1与c_2的最短路径。

定义3(最近共同祖先概念) 如果c_r在为c_1和c_2的最短路径上，并且同时是c_1和c_2的祖先概念，则c_r被称为是c_1和c_2的最近共同祖先概念。

定义4 (非继承性共同属性) 当c_1和c_2同时具有p_r属性，并且p_r属性并非继承自c_1和c_2的最近共同祖先概念c_r，那么p_r就被称为是c_1和c_2的非继承性共同属性。

2.2 基于图理论的方法

基于图理论的概念间语义度量方法是将本体的概念、属性和实例转化为有向树形图，再根据概念之间的路径距离长短来判断其语义相似程度。

假设要度量的2个概念分别为c_1和c_2，文献[1]提出利用概念间边计算应考虑的因素。文献[2]提出利用WordNet本体度量概念间语义相似度的语义距离应考虑的3个因素：(1)c_1和c_2的最近共同祖先节点的深度；(2)c_1和c_2的概要描述的重叠率；(3)c_1和c_2之间的最短距离。这种度量方法的公式为

$$sim(c_1, c_2) = \exp(-\frac{dist(c_1, c_2)}{b}) \quad (1)$$

基金项目：北京市优秀人才培养基金资助项目(110105197712102924)
作者简介：史 斌(1980－)，男，博士研究生，主研方向：语义搜索，Web信息获取，信息集成；闫健卓，副教授；王 普，教授、博士生导师；方丽英，博士
收稿日期：2009-02-25 **E-mail：**shibin@emails.bjut.edu.cn

其中，$sim(c_1, c_2)$表示 c_1 和 c_2 之间的语义相似度；$dist(c_1, c_2)$表示 c_1 和 c_2 之间的语义距离，计算方法为

$$dist(c_1, c_2) = \arg\min \begin{bmatrix} pl(c_1, c_2) \\ d_{nca}(c_1, c_2) \\ (1 + gloss(c_1, c_2)) \end{bmatrix} \tag{2}$$

其中，$pl(c_1, c_2)$代表 c_1 和 c_2 之间的最短路径长度；$d_{nca}(c_1, c_2)$表示 c_1 和 c_2 的最近共同祖先节点的深度；$gloss(c_1, c_2)$表示 c_1 和 c_2 在 WordNet 中查询到的概要描述的重复率。

基于语义距离的度量方法的前提为概念与概念之间连接的边具有统一的语义距离或对每条边进行加权，但由于加权的系数均为实验过程中所得，因此本文提出基于信息量的语义度量方法。

2.3 基于信息量的方法

信息量是信息论中表示涵盖信息多少的概念。在信息论中对信息量的定义为

$$IC(c) = -\mathrm{lb}\, p(c) \tag{3}$$

基于信息量的语义距离度量方法的原理是：如果 2 个概念共享的信息量越大说明 2 个概念就越相似。文献[3-4]提出利用信息量来度量语义距离的方法。在本体的树形结构中，叶子节点的信息量最大，节点的信息量随着节点的深度变小而变小，即越是抽象的概念信息量越小。如果全局存在根节点，那么根节点的信息量为 0。文献[5]提出的基于信息量度量语义距离的公式为

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\mathrm{lb}\, p(c)) \tag{4}$$

其中，$sim(c_1, c_2)$表示在本体数形结构中 c_1, c_2 的共同祖先节点的集合。也就是用 2 个概念间所有共同祖先概念中信息量最大的祖先概念的信息量来度量它们的语义相似度。

3 基于距离和信息量的语义相似度

文献[4]结合了语义距离和信息量，提出一种基于层次信息量和属性信息量的 HIC-AIC 算法，认为在基于语义距离和基于信息量的方法中都是利用了本体的自身的层次结构来计算相似度，但是忽略了很多重要的信息，例如本体的属性。因此，提出 2 个概念之间的相似度应该由本体的层次信息以及属性信息量来衡量，其公式为

$$sim(c_1, c_2) = HIC(c_1, c_2) + AIC(c_1, c_2) \tag{5}$$

其中，$HIC(c_1, c_2)$表示 c_1, c_2 之间的本体层次信息，由式(6)计算而得：

$$HIC(c_1, c_2) = -\mathrm{lb}\, P(c_r) \tag{6}$$

其中，c_r 表示 c_1, c_2 的最近共同祖先概念。通过式(6)可知本体的层次信息是通过 2 个概念的最近共同祖先概念的信息量来计算的。式(5)中的属性信息量由式(7)、式(8)计算而得：

$$AIC(c_1, c_2) = -\mathrm{lb}\, Q(c_1, c_2) \tag{7}$$

$$Q(c_1, c_2) = \frac{(M - N_i)}{M} \tag{8}$$

其中，$Q(c_1, c_2)$表示 c_1, c_2 的非继承性共同属性的概率；M 为所有属性的总数；N_i 为 c_1, c_2 的非继承性共同属性的数量。

这种方法利用了本体的属性这一自身信息，对于概念的语义相似度方法来说考虑更加全面，但是由定义可知，概念的非继承共同属性是不在 c_1 和 c_2 的最短距离路径上的，在之前的语义距离衡量方法中，都是利用 2 个概念的最近共同祖先来衡量，而这个方法则考虑了非最短路径代表的 2 个概念之间的语义共享信息。

图 1 是一部份本体概念层次图，表示多个概念之间的连接关系(主要为 IS-A)、概念在辞典中出现的概率及其信息量。

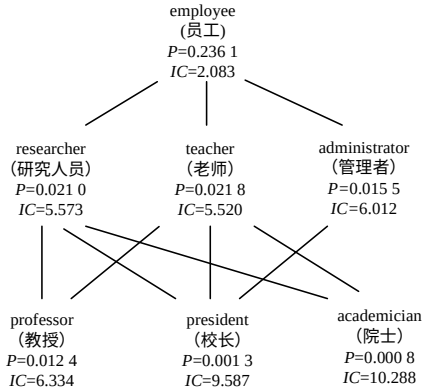

图 1 部分本体概念层次图

可以看出，2 个概念之间可以通过多条路径连接，校长与院士同时都是研究人员和老师，那么根据信息量度量方法，校长与院士之间的共享信息最大的路径是校长-研究人员-院士，这很符合人的直觉的结果。但是教授与院士与校长之间共享信息最大的概念是科研人员，并且共享的信息量与任课教师几乎相等，仅仅把科研人员作为 3 个概念的共享信息显得很片面，这就体现出文献[4]中描述的非继承性属性问题，也就是非最短路径的信息量问题。

为了将非最短路径的信息量考虑到概念之间的语义相似度度量方法中，本文提出采用概念之间的加权信息量总和来衡量概念之间信息量。

$$sim(c_1, c_2) = \sum_{c_{ri} \in S(c_1, c_2)} W_i \tag{9}$$

其中，W_i 为分别按各条路径连接 2 个概念的权重。W_i 的取值受到多种因素影响，包括 2 个概念之间的路径长度、2 个概念的局部密度，以及在连接 2 个概念之间的路径上连接关系的连接力度。

2 个概念之间的路径长度可以通过概念的深度与共同祖先概念的深度差来衡量，考虑到 2 个概念之间存在多条路径，则得到的路径长度通过 2 个概念之间的路径长度总和进行标准化。路径长度公式为

$$dis(c_1, c_2) = \frac{dep(c_1) + dep(c_2) - 2dep(c_r)}{\sum_{c_r \in S(c_1, c_2)} dep(c_1) + dep(c_2) - 2dep(c_r)} \tag{10}$$

其中，$dep(c)$表示概念 c 的深度；$dep(c_r)$表示 2 个概念的共同祖先的深度。

概念的局部浓度表示概念的孩子节点的连接数量。一般认为，2 个概念的共同祖先的局部密度越高，2 个概念的语义距离就越近。通过全体概念的局部密度来对 2 个概念的共同祖先的局部密度进行标准化，公式为

$$den(c_1, c_2) = \frac{den(c_r)}{\overline{den(c)}} \tag{11}$$

其中，$den(c_r)$表示 2 个概念的共同祖先的局部密度；$\overline{den(c)}$ 表示所有概念的局部密度的平均值。

2 个概念的连接力度标志了 2 个概念之间的关系的远近。笔者认为虽然同是 IS-A 的连接关系，但孩子概念和父亲概念的距离并不相等。通过这种方式解决了基于图理论的边的权重都相等的问题。连接力度可以通过 2 个概念的条件概率来计算，即当孩子概念 c 出现后，它表示了父亲概念 p 代表的意义的概率为 $p(c_i|p)$，并且有：

$$P(c_i \mid p) = \frac{P(c_i, p)}{P(p)} = \frac{P(c_i)}{P(p)} \tag{12}$$

$$IC(c_i \mid p) = -\mathrm{lb}(\frac{P(c_i)}{P(p)}) = IC(c_i) - IC(p) \quad (13)$$

2 个概念的连接力度可以通过 2 个概念路径上所有概念与父亲概念之间的信息量的差与深度的乘积来表示。

为了突出 2 个概念之间共享信息量最大的概念在度量 2 个概念之间的语义相似度计算中的作用，每条路径的权重计算公式为

$$W_p = \left(\frac{IC(C_{ri})}{\overline{IC(C_r)}} \right)^{\alpha} \quad (14)$$

其中，α是可调整的参数，代表该条路径在计算中的重要程度；$\overline{IC(C_r)}$ 表示所有共同属性的平均信息量。

综合以上所有因素，W_i的计算公式为

$$W_i = W_p dis(c_1, c_2)^{-\beta} (1-\gamma) den(c_1, c_2) \sum_{c_i \in S(c_1, c_2)} IC(c_i \mid p) \quad (15)$$

其中，β 是表示语义距离权重的参数；γ 是可调整的参数，标志该条路径上 2 个概念的共同祖先局部密度的重要程度。

4 实验结果与分析

对于概念之间的语义相似度度量，目前尚没有统一的检验方法，但 Miller 和 Charles 提出的 30 对名词对以及给出的相似度标准被大多数研究者认可，利用本文提出的方法应用到该数据集的部分结果如表 1 所示，其中，参数α取值 2；β 取值 1；γ 取值 0.5。

表 1　在 Miller 和 Charles 数据集上的部分实验结果

名词对	直觉相似度	基于距离和信息量方法
car – automobile	3.920	9.252
gem – jewel	3.840	11.332
journey – voyage	3.840	13.181
boy – lad	3.760	12.839
coast – shore	3.700	10.832
asylum – madhouse	3.610	8.320
magician – wizard	3.500	11.839
midday – noon	3.420	9.233
furnace – stove	3.110	16.930
food – fruit	3.080	13.283

实验结果表明，本文提出的方法比较符合人的直觉，在判断 2 个概念之间的语义相似度方面取得了较好的效果。

5 结束语

本文提出一种基于本体的概念相似度算法，算法考虑了 2 个概念之间连接路径的多种因素对相似度计算的影响，使概念之间的相似度计算结果与人的直觉更加相近。但是该方法只考虑了最常用的“IS-A”语义关系，在今后的研究中，将针对多种概念之间的语义关系进行相似度计算。

参考文献

[1] Li Yuhua, Bandar Z A, Mclean A D. An Approach for Measuring Semantic Similarity Between Words Using Multiple Information Sources[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 871-882.

[2] Alvarez M A, Lim S. A Graph Modeling of Semantic Similarity Between Words[C]//Proc. of ICSC'07. Irvine, California, USA: IEEE Computer Society, 2007.

[3] Cho M, Choi C, Kim W, et al. Comparing Ontologies Using Entropy[C]//Proc. of International Conference on Convergence Information Technology. Washington D. C., USA: IEEE Computer Society, 2007.

[4] Wang Guohua, Wang Yadong, Guo Maozu. An Ontology-based Method for Similarity Calculation of Concepts in the Semantic Web[C]//Proc. of the 5th International Conference on Machine Learning and Cybernetics. Dalian, China: [s. n.], 2006.

[5] Resnik P. Semantic Similarity in a Taxonomy: An Information-based Measure and Its Application to Problems of Ambiguity in Natural Language[J]. Journal of Artificial Intelligence Research, 1999, 11: 95-130.

编辑　顾姣健

（上接第 82 页）

3.3 删除点对 RNN 查询的影响

若从数据集 S 中删除已有的数据点，根据删除点的位置，基于算法 PRNN(q , S)，可得到解决删除点后对反向最近邻的影响问题的算法。

算法 Delete Point RNN, DEPRNN(S,q,P)

输入 数据点集 S，查询点 q，删除点 p ($p \in S$)

输出 新的反向最近邻集

Step1 调用算法 PRNN(q , S)，得到查询点 q 的反向最近邻集 $RNN[l]$。

Step2 若 $p \in RNN[l]$，则在 $RNN[l]$ 中删除 p，转 Step3；若 $p \notin RNN[l]$，则转 Step3。

Step3 找出 S 中以 p 为最近邻的点 p_i，做 $RNN[l]$ 中点 p_j 与 q 的垂直平分线，若 p_i 位于所有平分线的 p_j 侧，则输出查询点 q 的新反向最近邻集为 $RNN[l]$；若 p_i 位于所有平分线的 q 侧，则找出所有与以 p_i 为中心的 Voronoi 多边形邻接的 Voronoi 多边形的中心，做 q 与这些中心的垂直平分线，若 p_i 均位于所有平分线的 q 侧，则将 p_i 加入 $RNN[l]$ 中；否则输出查询点 q 的新反向最近邻集为 $RNN[l]$。

4 结束语

本文方法较适合处理平面及复杂曲面上数据点反向最近邻查询问题，与文献[4]中基于 Voronoi 图的反向最近邻查询相比，对多个查询点的问题有较大优势，下一步的研究重点主要集中在有障碍物的环境下移动对象的反向最近邻查询。

参考文献

[1] Korn F, Muthukrishnan S. Influence Sets Based on Reverse Nearest Neighbor Queries[C]//Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data. New York, USA: ACM Press, 2000: 201-212.

[2] Yang Congyun, Lin K I. An Index Structure for Efficient Reverse Nearest Neighbor Queries[C]//Proc. of the 17th International Conference on Data Engineeting. Heidelberg, Germany: [s. n.], 2001: 485-492.

[3] 周培德. 计算几何算法设计与分析[M]. 北京: 清华大学出版社, 2005.

[4] 李　松, 郝忠孝. 基于 Voronoi 图的反向最近邻查询方法研究[J]. 哈尔滨工程大学学报, 2008, 29(3): 261-265.

编辑　任吉慧