

基于树状结构的语义相似度计算方法分析

韩 欣^{1,2}, 樊永生¹, 马春森², 杨和平²

(1 中北大学 电子与计算机科学技术学院, 山西 太原 030051; 2 中国农业科学院 植物保护研究所, 北京 100193)

摘 要:文中提出了基于树状结构的语义相似度计算方法, 结合概念节点之间的语义关系、语义距离、概念节点的深度、密度对语义相似度的影响, 利用树的层次关系来表达概念节点之间的语义信息, 并对概念节点密度的计算进行改进, 加入了可调节的参数, 以适应不同的情景. 通过实验验证了该算法在查准率方面具有较强的优越性.

关键词: 语义相似度; 语义关系; 语义距离; 树状结构

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)05-0038-04

Analysis on Method of Semantic Similarity Computation Based on Tree Structure

HAN Xin^{1,2}, FAN Yong-sheng¹, MA Chun-sen², YANG He-ping²

(1 School of Electronics and Computer Science and Technology, North University of China, Taiyuan 030051, China;

2 Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China)

Abstract: The paper proposes a new semantic similarity calculation method which is based on tree structure. It combines factors that affect semantic similarity such as the semantic relation between concept nodes, semantic distance, the depth of concept nodes and node density. It uses the tree hierarchy to express the semantic information between concept nodes, and improve the traditional method to compute the density of concept nodes, also joined the adjustable parameters to adapt to different scenarios. At last it is verified this algorithm has a strong advantage by experiments.

Key words: semantic similarity; semantic relation; semantic distance; tree structure

1 引言

语义相似度的计算是语义检索的关键技术, 通过提高语义相似度的计算精度来尽可能地提高信息的查全率和查准率, 从而满足用户查询的需求.

目前, 对于语义相似度计算方法的研究已经比较成熟, 文献[1]提出广义相似度的定义, 并且给出了比较合理的计算方法和公式. 文献[2]认为语义相似度就是两个词语在不同的上下文中可以互相替换使用并且不改变文本句法语义结构的程度, 并对语义相似度进行量化, 指出相似度是一个数值, 取值范围在 0 到 1 之间. AH Doan^[3]采用联合分布概率统

计的方法来确定概念之间的语义相似度, 这种方法在概念实例的范畴划分方面, 很容易忽略一些概念的相似度. AKY Wong^[4]用子概念相似度来计算概念相似度, 但是在这种方法中子概念的权重很难确定. 李鹏、陶兰^[5]等提出语义相似度随语义距离增加而减小, 但在计算语义相似度时并没有考虑语义关系对语义相似度的影响, 这将影响语义相似度的计算精度.

针对上述不足, 本文选用树状结构模型, 利用概念节点的层次关系分析概念节点之间的语义关系、语义距离及概念节点的深度与密度对语义相似度值的影响, 来计算概念节点之间的语义相似度.

收稿日期: 2011-09-20; 修回日期: 2011-10-27

基金项目: 国家自然科学基金项目(50475176, 50675015); 北京市自然科学基金项目(4102020)

2 语义相似度

如果两个实体概念所拥有的相同属性越多,则说明它们的语义相似度越高,同样对于两个概念的某一个相同属性,如果其拥有越多的相同属性值,其语义相似度也就越高。

对于两个实体元素 x, y , 用 $\text{sim}(x, y)$ 表示 x, y 之间的语义相似度,形式上,语义相似度计算应满足^[3]:

- (1) 语义相似度的值为 $[0, 1]$ 区间中的一个实数,即 $\text{sim}(x, y) \in [0, 1]$;
- (2) 如果两个概念是完全相似的,则语义相似度为 1,即 $\text{sim}(x, y) = 1$ 当且仅当 $x = y$;
- (3) 如果两个概念没有任何共同特征,那么语义相似度为 0,即 $\text{sim}(x, y) = 0$;
- (4) 相似关系是对称的,即 $\text{sim}(x, y) = \text{sim}(y, x)$ 。

Dekang Lin^[1]认为任何两个词语的语义相似度取决于它们的共性(Commonality)和个性(Differences),然后从信息论的角度给出了语义相似度的计算公式:

$$\text{Sim}(A, B) = \frac{\log p(\text{Common}(A, B))}{\log p(\text{Description}(A, B))}$$

其中,分子表示描述 A, B 共性所需要的信息量,分母表示完整地描述 A, B 所需要的信息量。

刘群、李素建^[2]基于实例的机器翻译,提出两个词语之间的语义相似度的解释,即两个词语,如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大,二者的相似度就越高,否则相似度就越低。

3 影响树状结构语义相似度计算的因素

树的层次结构可以清晰地显示节点所处的深度及节点的密度,若用节点代表概念,概念之间的语义关系便一目了然。在树中,任意两个节点之间有且只有一条最短路径,这条最短路径就可以作为这两个节点之间的语义距离的一种度量。随着概念节点所处的深度变深,概念所包含的语义信息就越丰富,越能准确地描述概念的性质,因此,概念节点的深度对语义相似度起着决定作用。树中任意两个概念节点之间的边是等长的,所以任意两个概念节点之间的距离可以通过这两个概念节点之间边的数目来表示。

在这种树状结构模型中,用节点来表示实体概

念,连接各节点之间的边来表示实体概念之间的关系。影响语义相似度计算的主要因素包括:语义关系^[6]、语义距离、节点密度、节点深度。

4 基于树状结构的语义相似度计算方法

基于参考文献中所提到的语义距离、节点密度、节点深度这三个影响因素,对节点密度的计算方法进行改进,融入语义关系对语义相似度的影响,提出如下计算语义相似度的公式:

$$\text{Sim}(A, B) = \lambda \text{Relation}(A, B) + \alpha(A, B) + \beta \text{Depth}(A, B) + \gamma \text{Density}(A, B)$$

(1) $\text{Relation}(A, B)$ 表示两个节点概念之间的语义关系,语义关系通过概念节点之间的语义类型来计算,概念节点之间的语义类型主要有 *Synonym*、*Is-a*、*Part-a*。其中 *Synonym* 表示同义关系, *Is-a* 表示父子关系, *Part-a* 表示部分关系。一般主要考虑 *Is-a*、*Part-a* 这两种语义类型,而且 *Is-a* 比 *Part-a* 对概念节点的语义贡献大。且 A 与 B 之间的语义关系为:

$$\text{Relation}(A, B) = \frac{\sum_{\theta \in C_A \cap C_B} (S_A(\theta) + S_B(\theta))}{\sum_{\theta_A \in C_A} S_A(\theta_A) + \sum_{\theta_B \in C_B} S_B(\theta_B)}$$

其中, C_A 表示 A 与 A 的所有祖先节点所组成的集合, S_A 表示 A 的祖先节点对 A 的语义贡献,距离 A 较近的节点对 A 的语义贡献大,而距离 A 较远的节点对 A 的语义贡献小,且对任意一个属于 C_A 的节点 θ 的 $S_A(\theta)$ 定义如下:

$$\begin{cases} S_A(A) = 1 \\ S_A(\theta) = \text{AVG}\{\sigma * S_A(\theta_c) \mid \theta_c \in \text{childrenof}(\theta)\} \theta \neq A \end{cases}$$

式中, σ 表示节点之间语义联系类型的贡献因素,一般 $0 < \sigma < 1$ 。若两个节点之间是父子关系时 $\sigma = \text{Is-a}$; 为部分关系时 $\sigma = \text{Part-a}$ 。由此可知:若 S_A 与 S_B 的差别越小,则说明 A 与 B 的相似度越高。

(2) $\text{Dist}(A, B)$ 表示两个概念节点之间的语义距离,由于语义距离跟语义相似度成反比关系,所以构建如下计算方法:

$$\text{Dist}(A, B) = e^{-\text{dist}(A, B)}$$

两个概念之间的语义距离,是指在树状结构中连接两个概念节点的最短路径所跨的边数,语义距离是衡量概念之间相似度的条件之一。语言学家研究认为,两个词语的语义距离越大,其相似度越低。特别地当两个词语距离为 0 时,其相似度为 1; 当两个词语距离为无穷大时,其相似度为 0。可见,语义

相似度与语义距离的关系是反比关系,而且两个概念之间的语义距离表达式为:

$$\text{dist}(A,B) = \text{dist}(A,C) + \text{dist}(B,C)$$

式中, $\text{dist}(A,B)$ 表示概念 A 与 B 之间的语义距离, C 为 A,B 的最近共同祖先节点.

(3) $\text{Depth}(A,B)$ 表示两个概念节点的节点深度之间的关系,表达式为:

$$\text{Depth}(A,B) = \frac{|\text{depth}(A) - \text{depth}(B)| + 1}{\text{depth}(A) + \text{depth}(B)}$$

节点深度是指概念节点与根节点的最短路径中所包含的边数. 在树状结构中,每一层概念都是对上一层概念的细化,所以在语义距离相同的前提下,两个节点的深度差越大,概念之间的相似度越小;两个节点的深度差越小;概念之间的相似度越大. 概念所处的节点深度为:

$$\text{depth}(A) = s + 1$$

式中, $\text{depth}(A)$ 表示概念 A 的深度, s 表示概念 A 与根节点的最短路径中所包括的边的数目,根节点的深度为 1.

(4) $\text{Density}(A,B)$ 表示两个概念节点之间的节点密度,节点密度是指两个概念最近共同祖先的子节点的密度. 树状结构中不同地方节点的密度是不同的,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低;反之越高. 节点密度的表达式为:

$$\text{Density}(A,B) = \frac{n}{m}$$

式中, $\text{Density}(A,B)$ 表示 A 与 B 的最近共同祖先节点的节点密度, n 表示概念 A,B 的最近共同祖先的直接子节点的个数, m 表示以概念 A,B 的最近共同祖先为根节点所组成的子图中的所有节点数.

其中, $\lambda, \alpha, \beta, \gamma$ 为调节因子,且须满足 $\alpha + \beta + \gamma + \lambda = 1$. 一般情况下,语义距离对语义相似度的影响大于节点深度、语义关系和节点密度对其的影响,所以 $\alpha > \beta, \alpha > \gamma, \alpha > \lambda$.

5 实验分析

选取图 1 一个简单的树状结构来计算语义相似度,如图 1 所示.

选取该树状结构中的“显示器”概念节点来计算其与该树状结构中其他概念节点之间的语义相似度值,选参数 $Is-a=0.8, Part-a=0.5$ 及设定 $\lambda=0.1, \alpha=0.5, \beta=0.2, \gamma=0.2$ 来计算所查概念与其它概念之间的语义相似度值,如表 1 所示.

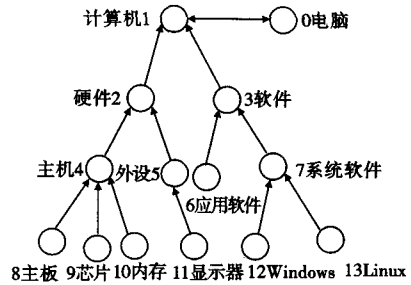


图 1 简单的树状结构实例

表 1 语义相似度实验结果比较

序号	查询概念	概念名称	本文中算法 的计算结果	文献[7]中算 法的结果
1	显示器	外设	0.479085	0.571
2	显示器	硬件	0.453373	0.380
3	显示器	主机	0.426221	0.301
4	显示器	计算机	0.348835	0.285
5	显示器	软件	0.267711	0.238
6	显示器	主板	0.18914	0.205
7	显示器	芯片	0.18914	0.205
8	显示器	内存	0.18914	0.205
9	显示器	系统软件	0.178515	0.205
10	显示器	应用软件	0.178515	0.205
11	显示器	Windows	0.120624	0.184
12	显示器	Linux	0.120624	0.184

通过表 1 可以看出:在文献[7]的结果中,应用软件、系统软件、主板、芯片、内存这 5 个概念与显示器都有相同语义相似度值,这显然与实际情况有所偏差,而本文计算方法所得到的结果可以很清楚地表现出这五个概念与显示器之间的语义相似度还是有明显区别的,同时可以看出其他概念与所查询概念之间的语义相似度值也比较准确,也符合人的主观判断,该方法除了文献[7]中所考虑的语义距离,节点密度,节点深度这三个因素的基础上,还将语义关系也作为一个考虑因素,并对节点密度的计算进行了改进,最后通过实验验证该方法有较好的准确性.

6 结束语

本文针对树状结构的结构特点,利用树状结构的层次关系,提出了树状结构中概念之间的语义相似度计算方法. 该计算方法综合考虑了语义关系、语

义距离、概念节点深度及概念节点密度等因素,并加入了调节因子,通过调节因子来表现出这些因素对计算语义相似度的权重及影响程度,且其中调节因子的设置可以使该计算方法在不同的情景的灵活使用。但是,这种计算方法也有一定的局限性,例如:在这种树状结构中,语义关系只是考虑该概念节点的直接父节点及同一层中与其相邻的概念,并没有考虑其他概念,所以在这个方面该计算方法还需要继续改进。

参考文献:

- [1] Lin D. An information-theoretic definition of similarity [C]//Proceedings of 15th International Conference on Machine Learning. USA San Francisco, 1998: 296-304.
- [2] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics Chinese Language Processing, 2002, 7(2): 59-76.
- [3] Doan A H, Madhavan J, Domingos P. Learning to map between ontologies on the semantic web [C] // Proceedings of the 11th International Conference on World Wide Web. USA: New York, ACM Press, 2002: 662-673.
- [4] Wong A K Y, Ray P, Waran N P. Ontology mapping for the interoperability problem in network management [J]. IEEE Journal on Selected Areas in Communication, 2005, 23(10): 2058-2068.
- [5] 李鹏,陶兰,王弼左. 一种改进的语义相似度计算及其应用[J]. 计算机工程与设计, 2007, 28 (1): 227-229.
- [6] 魏韡,向阳. 一种基因本体术语间的语义相似度计算方法[J]. Computer Engineering and Applications 计算机工程与应用, 2010, 46(26): 114-116.
- [7] 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.

作者简介:



韩欣女, (1986—), 硕士研究生。研究方向为计算机应用技术。

(上接第37页)

- [6] Wei Ye, John Heidemann, Deborah Estrin. An energy-efficient MAC protocol for wireless sensor networks [C]//Proceedings of Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. New York: Proceedings of the IEEE, 2002: 1567-1576.
- [7] Michael Buettner, Gary Yee, Eric Anderson, et al. X-MAC: A short preamble MAC protocol for duty-cycled wireless sensor networks [C]//Proceedings of the 4th international conference on Embedded networked sensor systems. New York: ACM Association for Computing Machinery, 2006: 307-320.
- [8] Guo S, Gu Y, Jiang B, et al. Opportunistic flooding in low-duty-cycle wireless sensor networks with unreliable links [C]//Proceedings of the 15th annual international conference on Mobile computing and networking. New York: ACM Association for Computing Machinery, 2009: 133-144.
- [9] Zhao Y, Govindan J, Estrin R D. Residual energy scan for monitoring sensor networks [C] // Wireless Communications and Networking Conference. Orlando: Proceedings of the IEEE, 2002: 356-362.
- [10] Texas Instruments. 2.4 GHz IEEE 802.15.4/ZigBee-ready RF transceiver (Rev. B) [EB/OL]. [2007-03-20]. <http://focus.ti.com/docs/prod/folders/print/cc2420.html>.

作者简介:

王浩男, (1985—), 硕士研究生。研究方向为无线传感器网络。

杨秋翔男, (1969—), 教授, 硕士生导师。研究方向为软件设计方法及应用、网络技术应用。

杨剑男, (1979—), 博士研究生。研究方向为计算机视觉及图像处理。

伍玉良男, (1985—), 硕士研究生。研究方向为信息隐藏、嵌入式系统。