基于知网的词语语义相似度改进算法

李 蕾,杨丽花
(南京邮电大学 江苏省无线通信重点实验室,江苏 南京 210003)

摘 要:词语语义相似度计算在很多领域都有广泛应用,而目前常用的基于知网的词语语义相似度计算方法由于未深入考虑同一棵树中的两个不同义原的可达路径上所有义原节点的密度对义原距离的影响,或未考虑义原深度与义原密度的主次关系,导致计算结果不够精确,从而使其应用受限。针对该问题,给出了一个新的节点间边权重函数,通过在边权重函数中引入两义原可达路径上所有义原节点的密度,并利用权重因子来调整义原深度和义原密度对义原距离的影响,从而提出一种改进的基于知网的词语语义相似度计算方法。实验结果表明,该方法可以更有效地提高词语语义相似度计算精度,比现有方法更具有实用性。

关键词:知网;词语语义相似度;义原密度;义原深度;义原距离
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2019)04-0042-05
doi:10.3969/j.issn.1673-629X.2019.04.009

Improved Algorithm of Word Semantic Similarity Based on HowNet

LI Lei, YANG Li-hua
(Key Laboratory of Wireless Communication of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Semantic similarity of words has been widely used in many fields. The current word semantic similarity calculation method based on HowNet does not deeply consider the influence of the density of all the semantically original nodes on the reachable path of two different sememes in the same tree, or also does not consider in depth with the primary and secondary relationship between the density and the depth of sememe, which causes the calculation result to be inaccurate. To solve the problem, we propose a new method using the density of all sememe nodes on the reachable path in the edge weight function, and the proposed method employs the weight factor to adjust the influence of the depth of sememe and the density of sememe on the distance of sememe. The simulation shows that the proposed algorithm can effectively improve the accuracy of the semantic similarity calculation of words, and is more practical than the existing methods.

Key words: HowNet; semantic similarity of words; density of sememe; depth of sememe; distance of sememe

0 引 言

词语语义相似度计算在信息检索[1]、基于实例的机器翻译[2]以及数据相似度检测等领域有着广泛的应用。目前常用的基于知网的词语语义相似度计算方法大致可分为两类:一类是利用大规模语料统计词语的相关性,即基于统计的方法;另一类是根据某种世界知识计算相似度的方法,即基于世界知识的方法[3]。其中,基于统计的方法是根据词汇上下文信息的概率分布计算词语语义相似度,该方法计算得到的结果精确度较高,但是需要依赖于训练所用的语料库,计算量比较大,计算方法也比较复杂。此外,由于数据稀疏和数据噪声等因素对基于统计的方法干扰较大,故该方法一般很少使用[4]。基于世界知识的方法通常是基于某个知识完备的语义词典中的层次结构关系进行计算,该方法简单有效,不需要用语料库进行训练,也比较直观和易于理解,但这种方法受人的主观意识影响较大,有时并不能准确反映客观事实[5]。

知网是一个以汉语和英语词语所代表的概念为描述对象、以揭示概念与概念之间以及概念所具有属性之间的关系为基本内容的常识库和知识库。基于知网

收稿日期:2018-05-28 **修回日期**:2018-09-29 **网络出版时间**:2018-12-20
基金项目:国家自然科学基金(61401232,61671251,61501254);江苏省自然科学基金(BK20140894)
作者简介:李 蕾(1994-),男,硕士研究生,研究方向为数据预处理方法、语义分析等;杨丽花,讲师,硕导,通讯作者,研究方向为高速移动通信、大数据、物联网频谱等。

的词语语义相似度计算最终可以归结于义原相似度计算的层面上。例如,文献[6]提出了一种根据义原距离计算词语语义相似度计算方法;文献[7]在考虑义原距离的基础上进一步考虑了义原深度对词语语义相似度的影响;文献[8]提出了一种同时考虑义原深度和义原密度的方法;文献[9]考虑了义原间的反义对义关系及文本情感色彩对词语语义相似度的影响;文献[10]提出一种考虑了义原的公共节点个数和义原深度的词语语义相似度计算方法;文献[11]根据差异以及共有信息进行词语语义相似度的计算;文献[12]提出了一种考虑词语词性因素的词语语义相似度计算方法;文献[13]将关系义原和关系符号义原进行加权合并,提出补充义原是对基本义原的语义补充,并且考虑了最小公共父节点的影响[14];文献[15]深入考虑义原之间的距离和义原层次深度的主次关系;文献[16]提出结合知网与同义词词林两个知识库的词语语义相似度算法。

然而,目前常用的基于知网的词语语义相似度计算方法未深入考虑同一棵树中的两个不同义原的可达路径上所有义原节点的密度对义原距离的影响,且也未考虑义原深度与义原密度的主次关系。对此,文中提出了一种改进的基于知网的词语语义相似度算法。

1　知网简介

知网中包含着丰富的词语语义知识和本体知识,揭示了概念与概念之间以及概念所具有的属性之间的关系。

知网中主要包含义项和义原两个概念,义项是对词语语义的一种描述,每一个词可以表达为几个义项,它是用一种知识表示语言来描述的。在知网中,每个汉语词语的一个义项由一个四元组构成:〈W_X=词语 E_X=词语例子 G_X=词语词性 DEF=概念定义〉,其中 DEF(语义表达式)是义项的主体,由一个个结合知识描述符号的基本义原组成,每个义原使用逗号隔开,例如义项"男人"的 DEF="human|人,family|家,male|男"。

知网中义原对于义项的描述是通过一种结构化的知识描述语言进行定义的,这种知识描述语言所用的词汇叫做义原,义原是用于描述一个义项的最小意义单位,它是从所有词语中提炼出的,并且可以用来描述其他词语的不可再分的基本元素。义项与义原之间的结构关系如下所示[6]

$$
\text{义项}\left[\begin{array}{l}
\text{第一基本义原描述} = \text{基本义原}_a \\
\text{其他基本义原描述} = \{\text{基本义原}_b,\text{基本义原}_c \cdots\cdots\} \\
\text{关系义原描述} = \left[\begin{array}{l}
\text{关系义原}_1 = \text{基本义原}_x \mid \text{具体词}_x \\
\text{关系义原}_2 = \text{基本义原}_y \mid \text{具体词}_y \\
\cdots\cdots
\end{array}\right] \\
\text{关系符号描述} = \left[\begin{array}{l}
\text{关系符号}_1 = \{\text{义原}_u \mid \text{具体词}_u,\text{义原}_v \mid \text{具体词}_v \cdots\cdots\} \\
\text{关系符号}_2 = \{\text{义原}_s \mid \text{具体词}_s,\text{义原}_t \mid \text{具体词}_t \cdots\cdots\} \\
\cdots\cdots
\end{array}\right]
\end{array}\right]
$$

2　词语语义相似度计算

2.1　基本定义

词语语义相似度不能根据明确的客观标准进行衡量,因此它是一个主观性比较强的概念。文中采用文献[6]所理解的词语语义相似度含义,即两个任意的词语如果在不同的上下文中可以互相替换且不改变文本语义的可能性越大,那么两者之间的相似度就越高,否则相似度就越低。

在知网中,一个词语可以表达为几个义项,比如:对于两个汉语词语 w_1 和 w_2,若 w_1 有 n 个义项,即 s_{11},s_{12},…,s_{1n};w_2 有 m 个义项,即 s_{21},s_{22},…,s_{2m},则词语 w_1 和 w_2 的语义相似度可以表示为:

$$\text{sim}(w_1,w_2) = \max_{i=1,2,\cdots,n,j=1,2,\cdots,m} \text{sim}(s_{1i},s_{2j}) \qquad (1)$$

其中,$\text{sim}(w_1,w_2)$ 表示词语之间的相似度;$\text{sim}(s_{1i},s_{2j})$ 表示义项之间的相似度。

根据式 1 可知,两个词语之间的语义相似度可以由两个义项之间的相似度的最大值表示。而所有的义项最终是用义原来表示,因此义项之间的相似度可表示为[7]:

$$\text{sim}(s_1,s_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} \text{sim}_j(p_1,p_2) \qquad (2)$$

其中,$\beta_1(1 \leqslant i \leqslant 4)$ 是可调节参数,且满足:$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$,$\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4$。

由于义项相似度是由义原相似度计算来的,因此,基于知网的词语语义相似度计算最终可以归结于义原相似度计算的层面上。

2.2　义原相似度计算

义原中提供了事件类、属性类、实体类、属性值类、次要特征类、数量类、数量值类、语法类、动态角色类与动态属性类等十棵义原层次树,它们相互之间不存在交集。对于处于同一棵树的两个不同义原,有且仅有一条长度为 N 的可达路径,并且不同树的义原之间没

有可达路径,在此定义两义原间路径的长度为义原的语义距离,即义原距离。一般而言,义原距离越小,义原相似度越大。

文中选取了一个以“实体”为根节点的义原层次体系树的分支(如图 1 所示),并根据该图,对三种常见的义原相似度计算方法进行分析。

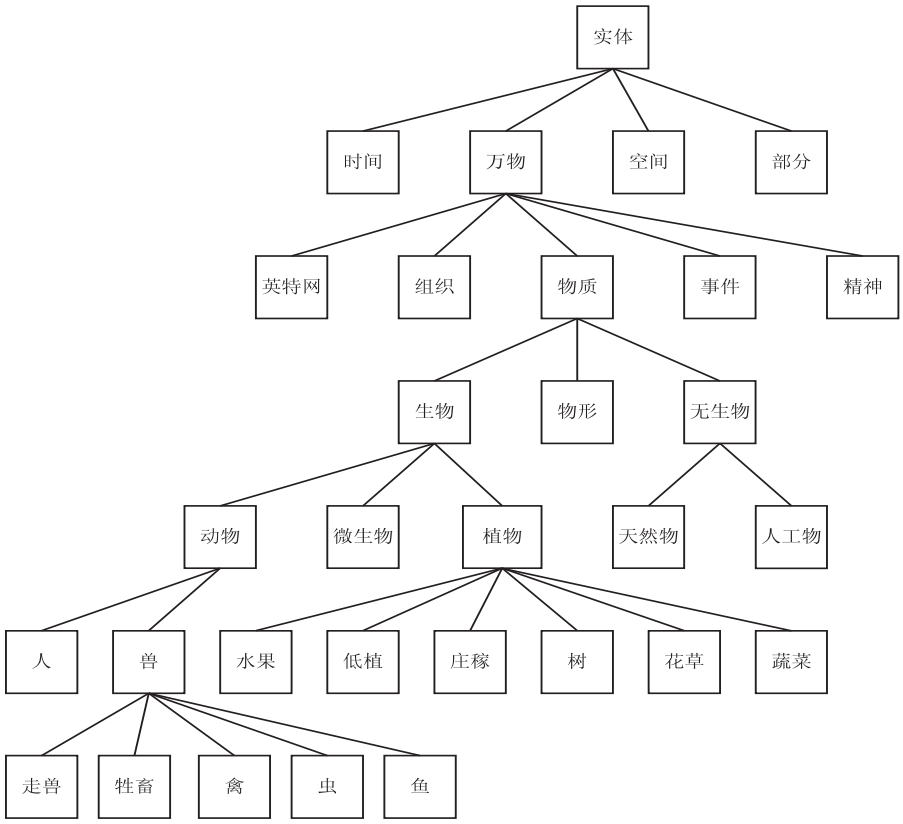

图 1 以“实体”为根点的树分支示意

刘群等提出通过计算两义原节点之间的义原距离计算两个义原之间的相似度[6],即

$$\mathrm{sim}(p_1,p_2)=\frac{\alpha}{\alpha+\mathrm{dis}(p_1,p_2)} \tag{3}$$

其中,p_1 和 p_2 表示义原,$\mathrm{dis}(p_1,p_2)$ 为 p_1 和 p_2 在义原层次树中的义原距离,当 p_1 和 p_2 处于不同的树时,$\mathrm{dis}(p_1,p_2)$ 将取一个较大常数;α 为可调参数。

该方法根据知网中义原之间的上下位关系对词语语义相似度进行计算,但是计算得到的词语语义相似度结果过于粗糙,准确度不高。

文献[13]提出一种根据义原深度和义原密度来计算义原相似度的方法。其中义原深度是指义原层次体系树上的根节点到该义原节点的路径长度,义原深度越小,义原表达的概念越抽象,义原深度越大,义原表达的概念越具体;义原密度是指该义原节点的兄弟节点的个数(含自身),义原密度越大,意味着分类越细致,描述越详细,其携带的语义信息越丰富,位于高密度区域的节点对义原距离小。利用义原深度和义原密度获得的义原相似度计算公式为:

$$\mathrm{sim}(p_1,p_2)=\frac{1}{2}\times\frac{\alpha}{\alpha+\sum_{i=1}^{N}\mathrm{weight}(\mathrm{level}(i))}+\frac{1}{2}\times\frac{2\times\log f(\mathrm{LCN})}{\log f(p_1)+\log f(p_2)} \tag{4}$$

其中,N 是义原 p_1 和 p_2 之间可达路径上的长度;$\mathrm{level}(i)$ 为两个义原可达路径上边 i 在义原层次树中的层次;LCN 是两个义原在树中的最小公共父节点[14];$f(\cdot)$ 函数反映了当前义原所在树中的密度信息,其值为当前义原的兄弟节点的个数(含自身)与树的总节点个数的比值;weight 函数是一个随层数 k 递增而单调递减的函数,表示每一条边的权重,定义为:

$$\mathrm{weight}(k)=\frac{2\times(\mathrm{depth}-k)}{\mathrm{depth}\times(\mathrm{depth}+1)} \tag{5}$$

其中,depth 为当前义原层次树的树高。

文献[13]虽然在义原距离的计算公式中引入了一个随层数递增而单调递减的边权重函数,但该函数采用的是线性递减策略,顶部边权重衰减过快,不符合知网层次结构的特点。为了避免该现象,文献[16]引入了一个正弦三角函数,即

$$\mathrm{weight}(k)=\frac{\mathrm{depth}-1-k}{\mathrm{depth}-1}\times(1+\sin(\theta*k*\pi/180)) \tag{6}$$

其中,θ 是调节参数,与树高 depth 成反比。

文献[16]虽然改进了义原距离计算公式中的边

权重函数,但却未考虑义原密度对义原相似度的影响。

3　提出的语义相似度计算方法

现有的义原相似度计算方法虽然考虑了义原树中义原深度和(或)密度对义原相似度的影响,但仅仅只考虑了当前所要计算的义原节点以及它们的最小公共父节点的密度对义原相似度的影响,而未考虑两个义原节点可达路径上所有节点的密度对义原距离的影响。为此,文中提出一种基于知网的词语语义相似度改进方法,该方法基于知网语义词典,通过将义原深度和义原节点间所有节点密度进行联合,并利用权重因子来权衡义原深度和义原密度对义原相似度的影响,获取新的义原相似度计算公式,根据义原、义项与词语之间的关系,最终得到改进的词语语义相似度计算表达式。

在义原树中,影响义原距离的因素有义原深度和义原密度,一般而言,义原深度越大,义原距离越小;义原密度越大,义原距离越小。该方法通过在边权重函数中引入义原可达路径上所有义原节点密度对义原距离的影响,给出了一个新的边权重函数表达式,即

$$\text{weight}(i_{p,q}) = c_1 \times \frac{\text{depth} - 1 - k_p}{\text{depth} - 1} \times (1 + \sin(\theta * k_p * \pi/180)) + c_2 \times (1 - \frac{\log f(p)}{\log \max}) \quad (7)$$

其中,$i_{p,q}$ 为义原 p 与 q 之间的边,p 表示当前义原节点,q 是当前义原节点 p 的上一层父节点;k_p 为当前义原节点 p 所在层的编号;θ 是调节参数,与树高 depth 成反比,文中取 $\theta = 4$;$f(\cdot)$ 函数反映了当前义原所在树中的密度信息,其值为当前义原的兄弟节点的个数(含自身);max 表示当前义原树中所有义原节点的总个数;c_1 和 c_2 为权重因子,其主要是用来权衡义原深度和义原密度对义原距离的影响。

利用式 8 给出的新的边权重函数,得到义原 p_1 与 p_2 之间的距离为:

$$\text{dis}(p_1, p_2) = \sum_{p=p_1, p \neq G}^{p_2} \text{weight}(i_{p,q}) \quad (8)$$

其中,G 是义原 p_1 与 p_2 的公共父节点。再分别根据式 3、式 2 和式 1,从而最终可获得两词语的语义相似度。

4　实验与分析

考虑到义原深度和义原密度对义原距离的影响不同,文中方法引进了权重因子 c_1 和 c_2 来权衡二者的影响。在此设置 4 组不同的权重因子组合来计算义原距离,通过比较义原距离,选取出符合该方法中的最佳权重因子 c_1 和 c_2,实验结果如表 1 所示。在表 1 中,当($c_1 = 0.4$, $c_2 = 0.6$)和($c_1 = 0.5$, $c_2 = 0.5$)时,“兽”和“人”的义原距离要大于“动物”和“植物”之间的义原距离,这显然和实际情况不相符合,在以“属性和属性值”为根节点的义原层次体系树中,“味道”和“气味”的义原距离要小于“酸”和“甜”之间的义原距离,这与实际情况是相符合的,但是当权重因子分别取($c_1 = 0.4$, $c_2 = 0.6$)和($c_1 = 0.5$, $c_2 = 0.5$)时,两者的义原距离都相差比较大,而权重因子取($c_1 = 0.7$, $c_2 = 0.3$)时的义原距离比较适中,在以“实体”为根节点的义原层次体系树中(如图 1 所示),“走兽”和“牲畜”、“花草”和“树”这两对义原对同是树中的叶子节点,因此这两对义原对的义原距离相差不大。因此,当 c_1 和 c_2 分别 0.7 和 0.3 时,得到的义原相似度更符合实际情况。

表 1　不同权重因子情况下的义原距离比较

义原对		权重因子(c_1, c_2)			
		(0,4,0.6)	(0,5,0.5)	(0.7,0.3)	(0.8,0.2)
动物	植物	1.765	1.780	1.812	1.827
无生物	生物	1.793	1.815	1.860	1.883
天然物	人工物	1.831	1.835	1.845	1.849
走兽	牲畜	1.60	1.608	1.626	1.634
兽	人	1.794	1.789	1.780	1.775
花草	树	1.615	1.641	1.691	1.716
人	动物	0.897	0.895	0.890	0.888
生物	动物	0.882	0.890	0.906	0.914
关系	归属	0.808	0.838	0.898	0.928
次序	有序	0.929	0.935	0.947	0.952
味道	气味	1.503	1.582	1.739	1.818
酸	甜	1.710	1.745	1.819	1.855

为了验证该方法,将其与现有词语语义相似度计算方法进行了仿真验证,如表 2 所示。

在表 2 中,方法 1 是文献[6]所提方法,方法 2 是文献[13]所提方法,方法 3 是文献[16]所提方法。在仿真中,$\alpha = 1.6$, $\beta_1 = 0.5$, $\beta_2 = 0.2$, $\beta_3 = 0.17$, $\beta_4 = 0.13$, $\gamma = 0.2$, $\delta = 0.2$。可看出,方法 1 中“男人”(取义项“human|人,family|家,male|男”)与“女人”(取义项“human|人,family|家,female|女”)和“和尚”(取义项“human|人,religion|宗教,male|男”)的词语语义相似度是相同的,这是因为方法 1 中没有考虑义原层次树中节点的层次深度和密度对义原相似度的影响,但在实际情况中,义原“男”和“女”与“家”和“宗教”的相似度显然是不同的,所以方法 2、方法 3 和文中方法比方法 1 更能区别不同词语之间的语义相似度,但是其中有些词语语义相似度的计算结果也不太合理,如方法 2 中“女人”和“男人”的相似度大于“和尚”和

“男人”的相似度,这与人的直觉是不相符合的,因为“和尚”和“男人”都为男性,它们之间的相似度应该比“女人”和“男人”的相似度要高,并且根据方法 2 和文中方法中所定义的义原相似度计算方法,义原“家”和“宗教”的相似度比义原“男”和“女”的相似度大,所以“男人”和“女人”的相似度应该比“男人”和“和尚”的相似度要小。另外,从 6-11 这 6 组词语对的相似度计算结果中可以看出,方法 2 的结果要远大于方法 1、方法 3 和文中方法的结果,而在方法 1、方法 3 和新方法的词语语义相似度的计算结果中,6-11 这 6 组词语对的词语语义相似度计算结果差距较小,较为合理。因此,文中方法计算的词语语义相似度结果更加合理和准确。

表 2 词语语义相似度计算结果比较

词语 1	词语 2	方法 1	方法 2	方法 3	文中方法
男人	女人	0.861	0.915	0.863	0.865
男人	父亲	1.0	1.0	1.0	1.0
男人	母亲	0.861	0.915	0.863	0.865
男人	和尚	0.861	0.931	0.861	0.874
男人	经理	0.602	0.548	0.603	0.613
男人	高兴	0.048	0.028	0.048	0.049
男人	收音机	0.108	0.338	0.116	0.119
男人	鲤鱼	0.209	0.411	0.229	0.229
男人	苹果	0.171	0.388	0.183	0.186
男人	工作	0.113	0.318	0.117	0.121
男人	责任	0.126	0.362	0.131	0.136

5 结束语

利用义原深度与两义原间可达路径上所有义原节点的密度提出了一种改进的词语语义相似度计算算法,并通过一权重因子来调整义原深度与义原密度的主次关系。实验结果表明,该方法所计算的词语语义相似度结果更加合理和准确。

参考文献:

[1] TAIEBMA H, AOUICHAM B, BOUROUIS Y. FM3S: features-based measure of sentences semantic similarity[C]// International conference on hybrid artificial intelligent systems. [s. l.]:[s. n.],2015:515-529.

[2] LI Ying. Study and implementation on key techniques for an example based machine translation system [C]//Second IITA international conference on geoscience and remote sensing. Qingdao, China: IEEE, 2010:316-320.

[3] 吴旭东,成卫青,黄卫东. 改进的主客观结合的词语语义相似度算法[J]. 计算机技术与发展,2012,22(9):45-49.

[4] SLIMANI T. Description and evaluation of semantic similarity measures approaches[J]. International Journal of Computer Applications, 2013, 80(10):25-33.

[5] KERMICHE S, SAIDI M L, ABBASSI H A. Gradient descent adjusting Takagi-Sugeno controller for a navigation of robot manipulator[J]. Journal of Engineering and Applied Science, 2006, 1(1):24-29.

[6] 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.

[7] 李 峰,李 芳. 中文词语语义相似度计算-基于《知网》2000[J]. 中文信息学报,2007,21(3):99-105.

[8] 袁晓峰.《知网》义原相似度计算的研究[J]. 辽宁大学学报:自然科学版,2011,38(4):358-361.

[9] 江 敏,肖诗斌,王弘蔚,等. 一种改进的基于知网的词语语义相似度计算[J]. 中文信息学报,2008,22(5):84-89.

[10] 张振幸,李金厚. 一种基于义原重合度的词语语义相似度计算[J]. 信阳师范学院学报:自然科学版,2010,23(2):296-299.

[11] 刘青磊,顾小丰. 基于《知网》的词语相似度算法研究[J]. 中文信息学报,2010,24(6):31-36.

[12] 王小林,王 义. 改进的基于知网的词语相似度算法[J]. 计算机应用,2011,31(11):3075-3077.

[13] 葛 斌,李芳芳,郭丝路,等. 基于知网的词汇语义相似度计算方法研究[J]. 计算机应用研究,2010,27(9):3329-3333.

[14] LIN Dekang. An information-theoretic definition of similarity semantic distance in WordNet[C]//Proceedings of the 15th international conference on machine learning. San Francisco: Morgan Kaufmann Publishers Inc, 1998:296-304.

[15] 张沪寅,刘道波,温春艳. 基于《知网》的词语语义相似度改进算法研究[J]. 计算机工程,2015,41(2):151-156.

[16] 朱新华,马润聪,孙 柳,等. 基于知网与词林的词语语义相似度计算[J]. 中文信息学报,2016,30(4):29-36.