文章编号：1003-0077(2016)04-0029-08

基于知网与词林的词语语义相似度计算

朱新华，马润聪，孙 柳，陈宏朝

（广西师范大学 计算机科学与信息工程学院，广西 桂林 541004）

摘 要：该文提出了一种综合知网与同义词词林的词语语义相似度计算方法。知网部分根据义原层次结构的特征，采用了顶部平缓而底部陡峭的曲线单调递减的边权重策略，改进了现有的义原相似度算法；词林部分采用以词语距离为主要因素、分支节点数和分支间隔为微调节参数的方法，改进了现有的词林词语相似度算法。然后再根据词语的分布情况，采用综合考虑知网与同义词林的动态加权策略计算出最终的词语语义相似度。该方法充分利用了词语在知网与词林中的语义信息，极大地扩充了可计算词语的范围，同时也提高了词语相似度计算的准确率。

关键词：语义相似度；知网；同义词词林；语义距离

中图分类号：TP391 **文献标识码**：A

Word Semantic Similarity Computation Based on HowNet and CiLin

ZHU Xinhua, MA Runcong, SUN Liu, CHEN Hongchao

(College of Computer Science & Information Technology, Guangxi Normal University, Guilin, Guangxi 541004, China)

Abstract: A word semantic similarity computation method based on the HowNet and CiLin is proposed in this paper. First, according to the characteristics of sememe hierarchical structure, an edge weighting strategy of monotonic decreasing curve with flat top and steep bottom is used in the HowNet part. In the CiLin part, a special method of taking the distance between words as the main factor and the branch node quantity and branch interval as micro-adjustable parameters is used. Then, according to the distribution of words, a dynamic weighting strategy of considering both HowNet and CiLin is used to calculate the final similarity, which greatly expands the computable range of words and improves the computation accuracy of word similarity.

Key words: semantic similarity; HowNet; CiLin; semantic distance

1 引言

词语语义相似度的计算在信息检索、文本聚类、机器翻译、词义消歧和智能教学等领域有着广泛的应用。当前词汇语义相似度计算方法大致可分为两类：一类利用大规模语料库进行统计，依据词汇上下文信息的概率分布进行计算；另一类基于某种世界知识来计算，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[1]。无论是基于本体知识还是基于大规模语料库都有自己的优劣，具体要看应用环境才能选出最佳方案。基于世界知识的方法简单有效，无需用语料库进行训练，也比较直观，易于理解，但这种方法得到的结果受人的主观意识影响较大，有时并不能准确反映客观事实[2]。基于语料库的方法比较客观，综合反映了词语在句法、语义、语用等方面的相似性和差异。但是，这种方法比较依赖于训练所用的语料库，计算量大，计算方法复杂，另外，受资料稀疏和资料噪声的干扰较大[2]。在信息检索和文本聚类中一般用语料库的方法，机器翻译以及智能教学中一般采用基于世界知识的方法。

1.1 知网简介

知网是董振东先生花了数十年时间建设的一个汉语常识库，其设计目标是通过汉语词语意义的描述实现中英文机器翻译，目前仍在发展更新中。《知网》中与词语意义相关的概念有：义原、义项、语义

收稿日期：2014-06-25 **定稿日期**：2014-10-27

基金项目：国家自然科学基金(61363036)

表达式。义原是描述“概念”的基本单位[2]，也可以说是原子概念，其作用是用来对其他非义原“概念”进行描述。一个词语的一项解释叫作义项，一般的词语都会有多个义项，义项也可以叫作“概念”。在知网中每个汉语词语的一个义项由一个四元组构成[3]：<W_X=词语，E_X=词语例子，G_X=词语词性，DEF=概念定义>。

DEF(语义表达式)是义项的主体，它由一个个结合知识描述符号的基本义原组成，每个义原用逗号隔开，例如，义项“雇员”的 DEF=“human|人，$employ|雇用”，其含义为“可以被他人雇用的人称为雇员”。知网建设的初衷是为了解决机器翻译这一难题，因此义原的基本形式为“英语单词|汉语词”。义原在知网中分为事件、实体、属性、属性值、数量、数量值、次要特征、语法、动态角色与动态属性等十大类，共计 1 500 多个(2000 版)，义原根据上下位关系构建出树状结构，如图 1 所示[4]。

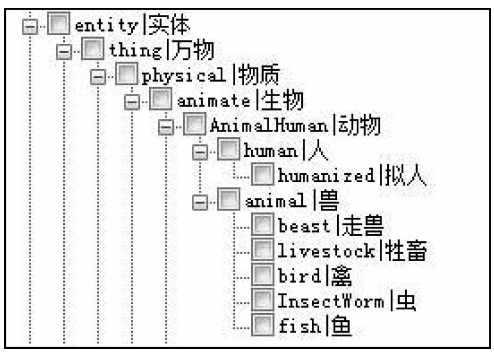


图 1 义原的树状层次结构

知网的建设还在不断的进行中，本文所说的知网无特殊说明，均指目前可在知网官方网站下载到的免费版本，主要部分是 2000 年版。

1.2 同义词词林简介

同义词词林是由梅家驹[5]等人于 1983 年编撰的可计算汉语词库，其设计目标是实现汉语同义词和同类词的划分和归类。同义词词林经哈尔滨工业大学信息检索研究室的扩展后，目前共有七万多个词语，这些词语被分为了 12 个大类，94 个中类，1 428 个小类，小类下方进一步划分为词群和原子词群两级[6]。这样，同义词词林的扩展版就具备了五层的树状结构。与知网中的树形结构不同，知网中每一个节点都是一个义原，同义词词林中上面四层的节点都代表抽象的类别，只有最底层的叶子节点才是一个个的词条，也有研究者称之为义项[7]，同一个词条可能在不同的类别中同时存在，也就是说词条的编码不是唯一的。第一至三大类多属名词，数词和量词在第四大类中，第五类多属形容词，第六至十类多是动词，十一类多属虚词，十二类是难以被分到上述类别中的一些词语。大类和中类的排序遵照从具体概念到抽象概念的原则[5]。

关于词条的编码如表 1 所示。第八位编码只有三种情况，“=”代表“相等”、“同义”。“#”代表“不等”、“同类”，属于相关词语。“@”代表“自我封闭”、“独立”，它在词典中既没有同义词也没有相关词[5]。前七位编码就可以唯一确定一条编码，即不存在这种情况：前七位编码相同而第八位不相同的多条编码同时存在。当前七位编码确定以后，第八位就是固定的，要么是“=”，要么是“#”，要么是“@”。例如，(导体，半导体，超导体)这一组同义词在词林中的编码为“Ba01B10#”。

表 1 词林中词语的编码结构

编码位	1	2	3	4	5	6	7	8
符号举例	A	a	0	1	B	0	2	=/#/@
性质	大类	中类	小类		词群	原子词群		
级别	第一层	第二层	第三层		第四层	第五层		

本文所使用的同义词词林来源于《哈工大信息检索研究室同义词词林扩展版》的 1.0 版本。

2 词语语义相似度的计算

2.1 改进的知网义原相似度计算

基于知网的词语相似度计算的是对两个词语的意义进行比较，其总体方法为：将词语相似度的计算转换为对词语义项语义表达式(DEF)的相似度计算，而义项 DEF 相似度的计算又可转换为对其中的义原进行相似度计算[2]，因此义原相似度是词语相似度计算的基础。

在义原树中影响义原相似度的因素有：义原距离、节点层次、节点密度等语义信息。义原距离和节点层次与义原相似度成反比，而节点密度与相似度成正比，即二个义原距离越大相似度越低；在路径长度相同的情况下，节点对所处层次越高，差异性越大，相似度也就越低；密度越大的地方说明分类越细，其同距离路径的语义距离也就越小[1]。

刘群等[2]提出了将义原距离转化为相似度的计算如式(1)所示。

$$sim(s_1,s_2)=\frac{\alpha}{dis(s_1,s_2)+\alpha} \quad (1)$$

其中 s_1 和 s_2 代表两个义原，$dis(s_1,s_2)$ 为 s_1 和 s_2 的语义距离，其值等于 s_1 和 s_2 在义原层次体系中的路径长度，α 为相似度约为 0.5 时的义原距离，在文献[2]中 α 取值为 1.6。在式(1)中，连接所有层次的边的权重都设为 1，因此没有考虑节点的层次与密度对相似度的影响。

为提高义原距离计算的合理性，文献[1]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，造成义原距离的计算结果与文献[2]的偏离过大，同时也不符合知网层次结构的特点。

在知网的义原层次结构中，顶部层都为大类且节点密度都相对低，而底部层都为小类且节点密度都相对高。根据该层次结构的特征，本文在加权距离算法中采用了顶部平滑而底部陡峭的曲线单调递减的边权重函数，如式(2)、式(3)所示。

$$dis(s_1,s_2)=\sum_{i=1}^{n} weigth(level(k)) \quad (2)$$

在式(2)中，设 s_1 和 s_2 的最短可达路径上共有 n 条边，$level(k)$ 代表第 k 条边上父节点在树形结构中的层次编号，并设根节点的层次编号为 0。

本文在边权重函数中，引入了一个正弦三角函数来修正文献[1]中顶部边权重随层数递增而衰减过快的现象，如式(3)所示。

$$weight(i)=\frac{m-1-i}{m-1}\cdot(1+\sin(\theta*i*\pi/180)) \quad (3)$$

其中，m 代表树的层次数，在知网中 m=14，即义原树层高为 14；θ 为一个与层高 m 成反比的调节参数，在不同的层高下，θ 的取值必须在不同的范围之内，以确保边权重函数的单调递减性，经测试，当 m=14 时，θ 取 4 比较理想；i 为一个正整数，代表节点的层次编号，$0\leqslant i\leqslant m-2$；$\pi$ 为圆周率。$weight(i)$ 代表的是第 i 层节点与第 i+1 层节点连接边的权重。

经实验对比，使用上式可以得到比文献[1]中顶部衰减更为平缓的边权重单调递减，如表 2 所示。

表 2　式(3)与文献 1 的对比(m=14，θ=4，结果取二位有效数)

层次编号 i	0	1	2	3	4	5	6	7	8	9	10	11	12
本文公式边权重	1	0.99	0.96	0.93	0.88	0.83	0.76	0.68	0.59	0.49	0.38	0.26	0.13
文献 1 最大值归 1 后的边权重	1	0.93	0.86	0.79	0.71	0.64	0.57	0.50	0.43	0.36	0.29	0.21	0.14

注：表 2 所列的边权重是指在同一颗义原大类树中连接不同层次节点的边的权重，当两个义原不在一颗大类子树中时，本文直接将两个义原的距离处理为 20。

2.2　基于知网的词语相似度计算

根据文献[2]的思想与方法，在知网中词语相似度的计算可以转换为对词语语义表达式(DEF)的相似度计算。刘群[2]将义项的语义表达式 DEF 划分为四个部分。排在最前面的是义项的第一基本义原，它刻画的是义项的本质属性。以符号如“～!@#$%&.*”等开头的是关系符号义原描述式。包含“=”号的是关系义原描述式。剩余的就是其他基本义原构成的描述式集合。江敏[8]把第一基本义原和其他基本义原合并在一起称为独立义原，本文借用该思想，将义项相似度的计算转化成对独立义原集合、关系义原特征结构与关系符号义原特征结构的相似度计算，具体方法为：

(1) 独立义原构成集合，其相似度的计算以 2.1 中所描述的义原相似度算法为基础，利用文献[1]和文献[9]中的二部图最大权匹配方法，算出其相似度。本部分记为 $sim_1(C_1,C_2)$。

(2) 关系义原是特征结构[2]，其计算的核心思想是先按一定规则配对，然后分别算出配对义原的相似度，再求平均值。关系义原以等号左边英文单词相同的进行配对，最后剩余的未配对义原都虚拟一个空值与之配对。关系义原特征结构相似度记为 $sim_2(C_1,C_2)$。

(3) 关系符号义原也是特征结构[2]，其计算的核心思想也是先按一定规则配对，然后分别算出配对义原的相似度，再求平均值。关系符号义原以相同符号开头的进行配对，最后剩余的未配对义原都虚拟一个空值与之配对。关系符号特征结构相似度记为 $sim_3(C_1,C_2)$。

义原或者具体词与空值的相似度都处理为一个较小的常数 δ。具体词指的是知网中尚未给出定义的词条，在 DEF 中一般用括号括起来。具体词与义原的相似度均处理为另一个较小的常数 γ。具体词与具

体词间相同相似度处理为1,不相同则处理为0。

将DEF的上述三部分相似度组合起来就可以得到义项的相似度,计算公式如式(4)所示[1-2,8]。

$$sim(C_1,C_2)=\sum_{i=1}^{3}\beta_i\prod_{j=1}^{i}sim_j(C_1,C_2) \quad (4)$$

其中,参数$\beta_i(1\leqslant i\leqslant 3)$是可调节的,且满足:$\beta_1+\beta_2+\beta_3=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3$。本文的实验中$\beta_1$取0.7,$\beta_2$取0.17,$\beta_3$取0.13。式(4)采用多个$sim$连乘的目的,主要是为了使用前面主要部分的相似度值来抑制后面次要部分的相似度所起的作用,从而避免出现当主要部分的相似度值过低时,因次要部分的相似度太高而导致整体相似度过高的不合理现象的出现[2]。

考虑到有的词语会有多个义项,两个词语的最终相似度取所有义项组合中相似度最大的值为有效值,公式如式(5)所示[2]。

$$sim(W_1,W_2)=\max_{i=1\cdots m,j=1\cdots n}\{sim(C_{1i},C_{2j})\} \quad (5)$$

2.3 改进的同义词词林词语相似度计算

同义词词林的整体构造是一个五层树形结构(图2),因此两个词语在词林树中的连接路径是影响词语相似度的主要因素。词林的第一层是大类,本文将不属于同一个大类的词语间的距离都处理为18,同时按从底层到高层的顺序,将连接上、下两层的四类边分别赋予一个权重$W_i(1\leqslant i\leqslant 4)$,且满足:$0.5\leqslant W_1\leqslant W_2\leqslant W_3\leqslant W_4\leqslant 5$,$W_1+W_2+W_3+W_4\leqslant 10$,如图2所示。在本文实验中,这四类边的权重分别取0.5、1、2.5、2.5,由于词语编码均在第五层叶子节点上(图2中的实心节点),于是词语编码距离d可取1、3、8、13、18这几个离散值。

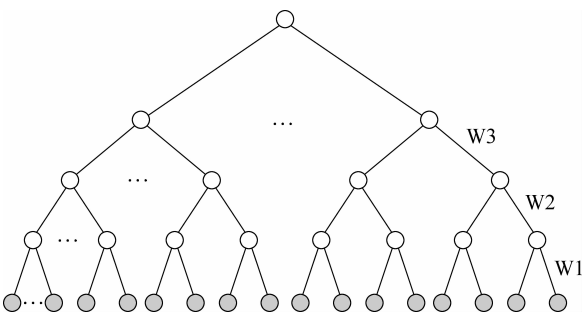


图2 同义词词林的5层树形结构

在词林中,影响词语相似度的还有两个次要因素:两个词语最近公共父节点的直接孩子的个数,也叫分支层节点总数n,以及在最近公共父节点中,两个词语所在分支的间隔距离k,比如"人"编码Aa01A01和"每人"编码Aa01A08这两个词语的n=9,k=7。分支层节点总数n反映了公共父节点的密度,因此与词语相似度成正比。在同一层中,词林是按一定的语义顺序对词语进行分类与排列的,因此分支间隔k与相似度成反比关系。

文献[10]提出了一个基于层分支的词林义项相似度计算公式,该公式是以分支节点数n和分支间隔k为主要考虑因素,因此会出现许多距离近的词语因分支间隔远而算出相似度过低的不合理现象。为解决这一问题,本文提出了一个以词语距离d为主要影响因素、分支节点数n和分支间隔k为调节参数的同义词词林词语相似度计算公式,如式(6)所示。

$$sim(C_1,C_2)=(1.05-0.05dis(C_1,C_2))\sqrt{e^{\frac{-k}{2n}}} \quad (6)$$

其中,$dis(C_1,C_2)$是词语编码C_1和C_2在树状结构中的距离函数,等于词语对的连接路径中各边的权重之和,可取值$2*W_1$、$2*(W_1+W_2)$、$2*(W_1+W_2+W_3)$、$2*(W_1+W_2+W_3+W_4)$。

式(6)设计的基本思路为:首先为词语对的相似度赋予一个根据词语距离计算出的初值;然后再根据词语对的最近公共父节点的密度n与词语对所在分支的间距k,对该初值进行向下修正,且要求该修正只能是微调,修正幅度不能超过25%。式(6)中,将n和k的表达式作为e的负指数,以及对其开平方,都是为了降低公式对n和k这二个参数的敏感度,避免出现修正幅度过大的现象。

当两个词语在编码的同一个"="后面时,相似度处理为1;在编码的同一个"#"后面时相似度处理为0.5。当两个词语不在一个大类中时词语间的距离都处理为18。当一个词语对应多个编码时,与知网中词语对应多个义项的处理方法类似,计算出所有的编码组合的相似度,取最大的相似度作为词语的相似度。

2.4 综合知网和词林的词语相似度计算

综合考虑知网和词林的词语相似度计算的总体思想为:对于任意两个词语W_1和W_2,根据它们在知网和词林中的分布情况,按照一定的策略综合利用知网和同义词词林分别计算出词语的两个相似度,记作s_1和s_2,同时为这两个相似度分别赋予权重λ_1和λ_2,且满足:$\lambda_1+\lambda_2=1$,然后按照式(7)计算出综合知网和词林的词语语义相似度。

$$s = \lambda_1 s_1 + \lambda_2 s_2 \tag{7}$$

词语在知网和词林中的分布情况分类如图 3 所示。I 代表所有的词语构成的全集；A 代表知网中特有的词语，即知网中收录，词林中未收录的词语，共有 19 296 个[7]；B 代表词林中特有的词语，即词林中收录，知网中未收录的词语，共有 21 330 个[7]；C 代表知网和词林中同时收录的词语，共有 30 926 个[7]。由于知网与词林都在不断建设中，上述数据也在不断变化。

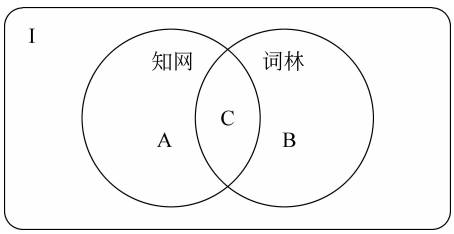


图 3　词语在知网与词林中的分布图

根据图 3 中词语的分布情况，式(7)采用如下综合考虑知网和同义词词林的动态加权计算策略：

(1) 当 $W_1 \in C, W_2 \in C$ 时，同时使用知网和词林分别计算 W_1 和 W_2 的相似度，分别记作 s_1 和 s_2。在不同的具体应用中，λ_1 和 λ_2 可以任意调节，本文实验中取 $\lambda_1 = 0.5, \lambda_2 = 0.5$。

(2) 当 $W_1 \in A, W_2 \in A$ 或者 $W_1 \in B, W_2 \in B$ 中时，单独对 W_1 和 W_2 进行基于知网或基于词林的相似度计算，记作 s_1 或 s_2。此时，λ_1 和 λ_2 一个为 1，另一个为 0。

(3) 当 $W_1 \in A, W_2 \in B$ 时，在词林中查找 W_2 的同义词集合，依次与 W_1 进行基于知网的相似度计算，取其中的最大值作为两个词语的相似度，记作 s_1；如果 W_2 在词林中无同义词，则取 $s_1 = 0.2$。此时取 $\lambda_1 = 1, \lambda_2 = 0$。

(4) 当 $W_1 \in A, W_2 \in C$ 时，首先对 W_1 和 W_2 进行基于知网的相似度计算，结果记作 s_1；然后在词林中查找 W_2 的同义词集合，依次与 W_1 进行基于知网的相似度计算，取其中的最大值作为 s_2；如果 W_2 在词林中无同义词，则取 $s_2 = s_1$。此时要求 $\lambda_1 > \lambda_2$，本文实验中取 $\lambda_1 = 0.6, \lambda_2 = 0.4$。

(5) 当 $W_1 \in B, W_2 \in C$ 时，首先对 W_1 和 W_2 进行基于词林的相似度计算，结果记作 s_2；然后在词林中查找 W_1 的同义词集合，依次与 W_2 进行基于知网的相似度计算，取其中的最大值作为 s_1；如果 W_1 在词林中无同义词，则取 $s_1 = s_2$。此时要求 $\lambda_2 > \lambda_1$，本文实验中取 $\lambda_1 = 0.4, \lambda_2 = 0.6$。

对于一个词语的所有同义词在知网中都不存在的情况暂不予考虑，对于知网和同义词同时未收录的词语目前使用本文的方法还不能计算。

3　实验与分析

3.1　对比实验

目前国际上对词语相似度算法的评价标准普遍采用 Miller&Charles 发布的英语词对集的人工判定值[11]。该词对集由十对高度相关、十对中度相关、十对低度相关共 30 个英语词对组成，然后让 38 个受试者对这 30 对进行语义相关度判断，最后取他们的平均值作为人工判定标准[12]。本文采用 Miller&Charles 发布的词对集及其人工判定值作为标准，通过计算各种方法与其的皮尔森相关系数(Pearson correlation coefficient)，将本文提出的方法分别与刘群[2]和田久乐[10]的方法进行对比。首先，将这 30 个英语词对按照同词性、意义最接近的原则翻译成对应的中文词对，然后采用各种方法对词对计算相似度(表 3)，最后计算出不同方法的结果与 miller 人工值的皮尔森相关系数(表 4)。为增加结果的可比性，表 4 还列出了四种英文方法的皮尔森相关系数。

表 3　不同方法对 Miller 词对集的计算结果

词语 1	词语 2	基于词林的相似度		基于知网的相似度		本文综合方法	Miller 人工判定值
		本文方法	田久乐[10]	本文方法	刘群[2]		
轿车	汽车	0.821 7	0.211 9	1	1	0.910 8	0.98
宝石	宝物	0.836 5	0.408 4	0.600 0	0.145 5	0.718 2	0.96
旅游	游历	1	1	1	1	1	0.96
男孩子	小伙子	0.807 1	0.272 2	1	1	0.903 6	0.94
海岸	海滨	0.939 4	0.957 7	1	1	0.969 7	0.925

续表

词语 1	词语 2	基于词林的相似度		基于知网的相似度		本文综合方法	Miller 人工判定值
		本文方法	田久乐[10]	本文方法	刘群[2]		
庇护所	精神病院	0.964 9	0.952 8	0.537 6	0.579 2	0.751 3	0.902 5
魔术师	巫师	0.845 5	0.897 8	0.654 1	0.676 0	0.749 8	0.875
中午	正午	1	1	1	1	1	0.855
火炉	炉灶	0.975 3	0.945 4	0.558 4	0.589 6	0.766 9	0.777 5
食物	水果	0.348 1	0.309 1	0.216 3	0.126 3	0.282 2	0.77
鸟	公鸡	0.636 0	0.704 4	1	1	0.818 0	0.762 5
鸟	鹤	0.643 0	0.736 4	1	1	0.821 5	0.742 5
工具	器械	0.530 5	0.171 7	1	1	0.765 3	0.737 5
兄弟	和尚	0.365 8	0.450 5	0.815 9	0.861 1	0.590 8	0.705
起重机	器械	0.530 5	0.171 7	0.393 7	0.369 2	0.462 1	0.42
小伙子	兄弟	0.392 9	0.630 7	0.733 3	0.800 0	0.563 1	0.415
旅行	轿车	0.132 4	0.1	0.074 1	0.074 1	0.103 2	0.29
和尚	圣贤	0.386 0	0.585 6	0.581 5	0.682 5	0.483 8	0.275
墓地	林地	0.593 0	0.461 9	0.200 5	0.122 1	0.396 8	0.237 5
食物	公鸡	0.353 0	0.343 4	0.204 8	0.111 6	0.278 9	0.222 5
海岸	丘陵	0.630 0	0.792 2	0.170 2	0.1	0.400 1	0.217 5
森林	墓地	0.146 9	0.1	0.189 5	0.111 6	0.168 2	0.21
岸边	林地	0.353 0	0.343 4	0.170 3	0.096 5	0.158 6	0.157 5
和尚	奴隶	0.353 0	0.360 4	0.549 1	0.661 1	0.451 1	0.137 5
海岸	森林	0.383 7	0.549 5	0.177 6	0.111 6	0.280 7	0.105
小伙子	巫师	0.379 1	0.540 6	0.418 1	0.600 0	0.398 6	0.105
琴弦	微笑	0.129 6	0.1	0.054 8	0.074 1	0.092 2	0.032 5
玻璃	魔术师	0.146 9	0.1	0.193 8	0.121 9	0.170 4	0.027 5
中午	绳子	0.146 9	0.1	0.123 5	0.099 9	0.135 2	0.02
公鸡	航行	0.132 4	0.1	0.074 1	0.074 1	0.103 2	0.02

表 4 不同方法与 miller 人工值的皮尔森相关系数

续表

词对语言	相似度方法	使用的语义词典	与 miller 人工值的皮尔森系数
英文	Resnik[13]	WordNet	0.795
	CP/CV[14]	WordNet	0.813 8
	OHIIC[15]	WordNet	0.820 3
	Mohamed[16]	WordNet	0.85
中文	本文词林方法	同义词词林	0.838 6
	田久乐[10]	同义词词林	0.530 1
	本文知网方法	HowNet	0.805 6
	刘群[2]	HowNet	0.699 1
	本文综合方法	同义词词林 & HowNet	0.888 4

3.2　扩展计算实例

根据图 3，可以得出基于知网的词语相似度的可计算词语范围为：A∪C＝50 222 个，基于词林的词语相似度的可计算词语范围为：B∪C＝52 256 个，本文提出的综合知网与词林的词语相似度方法的可计算词语范围为：A∪B∪C＝71 552 个。该综合方法对知网方法的可计算词语范围扩展了：B/(A∪C)＝42.47%，对词林方法的可计算词语范围扩展了：A/(B∪C)＝ 36.93%。表 5 给出了几个典型的扩展计算实例，其中不带括号的词语表示被知网和词林同时收录，带圆括号的词语表示仅被知网收录而未被词林收录，带方括号的词语表示仅在词林中收录而未被知网收录。

表 5　扩展计算实例

词语 1	词语 2	词林计算	知网计算	综合结果	加权策略
(视频)	音频	将音频替换成旋律	S_1＝0.896 1，S_2＝0.677 2	0.808 5	4
售票员	[保洁员]	S_2＝0.5，S_1＝0.5	无	0.5	5
(校花)	国花	将国花替换成国色天香	S_1＝0.214 7，S_2＝0.242 6	0.225 9	4
(养老金)	[保险金]	将保险金替换为保证金	S_1＝0.608 1	0.608 1	3
(养老金)	钱	将钱替换为钱财	S_1＝0.678 7，S_2＝ 0.678 7	0.678 7	4
[场址]	地址	S_2＝0.904 8，S_1＝0.904 8	无	0.904 8	5
[鸟窝]	[鸟巢]	S_2＝1	无	1	2
[蜂巢]	蜜蜂	S_2＝0.373 2　将蜂巢替换为蜂窝	S_1＝0.170 8	0.292 2	5

3.3　结果分析

通过上述实验与计算实例，可以得出以下结论：

(1) 从上述对比实验可以看出：效果最好的是本文综合知网和词林的词语相似度计算，该方法词语计算范围广，与 miller 人工值的皮尔森相关系数最高，达到了 0.888 4，与国外相关算法相较也是优秀的；其次的是本文改进的分别基于词林与基于知网的两种词语相似度计算，与 miller 人工值的皮尔森相关系数都有一定程度的提高，其值都超过了 0.8，达到了实用水平；田久乐[10]实现的基于词林的词语相似度计算，与 miller 人工值的皮尔森相关系数只有 0.530 1，效果最差。

(2) 从表 3 的结果还可以看出，在知网 2000 版中，由于知识描述语言的局限性，有一些词语的定义比较粗糙，在计算词语相似度时，无论在算法上如何改进，其相似度计算总是会出现不尽人意的地方，引入词林计算模块以后正好修正了这些粗糙点。词林编码中同一个“＃”后面的词语相似度全部处理为 0.5，这样会使得有些词语间的相似度偏低，知网计算模块可以改善这一点。

(3) 目前，国际上对英文词语的相似度计算，普遍是基于 WordNet[17]语义词典，主要是利用词节点之间上下位关系构成的最短路径来计算英文词语之间的相似度，并同时考虑两个词的公共祖先节点的最大信息量、概念层次树的深度与区域密等信息来调节词语相似度[13-16]，与本文改进的基于同义词词林的中文词语相似度算法的思想与效果基本相同。

(4) 田久乐的词林方法在计算“轿车”与“汽车”、“男孩子”与“小伙子”两个词对的相似度时，值都偏低，这主要是由于在这两对词语中分支间隔 k 与分支节点数 n 的比值都较大，而在其计算公式中，该比值与相似度是线性负相关的，因此，计算结果对该比值过于敏感，造成相似度偏差较大；而在本文改进的词林方法中，该比值与相似度是曲线负相关的，从而降低了对该比值的敏感度，提高了词语相似度的准确度。

(5) 表 3 中，所有方法在计算“食物”与“水果”词对的相似度时与 miller 人工值相比都偏低，这主要是在同义词词林分类结构中，将“食物”归为第二大类“物”的“粮食”中类而将“水果”归为“物”的“草木”中类，造成二者的共同父节点的层次过高；同样在知网中的义原分类结构中，二者的共同父节点为第三层的“物质”(“食物”为“无生物物质”、“水果”为“生物物质”)，共同父节点的层次也很高。词林与知网对于这二个词的分类法是否正确，还有待商榷。

4 结束语

本文所提出的词语语义相似度计算方法，结合了知网与同义词词林两个知识库，充分利用了词语在不同知识库中的语义信息，得到的相似度更为准确与合理。在实验中我们也发现有一些词语无论用那种方法在两个知识库中的计算结果均不理想，这种情况一般是义项定义不合理，或者词语在词林中的分类不合理造成的。因此，利用相似度的计算可以反过来检验词语的定义以及分类，修正知识库中的不合理之处。基于树形结构的词语相似度的计算，在算法方面基本已经考虑到了所有可利用信息。词语相似度的计算进一步工作还可以将词语的语用信息结合进来，这样得到的相似度具有更好地可靠性。

参考文献

[1] 葛斌，李芳芳，郭丝路，汤大权. 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，2010，09：3329-3333 .

[2] 刘群，李素建. 基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义研讨会，台北，2002.

[3] 董振东，董强. 知网[DB/OL]，http://www. keenage. com/zhiwang/c_zhiwang. html.

[4] 贾玉祥，俞士汶. 基于词典的名词性隐喻识别[J]. 中文信息学报，2011，25(03)：99-102.

[5] 梅家驹等编. 同义词词林[M]. 上海：上海辞书出版社，1996.

[6] 刘丹丹，彭成，钱龙华，周国栋.《同义词词林》在中文实体关系抽取中的作用[J]. 中文信息学报，2014，28(02)：91-99.

[7] 梅立军，周强，臧路，陈祖舜. 知网与同义词词林的信息融合研究[J]. 中文信息学报，2005，19(01)：63-70.

[8] 江敏，肖诗斌，王弘蔚，施水才. 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(05)：84-89.

[9] 朱征宇，孙俊华. 改进的基于《知网》的词汇语义相似度计算[J]. 计算机应用，2013，08：2276-2279，2288.

[10] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版)，2010，06：602-608.

[11] G A Miller，W G Charles. Contextual correlates of semantic similarity [J]. Language and Cognitive Processes，1991，6(1)：1-28.

[12] 刘宏哲. 文本语义相似度计算方法研究 [D]. 北京交通大学博士学位论文，2012.

[13] P Resnik. Semantic Similarity in Taxonomy：An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language[J]. Journal of Articial Intelligence Research，1999，11：95-130.

[14] J W Kim，K S Candan. CP/CV：concept similarity mining without frequency information from domain describing taxonomies[C]//Proceedings of the 15th International Conference on Information and Knowledge Management，2006：483-492.

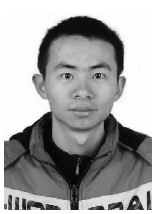
[15] S Bin，F Liying，Y Jianzhuo，W Pu，Z Zhongcheng. Ontology-Based Measure of Semantic Similarity between Concepts[C]//Proceedings of the World Congress on Software Engineering，2009，2：109-112.

[16] A H T Mohamed，B A Mohamed，A B Hamadou. Ontology-based approach for measuring semantic similarity[J]. Journal of Engineering Applications of Artificial Intelligence，2014，36：238-261.

[17] Princeton University. WordNet [DB/OL]，http://wordnet. princeton. edu/.

朱新华(1965－)，教授，研究生导师，主要研究领域为自然语言处理、信息抽取。
E-mail：zxh429@263. net

孙柳(1988－)，硕士研究生，主要研究领域为自然语言处理、信息抽取。
E-mail：515718167@qq. com



马润聪(1989－)，硕士研究生，主要研究领域为自然语言处理、信息抽取。
E-mail：maruncong@163. com