基于信息内容的词林词语相似度计算*

彭 琦[a]，朱新华[b†]，陈意山[c]，孙 柳[b]，李 飞[b]
(广西师范大学 a. 网络信息中心；b. 计算机科学与信息工程学院；c. 漓江学院，广西 桂林 541004)

摘 要： 针对哈尔滨工业大学《同义词词林》扩展版的层次结构不能有效反映词语之间信息内容含量差异性的问题进行了研究，进行了《同义词词林》作为词语相似度计算本体的结构改造，增加了原编码信息节点的语义，提出了一种较为适合改造后本体的相似度计算策略。经实验证明，修改后的本体更能体现词语在本体中信息内容含量的差异性，提出的相似度计算策略应用在改进后的本体上时，得出的相似度计算结果准确程度达到了较高水平，具有较好的实用价值。

关键词： 词林；词语相似度；信息内容
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695(2018)02-0400-05
doi:10.3969/j.issn.1001-3695.2018.02.018

IC-based approach for calculating word semantic similarity in CiLin

Peng Qi[a], Zhu Xinhua[b†], Chen Yishan[c], Sun Liu[b], Li Fei[b]
(*a. Network Center, b. College of Computer Science & Information Technology, c. Lijiang College, Guangxi Normal University, Guilin Guangxi* 541004, *China*)

Abstract: This paper resolved the problem that the hierarchical structure of CiLin provided by HIT (Harbin Institute of Technology) could not effectively reflect the difference of information content among the words, extracted a word for a node that only has the encoding, optimized the structure of CiLin, and proposed a semantic similarity computing strategy. The experimental results show that the optimized CiLin can better reflect the difference among the words. The proposed new strategy is applied to the optimized CiLin, the accuracy of the calculation results reach a high level, and has good practical value.

Key words: CiLin; semantic similarity; information content

词汇语义相似度计算是自然语言处理中一个十分重要的研究领域，一些自然语言处理的基本任务如歧义消解[1]、信息抽取[2]、同义词检测[3]等都依赖于词语相似度计算。词语相似度的计算方法有很多种，不同的计算方法根据其计算特点应用于不同领域。这些计算方法可以分为两种[4]：一种是利用大规模语料进行词语统计分析，将语料中词语出现的次数及分布情况作为词语相似度的计算依据[5]，这种方法的计算结果依赖于所选取的语料库[6]；另一种是通过计算本体中概念之间的共性或差异性来评估词语之间的相似度[7]。目前常用的计算相似度的本体有 WordNet[8]、知网[9]以及哈工大《同义词词林》扩展版[10](以下简称词林)等。在以英语为计算语言的词语相似度研究中，基于本体的相似度计算方法目前已形成较完备的理论体系，大多数研究以 WordNet 作为本体；在中文领域，常使用的本体为知网、词林。使用本体进行相似度计算的方法比较单一[11]，大多使用基于路径的计算方法[12~15]，而较少使用其他计算方法。本文旨在将英语词语相似度计算方法中常用的基于信息内容的计算方法引入到中文词语相似度计算中，并以词林作为本体，测试这种方法在中文本体上的可行性，在此基础上对词林的拓扑结构进行改进，并构建适合中文本体的信息内容相似度计算方法。

1 词语相似度的定义及计算方法

1.1 词语相似度的定义

判断两个词语之间的相似度，对人脑来说是一项简单的工作，能够通过直觉和经验完成判断。如何形式化人脑的判断过程并通过公式表现出来，是计算词语相似度需要解决的问题[16]。人脑对词语的理解，是指人对该词语在现实世界中所指代概念的理解[17]。两个词语之间的相似程度，反映出这两个词语所指代的概念在现实世界中的共性或差异性。

对此 Lin[18] 进行了大胆的假设，他认为，对于两个概念之间的相似度，存在下列三个定义。

定义 1 两个概念之间的相似程度与它们的共性有关，共性越多越相似。

定义 2 两个概念之间的相似程度与它们的差异性有关，差异性越多越不相似。

定义 3 当两个概念完全相同时，这两者之间的相似性达到最大值。

Lin[18] 认为，如果有一组公式，将以上几点完整地统一，那么这组公式就可以成为计算两个概念之间相似度的算法。

收稿日期：2016-10-20；修回日期：2016-12-13 基金项目：国家自然科学基金资助项目(61363036,61462010)；广西师范大学自然科学青年基金项目

作者简介：彭琦(1988-)，男，广西桂林人，助理研究员，硕士，主要研究方向为自然语言处理；朱新华(1965-)，男(通信作者)，广西桂林人，教授，主要研究方向为自然语言处理(zxh429@263.net)；陈意山，男，广西桂林人，副教授，主要研究方向为自然语言处理；孙柳(1988-)，女，河南周口人，硕士，主要研究方向为自然语言处理；李飞(1990-)，男，湖北郧西人，硕士，主要研究方向为自然语言处理.

1.2　词语相似度计算方法简介

基于本体的词汇语义相似度计算方法可以分为以下四种:

a) 基于路径的相似度计算方法[19]。这种方法主要依靠两个词语在本体中的最短路径长度来评定两个词语之间的相似度。该方法通常结合两个词语最近公共父节点深度[20]来提高相似度测量精度。

b) 基于信息内容的相似度计算方法[21]。该方法依靠信息内容函数,计算出两个概念及其最近公共父节点在本体中的信息内容含量,以此判断词语之间的相似度。一般来说抽象性词语的信息内容含量往往较少,而分支较细、指代明确的词语所包含的信息内容较多。

c) 基于特征的相似度计算方法[22]。根据词语在不同本体中的共同特征与不同特征来计算词语之间的相似度,如词语的词性、所属类别等。

d) 利用本体中词语的注释来计算词语之间的相似度。其主要的计算方法是量化两个概念之间注释的重叠内容,这是 WordNet 所特有的词语相似度计算方法。

在上述四种方法中,基于信息内容的相似度计算方法因其能够较好地解决大型本体中各学科分类的不均匀性问题[23],而在国际上成为目前的研究热点。

1.3　基于信息内容的词语相似度计算方法

使用本体和信息内容的方法来计算词语相似度可以追溯到 1995 年 Resnik[24] 的研究,他认为,一对词语的相似度由它们的共性决定。在一个具有良好分类层次结构的本体中,一对概念(一个词语可能包含多个概念,不同概念存在于本体中的不同节点上)的共性,就表现在它们有多少共享的信息。例如,如果概念 A 代表的是“苹果”,概念 B 代表的是“梨”,那么 A 和 B 的共性就表现在它们都是水果。在一个具有良好分类层次结构的本体中,概念“苹果”和“梨”的最近公共父节点应该是“水果”,因此在一些具有良好分类层次结构的本体中,可以用最近公共父节点的信息内容含量来表示 A 和 B 的共性。

基于以上考虑,Resnik[24] 定义两个概念之间相似度的计算公式如式(1)所示。

$$\mathrm{sim}_{\mathrm{Resnik}}(C_1,C_2)=IC(LCS(C_1,C_2)) \tag{1}$$

可以看出,Resnik 提出的公式反映了 Lin 提出的定义 1,其中 $LCS(C_1,C_2)$ 表示 C_1 和 C_2 的最近公共父节点,$IC(C)$ 表示概念 C 的信息内容含量(information content)。概念的信息内容含量[25]可用式(2)进行计算。

$$IC_{\mathrm{Resnik}}(C)=-\log P(C) \tag{2}$$

其中: $P(C)$ 为概念 C 所在测量样本总体中的或然率。在这一理论的基础上,发展出了能够在本体中使用的信息内容含量计算公式。比较典型的有 Seco 等人[26]提出的基于 WordNet 的信息内容含量计算方法,如式(3)所示。

$$IC_{\mathrm{Seco}}(C)=1-\frac{\log(\mathrm{hypo}(C)+1)}{\log(\mathrm{maxnodes})} \tag{3}$$

其中: hypo(C)是所要计算的概念在本体中的下位个数,maxnodes 为本体的节点总数。式(3)反映出,如果一个概念在本体中层次越高,即越一般和普遍的概念,它所含的信息内容越少; 相反,如果一个概念在本体中的下位个数越少,即它在本体中的层次越深,那么它所含的信息内容越多。这与信息论中的信息内容含量理论是相符的。

Jiang 等人[27]在 1997 年的研究中使用了与 Resnik 相类似的方法来计算相似度,他们所关注的是概念之间的差异性,即 Lin 提出的定义 2。Jiang 等人认为,如果两个概念之间的差异性越大,它们的相似度就会越小。其提出的计算两个概念之间差异性的方法如式(4)所示。

$$\mathrm{dis}_{JC}(C_1,C_2)=IC(C_1)+IC(C_2)-2IC(LCS(C_1,C_2)) \tag{4}$$

Lin 提出的词语相似度计算方法试图同时体现他提出的定义 1 和 2,具体的计算公式如式(5)所示。

$$\mathrm{sim}_{\mathrm{Lin}}(C_1,C_2)=\frac{2IC(LCS(C_1,C_2))}{IC(C_1)+IC(C_2)} \tag{5}$$

式(5)的分子部分代表的是 C_1 和 C_2 的共性,分母部分代表的是能够将 C_1 和 C_2 全部描述出来的信息,即 C_1 和 C_2 各自的信息内容含量之和。

Meng 等人[28]对 Lin 的计算方法进行了非线性改进,其计算方法如式(6)所示。

$$\mathrm{sim}_{\mathrm{Meng}}(C_1,C_2)=\mathrm{e}^{\mathrm{sim}_{\mathrm{Lin}}(C_1,C_2)}-1 \tag{6}$$

2　《同义词词林》简介与改进

2.1　《同义词词林》简介

《同义词词林》是由梅家驹[29]于 1983 年编撰的可计算汉语词库,其设计目的是实现汉语同义词和同类词的划分归类。《同义词词林》经哈尔滨工业大学社会计算与信息检索研究中心的扩展后,目前共有 7 万多个词语,9 万多个概念,这些概念被分为 12 个大类,95 个中类,1 428 个小类,4 026 个词群和 17 797个原子词群[30]。词林拓扑结构如图 1 所示。

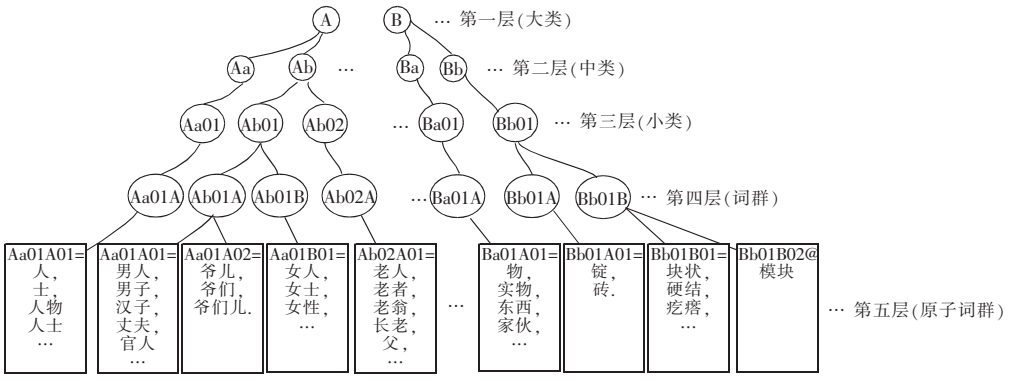

图 1　词林拓扑结构

词林的第一层,是按照概念类别划分的 12 个大类,大类编码 A ~ L; 第二层是 95 个中类,中类编码在大类编码后加一位小写英文字母共同表示; 第三层是小类,小类编码在中类编码后加两位十进制整数共同表示; 第四层是词群分类; 第五层是

原子词群分类，编码方式逐层累加。

词林以文本的方式存储在计算机内部，存储格式如表 1 所示。词林中每个原子词群为一行，包含一个或多个概念。在概念前的 8 位字符是该原子词群的编码，该编码包含了从大类到原子词群分类的全部类别信息。

表 1 词林文本存储格式

```
……
Ab03A03@ 社会青年
Ab03A04 = 少年人 少年 苗子 苗 年幼 未成年 未成年人
Ab03A05 = 芝兰 龙驹 千里驹
Ab03A06# 待业青年 务工青年
Ab03A07# 男孩子 少男
……
```

2.2 词林的改进

从词林的拓扑结构及存储方式可以看出如下不合理之处：第一层到第四层的分类节点只有分类编码没有具体概念，所有的概念只在原子词群上，这样的拓扑结构无法区分抽象的和具体的概念。例如“水果”这一指代某一类事物的概念与“苹果”“梨子”“柑橘”等表示具体水果的概念放置在同一个原子词群节点下，极有可能得出“苹果”与“水果”之间的相似度等价于“苹果”与“梨子”之间的相似度，这不利于它们之间的相似度计算。因此，本文参考 WordNet、知网等其他用于相似度计算的本体，以及《同义词词林》原著，将词林中具有抽象性的原子词群或概念提取到更高的分类节点，使一至四层分类节点不仅包含分类编码信息，并且包含代表整个类别的具体概念。

例如，将“人”“食物”“水果”与“动物”等抽象性概念提取出来，放置在大类、中类或者小类的节点当中，经本文修订后的词林拓扑结构如图 2 所示。

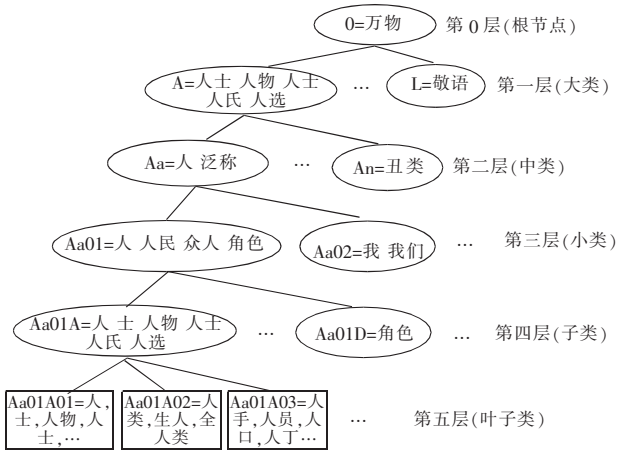

图 2 修订后词林拓扑结构

本文修订后的词林在计算机中的存储片段如表 2 所示。

表 2 修订后词林文本存储片段

```
A = 人
Aa = 泛称
Aa01 = 人 人民 众人
Aa01A = 人 士 人物 人士 人氏 人选
Aa01A01 = 人 士 人物 人士 人氏 人选
Aa01A02 = 人类 生人 全人类
Aa01A03 = 人手 人员 人口 人丁 口 食指
Aa01A04 = 劳力 劳动力 工作者
……
```

3 基于信息内容和词林本体的词语相似度计算

3.1 基于词林的信息内容相似度计算公式

词林是一个分类严格的层次结构树，而不是网状结构。对于一个树型结构来说，节点之间的关系体现在：任何一个非叶子节点可以拥有一个或多个下位节点，但所有非根节点都只能拥有一个上位节点，这与 WordNet 允许一个节点拥有多个上位节点的网状结构有本质的不同。对于词林这样一棵分类层次结构树来说，更能反映词语之间的差异性，而较少反映词语之间的共性。因此本文主要使用差异性计算公式来计算词林中概念的相似度。在词林中，完全相同的两个概念或同义词之间的差异性最小，记为 MinDIFF。选取词林中任意两个同义词或完全相同的两个概念 C_1、C_2，即 C_1 等价于 C_2，由于它们的最近公共父节点就是它们自身，所以它们自身的信息内容含量与它们最近公共父节点的信息内容含量完全相同。信息内容含量计算公式使用 Seco 等人提出的式(3)，差异性计算公式使用 Jiang 提出的式(4)，得出两个同义词或完全相同的两个概念之间的差异性为 0，因此 MinDIFF 值也为 0。

在词林中差异性最大的两个概念可定义为：处于本体边缘的两个叶子概念，且这两个概念的最近公共父节点为整个分类树的根节点，则这两个概念的差异性最大，记做 MaxDIFF。例如词“公鸡”与“远行”，这两个概念都处于词林的叶子节点上，“公鸡”属于“物”这个大类，大类编码为 B；“远行”属于“活动”这个大类，大类编码为 H。因此，这两个概念的最近公共父节点是整个分类树的根节点。由于所有叶子节点的下位数为 0，根节点的下位数为词林中的所有节点个数(maxnodes = 90114)，根据式(3)可得出词林中叶子节点的 IC 值(IC(叶子))与根节点的 IC 值(IC(根))。

$$IC(\text{叶子}) = 1 - \frac{\log(1)}{\log(\text{maxnodes})} = 1$$

$$IC(\text{根}) = 1 - \frac{\log(\text{maxnodes}+1)}{\log(\text{maxnodes})} \approx 1$$

将 IC(叶子)及 IC(根)代入式(4)进行计算，得出 MaxDIFF 为 2。

$$\text{MaxDIFF} = IC(\text{叶子}) + IC(\text{叶子}) - 2 \times IC(\text{根}) = 2$$

任何两个概念之间的差异性，都应该在差异性的最大值与最小值之间，且差异越大，离 MaxDIFF 越近；差异越小，离 MinDIFF越近。

综上所述，本文提出如下的基于信息内容词语相似度计算公式：

$$\text{sim}(C_1, C_2) = \frac{\text{MaxDIFF} - \text{dis}(C_1, C_2)}{\text{MaxDIFF} - \text{MinDIFF}} \tag{7}$$

其中：C_1 和 C_2 为词林中任意两个概念，$\text{dis}(C_1, C_2)$ 的计算公式为 Jiang 等人提出的式(4)。式(7)完全满足 1.1 节中 Lin 提出的关于概念相似度算法的三个定义，证明过程如下：

a) 式(7)反映了差异性越大的两个概念越不相似，即 $\text{sim}(C_1, C_2)$ 与 $\text{dis}(C_1, C_2)$ 成反比，因此满足 Lin 提出的定义 2。

b) 将式(4)代入式(7)，得到式(7)的一个变异：

$$\text{sim}(C_1, C_2) = \frac{2IC(LCS(C_1, C_2)) + \text{MinDIFF} - IC(C_1) - IC(C_2)}{\text{MaxDIFF} - \text{MinDIFF}} \tag{8}$$

通过式(8)可发现，任意两个概念之间的相似度与其最近公共父节点的 IC 值成正比。由于最近公共父节点的 IC 值反映两个概念之间的共性，所以式(7)也满足 Lin 提出的定义 1。

c) 据前文所述，词林中完全相同的两个概念之间的差异性最小，其值为 0，代入式(7)，得到它们之间相似度为最大值 1，因此式(7)满足 Lin 提出的定义 3。

考虑到同一词语可能会代表多个概念，即同一个词语存在

于词林中的多个节点上,因此,两个词语 W_1 与 W_2 之间的相似度计算公式如式(9)所示。即在计算过程中若遇一个词语包含多个概念,计算出所有概念组合的相似度后,取相似度最大值作为词语 W_1 和 W_2 的相似度。

$$\mathrm{sim}(W_1,W_2)=\max_{(C_1,C_2)\in \mathrm{syn}(W_1)\times \mathrm{syn}(W_2)}\mathrm{sim}(C_1,C_2) \tag{9}$$

3.2　对比实验

本文采用国际上广泛使用的由 Miller & Charles(MC)[31] 发布的 30 对词语的数据集以及 Rubenstein & Goodenough(RG)[32] 发布的 65 对词语数据集作为测试集,这两组数据集分别由高度相关、中度相关与低度相关的英语词对组成。本文将这两个数据集中的英语词对按照意义最接近的原则翻译成对应的中文词对,以词林作为本体,对文中提及的各个公式进行实验对比。

3.2.1　实验 1

首先使用词林作为本体进行基于信息内容的相似度计算测试。实验使用在前文中介绍的式(1)(5)(6)以及本文提出的式(7)作为相似度计算公式。词语信息内容的计算公式统一使用 Seco 等人提出的信息内容计算式(3)。

得出在 MC30 数据集上的相似度测量结果如表 3 所示。由于 RG65 数据集词对较多,本文不一一列出。

表 3　词林 MC30 数据集实验结果

词语 1	词语 2	Resnik 式(1)	Lin 式(5)	Meng 式(6)	本文公式	MC 人工判定值(参考值)
轿车	汽车	0.665	0.665	0.643	0.665	0.98
宝石	宝物	0.793	0.793	0.824	0.793	0.96
旅行	远行	0.771	0.771	0.791	0.771	0.96
男孩子	小伙子	0.793	0.793	0.824	0.793	0.94
海岸	海滨	0.840	0.840	0.896	0.840	0.925
庇护所	精神病院	0.793	0.793	0.824	0.793	0.902 5
魔术师	巫师	0.745	0.745	0.753	0.745	0.875
中午	正午	0.931	0.931	1.000	0.931	0.855
火炉	炉灶	0.753	0.753	0.764	0.753	0.777 5
食物	水果	0.144	0.144	0.106	0.144	0.77
鸟	公鸡	0.377	0.377	0.312	0.377	0.762 5
鸟	鹤	0.377	0.377	0.312	0.377	0.742 5
工具	器械	0.349	0.349	0.284	0.349	0.737 5
兄弟	和尚	0.254	0.254	0.197	0.254	0.705
起重机	器械	0.349	0.349	0.284	0.349	0.42
小伙子	兄弟	0.254	0.254	0.197	0.254	0.415
旅行	轿车	0.000	0.000	0	0.000	0.29
和尚	圣贤	0.254	0.254	0.197	0.254	0.275
墓地	林地	0.370	0.370	0.304	0.370	0.237 5
食物	公鸡	0.144	0.144	0.106	0.144	0.222 5
海岸	丘陵	0.516	0.516	0.459	0.516	0.217 5
森林	墓地	0.000	0.000	0	0.000	0.21
岸边	林地	0.144	0.144	0.106	0.144	0.157 5
和尚	奴隶	0.254	0.254	0.197	0.254	0.137 5
海岸	森林	0.144	0.144	0.106	0.144	0.105
小伙子	巫师	0.254	0.254	0.197	0.254	0.105
琴弦	微笑	0.000	0.000	0	0.000	0.032 5
玻璃	魔术师	0.000	0.000	0	0.000	0.027 5
中午	绳子	0.000	0.000	0	0.000	0.02
公鸡	远行	0.000	0.000	0	0.000	0.02

3.2.2　实验 2

实验 2 使用与实验 1 相同的公式与方法,但本体使用本文修订后的词林,得出在 MC30 数据集上的相似度测量结果如表 4 所示。

3.2.3　对比汇总

通过前后两个实验的对比,可以看出经过本文修改后的词林使用式(5)(6)以及本文公式进行计算时,一些词语的相似度值变大了,这是因为经过本文修订后的词林中一些概念的层次提高了,具有了多个下位节点(在使用词林作为本体的实验中,所有的概念都处在叶子节点的位置)。表 5 展示了一些相似度变化较大的概念在下位节点数量上的变化。

表 4　修订后词林 MC30 数据集实验结果

词语 1	词语 2	Resnik 式(1)	Lin 式(5)	Meng 式(6)	本文公式	MC 人工判定值(参考值)
轿车	汽车	0.665	0.665	0.583	0.665	0.98
宝石	宝物	0.793	0.885	0.876	0.897	0.96
旅行	远行	0.771	0.871	0.856	0.886	0.96
男孩子	小伙子	0.793	0.793	0.746	0.793	0.94
海岸	海滨	0.840	0.840	0.811	0.840	0.925
庇护所	精神病院	0.793	0.793	0.746	0.793	0.902 5
魔术师	巫师	0.745	0.788	0.739	0.800	0.875
中午	正午	0.931	0.964	1	0.966	0.855
火炉	炉灶	0.753	0.859	0.839	0.877	0.777 5
食物	水果	0.144	0.219	0.151	0.484	0.77
鸟	公鸡	0.377	0.467	0.367	0.570	0.762 5
鸟	鹤	0.377	0.511	0.411	0.639	0.742 5
工具	器械	0.349	0.386	0.29	0.445	0.737 5
兄弟	和尚	0.254	0.287	0.205	0.369	0.705
起重机	器械	0.349	0.349	0.257	0.349	0.42
小伙子	兄弟	0.254	0.287	0.205	0.369	0.415
旅行	轿车	0.000	0.000	0	0.114	0.29
和尚	圣贤	0.254	0.277	0.196	0.334	0.275
墓地	林地	0.370	0.370	0.276	0.370	0.237 5
食物	公鸡	0.144	0.167	0.112	0.282	0.222 5
海岸	丘陵	0.516	0.516	0.416	0.516	0.217 5
森林	墓地	0.000	0.000	0	0.000	0.21
岸边	林地	0.144	0.144	0.096	0.144	0.157 5
和尚	奴隶	0.254	0.269	0.19	0.309	0.137 5
海岸	森林	0.144	0.144	0.096	0.144	0.105
小伙子	巫师	0.254	0.269	0.19	0.309	0.105
琴弦	微笑	0.000	0.000	0	0.000	0.032 5
玻璃	魔术师	0.000	0.000	0	0.103	0.027 5
中午	绳子	0.000	0.000	0	0.034	0.02
公鸡	远行	0.000	0.000	0	0.000	0.02

表 5　修订后词林词语层次变化

概念名称		概念所在层次		概念所含下位节点		最近公共父节点所含下位数
概念 1	概念 2	概念 1	概念 2	概念 1	概念 2	
宝石	宝物	5	4	1	7	7
旅行	远行	4	5	9	0	9
火炉	炉灶	3	5	11	0	11
食物	水果	3	3	15	57	5 503
鸟	公鸡	4	5	48	0	528
鸟	鹤	4	4	48	3	528
工具	器械	4	5	6	0	703
兄弟	和尚	4	5	9	0	1 817
食物	公鸡	3	5	15	0	5 503

将两组数据集上得出的实验结果与人工判定值进行 Pearson 相关系数分析,并结合对比本文所介绍的公式在以 WordNet 为本体的计算中得到的相似度数据,对比结果如表 6 所示。

4　结果分析

通过上述对比实验,可以得出以下结论:

a) 从表 3、4 可以看出,在使用相同公式分别在词林和本文修订后的词林两个本体上进行测试时,本文修订后的词林能够取得较好的结果,证明了本文修订后的词林能够更好地反映词语之间的差异性,概念分布更为科学。

b) 从表 6 可以看出,所有公式使用词林作为本体计算出的相似度与人工判定值的 Pearson 相关系数都在 0.8 之上,超

过了 WordNet 在词语相似度测量中的表现，说明词林在计算词语相似度方面的应用具有较高的价值。

c) 根据 Renisk 的研究成果，人与人之间词语相似度判定的 Pearson 相关系数为 0.901 5[14]，得出了计算机词语相似度测量与人工判定之间相关度的理论上限为 0.90。本文公式与本文修订后的词林相结合的词语相似度测量，在 MC30 数据集中与人工判定值相比取得了 0.899 的 Pearson 相关系数，该值已非常逼近计算机词语相似度测量的理论上限，从而证明了本文提出的相似度计算公式能较好地利用修订后的词林分类结构。

表 6 Pearson 相关系数分析

方法名称	公式名称	使用本体	RG65 中的 Pearson 相关系数	MC30 中的 Pearson 相关系数	测试来源
Resnik [24]	式(1)	WordNet	0.72	0.72	文献[19]
Lin [18]	式(5)	WordNet	0.72	0.70	文献[19]
Jiang [27]	式(6)	WordNet	0.75	0.73	文献[19]
Resnik [24]	式(1)	词林	0.802 3	0.830 4	本文
Lin [18]	式(5)	词林	0.802 3	0.830 4	本文
Meng [28]	式(6)	词林	0.813 2	0.818 0	本文
本文公式	式(7)	词林	0.820 1	0.830 1	本文
Resnik [24]	式(1)	本文修订后的词林	0.820 3	0.830 4	本文
Lin [18]	式(5)	本文修订后的词林	0.845 4	0.856 8	本文
Meng [28]	式(6)	本文修订后的词林	0.840 2	0.841 2	本文
本文公式	式(7)	本文修订后的词林	0.870 1	0.899 0	本文

5 结束语

本文将基于信息内容的词语相似度计算方法运用到词林当中，并对词林的拓扑结构进行了修改，将一些具有抽象含义的概念放置到较高的节点中，使之更适应于信息内容相似度算法。经实验证明，使用本文修订后的词林配合本文所提出的信息内容相似度算法计算出的词语相似度与人工判定值高度相似，说明了本文修订后的词林具有较合理的拓扑结构及较高的实用价值，也为今后将词林运用到更多自然语言处理领域开阔了道路。

参考文献:

[1] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation [C]//Lecture Notes in Computer Science, vol 2588. 2003: 241-257.

[2] Atkinson J, Ferreira A, Aravena E. Discovering implicit intention-level knowledge from natural-language texts [J]. Knowledge-Based Systems, 2009, 22(7): 502-508.

[3] Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective [J]. Journal of Biomedical Informatics, 2011, 44(5): 749-759.

[4] 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.

[5] 石静，吴云芳，邱立坤，等. 基于大规模语料库的汉语词义相似度计算方法 [J]. 中文信息学报，2013，27(1)：1-6.

[6] Li Yuhua, Bandar Z A, McLean D. An approach for measuring semantic similarity between words using multiple information sources [J]. IEEE Trans on Knowledge & Data Engineering, 2003, 15(4): 871-882.

[7] 梅立军，周强，臧路，等. 知网与同义词词林的信息融合研究 [J]. 中文信息学报，2005，19(1)：63-70.

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

[9] 董振东，董强.《知网》[DB/OL]. http://www.keenage.com.

[10] 哈工大社会计算与信息检索研究中心同义词词林扩展版 [EB/OL]. (2015-09-13). http://www.datatang.com/data/42306/.

[11] 于江生，俞士汶. 中文概念词典的结构 [J]. 中文信息学报，2002，16(4)：12-20.

[12] 刘群，李素建. 基于《知网》的词汇语义相似度计算 [D]. 北京：中国科学院计算技术研究所，2002.

[13] 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，2010，28(6)：602-608.

[14] 张亮，尹存燕，陈家骏. 基于语义树的中文词语相似度计算与分析 [J]. 中文信息学报，2010，24(6)：23-30.

[15] 吴佐衍，王宇. 基于 HNC 理论的词语相似度计算 [J]. 中文信息学报，2014，28(2)：37-43.

[16] 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.

[17] 许嘉璐. 现状和设想——试论中文信息处理与现代汉语研究 [J]. 中文信息学报，2001，15(2)：1-8.

[18] Lin Dekang. An information-theoretic definition of similarity [C]//Proc of the 15th International Conference on Machine Learning. [S. l.]: Morgan Kaufmann Publishers Inc, 1998: 296-304.

[19] Hadj T M A, Ben A M, Ben H A. Ontology-based approach for measuring semantic similarity [J]. Engineering Applications of Artificial Intelligence, 2014, 36(C): 238-261.

[20] Wu Zhibiao, Palmer M. Verb semantics and lexical selection [C]//Proc of Annual Meeting on Association for Computational Linguistics. 1995: 133-138.

[21] Hadjtaieb M A, Aouicha M B, Hamadou A B. A new semantic relatedness measurement using WordNet features [J]. Knowledge and Information Systems, 2014, 41(2): 467-497.

[22] Sánchez D, Solé-Ribalta A, Batet M, *et al*. Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain [J]. Journal of Biomedical Informatics, 2012, 45(1): 141-155.

[23] Zhou Zili, Wang Yanna, Gu Junzhong. A new model of information content for semantic similarity in WordNet [C]//Proc of International Conference on Future Generation Communication and Networking Symposium. 2008: 85-89.

[24] Resnik P. Using information content to evaluate semantic similarity in a taxonomy [C]//Proc of International Joint Conference on Artificial Intelligence. [S. l.]: Morgan Kaufmann Publishers Inc, 1995: 448-453.

[25] Cover T M, Thomas J A. 信息论基础 [M]. 阮吉寿，张华，译. 北京：机械工业出版社，2005.

[26] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet [C]//Proc of Eureopean Conference on Artificial Intelligence. 2004: 1089-1090.

[27] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy [C]//Proc of the 10th International Conference on Research in Computational Linguistics. 1997.

[28] Meng Lingling, Gu Junzhong, Zhou Zili. A new model of information content based on concept's topology for measuring semantic similarity in WordNet [J]. International Journal of Grid & Distributed Computing, 2012, 5(3): 81-94.

[29] 梅家驹. 词林 [M]. 上海：上海辞书出版社，1983.

[30] 刘丹丹，彭成，钱龙华，等.《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(2)：91-99.

[31] Miller G A, Charles W G. Contextual correlates of semantic similarity [J]. Language Cognition & Neuroscience, 1991, 6(1): 1-28.

[32] Rubenstein H, Goodenough J B. Contextual correlates of synonymy [J]. Communications of the ACM, 1965, 8(10): 627-633.