基于多重继承与信息内容的知网词语相似度计算

作者	张波，陈宏朝，朱新华，吴田俊
机构	贺州学院 数学与计算机学院；广西师范大学 计算机科学与信息工程学院
基金项目	国家自然科学基金资助项目（61462010，61363036）
预排期卷	《计算机应用研究》 2018 年第 35 卷第 10 期
摘要	针对目前中文词语语义相似度方法中，基于信息内容的算法研究不足的问题，对知网信息模型上使用基于信息内容的中文词语相似度算法进行了研究。根据知网采用语义表达式表示知识而缺乏完整概念结构的特点，通过抽取知网语义表达式中的抽象概念，结合原知网义原树构建具有多重继承特征的知网义项网作为基于信息内容的计算本体。根据该义项网，对基于信息内容的词语相似度算法进行了改进，提出了新的信息内容含量计算方法。经过Miller&Charles（MC30）基准平台的测试，验证了基于信息内容方法在计算中文语义相似度方面的可行性，也证明了本文的计算策略和改进算法的合理性。
关键词	词语相似度；知网；多重继承；描述逻辑
作者简介	张波（1983-），男，山西长治人，讲师，硕士，主要研究方向为自然语言处理、远程教育技术 ;陈宏朝(1963-),男,广西玉林人,副教授,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）；吴田俊（1992-），男，硕士研究生，主要研究方向为自然语言处理、文本相似度计算 .
中图分类号	TP391.1
访问地址	http://www.arocmag.com/article/02-2018-10-021.html
发布日期	2017 年 9 月 27 日
引用格式	张波，陈宏朝，朱新华，吴田俊. 基于多重继承与信息内容的知网词语相似度计算[J/OL]. 2018, 35(10). [2017-09-27]. http://www.arocmag.com/article/02-2018-10-021.html.

基于多重继承与信息内容的知网词语相似度计算 *

张 波[1]，陈宏朝[2]，朱新华[2†]，吴田俊[2]

(1. 贺州学院 数学与计算机学院，广西 贺州 542899; 2. 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：针对目前中文词语语义相似度方法中，基于信息内容的算法研究不足的问题，对知网信息模型上使用基于信息内容的中文词语相似度算法进行了研究。根据知网采用语义表达式表示知识而缺乏完整概念结构的特点，通过抽取知网语义表达式中的抽象概念,结合原知网义原树构建具有多重继承特征的知网义项网作为基于信息内容的计算本体。根据该义项网,对基于信息内容的词语相似度算法进行了改进,提出了新的信息内容含量计算方法。经过Miller&Charles（MC30）基准平台的测试，验证了基于信息内容方法在计算中文语义相似度方面的可行性，也证明了本文的计算策略和改进算法的合理性。

关键词：词语相似度；知网；多重继承；描述逻辑

中图分类号：TP391.1

Based on multi-inheritance and IC approach for calculating word semantic similarity in Hownet

Zhang Bo[1], Chen Hongchao[2], Zhu Xinhua[2†], Wu Tianjun[2]

(1. *School of Mathematics & Computer Science*, *Hezhou University*, *Hezhou Guangxi* 542899; *China*; 2. *College of Computer Science & Information Technology Guangxi Normal University*, *Guilin Guangxi* 541004, *China*)

Abstract: For the issue of insufficient research on IC-based Chinese word semantic similarity approach, this paper researched the IC-based Chinese word semantic similarity measurement on Hownet information model. According to the characteristics of the Hownet, which takes the concept semantic expression to represent knowledge and lacks of complete conceptual structure, this paper constructed a multi-inheritance-based Hownet concept net as the ontology of IC-based measurement by extracting the abstract concepts from the concept semantic expression, and extending the original Hownet sense structure. Based on it, this paper studied the given IC-based Chinese word semantic similarity, and then improved the IC approach. The test results on MC30 show that the IC-based Chinese word semantic similarity measurement is feasible. The test results also verifie the validity of this method on Chinese semantic similarity measurement.

Key Words: word similarity; HowNet; Multi-Inheritance; Description logic

0 引言

词语相似度计算是自然语言处理领域非常重要的基础研究问题之一，这个问题的研究直接关系到文本分类、信息提取、自动摘要、词义消岐等研究，被广泛应用于问答系统、机器翻译、数据挖掘和智能教育等领域。目前，词语相似度的计算主要包括两种方法论：其一是使用统计学的方法把词语相似度问题转换为词语在语料库中上下文信息概率分布问题，根据词语之间在语料库中出现的次数和分布情况进行计算；其二是基于某个世界知识的本体模型，通过计算词语在本体概念中的共性和差异性进行计算[1]。目前研究词语相似度所使用的本体模型包括英文的 WordNet[2]和中文的《知网》[3]（英文名称为 HowNet）与哈工大《同义词词林》[4]。在目前相似度计算研究中，基于 WordNet 的英文词语相似度研究的理论体系已经日臻完善，其主流方法是基于信息内容的计算方法[5]，而基于知网和同义词词林的中文词语相似度的计算方法还比较单一，主要是基于路径和基于意义相似度算法，鲜有基于信息内容的计算方法研究。本文将基于信息内容(information content,IC)的词语相似度计算方法引入以知网为本体的相似度计算中，并根据知网采用多维知识表示形式、缺乏义项树的特点，通过抽取抽象概念构建由

基金项目：国家自然科学基金资助项目（61462010，61363036）

作者简介：张波（1983-），男，山西长治人，讲师，硕士，主要研究方向为自然语言处理、远程教育技术；陈宏朝（1963-），男，广西玉林人，副教授，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）；吴田俊（1992-），男，硕士研究生，主要研究方向为自然语言处理、文本相似度计算.

义原和抽象概念组成的义项树，并把所有义项连接为义项树的叶子节点而形成一棵具有多重继承特征的义项网，以此为基础提出了一个适合知网本体的信息内容含量计算方法以改进中文词语相似度计算。

1　研究背景

1.1　知网简介

知网[3]是董振东先生历经数十年建设的一个中英文常识库，目前依然在不断扩展中。在知网中，义原是描述义项的最基本单位，分为事件、实体、属性、属性值、数量、数量值、次要特征、语法、动态角色和动态属性等 10 大类，2000 版的知网共有 1618 个义原。通过上下位关系，所有的知网义原组成一个树状义原层次体系，如图 1 所示。

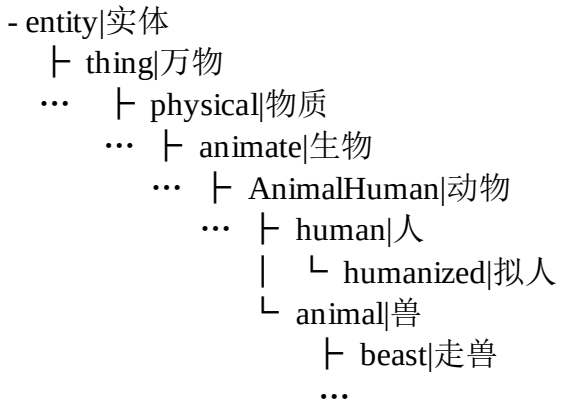

图 1　义原的树状层次结构

与一般的语义词典（如《同义词词林》或 Wordnet）不同，知网中所有的义项（又称“概念”）并不是组织为树状概念结构，而是采用“义原”对义项进行描述定义。每一个词语的义项以<W_X=词语，E_X=词语例子，G_X=词语词性，DEF=概念定义>的四元组进行描述。其中 DEF 项又称为语义表达式，是知网信息模型的核心，它给定了义项的定义描述。除了义项和语义表达式外，另一个核心概念是义原，它是描述一个义项的最小意义单位。通过上下位关系，所有义原被组织成一个树状层次结构。每一个语义表达式由一到多个用逗号隔开的“语义描述式”组成，按照刘群的划分方法，语义描述式分为以下三类：

a）独立义原描述式：其内容可以是义原，也可以是“(具体词)”的形式，根据义原在语义表达式中的排列位置又可以细分为第一独立义原和其他独立义原。

b）关系义原描述式：以“关系义原=义原”或者“关系义原=(具体词)”或者“(关系义原=具体词)”的形式表示，用于定义义项之间的关系。

c）符号义原描述式：以“<关系符号>义原”或者“<关系符号>(具体词)”的形式表示，用于定义多个义项之间的关系。其中主要的关系符号及含义如表 1 所示。

比如词语“编程”的一个义项的语义表达式为“DEF=compile|编辑,ContentProduct=software|软件,#computer|电脑”。该表达式由 3 个语义描述式组成，其中“compile|编辑”为独立义原描述式，“ContentProduct=software|软件”是关系义原描述式，“#computer|电脑”是符号义原描述式。该义项表达式的基本含义是:“编程”是一种与电脑相关的软件产品的编辑。

表 1　《知网》中的常用符号及其含义

符号	含义
,	表示“和”的关系
#	表示“与其相关”
%	表示“是其部分”
$	表示“可以被该‘V’处置，或是该“V”的受事，对象，领有物，或者内容
*	表示“会‘V’或主要用于‘V’，即施事或工具
+	对 V 类，它表示它所标记的角色是一种隐性的，几乎在实际语言中不会出现
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空间或时间
?	表示可以是“N”的材料，
{ }	(1) 对于 V 类，置于[ ]中的是该类 V 所有的“必备角色”。(2)表示动态角色，如介词的定义
( )	置于其中的应该是一个词表记
^	表示不存在，或没有，或不能
!	表示某一属性为一种敏感的属性，
[ ]	标识概念的共性属性

1.2　基于知网的词语相似度的研究现状

在计算两个词语的相似度的时候，一般的计算方法是计算两个词语所具有的义项之间的相似度的最大值，把词语相似度计算问题转换为义项相似度的计算问题。假设计算词语 W1 和 W2 的相似度，设定 W1 包含 n 个义项 S_{11}，S_{12}，……，S_{1n}，W2 包含 m 个义项：S_{21}，S_{22}，……，S_{2m}，则 W_1 和 W_2 的相似度计算公式为

$$\mathrm{Sim}(W_1,W_2)=\max_{i=1..n,\,j=1..m}\{\mathrm{Sim}(S_{1i},S_{2j})\} \tag{1}$$

基于知网的词语相似度计算方面，影响最大的是刘群等人[6]在 2002 年提出的计算方法。该方法根据知网义项是通过语义表达式（DEF）描述的特点，把义项相似度计算转换为语义表达式的相似度计算问题。他们采取对四类语义表达式分别计算相似度并进行加权求和的方式取得义项相似度，计算公式如下：

$$\mathrm{Sim}(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}\mathrm{Sim}_j(S_1,S_2) \tag{2}$$

其中:$Sim_1(S_1,S_2)$表示两个义项的第一独立义原描述式的相似度，$Sim_2(S_1,S_2)$表示其他独立义原描述式的相似度，$Sim_3(S_1,S_2)$表示关系义原描述式的相似度，$Sim_4(S_1,S_2)$表示符号义原描述式的相似度。该公式包括 4 个调节参数 $\beta_1,\beta_2,\beta_3,\beta_4$，且：$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。义原描述式的相似度是通过

计算所包含的义原之间的相似度而得。由于知网的义原根据上下位关系构成了一个树状结构，义原相似度采用了基于路径的相似度算法。计算公式如下：

$$\mathrm{Sim}(p_1,p_2)=\frac{a}{d+a} \tag{3}$$

其中: p_1 和 p_2 表示义原，d 是 p_1 和 p_2 在义原层次体系中的最短路径长度。α 是一个可调节的参数。在刘群的研究中，调节参数的值分别是：α= 1.6, β_1 = 0.5, β_2 = 0.2, β_3 = 0.17,β_4 = 0.13。

此后基于知网词语相似度的研究大都沿用了刘群的计算思维并进行了算法改进。比如张敏[7]改进了调节参数 β_i（1≤i≤4）的设置方法减少了计算时间并提高了计算结果的合理度。朱征宇[8]增加了位置相关的权重分配策略，用二部图最大权匹配进行计算，以提高计算结果的 F 值(F-measure)。葛斌[9]的算法考虑了义原树的深度、区域密度等因素改进了义原的相似度算法，该算法提高了微小语义之间的差异性。王小林[10]在刘群算法的基础上结合了基于信息量的计算方法，并考虑了词语密度，以去除相似度较低的义项组合对词语语义相似度的不利影响。魏韡和刘杰[11, 12]等人研究了新版知网文件的相似度计算问题，其计算方法依然沿用了刘群的思想。这些研究都一定范围内提高了计算的精度，改善了刘群算法的准确度，但是本质上依然是基于路径的相似度计算方法。

1.3　基于信息内容的词语相似度计算方法

最早提出基于信息内容的相似度算法的是 1995 年 Resnik[13]的研究，他认为，两个概念（义项）的相似度取决于两个概念共享信息的多少。在一个层次结构良好的本体中，两个概念共享信息可以通过两个概念的最近公共父节点的信息含量进行计算。因此，该方法的计算过程取决于两个子计算方法：每一个概念的信息内容含量（IC）的计算方法、基于信息内容含量的词语相似度（Sim）计算方法。

Resnik 提出的基于信息内容含量词语相似度计算公式为

$$\mathrm{Sim}_{\mathrm{Resnik}}(C_1,C_2)=IC\left(LCS(C_1,C_2)\right) \tag{4}$$

其中:C1 和 C2 表示概念，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。

Jiang 和 Conrath[14]于 1997 年结合式（4）提出了根据概念差异性计算相似度的方法，他们认为两个概念的相似度应该关注概念之间的差异性，差异性越大相似度越小。他们提出的差异性计算公式为

$$\mathrm{dis}_{JC}(C_1,C_2)=IC(C_1)+IC(C_2)-2IC\left(LCS(C_1,C_2)\right) \tag{5}$$

随后，Lin[15]在此研究基础上提出了一种兼顾共性和差异性的计算方法，公式如下：

$$Sim_{Lin}(C_1,C_2)=\frac{2\times IC\left(LCS(C_1,C_2)\right)}{IC(C_1)+IC(C_2)} \tag{6}$$

其中:分子部分表示 $\mathbf{C_1}$ 和 C_2 的共性，分母部分通过计算 C_1 和 C_2 各自的信息内容含量之和的方式体现 C_1 和 C_2 的差异性。

Resnik[13]提出的计算概念的信息内容(IC)的含量公式为

$$IC_{\mathrm{Resnik}}(C)=-\log P(C) \tag{7}$$

其中:P(C)为概念 C 所在测量样本总体中的或然率，负号表示概率越大所具有的信息含量越小。在这一理论的基础上，Seco 等人[16]提出的基于 WordNet 的 IC 计算公式：

$$IC_{\mathrm{Seco}}(C)=1-\frac{\log\left(|hypo(C)|+1\right)}{\log(\max\_nodes)} \tag{8}$$

其中:max_nodes 表示本体的最大节点数，|hypo(C)|表示概念 C 在本体层次结构中的所有下位节点数，下位节点总数加 1 是为了解决叶子节点不存在下位节点的问题。根据该公式，概念层次越高，说明概念越普遍，所含的信息内容含量越小，反之同理。

Zhou 等人[17]的研究在 Seco 公式的基础上提出了一个新的方法，该方法考虑了概念所处的深度因素，解决了式（8）在不同深度的概念的下位节点数总数相同的偏差，提高了计算精度。Zhou 提出的计算公式为

$$\begin{aligned}IC_{\mathrm{zhou}}(C)=k\left(1-\frac{\log\left(|hypo(C)|+1\right)}{\log(\max\_\mathrm{nodes})}\right)+\\(1-k)\left(\frac{\log(\mathrm{depth}(C))}{\log(\max\_\mathrm{depth})}\right)\end{aligned} \tag{9}$$

其中:depth(c)表示概念 C 在本体层次结构中的深度，max_depth 表示本体层次结构的最大深度，k 是调整权重因子，在 Zhou 的基于 WordNet 词语相似度计算中取 0.5。

2　基于多重继承与信息内容的知网词语相似度计算

2.1　基于义原树与多重继承的义项网

根据董振东对知网的定义，知网是一个用“知网知识系统描述语言”定义的常识知识库，每一个用于定义义项的语义表达式本质上是一种简化的描述逻辑（DL）[18]。在义项的语义表达式中，独立义原描述式用于定义义项的主要语义特征与上位义原，关系义原描述式和符号义原描述式均是对其前面的语义描述式的语义限定。比如：“编程”的一个义项的语义表达式“DEF=compile|编辑,ContentProduct=software|软件,#computer|电脑”，相当于描述逻辑中对“编程”的蕴涵公理定义：

$$\text{编程}\sqsubseteq \text{compile|编辑}\cap\exists\text{ContentProduct.software|软件}\cap\exists\text{\#.computer|电脑}$$

在该公理定义中，由于知网仍在不断发展中，义项的定义还在不断完善，所以公理采用概念蕴涵或特化（$\sqsubseteq$）而不是概念定义。在描述逻辑的语法中包括的算子有：$\neg$C（补）、C∩D（交）、C∪D（并）、$\exists$R.T（存在约束）和$\forall$ R.T（全称约束）。在上例中，“$\exists$ContentProduct. software|软件”是对独立义原“compile|编辑”的限定，表示其编辑内容为某种软件。在知网中，关系符号‘#’表示“与其相关”，在本例中，“$\exists$#. computer|电脑”是对“compile|编辑 ∩ $\exists$ContentProduct. software|软件”的限定，表示其前面的语义描述式存在与‘电脑’相关的关系

约束。本语义表达式的三个义原描述式组合起来，形成了“编程”的上位概念，它们定义了“编程”的语义为：属于一种与电脑相关的软件产品的编辑。

在本文中，为了便于建立计算词语相似度所需的义项分类结构，将知网中现存的义项和义原统称为“实概念”，而义项的语义表达式中的独立义原与起限定作用的非独立义原通过交运算（∩）组成的概念称为“抽象概念”。实概念表示该概念在现实中真实存在，而抽象概念表示该概念仅用于对其他概念的解释和分类，在现实中并不存在具体的词语与其对应。在上例中，“编辑”、“软件”和“电脑”是实概念，“compile|编辑∩∃ContentProduct. software|软件”和“compile|编辑∩∃ContentProduct. software|软件∩∃#. computer|电脑”为虚（抽象）概念。

由于知网义项的语义表达式中可能存在多个独立义原描述式，每一个独立义原与其紧接的非独立义原通过交运算（∩）形成概念，所以一个义项可能有多个概念作为其上位概念。每一个上位概念通过并运算（∪）形成对该义项的蕴涵公理定义。比如义项“国防部长”的蕴涵公理定义为：

$$\text{国防部长} \subseteq (\text{human}|\text{人} \cap \#.\text{occupation}|\text{职位}) \cup (\text{official}|\text{官} \cap *.\text{manage}|\text{管理} \cap \#.\text{country}|\text{国家}) \cup (\text{military}|\text{军})$$

此例中，抽象概念“human|人∩#.occupation|职位”、“official|官∩*.manage|管理∩#.country|国家”，以及实概念“military|军”均是义项“国防部长”的上位概念。

在知网中，通过义原之间的上下位关系，已经存在一个树状义原层次体系，该层次体系中包括所有的义原。根据义项的表达式可以提取抽象概念和实概念，并形成义项的蕴涵公理定义表达式。由于原知网的义原层次体系只包括实概念而不包括抽象概念，需要为义项公理定义表达式中的抽象概念逐层找到其直接上位概念，并挂在原有的义原树上，形成一个包含知网所有概念的义项结构。由于义项的语义表达式由一到多个概念通过并运算（∪）组成，所以义项可能有多个上位概念，而由义项组成的义项结构是一个具有多重继承特征的义项网。在本文所形成的义项网中，每一个抽象概念的直接上位概念必须是与该概念具有相同的独立义原和更少的语义限定，且该上位概念必须是独立义原或是独立存在于其他义项定义中的抽象概念。比如“official|官∩*.manage|管理∩#.country|国家”的上位概念是“official|官∩*.manage|管理”，该抽象上位独立存在于“把头”等义项的定义中；而由于在任何义项定义中均不独立存在“police|警∩*.check|查”抽象概念，所以“police|警∩*.check|查∩#.crime|罪”的上位概念是“police|警”而不是“police|警∩*.check|查”。设当前所处理的抽象概念所包含的约束条件的个数为 N，则本文的基于义原树的义项网具体的计算步骤为：

a）对每一个义项的语义表达式进行交并处理：把义项的每一个独立义原与其后的所有紧接的非独立义原通过交运算（∩）结合为抽象概念，然后将每一个抽象概念或独立义原通过并运算（∪）形成义项的蕴涵公理定义。比如：“国防部长”的蕴涵公理定义是：(human|人∩#.occupation|职位)∪(official|官∩*.manage|管理∩#.country|国家)∪(military|军)。

b）遍历所有义项蕴涵公理，找出无任何约束的单一独立义原，作为义项的一个直接上位，在义原树中，直接将义项挂在该独立义原下，此时 N=0。

c）遍历所有义项蕴涵公理，找出带有一个约束条件的抽象概念，然后在义原树中，将该抽象概念挂在其所包含的独立义原下，义项挂在该抽象概念下，此时 N=1。

d）重复步骤 c)，置 N=N+1，遍历所有义项蕴涵公理，找出带有 N 个约束条件的抽象概念 CA_{n+1}，然后判断减少一个约束条件的抽象概念是否在义项网中存在，若存在，则将抽象概念 CA_{n+1} 挂在该概念下，义项挂在抽象概念 CA_{n+1} 下；否则进一步减少一个约束条件，并作相同的判断与处理，直到无任何约束的独立义原。

e）重复步骤 d)，直到将所有义项蕴涵公理中的抽象概念挂到基于义原树的义项网中，同时将义项挂在其蕴涵公理中的所有抽象概念下，形成义项的多重继承，如图 2 所示。

新构建的义项网的枝节点由义原和抽象概念组成，每一个枝节点只有一个上位，所以所有枝节点构成的是标准树状结构。义项网的叶子节点由知网中所有的义项组成，如图 2 所示。由于一个词语可能有多个义项，每一个义项可能有多个上位节点（抽象概念或实概念），所以体现了义项网的多重继承特征，但这种多重继承仅表现在叶子节点（义项）上。

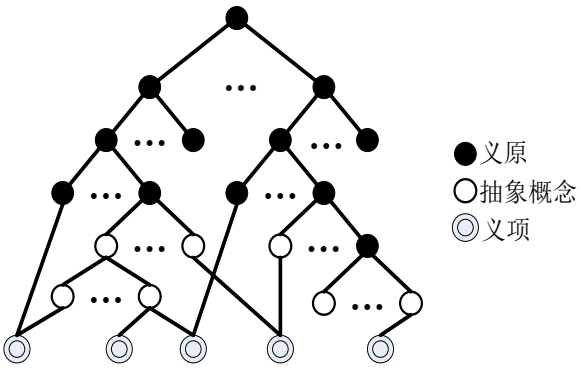

图 2　义项网示意图

该义项网的节点总数为 67883，其中实概念有 57153 个，包括原知网义原树的 1618 个义原和 55535 个义项，抽象概念有 10730 个，最大深度为 17。

2.2　信息内容的计算方法的改进

通过式（1），可将计算两个词语相似度问题转换为词语概念的相似度计算，本章节重点考虑计算两个概念的相似度方法。基于信息内容的相似度算法主要取决于概念信息内容含量（IC）的计算和基于 IC 的概念相似度（Sim）计算两个方法。

本文采用 Lin 提出的式（6）作为 Sim 计算方法。根据式（6），两个概念 C_1 和 C_2 的相似度取决于其最近公共父节点（LCS）的信息内容含量和两个概念本身的信息内容含量。

在概念的 IC 计算方法上，本文采用了 Zhou 的算法的基本

思想，并在式（8）的基础上进一步考虑了概念所处的深度与节点数的关系。由于在本文所构建的知网义项网中，随着概念深度的增加，概念的上位节点数呈指数级增加而其下位节点数呈指数级减少，即概念的信息内容含量与其深度之间是线性关系而不是对数关系。据此对式（9）作以下改进：

$$IC_{zhang}(C)=k\left(1-\frac{\log\left(|hypo(C)|+1\right)}{\log(\max\_nodes)}\right)+(1-k)\left(\frac{depth(C)}{\max\_depth}\right) \tag{10}$$

其中：depth(C)表示概念 C 在知网义项网中的深度，max_depth 表示知网义项网结构的最大深度，k 是权重因子。$\frac{depth(C)}{\max\_depth}$ 部分是在式（9）的基础上，将 $\frac{\log(depth(C))}{\log(\max\_depth)}$ 中的分子和分母分别换算为对应深度所包含节点总数的对数，同时分别采用以 e 为底数、深度为幂的指数来表示该深度所包含的节点总数，即通过 $\frac{\log e^{depth(C)}}{\log e^{\max\_depth}}$ 计算而来。经本文基于 MC30 数据集测试，k 在 0.75-0.85 之间采用式（10）计算相似度与人工判定值之间的皮尔森相关系数最高，如图 3 所示，本文设定 k 值为 0.8。

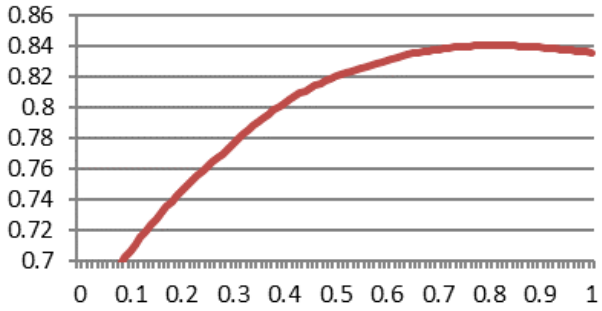

图 3　皮尔森相关系数随权重因子 k 的变化图

3　实验与分析

3.1　实验比较

本文采用国际上广泛使用的由 Miller&Charles（MC）[19]发布的 30 对词语的数据集作为测试集。MC30 测试集是目前国际普遍采用的词语相似度的测试平台，有着规范的人工判定标准。MC30 测试集是由 10 对高度相关、10 对中度相关、10 对低度相关的词对组成，词对的人工判定相似度由 38 个受试者的人工语义相似度判断取平均值计算而来。首先将 MC30 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，然后分别采用刘群算法（式（2））、采用 Seco 的 IC 算法（式(8）)的 Lin 的相似度算法、采用 Zhou 的 IC 算法（式(9)）的 Lin 的相似度算法，以及采用本文提出的改进 IC 式(10)的 Lin 的计算方法，并分别计算这些相似度方法计算结果与 MC30 人工判定之间的皮尔森相关系数。本文还对比了不使用多重继承的知网义项网的基于信息内容相似度计算结果的差别。即把语义表达式中每一个义原描述式直接挂在义原树上，而不采用抽象概念的方式，直接计算基于信息内容的语义相似度的计算结果。计算数据如表 2 所示。

为了对比基于多重继承的知网义项网在基于信息内容的计算词语相似度方面的优越性，本文比较了若干基于 WordNet 的基于信息内容的 MC30 计算结果和典型的基于知网和同义词词林的计算结果。如表 3 所示。

3.2　实验结果分析

对比表 2，可以发现在基于信息内容的相似度计算中，采用多重继承的知网义项网的结果数据均优于使用知网原来的义原树的计算结果，使用多重继承的知网义项网的计算值更接近人的主观判断，这说明了基于多重继承的知网义项网的基于信息内容方法是比较真实可信的。同时发现基于信息内容的知网语义相似度算法结果优于刘群提出的四类语义表达式分别计算相似度并进行加权求和的方式的计算结果，即便是使用知网原来的义原树进行相似度计算，MC30 测试集计算结果与人工判定值的皮尔森相关系数均高于刘群的算法，这也说明基于信息内容相似度计算的优势，基于信息内容计算相似度的方法在中文知网计算相似度方面同样有比较好的表现。本文提出的式(10）是根据知网多重继承义项网的特点在 Zhou 的式（9）的基础上进行的改进，测试结果表示，同样的 Lin 的相似度计算方法，采用本文提出的 IC 方法比采用 Zhou 的 IC 方法与人工判定的皮尔森相关系数更高。在表 2 中，采用知网多重继承义项网的基于信息内容的相似度与人工判定值的皮尔森相关系数均高于 0.8，说明知网多重继承义项网在计算词语相似度方面具有较高的应用价值。对比表 3 中，本文提出的 IC 改进算法在基于多重继承知网义项网上采用 Lin 的相似度计算结果高于大部分基于知网和基于同义词林的算法，也优于与采用 WordNet 为语义词典的各种方法，说明了本文提出的算法是成功的。

4　结束语

本文将基于信息内容的词语相似度计算方法引入到以知网为本体的词语相似度计算中，并根据知网义项定义的特点，通过抽取知网语义表达式的抽象概念，把抽象概念加入到义原树中，与义项一起组成基于多重继承的知网义项网作为基于信息内容的计算本体。并根据构建的义项网的特点改进了基于信息内容词语相似度计算方法。实验表明，相对于其他基于知网的计算方法，本文提出的以知网为本体的基于信息内容的中文词语相似度计算方法有较好的表现，说明本文构建的基于多重继承的知网义项网有较为合理的义项结构，提出的改进方法也有较好的应用价值。

表 2　Lin 的方法在 MC30 数据集中使用改进的知网相似度测量结果对比

词语 1	词语 2	人工判定值	刘群	原知网义原树			多重继承的义项网		
				Seco	Zhou	改进 IC	Seco	Zhou	改进 IC
轿车	汽车	0.98	1	0.477	0.685	0.482	0.566	0.734	0.543
宝石	宝物	0.96	0.146	0.701	0.802	0.649	0.630	0.818	0.663
旅游	游历	0.96	1	0.585	0.759	0.593	0.687	0.814	0.664
男孩子	小伙子	0.94	1	0.494	0.629	0.446	0.528	0.651	0.465
海岸	海滨	0.925	1	0.471	0.660	0.482	0.698	0.777	0.614
庇护所	精神病院	0.9025	0.579	0.407	0.574	0.376	0.340	0.554	0.394
魔术师	巫师	0.875	0.676	0.207	0.497	0.253	0.240	0.500	0.255
中午	正午	0.855	1	0.721	0.695	0.602	0.734	0.705	0.608
火炉	炉灶	0.7775	0.590	0.505	0.709	0.516	0.338	0.587	0.347
食物	水果	0.77	0.126	0.067	0.272	0.091	0.129	0.309	0.130
鸟	公鸡	0.7625	1	0.504	0.689	0.491	0.441	0.710	0.511
鸟	鹤	0.7425	1	0.504	0.689	0.491	0.541	0.710	0.511
工具	器械	0.7375	1	0.447	0.522	0.383	0.464	0.554	0.390
兄弟	和尚	0.705	0.861	0.394	0.629	0.446	0.516	0.643	0.455
起重机	器械	0.42	0.369	0.242	0.515	0.269	0.296	0.539	0.301
小伙子	兄弟	0.415	0.8	0.494	0.629	0.446	0.516	0.643	0.455
旅行	轿车	0.29	0.074	0.038	0.167	0.055	0.099	0.200	0.094
和尚	圣贤	0.275	0.683	0.200	0.497	0.235	0.140	0.521	0.255
墓地	林地	0.2375	0.122	0.155	0.385	0.174	0.176	0.382	0.180
食物	公鸡	0.2225	0.112	0.067	0.268	0.091	0.129	0.305	0.130
海岸	丘陵	0.2175	0.1	0.453	0.660	0.450	0.081	0.048	0.068
森林	墓地	0.21	0.112	0.155	0.385	0.174	0.129	0.305	0.130
岸边	林地	0.1575	0.097	0.453	0.660	0.450	0.081	0.047	0.068
和尚	奴隶	0.1375	0.662	0.200	0.497	0.235	0.240	0.521	0.255
海岸	森林	0.105	0.112	0.067	0.268	0.091	0.081	0.048	0.068
小伙子	巫师	0.105	0.6	0.200	0.497	0.235	0.240	0.513	0.255
琴弦	微笑	0.0325	0.074	0	0	0	0	0	0
玻璃	魔术师	0.0275	0.122	0.067	0.272	0.091	0.129	0.297	0.130
中午	绳子	0.02	0.100	0.023	0.015	0.031	0.081	0.052	0.068
公鸡	航行	0.02	0.074	0	0	0	0	0	0
与人工判定之间的皮尔森相关系数			0.699	0.738	0.701	0.745	0.823	0.804	0.840

表 3　不同方法计算的 MC30 词语相似度与人工值之间的皮尔森相关系数

方法	类型	语义词典	相关度	评测文献
采用 Resnik 的 IC 算法的 Lin 方法	基于信息内容	WordNet	0.70	[20]
采用 Resnik 的 IC 算法的 Jiang and Conrath 方法	基于信息内容	WordNet	0.73	[20]
采用 Seco 的 IC 算法的 Lin 方法	基于信息内容	WordNet	0.81	[21]
采用 Seco 的 IC 算法的 Jiang and Conrath 方法	基于信息内容	WordNet	0.84	[21]
采用 Zhou 的 IC 算法的 Resnik 方法	基于信息内容	WordNet	0.82	[22]
采用 Zhou 的 IC 算法的 Lin 方法	基于信息内容	WordNet	0.82	[22]
采用 Zhou 的 IC 算法的 Jiang and Conrath 方法	基于信息内容	WordNet	0.82	[22]
刘群方法	基于意义的相似性	原始知网	0.699	[6]
李峰方法	基于意义的相似性	原始知网	0.793	[23]
Hao 方法	基于深度与路径	同义词词林	0.825	[24]
Liu 方法[25]	基于深度与路径	同义词词林	0.809	[25]
采用 Seco 的 IC 算法的 Lin 方法	基于信息内容	原始知网	0.738	本文
采用 Zhou 的 IC 算法的 Lin 方法	基于信息内容	原始知网	0.701	本文
本文方法	基于信息内容	原始知网	0.745	本文
采用 Seco 的 IC 算法的 Lin 方法	基于信息内容	改进知网	0.823	本文
采用 Zhou 的 IC 算法的 Lin 方法	基于信息内容	改进知网	0.804	本文
本文方法	基于信息内容	改进知网	0.840	本文

参考文献：

[1] 石静, 吴云芳, 邱立坤. 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.

[2] Princeton University. WordNet [DB/OL]. (2015-3-17) [2016-12-20]. http://wordnet. princeton. edu/.

[3] 董振东, 董强. 《知网》 [DB/OL]. (2000-10-5) [2017-3-1]. http://www. keenage. com.

[4] 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]. (2012-07-02) [2016-10-01]. http://ir. hit. edu. cn/.

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

[6] 刘群, 李素建. 基于《知网》的词汇语义相似度的计算 [C]// 第三届汉语词汇语义学研讨会论文集. 2002.

[7] 张敏, 王振辉, 王艳丽. 一种基于《知网》知识描述语言结构的词语相似度计算方法 [J]. 计算机应用与软件, 2013, 30 (7): 225-267.

[8] 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, 33 (8): 2276-2279.

[9] 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.

[10] 王小林, 王东, 杨思春, 等. 基于《知网》的词语语义相似度算法 [J]. 计算机工程, 2014, 40 (12): 177-181.

[11] 魏韡, 向阳. 基于 2008 版《知网》的词语相似度计算方法 [J]. 计算机工程, 2015, 40 (9): 215-219.

[12] 刘杰, 郭宇, 汤世平, 等. 基于《知网》2008 的词语相似度计算 [J]. 小型微型计算机系统, 2015, 36 (8) .

[13] Resnik P. Using information content to evaluate semantic similarity in a taxonomy [C]// Proc of the 14th International Joint Conference on Artificial Intelligence. 1999: 448-453.

[14] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy [C]// Proc of International Conference Research on Computational Linguistics. 1997: 115-120.

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

[16] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in wordnet [C]// Proc of the 16th European Conference on Artificial Intelligence. 2004: 1089-1090.

[17] Zhou Z, Wang Y, Gu J. A new model of information content for semantic similarity in WordNet [C]// Proc of International Conference on the Future Generation Communication and Networking Symposia. 2008: 85–89.

[18] Bernerslee T, Hendler J, Lassila O. The semanticWeb [J]. Scientific American, 2001, 284 (5): 34-43.

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

[20] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts [C]// Bringing Computational Linguistics and Psycholinguistics Together. 2006: 1–8.

[21] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet [C]// Proc of the 16th European Conference on Artificial Intelligence. 2004: 1089–1090.

[22] Sanchez D, Batet M, Isern D. Ontology-based information content computation [J]. Knowledge-Based Systems, 2011, 24 (2): 297–303.

[23] 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 (3): 99-105.

[24] Hao D, Zuo W L, Peng T, et al. An approach for calculating semantic similarity between words using wordnet [C]// Proc of the 2nd International Conference on Digital Manufacturing and Automation. 2011: 177–180.

[25] Liu X, Zhou Y, Zheng R. Measuring semantic similarity in WordNet [C]// Proc of the 6th International Conference on Machine Learning and Cybernetics. 2007: 3431–3435.