基于遗传算法的词语语义相似度计算研究

杨 泉
(北京师范大学,北京 100875)

摘 要:语义相似度计算就是把词语间语言学上的信息映射为0到1之间的数值。基于知识本体的语义相似度计算方法,利用知识本体提供的信息,建立词语关系和语义相似度之间的函数关系,该方法可解释性强、使用简单,成为语义相似度计算的一类重要方法。提出了一种基于《同义词词林》的语义相似度计算模型,该模型运用遗传算法探索了《同义词词林》语义编码与语义相似度之间的内在联系,建立了更符合《同义词词林》中所蕴含的语义相似信息的函数关系式。该方法使用遗传算法搜索知识与语义相似度的函数表达式,克服了先验模型中函数形式及调节参数的局限性,所得计算结果与人工判定结果的皮尔逊相关系数为0.864 5,为使用人工智能方法挖掘自然语言处理中的规律提供了一种新的思路和方法。

关键词:语义相似度;同义词词林;遗传算法;函数模型;知识本体;路径
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2021)02-0008-06
doi:10.3969/j.issn.1673-629X.2021.02.002

Research on Word Semantic Similarity Calculation Based on Genetic Algorithm

YANG Quan
(Beijing Normal University, Beijing 100875, China)

Abstract: Semantic similarity calculation is to map linguistic information between words to values between 0 and 1. The semantic similarity calculation method based on the ontology uses the information provided by the ontology to establish the functional relationship between the word relationship and the semantic similarity. This method has strong interpretability and simple use, which becomes an important method of semantic similarity calculation. We propose a semantic similarity calculation model based on CiLin. The model uses genetic algorithm to explore the deep relationship between semantic coding and semantic similarity of CiLin, then establishes a function more consistent with the semantic similarity information contained in CiLin. This method uses genetic algorithm to search the function expression of knowledge and semantic similarity, and overcomes the limitation of function form and adjusting parameter in prior model. The Pearson correlation coefficient between the calculated result and the artificial judgment result is 0.864 5. It provides a new way of thinking and method for explore the deep law in natural language processing by using artificial intelligence method.

Key words: semantic similarity; CiLin; genetic algorithm; function model; knowledge ontology; path

0 引 言

语义相似度是对给定的语言对象间语义相似程度的衡量,通常用[0,1]之间的数值来表示。语义相似度计算就是计算语义相似度具体数值的过程。语义相似度计算对象的层级可分为词、短语、句子、篇章,该文主要研究词层级上两个词之间的语义相似度计算问题。

语义相似度计算目前在机器翻译、人机问答、情感计算、信息提取等很多领域中都有着广泛的应用[1]。语义相似度计算方法主要分为两类:一类是在大规模语料的基础上直接统计和计算的方法;另一类是根据某种已有知识本体(ontology)或分类体系(taxonomy)来计算的方法[2-3]。基于语料库的方法对语料的依赖性较大,需要在大规模精确标注语料的基础上进行,但语料的规模、内容、范围以及标注的标准和规范难以统一,而且可解释性较差;而基于知识本体或分类体系的方法在这些方面就显示出了其优越性,越来越多的专家学者都进行了有效的尝试。

用于语义相似度计算的汉语知识本体目前主要有《知网》[4]和《同义词词林》[5]。前人研究中有很多利用《知网》的树状结构或概念义原来进行语义相似度计算,如文献[6]介绍了一种基于《知网》树状结构的

收稿日期:2020-03-26 **修回日期**:2020-07-28
基金项目:国家语委科研项目(YB135-91)
作者简介:杨 泉(1977-),女,副教授,硕导,博士,研究方向为计算语言学。

语义相似度计算方法;文献[7]在综合考虑《知网》义原距离、义原深度、义原宽度、义原密度和义原重合度的基础上,利用多特征结合的方法计算词语相似度;文献[8]基于对《知网》中词语、义项和义原三个层次概念的研究,针对词语相似度计算中结果合理性的问题,提出了一种结合信息论研究中熵的概念的新的词语相似度计算方法。但是与《知网》相比较而言,《同义词词林》内部结构比较清楚,可以较为容易地转化成树形图来计算词语的深度和路径,国内也有很多研究人员利用《同义词词林》计算词语之间的语义相似度,文献[6,9]利用《词林》的编码及结构特点,结合词语的相似性和相关性,计算语义相似度。文献[10]提出了一种综合《知网》与《同义词词林》的计算方法。《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数为微调节参数的方法计算语义相似度。文献[11]根据《词林》提出了一种基于路径与深度的算法。该方法通过两个词语义项之间的最短路径以及它们的最近公共父节点在层次树中的深度计算出两个词语义项之间的相似度。在计算过程中为分类树中不同层之间的边赋予不同的权值,同时通过两个义项在其最近公共父节点中的分支间距动态调节词语义项间的最短路径。文献[12]提出了一种基于路径与《同义词词林》编码相结合的语义相似度计算方法。该方法认为《词林》编码体系是按从左到右依次递增的关系排列分支,距离越近的概念分支间隔越小,编码距离也越近,由此根据每个分类节点下面的分支节点顺序及编码规律设计了计算模型。

以上这些模型都是根据经验建立语义相似度的函数表达式,主要从两个方面提高计算语义相似度的准确性:一是如何使用知识本体中的知识并进行量化;二是如何选择更合适的函数表达式。由于《同义词词林》的内部结构清晰简洁,使用深度、距离和节点分支数作为基础知识进行相似度计算已经成为共识。因此如何突破已有经验的局限性,寻找并建立更加合理的相似度函数表达式是进一步完善基于《同义词词林》的语义相似度计算方法的主要途径。

最近机器学习方法在各个应用领域都取得了突破性进展,其一般学习模型框架可表示为寻找并建立函数表达式的过程。将已知的数据表示为训练集,$S=\{(x_1,y_1),\cdots,(x_m,y_m)\}$,其中 x_i,y_i 分别属于领域集和标签集 y。机器学习的目标是使用学习器从规则集 F 中输出一个规则 $f:x\to y$,该规则一般情况下可以用函数表示,常称为预测器或分类器。该函数是未知的,因此实际应用中用 $\tilde{f}(s)$ 表示学习算法在给定训练集 s 的情况下,获得的学习函数。如果用机器学习的方法在知识本体的基础上计算两个词语之间的相似度,其框架一般可以描述为:在训练集 s 的基础上,从函数集 F 中选择一个适当函数 $\tilde{f}(s)$ 来表示该知识本体中蕴含的词语相似度规则。理论上,大部分机器学习算法都可以用于解决这个问题。近年来,使用基于遗传算法学习复杂系统规律的模型取得了很大进展,其思想借鉴了自然界生物进化理论和遗传原理,是一种自动随机产生搜索程序的方法。该算法简单通用、鲁棒性强,并且对非线性复杂问题显示出很强的求解能力,因而被成功应用到了许多不同的研究领域[13]。在此方法的基础上,提出了用于搜索语义相似度函数的遗传算法。该方法将所有可能计算相似度的函数系统看作一个生物种群,在给定一定数量初始个体的基础上,通过这个群落不断进化,最终选取优胜的个体,此个体所对应的函数在描述语义相似度方面具有优良的结果。该文拟在《同义词词林》的基础上,参考人工判别结果,使用遗传算法建立描述语义相似度的关系模型,以期突破根据先验经验建立函数模型的局限性。

1　《同义词词林》简介

《同义词词林》是梅家驹等人 1983 年编撰的可计算汉语词库,后经哈工大信息检索研究室扩展编辑为《哈工大信息检索研究室同义词词林扩展版》(下文简称《词林》)。经统计《词林》共有 77 456 条词语,分为 12 个大类;95 个中类;1 428 个小类;4 026 个词群和 17 817个原子词群。前面四个层级的节点都代表词语的类别,第五层叶子节点上是原子词群,每个原子词群可用一个 8 位编码唯一表示。表 1 展示了《词林》中的义项编码。

表 1　《词林》义项编码

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

第八位编码只有三种情况:其中“=”代表“相等、同义”关系;“#”代表“不等、同类”关系;“@”代表“唯一、独立”关系。前七位编码确定后就可以唯一确定一条编码,不存在前七位编码相同而第八位不同的情况。

在大类中 A、B、C 类多为名词,D 类多为数词和量词,E 类多为形容词,F、G、H、I、J 类多为动词,K 类多为虚词,L 类是难以被分到上述类别中的一些词语,各大类编码具体含义如表 2 所示。

表 2 《词林》大类编码含义

A	B	C	D	E	F	G	H	I	J	K	L
人	物	时空	抽象	特征	动作	心理	活动	状态	关联	助语	敬语

《词林》结构安排中大类和中类的排序遵照从具体到抽象的原则[5],每个大类都可以转化为一个树形结构图,比如 E 大类下面分为 6 个中类,从“外形”到“境况”,详见图 1。

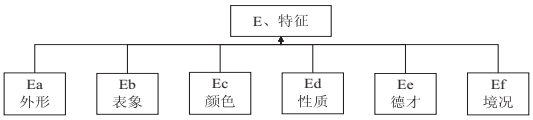

图 1 《词林》E 大类语义场

通过上文对《词林》整体架构的分析,其义项编码可以直接映射为一个树形结构图,所有的词语都可以对应到叶子节点的词群里。实际上这个树形结构图就是使用的知识本体,而每个知识本体反映的都是作者对于世界知识的认识,语义相似性是世界知识很重要的一个组成部分,作者在编著《同义词词林》时就已经融入了语义相似信息,只是没有把这种相似性信息数量化、数值化。因此基于《词林》的两个词语之间的语义相似度计算,实际上就是解析蕴含于知识本体中的语义相似信息,将其形式化后转化为可计算的函数表达式,最终计算出量化的数值。

2 基于遗传算法的语义相似度计算模型

表 1 说明《词林》中共有五个层级,为便于计算,该文在第一层级上面再引入一个虚拟层级,称为第 0 层,对应树形结构图中的根节点,记为 R。在此情况下《词林》共有六层节点、五层边,所有词语都落在树形结构图最底层的叶子节点上,所有叶子节点都是一个原子词群。在该树形结构中将两个节点之间最小的边数称为两个节点之间的路径长度或距离。将各非根节点到根节点 R 的距离称为该节点的深度。

计算语义编码分别对应不同的叶子节点的词语 s_1 与 s_2 的语义相似度 S,根据《词林》编码规则,这两个词语在其最近公共父节点处分离,分属不同类别。将其公共父节点记为 F,将 F 的深度记为 D。从《词林》体系中可以直观地看出,F 在《词林》体系中所处层级越高,则 D 的取值越小,此时 s_1 与 s_2 分离得越早,相似度就低;相反 F 在《词林》中所处层级越低,D 的取值越大,则 s_1 和 s_2 分开得越晚,其相似度就高。因此 D 的取值与 S 成正比关系;而 F 的位置与 S 成反比关系。这从语言学角度也很容易理解,当两个词语所处的分支层的公共父节点越低,说明这两个词语所在的类别距离越近,两个词语的语义相似程度就越高;相反当两个词语所处的分支层的公共父节点越高,说明这两个词语所在的类别距离越远,两个词语的语义相似程度就越低。上述分析表明在《词林》所表示的知识本体中,两个词语 s_1 与 s_2 的最近公共父节点的深度对其相似度起决定性作用。例如“我们”的语义编码为“Aa02B01 =”,“你”的语义编码为“Aa03A01 =”,“消毒剂”的语义编码为“Br13D04#”。“我们”与“你”的语义类别在同一个大类 A 中,而“我们”与“消毒剂”的语义类别分别在 A 和 B 两个大类中,因此前两者的语义相似度一定高于后两者。

在树形结构中还常用两个节点间的路径长度 H 来表示两个节点之间的关系。任意两个叶子节点之间的路径长度 H 就是它们到其最近公共父节点路径长度之和,根据《词林》中树形结构的特点:所有叶子节点到根节点 R 的路径长度相同,在此记为常数 C;叶子节点到其公共父节点的路径长度也相同。而叶子节点到根节点的路径长度等于叶子节点到其任意父节点的路径长度与该父节点到根节点路径长度之和。由此可以得出路径长度与深度之间的关系式:

$$D = C - \frac{1}{2}H \tag{1}$$

该结论说明路径长度和深度是两个能够相互表示的量,该文在计算相似度时选择将深度作为主要因素。文献[2]在总结基于 WordNet 的英语语义相似度计算方法中有一类使用路径和深度的计算方法,但由于 WordNet 与《词林》的组织架构不同,在 WordNet 中不同的词语可能具有不同的深度,这种叶子节点深度不均匀,义项遍布所有节点的组织方式与《词林》是截然不同的。

在《词林》体系中,词语按照类别逐级细分,在同一个类别中的排序遵照从具体到抽象的原则进行排列(如图 1 所示)。这说明在同一个类别层级中,意思接近的两个分类其排列的位置也会接近,对应到树形结构中,就是在同一个节点上排列的分支中,离得越近的分支其代表的意思也越接近。因此词语 s_1 与 s_2 的语义相似度除由其最近公共父节点的深度决定外,也会受到该父节点处两个叶子节点所在分支的位置关系以及最小公共父节点处分支总数的影响。将最近公共父节点所含分支总数记为 N,将 s_1 与 s_2 所在分支的间隔数记为 K。在《词林》框架体系下,对 s_1 与 s_2 两个待计算相似度的词语,根据前面分析和相关文献中的研究结果,整合为如下相似度关键信息 x:

$$x = D + K/N \tag{2}$$

其中,D 为最近公共父节点深度;N 为最近公共父节点处分支总数;K 为词语所在分支间隔数。则 s_1 与 s_2 之间的语义相似度 y 可以表示为关键信息 x 的函数:

$$y = F(x) \tag{3}$$

目前所有基于《词林》的语义相似度计算模型都属于这个框架,只不过不同模型使用了不同的函数。如果把一些计算语义相似度的函数放在一起,然后再制定一个评价这些相似度计算函数的规则来评价,则这些函数就可以看成是一个具有不同竞争优势的种群。借鉴遗传算法的思想,对由相似度函数构成种群进行生物进化方面的选择、交叉和变异等操作来使种群进行不断繁衍,从而得到新的种群即新的相似度计算函数。根据自然选择优胜劣汰的规律,有理由相信能够找到比单纯通过经验建立的更好的相似度计算函数。为实现这个目标,执行以下操作:

(1)函数编码。

首先需要将函数映射为便于使用遗传算法的表示形式。该文将函数用树的形式进行编码,目的是把函数转化为利于计算机操作的形式。这种方法将函数中包含的四则运算、复合运算作为树的中间节点,将自变量 x 作为树的叶子节点。例如对于具有如下形式的相似度计算函数:

$$y = F(x) = w_1x^2 + w_2R + w_3e^x + w_4\sin x \qquad (4)$$

其中, w_1, w_2, R, w_3, w_4 为常数,则可以表示为图 2 所示的树状结构。

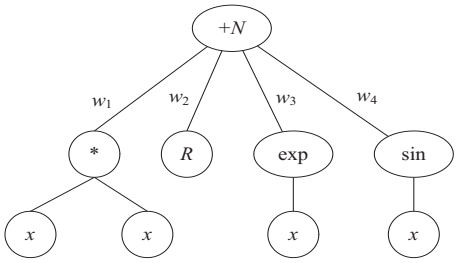

图 2　函数编码的树状结构

根据这种思想,语义相似度计算函数的自变量就是上面的《词林》信息 x ,将基本初等函数作为基本的函数集 $F = \{x, \sin x, \ln x, e^x, \arcsin x\}$,取四则运算为运算集 $H = \{+, -, \times, \div\}$ 。在生成函数种群时,只需从不同集合中选取元素填入相应节点,就可以生成不同的函数,反复操作 $2M$ 次即可生成一个含有 $2M$ 个函数的初始种群。

(2)适应度函数。

遗传算法应用达尔文的自然选择(适者生存)原则,从种群中确定胜出的那些个体,因此根据目标区分群体中个体好坏的函数称为适应度函数,也称为评价函数。该文采用目前应用较为广泛的“最小二乘标准”作为适应度函数。对于种群中的个体,语义相似度计算函数 $y = F(x)$,它计算 m 组词语对的相似度为 $\tilde{y} = [\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_m]$;设 $y = [y_1, y_2, \cdots, y_m]$ 为 m 组词语对已知的标准相似度,则关于个体 F 的适应度函数定义为:

$$R(F) = \frac{1}{m}\sqrt{\sum_{i=1}^{m}(\tilde{y}_i - y_i)^2} = \frac{1}{m}\|\tilde{y} - y\|_2 \qquad (5)$$

显然 $R(F)$ 越小,相似度函数 F 的计算结果与标准结果就越接近,该个体在种群中就越优秀,具有更强的竞争力。

(3)选择。

要完成种群的更新需要从父代群体中选取部分个体,以便生存和繁衍产生下一代群体,这种操作称为选择。该文采取优者胜出的选择方法,将当前种群中的 $2M$ 个函数按照适应度 $R(F)$ 从小到大进行排序,然后将适应度最好的 M 个函数保留,将较差的 M 个函数淘汰,以保留下来的 M 个函数为基础进行下面的操作形成下一代种群。

(4)交叉。

在遗传算法中交叉是利用父代个体形成子代个体的过程,该文研究的个体是函数,在将函数编码后,随机设置交叉点,然后在交叉点处进行断开和重组,完成基因交换,生成新的个体。具体过程如图 3 所示,左边为选择的两个个体,图中方框处为选择作为断点的节点位置,然后分别交换和重组后,得到右侧两个新生成的个体。

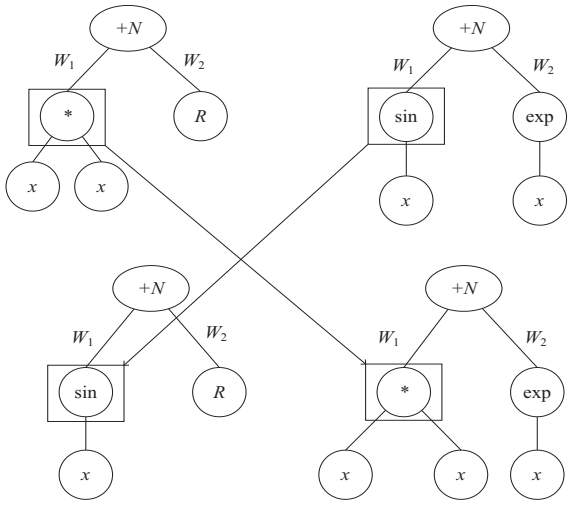

图 3　交叉生成新的个体

(5)变异。

遗传算法中的变异,是指将个体编码串中的某些基因用其他等位基因来替换,从而形成新个体的过程。例如图 4 中,左侧为选中的变异个体,其中方框处为选择的变异位置,右侧为该位置变异后生成的新个体。

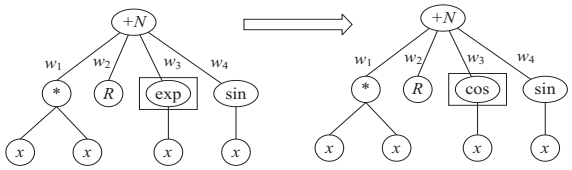

图 4　变异生成新的个体

以上过程描述了一种基于遗传算法的相似度函数

构建模型,该方法使用遗传算法的思想,随机生成一系列函数个体组成初始的“种群”,然后根据适应度函数来评价个体的适应度。若当前种群中的函数所计算的语义相似度都不能满足要求,则模拟生物进化中的基因变异、复制、删除等行为,繁衍生成新一代种群,经过不断迭代,寻找更好的语义相似度计算函数。下面根据遗传算法的思想为《词林》建立语义相似度计算模型,具体算法描述如下:

第 1 步:给定 m 组词语的《词林》信息 $\{x_1,x_2,\cdots,x_m\}$ 和标准相似度结果 $\{y_1,y_2,\cdots,y_m\}$,基本函数集 $F=\{x,\sin x,\ln x,e^x,\arcsin x\}$,运算符号集 $H=\{+,-,\times,\div\}$,最大进化代数 T。

第 2 步:随机生成包含 $2M$ 个计算语义相似度的函数初始种群:$\{F_1,F_2,\cdots,F_{2M}\}$。

第 3 步:当进化代数小于最大进化代数时,生成新的计算语义相似度函数种群,完成种群繁衍迭代。具体方法如下:

①选择:计算种群内全部语义相似度函数个体 $\{F_1,F_2,\cdots,F_{2M}\}$ 的适应度,保留 M 个适应度最好的语义相似度函数个体;

②交叉:随机选择两个语义相似度函数,通过交叉生成新的函数,重复四分之三 M 次,生成复四分之三 M 个新的语义相似度函数;

③变异:随机选取四分之一 M 个语义相似度函数,然后随机选取节点进行变异,生成四分之一 M 个新的语义相似度函数;

第 4 步:回到第 3 步继续进化,直到达到最大进化代数;

第 5 步:计算最终得到的种群中 M 个语义相似度函数的适应度,并将最优个体作为最终相似度计算函数。

该方法中采取了优者胜出的选择方法,每一代中的最优个体会保留到下一代中,随着种群的繁衍,该方法会得到越来越优秀的个体,即越来越好的相似度计算函数。如果达到最大繁衍代数后,得到的相似度计算函数还不够理想,可以适当增加种群大小,即增加迭代次数,甚至反复执行该方法,直到得到满意的相似度计算函数为止。

3 实验及结果分析

目前国际上对语义相似度算法的评价标准普遍采用 Miller & Charles 发布的 30 组英语词对集(简称 MC30)的人工判定值作为比较或学习的标准[14-15]。该文首先根据《词林》提供的关于这 30 组词对的信息计算其相应的词对信息值 x;然后使用遗传算法模型寻找关于 x 的相似度函数表达式;最后,使用新找到的模型重新计算词对相似度并与标准结果和相关结果进行对比。在试验中设定函数构成分量的长度为 3;此时函数关系式可表示为:

$$F(x)=w_1f_1(x)+w_2f_2(x)+w_3f_3(x) \tag{6}$$

初始种群的数量为 50,在遗传算法开始时随机产生 50 个函数 $\{F_i(x),I=1,2,\cdots,50\}$;此后每代种群的最大数量为 100,即有 100 个候选函数;种群的最大进化代数为 1 000 代。若达到最大进化代数,则选取最后一代中最优的函数作为相似度计算模型。经过运行模型算法,最终选定的函数模型为:

$$y=0.194\ 8x+0.006\ 5\frac{\cos x}{x}-0.041\ 1x \tag{7}$$

利用式(7)计算得到的语义相似度结果如表 3 所示。

表 3 语义相似度计算结果

序号	词项 1	《词林》代码	词项 2	《词林》代码	该文结果	MC30 结果
1	轿车	Bo21A04=	汽车	Bo21A26#	0.935 3	0.98
2	宝石	Ba08A06#	宝物	Ba08A01=	0.935 1	0.96
3	旅游	Hj48B01=	游历	Hj48B01=	0.954 7	0.96
4	男孩子	Ab03A07#	小伙子	Ab03A01=	0.935 2	0.94
5	海岸	Be03B03#	海滨	Be03B04#	0.920 7	0.925
6	庇护所	Dm06A04=	精神病院	Dm06A03=	0.913 1	0.902 5
7	魔术师	Ae18A03=	巫师	Ae18B01=	0.775 8	0.875
8	中午	Ca28B01=	正午	Ca28B01=	0.954 7	0.855
9	火炉	Bo19A01=	炉灶	Bo19A02=	0.909 6	0.777 5
10	食物	Br03A01=	水果	Bh07A01=	0.307 8	0.77
11	鸟	Bi11B01=	公鸡	Bi13A02=	0.455	0.762 5
12	鸟	Bi11B01=	鹤	Bi12D01@	0.442 3	0.742 5
13	工具	Bo01B01=	器械	Bo27A01=	0.665 4	0.737 5
14	兄弟	Aa02A07=	和尚	Am01B04=	0.389 1	0.705
15	起重机	Bo01A13=	器械	Bo27A01=	0.665 4	0.42
16	小伙子	Ab03A01=	兄弟	Aa03A03=	0.194 4	0.415

续表 3

序号	词项 1	《词林》代码	词项 2	《词林》代码	该文结果	MC30 结果
17	旅行	Hf04A01 =	轿车	Bo21A04 =	0.092 6	0.29
18	和尚	Am01B04 =	圣贤	Ak03C01 =	0.209 6	0.275
19	墓地	Cb18B01 =	林地	Cb07B05@	0.537 2	0.237 5
20	食物	Br03A01 =	公鸡	Bi13A02 =	0.293 6	0.222 5
21	海岸	Be03B03#	丘陵	Be04A06 =	0.466 1	0.217 5
22	森林	Bh01A03 =	墓地	Cb18B01 =	0.090 9	0.21
23	岸边	Be03B01 =	林地	Bn12A09#	0.293 6	0.157 5
24	和尚	Am01B04 =	奴隶	Af01B01 =	0.293 6	0.137 5
25	海岸	Be03B03#	森林	Bh01A03 =	0.214 8	0.105
26	小伙子	Ab03A01 =	巫师	Ae18B01 =	0.225 4	0.105
27	琴弦	Bp13A39#	微笑	Ic01A03 =	0.105	0.032 5
28	玻璃	Bm15C01 =	魔术师	Ae18A03 =	0.090 9	0.027 5
29	中午	Ca28B01 =	绳子	Bp25A01 =	0.090 9	0.02
30	公鸡	Bi13A02 =	航行	Hf03A03 =	0.092 6	0.02

遗传算法模型对 MC30 语义相似度的具体计算结果如表 3 所示,该文计算结果与皮尔逊相关系数为 r = 0.864 5。在实际应用中一般认为:当 $r \geq 0.8$ 时,两个变量间高度相关;当 $0.5 \leq r < 0.8$ 时,两个变量中度相关。以上结果说明,该文提出的语义相似度计算模型能够表达《词林》中包含的词语相似度关系,与人工值有较强的相关性。从表 3 中的相似度计算值中可以看出,仍然存在该文计算结果与 MC30 的人工判定值有较大差异的词对,比如第 10 个词对“食物(Br03A01 =)”与“水果(Bh07A01 =)”;第 14 个词对“兄弟(Aa02A07 =)”与“和尚(Am01B04 =)”。其差异的深层次主要原因是《词林》中对该词对的相似度判断标准与 MC30 的判断标准在语言学认识上的差异。这种差异既有不同判定者主观因素,也有不同语言之间在翻译时所带来的差异。

4　结束语

该文所提出的语义相似度计算方法是在《词林》体系中词语的深度、路径和分支节点信息基础上进行的,充分利用了人工智能遗传算法强大的搜索能力,所得相似度计算模型更为准确合理。在此研究过程中发现,已有的模型中有一些词语无论使用哪种方法,其计算结果均不理想,这种情况一般既有知识本体中义项定义或者词语分类不合理的原因,也有相似度计算模型不够完善的原因。为了克服前人研究中的不足,在知识方面充分利用《词林》已有的词语信息;在算法方面利用遗传算法从更大更广的函数空间中寻找函数模型,因此所得结论中既能得到较为理想的计算结果,也能更好地反映出语言知识层面的关系。

参考文献:

[1] 冉　婕,孙　瑜. 语义检索中的词语相似度计算研究[J]. 计算机技术与发展,2011,21(4):94-97.

[2] LASTRA-DÍAZ J J, GOIKOETXEA J, TAIEB M A H, et al. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art[J]. Engineering Applications of Artificial Intelligence, 2019, 85: 645-665.

[3] 刘　群,李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(8): 59-76.

[4] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用,1998(3):76-83.

[5] 梅家驹. 同义词词林[M]. 上海:上海辞书出版社,1983.

[6] 魏凯斌,冉延平,余　牛. 语义相似度的计算方法研究与分析[J]. 计算机技术与发展,2010,20(7):102-105.

[7] 张培颖,房龙云. 多特征结合的词语相似度计算模型[J]. 计算机技术与发展,2014,24(12):37-40.

[8] 王小林,陆骆勇,邰伟鹏. 基于信息熵的新的词语相似度算法研究[J]. 计算机技术与发展,2015,25(9):119-122.

[9] 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.

[10] 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.

[11] 陈宏朝,李　飞,朱新华,等. 基于路径与深度的同义词词林词语相似度计算[J]. 中文信息学报,2016,30(5):80-88.

[12] 王松松,高伟勋,徐逸凡. 基于路径与词林编码的词语相似度计算方法[J]. 计算机工程,2018,44(10):160-167.

[13] 郁　磊,史　峰,王　辉,等. Matlab 智能算法的 30 个案例分[M]. 第 2 版. 北京:北京航空航天大学出版社,2015.

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

[15] RUBENSTEIN H, GOODENOUGH J B. Contextual correlates of synonymy[J]. Communications of the ACM, 1965, 8(10): 627-633.