网络出版时间：2017-07-21 13:51:53
网络出版地址：http://kns.cnki.net/kcms/detail/51.1196.TP.20170721.1351.024.html

计算机应用研究 优先出版

原创性 时效性 就是科研成果的生命力
《计算机应用研究》编辑部致力于高效的编排
为的就是将您的成果以最快的速度
呈现于世

* 数字优先出版可将您的文章提前 8~10 个月发布于中国知网和万方数据等在线平台

一种改进的基于《知网》的词语语义相似度算法

作者	张小川，于旭庭，张宜浩
机构	重庆理工大学 计算机科学与工程学院
发表期刊	《计算机应用研究》
预排期卷	2018 年第 35 卷第 8 期
访问地址	http://www.arocmag.com/article/02-2018-08-012.html
发布日期	2017-07-21 11:41:10
引用格式	张小川，于旭庭，张宜浩. 一种改进的基于《知网》的词语语义相似度算法[J/OL]. [2017-07-21]. http://www.arocmag.com/article/02-2018-08-012.html.
摘要	词语语义相似度计算在信息检索、文本聚类、语义消歧等方面有着广泛的应用。针对《知网》中现有词语语义相似度计算方法未考虑义原距离与义原深度的主次关系进行了研究，通过约束义原深度因素来改进了义原相似度算法；另外，提出了以词语间第一基本义原相似度最高的概念组合为计算对象，并引入动态加权因子实现了对词语语义相似度算法的改进。对改进前后的算法分别进行了实验，结果表明改进后的算法提高了词语语义相似度的准确性和客观性。
关键词	词语语义相似度，义原距离，第一基本义原，加权因子
中图分类号	TP391.1
基金项目	国家自然科学基金资助项目（61502064）；重庆市“121”科技支撑示范工程项目（cstc2014fazktjcsf40009）

一种改进的基于《知网》的词语语义相似度算法 *

张小川，于旭庭*，张宜浩

(重庆理工大学 计算机科学与工程学院，重庆 400054)

摘　要：词语语义相似度计算在信息检索、文本聚类、语义消歧等方面有着广泛的应用。针对《知网》中现有词语语义相似度计算方法未考虑义原距离与义原深度的主次关系进行了研究，通过约束义原深度因素来改进了义原相似度算法；另外，提出了以词语间第一基本义原相似度最高的概念组合为计算对象，并引入动态加权因子实现了对词语语义相似度算法的改进。对改进前后的算法分别进行了实验，结果表明改进后的算法提高了词语语义相似度的准确性和客观性。

关键词：词语语义相似度；义原距离；第一基本义原；加权因子

中图分类号：TP391.1

Improved word semantic similarity algorithm based on HowNet

Zhang Xiaochuan, Yu Xuting*, Zhang Yihao

(College of Computer Science & Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: The word semantic similarity calculation has a wide range of applications in information retrieval, text clustering and semantic disambiguation, etc. The existing word semantic similarity method based on hownet doesn't consider the different importance level of distance and depth. Proposed a method of restricting the depth to improve the sememe similarity algorithm. In addition, Proposed a word semantic similarity algorithm, this method filtered the combination terms of the highest first basic sememe similarity value, and absorbed the dynamic weighting factor. And the experiment shows the modified algorithm improves the computational accuracy and objectivity of the similarity calculation.

Key Words: word semantic similarity;sememe distance;the first basic sememe; weighting factor

0　引言

互联网时代，网上的数据信息量每天都呈几何倍数地增长，其中，文本信息占有很大一部分比例。如何高效处理这些文本信息是信息检索、文本挖掘等自然语言处理领域的研究热点。

词语相似度计算是这一领域的基础，在信息检索、文本聚类、语义消歧等方面有着广泛的应用。目前，词语相似度计算主要有两种方法[1]：

a）基于统计词频的方法：该方法结合上下文背景，将文本看成独立词语的集合，最终转换成空间向量的形式进行词语间的相似度计算。例如，Lee[2]利用相关熵，Brown[3]采用平均互信息来计算词语之间的相似度。该方法只统计词语词频信息，对词语语义不进行深层分析，且依赖大规模语料库，计算过程比较复杂，计算结果也容易受训练数据的噪声影响。

b）基于语义词典的方法：该方法借助现有语义词典，把词语的相关概念组织在树型结构中分析词语的语义信息，对词语语义进行深层次的分析。该方法中，大规模的语义词典是词语相似度计算的基础：英文方面，荀恩东等[4]采用 *WordNet* 进行词语相似度计算；中文方面，王斌[5]提出基于《同义词词林》进行词语相似度计算，刘群等[6]提出基于《知网》进行词语相似度计算等。基于语义词典的方法简单有效，比较直观，但对词典依赖性较大，且易受主观意识影响，当前词汇语义相似度计算大多采用该方法。

《知网》是目前国内词语语义计算的主流工具。《知网》中，词语是由概念的集合语义描述，而概念又是由义原定义[7]。因此，基于《知网》计算词语语义相似度，可以先转换为计算概念组合间的相似度，然后计算义原相似度[8]。比较有代表性的是刘群等[6]提出的仅考虑义原之间距离因素的词汇语义相似度计算方法，李峰等[9]在此基础上提出考虑义原深度因素的计算方法，Lin Dekang[10]提出的基于信息论的计算方法，王小林[11]等提出的概念间变系数计算方法等。

针对现有词语语义相似度计算方法中未考虑义原距离与义原深度的主次关系的不足，提出一种改进的基于《知网》词语语义相似度计算方法。介绍了现有义原相似度、概念相似度以及词语相似度的计算方法。在此基础上，首先通过约束义原深度因素，然后通过筛选出符合条件的概念组合，并引入动态加

基金项目：国家自然科学基金资助项目（61502064）；重庆市“121”科技支撑示范工程项目（cstc2014fazktjcsf40009）

作者简介：张小川（1965-），男，重庆人，教授，主要研究方向为人工智能、计算机软件；于旭庭（1990-），女（通信作者），硕士研究生，主要研究方向为人工智能（1527593026@qq.com）；张宜浩（1982-），男，博士，主要研究方向为自然语言处理.

权因子实现对词语语义相似度算法的改进。在实验部分，验证了本文提出的方法有一定的改进效果。

1 基于《知网》的词语语义相似度计算方法

由《知网》结构可知，基于《知网》计算词语语义相似度，可以通过先计算义原相似度，再转换为计算概念组合间的相似度，最终得到对应词语的语义相似度[12]。本章列举了三者相似度的现有计算方法，并重点分析了现有的义原相似度和词语相似度计算方法。

1.1 义原相似度计算

在《知网》中，词语是由概念描述，而构成概念的基本单位是义原，故义原相似度计算是整个词语相似度计算的基础。因为义原处于一个树状的层次结构中，所以通过义原树中各个节点间的相互关系，实现义原相似度的计算[13]。影响最终义原相似度结果的因素主要有两个：a)义原距离,两个义原间最短路径上边的数目，是描述义原间相对关系的量，距离越大，义原相似性越低；b)义原深度,两个义原节点层次数的最小值，是描述义原在层次结构中绝对关系的量，深度越小，义原相似性越低[14]。现有义原相似度的计算方法主要是围绕着上述两个因素展开研究。

其中，刘群[6]提出利用义原的上下距离关系，即义原距离来计算义原相似度，如式(1)所示；文献[9]引入最小义原深度因素，如式(2)所示；Lin Dekang 认为节点包含信息量与层次深度成正比[10]，提出两个事物的相似度取决于它们的共性与个性，如式(3)所示。

下面列举几种常见的义原相似度计算方法。

$$Sim\left(S_1,S_2\right)=\frac{a}{a+dist\left(S_1,S_2\right)} \tag{1}$$

其中：S_1 和 S_2 表示 2 个义原；$dist\left(S_1,S_2\right)$ 是 S_1 和 S_2 在义原树中的距离；a 是调节参数，取值 1.6。

$$Sim\left(S_1,S_2\right)=\frac{a*\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)}{a*\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)+dist\left(S_1,S_2\right)} \tag{2}$$

其中：S_1 和 S_2 表示 2 个义原；$\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)$ 是 S_1 和 S_2 在义原树中较小的层次深度值；$dist\left(S_1,S_2\right)$ 是 S_1 和 S_2 在义原树中的距离；a 是调节参数，取值 0.5。

$$Sim\left(S_1,S_2\right)=\frac{\inf o\left(S_1\right)\cap\inf o\left(S_2\right)}{\inf o\left(S_1\right)\cup\inf o\left(S_2\right)} \tag{3}$$

其中：S_1 和 S_2 表示 2 个义原；$\inf o\left(S_1\right)$ 是 S_1 在义原树中拥有的信息量，$\inf o(S_2)$ 是 S_2 在义原树中拥有的信息量。

分析上述三个公式可知，式(1)仅仅以距离为衡量标准，计算结果比较粗糙；式(2)在此基础上，引入义原深度因素，提高计算结果的准确性，但是仅仅考虑了深度较小的义原，文献[15]指出若义原的最小深度因素过大，甚至超过距离因素，会影响义原相似度的计算结果，并给出实例分析，因此，必须保证距离因素对最终相似度计算结果的主导影响；式(3)则从信息论的角度，综合考虑了两个义原的深度因素，计算结果较合理。

为了保证义原距离对义原相似度计算结果的主导影响，本文通过约束义原深度因素的方法来实现该目的。其中，式(2)(3)以及文献[15]对本文研究工作有启发作用，本文将基于此改进义原相似度计算方法。

1.2 概念相似度计算

《知网》中对概念的描述分为第一基本义原、其他基本义原、关系义原和关系符号描述义原 4 类。在概念相似度计算方法中，刘群[6]将 4 个子类义原集合间的相似度合成概念相似度，如式(4)所示。

$$Sim\left(C_1,C_2\right)=\sum_{i=1}^{4}\beta_i Sim_i\left(C_1,C_2\right) \tag{4}$$

其中：C_1 和 C_2 表示两个概念总集合，$Sim_i\left(C_1,C_2\right)$ 表示第 i 个类型的义原集合间的相似度，β_i 为赋予给部分集合相似度的固定权重值，$\beta_1+\beta_2+\beta_3+\beta_4=1$ 且 $\beta_1\geq\beta_2\geq\beta_3\geq\beta_4$。

特别声明一点，本文对这部分不做改进。

1.3 词语相似度计算

对于两个汉语词语 W_1 和 W_2，如果 W_1 有 m 个概念：C_{11}、C_{12}、…、C_{1m}；W_2 有 n 个概念：C_{21}、C_{22}、…、C_{2n}，则词语 W_1 和 W_2 的相似度为各个概念之间的相似度的最大值，其计算公式如式(5)所示[6]。

$$Sim\left(W_1,W_2\right)=\max_{i=1,2,\ldots,m;\ j=1,2,\ldots,n}Sim\left(C_{1i},C_{2j}\right) \tag{5}$$

其中，文献[11]认为不同词性概念间的相似度对于词语相似度影响很小，在此基础上，提出一种新的计算方法：以词性为标准，组合 2 个词语中的概念，并取该词性相同的概念组合中最大相似度值为最终词语的相似度值，否则按照文献[6]中的方法计算。

文献[16]指出选取词性相同的概念组合等价于选取概念中第一基本义原相似度最大的组合。基于此，提出利用第一基本义原相似度最大的概念组合进行词语相似度计算，并用筛选后的组合算数平均值取代最大值，提高了词语相似度的准确性。

文献[11，16]提出的方法，对本文的研究思路有引导作用，2.2 节将基于这个思路进行算法改进。

2 改进的相似度计算方法

2.1 改进的义原相似度计算

经过 1.1 节分析可知，当义原最小层次深度较大时，义原

深度因素对最终相似度的计算结果影响过大，出现和客观事实不符的矛盾现象[17]。本文在综合考虑距离和深度两个因素的情况下，通过约束深度因素，保证距离因素对相似度计算结果的主导影响，并综合考虑了两个义原的层次深度因素。在文献[15]和公式(2)的基础上，提出令 $\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)$ 部分除以 $\lambda dist\left(S_1,S_2\right)$，其中 λ 取值为

$$\frac{\max\left(dep\left(S_1\right),dep\left(S_2\right)\right)}{\max\left(dep\left(S_1\right),dep\left(S_2\right)\right)+\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)}$$

综合考虑两个义原的深度，化简得到式(6)：

$$Sim\left(S_1,S_2\right)=\frac{a*\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)+\varepsilon}{a*\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)+\lambda dist^2\left(S_1,S_2\right)+\varepsilon} \tag{6}$$

其中：$Sim\left(S_1,S_2\right)$ 代表义原 S_1 和 S_2 的相似度，$dist\left(S_1,S_2\right)$ 是义原间距离，$\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)$ 是义原最小深度，$\max\left(dep\left(S_1\right),dep\left(S_2\right)\right)$ 是义原最大深度，$0.5\le\lambda\le1$，ε 为调节参数，取值为 2。

式(6)既考虑了义原层次深度又适当减小其对最终结果的影响力。与文献[15]相比，综合考虑了两个义原的深度因素的影响，获得更合理的结果。为更加直观地比较公式(1)、公式(2)和公式(6)，图 1、图 2 和图 3、图 4 分别展示了 3 个公式的函数图像。其中，$0\le dist\left(S_1,S_2\right)\le20$，本文认为不在同一颗义原树时，距离为 20。$0\le\min\left(dep\left(S_1\right),dep\left(S_2\right)\right)\le10$，$0\le Sim\left(S_1,S_2\right)\le1$[9]。

由于公式(6)含有 3 个自变量，为了达到图示化的目的，分别取 λ 为 0.5 和 1，得到图 3、4。

观察图 1，发现义原相似度仅在距离轴上有变化，这是因为公式(1) 仅考虑距离因素。比较图 2 和 3，发现当义原距离较小时，义原相似度值图像近似重合，当距离渐渐增大，相似度值才开始出现分歧。特别说明一点，图 3、4 分别是 λ 取 0.5 和 1 时式(6)对应的图像，可以看出其中义原距离对深度的拘束力是有变化的，取决于两个义原的深度。

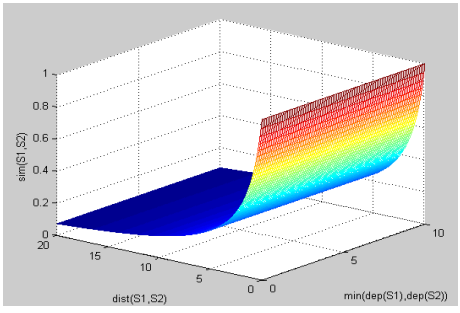

图 1 式(1)的函数图像

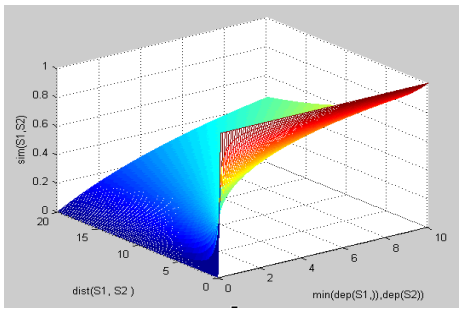

图 2 式(2)的函数图像

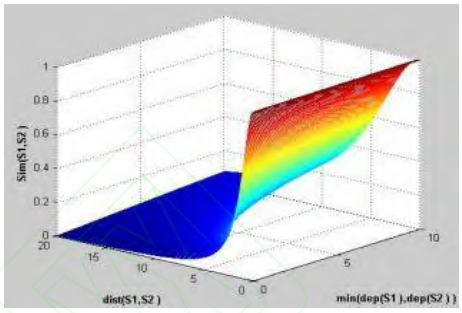

图 3 式(6-λ_1)的函数图像

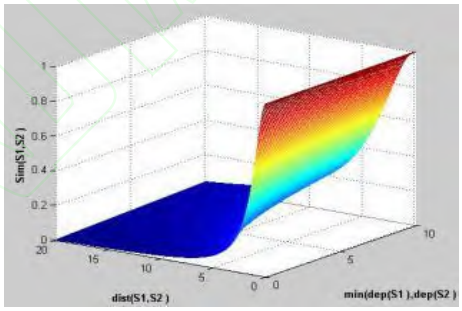

图 4 式(6-λ_2)的函数图像

特别说明一点，本节提出的义原相似度改进公式是针对除去第一基本义原之外的其他义原相似度计算方法而言，这是因为《知网》中的第一基本义原都是由单个义原描述的，并且距离根节点很近，故深度因素对第一基本义原影响作用很小[18]。

2.2 改进的词语相似度计算

经过 1.3 节分析可知，若 2 个概念第一基本义原相似度较低，则这 2 个概念深层次的其他距离就会很大，所以它们之间的相似度对最终词语相似度的影响微乎其微[16,18]。本文在文献[11，16]改进词语相似度计算方法的基础上，以第一基本义原相似度最大的概念组合为对象，引入动态加权因子进行词语相似度算法的改进。

本文在强调词语对应的第一基本义原重要性的基础上，也考虑了除第一义原之外的其他义原因素，这是通过概念组合中包含的义原个数反映出来。首先，计算词语对应的第一基本义原相似度，并通过第一基本义原最大值筛选其对应的概念组合，其次，计算筛选后概念对的相似度，并引入动态加权因子，用得到各组合相似度的加权平均值取代最大值，以此提高词语相似度计算的效率和客观性。

具体改进算法步骤如下：

a)针对两个词语 W_1， W_2，其对应的概念分别为 C_{11}、C_{12}、…、C_{1m} 和 C_{21}、C_{22}、…、C_{2n}，两两组合概念中的第一基本义原，并利用公式(1)计算第一基本义原之间的相似度。

b)以第一基本义原相似度最大的概念组合为标准，两两组合筛选后的概念对，作为参与计算最终词语相似度的对象，并按公式(7)计算 W_1， W_2 的最终相似度值。

$$Sim\left(W_1,W_2\right)=\frac{n_i Sim(C_{1i},C_{2i})}{\sum_{i=1}^{m} n_i} \tag{7}$$

其中: W_1 和 W_2 表示 2 个词语； C_{1i} 和 C_{2i} 为筛选后的第 i 个概念对， m 为筛选后的概念对总个数， n_i 为第 i 个概念对包含的义原总个数。

改进后的方法考虑了文献[16]提出的第一基本义原对最终的词语相似度计算结果的重要性，以第一基本义原相似度最大的概念组合作为参与最终词语相似度的计算对象，省去了原有方法中需要计算两个词语的所有概念组合的工作量，减少了这个过程中的计算次数，也在一定程度上提高了运算速度。另外，引入动态加权因子来反映参与最终计算的概念组合的重要程度，使得最终的计算结果更加合理。

为了更清楚直观地说明问题，以表 1 中的两个词语“教育”和“专业”为实例进行改进后算法的说明。其中，第 3 列、第 4 列和第 5 列分别是利用文献[6,11]和改进算法提出的词语相似度算法计算得到的结果，分别记为 sim1、sim2 和 sim3，如表 1 所示。

表 1　词语相似度比较表

教育	专业	sim1	sim2	sim3
affairs|事务，education|教育	aValue|属性值，attachment|归属，#occupation|职位，formal|正式	0.039		
affairs|事务，education|教育	affairs|事务，education|教育	**1.000**	**1.000**	**1.000**
affairs|事务，education|教育	affairs|事务，industrial|工	0.722	0.722	**0.783**
teach|教	aValue|属性值，attachment|归属，#occupation|职位，formal|正式	0.043		
teach|教	affairs|事务，education|教育	0.044		
teach|教	affairs|事务，industrial|工	0.044		

注：表中黑体数字为最终有效值。

分析表 1 中的第 3 列，可以看出“教育”和“专业”2 个词语间，若第一基本义原间相似度很小，则对整个词语相似度的贡献也很小。文献[6]需要 6 次组合计算才能完成，其中相似度较小的第一基本义原对应的概念进行了 4 次运算，但是这 4 次运算对最终的词语相似度并未起到作用。而文献[11]根据词性组合概念对，筛选后的组合仅需要 2 次运算。事实上，文献[16]也是需要 2 次概念组合间的运算（见表 1 中的 Sim2 筛选后的概念组合），与文献[11]不同的是，最终的相似度结果是取概念组合结果的算术平均值，而非其最大值。从表 1 中可以看出，改进算法也是通过 2 次运算完成，在提高运算速度的同时，并未降低概念间的计算精度，且在强调第一基本义原重要性的基础上，考虑了其他义原因素，使得计算结果更加合理。

另外，按照文献[6]和文献[11]的词语相似度计算方法，“教育”和“专业”的相似度值均为 1，这显然与人们的客观认识矛盾。而根据改进后的算法得到相似度值为 0.892，这是因为使用筛选后的概念组合间相似度的加权平均值取代了最大值，其中，参与最终计算概念组合的重要程度由它们包含的义原总个数决定，并通过加权因子量化表示，见公式(7)，使得结果更加符合实际。

3　实验结果与分析

3.1　义原相似度实验

为了验证式(6)的改进效果，本文与其他 3 种义原相似度计算方法进行比较。

其中，方法 1、方法 2 分别为式(1)(2)得到的义原相似度结果，方法 3 为式(6)中 λ 取定值 0.5[9]得到的义原相似度结果。

利用文献[9]的数据为实验对象，并将参数值 α 和 ε 分别设置为 1.6 和 2.0。实验结果如表 2 所示。

表 2　义原相似度比较表

行号	义原 1	义原 2	距离	义原深度	方法 1	方法 2	方法 3	方法 4
1	实体	实体	0	(0,0)	1.000	0.000	1.000	1.000
2	动物	植物	2	(4,4)	0.444	0.762	0.787	0.769
3	无生物	生物	2	(3,3)	0.444	0.706	0.744	0.720
4	车	电脑	3	(6,7)	0.348	0.762	0.702	0.668
5	物质	食物	3	(2,5)	0.348	0.516	0.483	0.352
6	植物	禽	4	(4,6)	0.286	0.615	0.481	0.407
7	生物	食物	4	(3,5)	0.286	0.545	0.420	0.336
8	车	雨雪	7	(6,7)	0.186	0.615	0.332	0.270
9	车	物质	7	(2,7)	0.186	0.314	0. 146	0.150
10	车	分钟	20	(1,7)	0.074	0.359	0. 057	0.011
11	车	高度	20	(2,7)	0.074	0.138	0.021	0.005

下面对义原相似度计算结果进行分析：

a)表 2 中义原距离，随着行号的增加整体上呈变大的趋势，方法 1、方法 3、方法 4 的义原相似度结果整体上是递减的，而方法 2 的义原相似度计算结果则出现多次波动。

b)比较第 2 行与第 3 行、第 4 行与第 5 行、第 6 行与第 7 行、第 8 行与第 9 行、第 10 行与第 11 行，这 5 组义原对通过方法 1 得到的相似度在组内并无区分度，这是因为方法 1 只考虑了义原距离因素；而其他方法得到的相似度则有差异，这是因为其他方法引入了义原深度因素的结果。

c)比较第 3 行与第 4 行、第 7 行与第 8 行、第 9 行与第 10 行，方法 2 得到的相似度计算结果与义原距离成正比，即距离大的义原对，相似度结果反而越大，而另外 3 种方法则相反。比如，方法 2 中计算得到的第 10 行中“车”与“分钟”的相似度比第 9 行中“车”与“物质”的相似度还要高，然而事实上，前者对应的义原距离为 20(即 2 个义原不在一棵义原树上)，而后者义原距离为 7(即 2 个义原位于一棵义原树)，显然，方法 2 计算结果是不合理的。究其原因，是因为方法 2 放大了义原层次深度，导致最终的计算结果受到层次深度因素过多地影响。而方法 4 在此基础上，拘束了义原层次深度，故得到的相似度结果较之合理。

d)方法 3 和方法 4 均对方法 2 进行了改进，唯一不同的是前者对义原层次深度进行静态制约程度，即 λ 取固定值。而后者 λ 为动态取值，相比方法 3，它更加综合地考虑了两个义原的层次深度信息，因此，方法 3 和方法 4 的义原相似度结果有差异。方法 4 与方法 1、方法 2 相比，义原相似度结果显示出向两端扩散的现象，使计算的结果更加合理细腻。

为了更加直观地分析表 2，图示化表中数据，得到图 5。

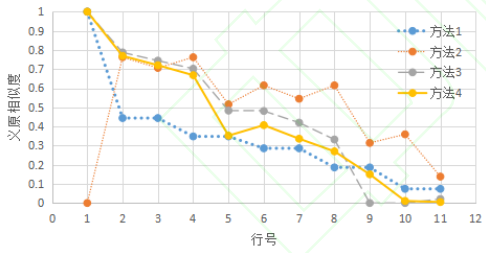

图 5 四种方法义原相似度结果比较图

表 2 中随着序号的增大，义原距离也增大。图 5 可以看出，方法 1 的义原相似度随着距离因素的增大而变小，方法 4 和方法 2 在义原距离较小时，相似度值接近相等。而方法 2 的相似度值在义原距离增大的过程中，出现多次起伏波动，这是因为层次深度因素比义原距离因素过大地影响最终的相似度结果。

3.2 词语相似度实验

为了验证本文提出的词语相似度计算方法的改进效果，本文与其他 2 种词语相似度计算方法进行比较。利用文献[6]的常用的词语和新添加的词语为实验对象。

其中，方法 1、方法 2 和方法 3 分别是利用文献[6,9]和本文中提出的义原相似度和词语相似度计算方法得到的结果，实验结果如表 3 所示。

表 3 词语相似度对比表

词语 1	词语 2	方法 1	方法 2	方法 3
男人	苹果	0.171	0.313	0.303
男人	经理	0.630	0.530	0.579
男人	工作	0.164	0.148	0.120
男人	鲤鱼	0.208	0.357	0.209
跑	跳	0.444	0.762	0.866
发明	创造	0.615	0.849	0.955
学校	实验室	0.575	0.640	0.591
学校	图书馆	0.575	0.618	0.577
本科	必修课	1.000	0.785	0.501
考	考核	1.000	1.000	1.000

下面对词语相似度计算结果进行分析：

a)从整体上看，方法 2 与方法 3 的结果明显比方法 1 更符合实际情况，这是因为方法 1 仅仅考虑了词语中义原距离因素，结果比较粗糙；而方法 2 和方法 3 则在此基础上，不同程度地考虑了义原深度因素。

b)方法 3 与方法 2 比较，大部分数据有所降低，从词语相似度计算方法的角度分析，是因为方法 3 用概念集合的加权平均值代替了最大值，使得最终结果更加客观。

c)“跑”与“跳”、“发明”与“创造”两组词语在方法 3 得到的结果高于方法 2，结果更加符合实际。这是因为这两组词语均不含相同的第一基本义原，实质上，方法 3 提出的加权平均值就等同于最大值，主要是改进的义原相似度算法导致出现这一结果。

d)“学校”与“实验室”、“学校”与“图书馆”两组词语在方法 1 中的计算结果相同，而方法 2 与方法 3 对这两组词语的相似度有不同程度区分，均为前者的相似度较大。从义原相似度角度分析原因，是因为前者词语中的义原“学”与“研究”的最小层次深度要大于“教”与“借入”。

e)“本科”与“必修课”词语对在方法 1 中的相似度为 1，这显然不符合客观实际。从词语相似度角度分析，是因为按照方法 1 的方法得到词语中概念相似度的最大值为 1，故词语的相似度也为 1。而根据本文提出概念组合的加权平均值取代最大值的方法，得到词语的相似度值分别为 0.501，显然后者比较符合客观实际。

f)“考”与“考核”词语对在三种方法中的相似度值均为 1，这是由于这两个词语的概念完全相同，从《知网》的角度分析，实质上是同一个词语间的相似度计算，因此，三种方法的结果均为 1。

因此，通过第一基本义原筛选出的词语概念组合，计算结果既不影响计算的精度，而且大大提高了运算速度，尤其对于那些在《知网》层次结构中解释概念较多和同词性概念较多的词语，这样的速度优势更明显。

4 结束语

为了解决现有词语语义相似度计算方法未考虑义原距离与义原深度的主次关系，通过距离约束最小层次深度因素，并且综合考虑两个义原的层次深度，改进义原相似度计方法；另外，通过筛选出词语对应的第一基本义原相似度最高的概念组合，再引入动态加权因子，用组合间概念相似度值的加权平均值取代现有方法的最大值，以此提高词语相似度的准确性和客观性。尽管改进后的方法有较好的效果，但由于汉语词汇表达的复杂性、词汇语义概念较强的主观性、具体应用领域的专业性等因素影响，词汇相似度计算仍有很大的研究空间，这也是后续的研究方向。

参考文献:

[1] 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.

[2] Lee L. Similarity based approaches to natural language processing[D]. Cambridge: Harvard University, 1997.

[3] Brown P. Word sense disambiguation using tactical methods[C]//Proc of the 29th Meeting of Association for Computational Linguistics, 1991.

[4] Floreano D, Monidada F. Evolutionary neuro-controller for autonomous mobile robots[J]. Neural Networks, 1998, 11(7/8): 1461-1478.

[5] 王斌. 汉英双语语料库自动对齐研究[D]. 北京: 中国科学院计算技术研究所, 1999.

[6] 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学研讨会论文集. 2002: 59-76.

[7] 金玉, 范学峰. 基于《知网》的中文 Deep Web 模式匹配算法研究[J]. 计算机应用研究, 2009, 26(10): 3750-3753.

[8] 程传鹏, 吴志刚. 一种基于知网的句子相似度计算方法[J]. 计算机工程与科学, 2012, 34(2): 172-175.

[9] 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.

[10] Lin Dekang. An information-theoretic definition of similarity semantic distance in WordNet[C]//Proc of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, Inc. , 1998: 296-30.

[11] 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.

[12] 程传鹏, 吴志刚. 一种基于知网的句子相似度计算方法[J]. 计算机工程与科学, 2012, 34(2): 172-175.

[13] 李湘东, 曹环, 丁丛, 等. 利用《知网》和领域关键词集扩展方法的短文本分类研究[J]. 现代图书情报技术, 2015, 2(255): 31-37.

[14] 廖志芳, 周国恩, 李俊锋, 等. 中文短文本语法语义相似度算法[J]. 湖南大学学报: 自然科学版, 2016, 43(2): 135-140.

[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] 王小林, 王东. 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.

[17] 王义, 王小林. 基于改进的义原关联度算法的词语相关度计算[J]. 情报学报, 2012, 31(12): 1271-1275.

[18] 张亮, 尹存燕. 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.