·人工智能及识别技术· 文章编号:1000-3428(2015)09-0215-05 文献标识码:A 中图分类号:TP391

基于 2008 版《知网》的词语相似度计算方法

魏 韡[1,2],向 阳[2]

(1. 井冈山大学电子与信息工程学院流域生态与地理环境监测国家测绘地理信息局重点实验室,江西 吉安 343009;
2. 同济大学电子与信息工程学院,上海 201804)

摘 要: 词语相似度的计算是自然语言处理领域的重要问题,在机器翻译、信息检索、文本分类等领域有广泛的应用。分析和利用新版语义词典 2008 版《知网》,从概念的主类义原和概念的特征描述 2 个方面综合计算词语相似度。运用义原树的树形层次结构,得到义原的深度信息量,再考虑义原的路径计算得到义原相似度。通过层次特征类型匹配计算概念特征描述的相似度。综合主类义原相似度、概念特征描述相似度以及义原之间的对义、反义关系计算得到词语相似度。实验结果表明,该方法得到的词语相似度计算结果与人的主观认识趋于一致。

关键词: 词语相似度;2008 版《知网》;义原;深度信息量;路径;特征描述

中文引用格式: 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,2015,41(9):215-219.
英文引用格式: Wei Wei, Xiang Yang. Method of Word Similarity Computation Based on HowNet 2008[J]. Computer Engineering, 2015, 41(9): 215-219.

Method of Word Similarity Computation Based on HowNet 2008

WEI Wei[1,2], XIANG Yang[2]

(1. Key Laboratory of Watershed Ecology and Geographical Environment Monitoring, College of Electronics and Information Engineering, Jinggangshan University, Ji'an 343009, China;
2. College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

【Abstract】 Word similarity computing is a key issue in natural language processing, which is widely used in machine translation, information retrieval and text classification. Based on lexical taxonomy new HowNet (2008), this paper proposes a new method to analyze and compute Chinese word similarity from two dimensions: the main sememe of the concept and the concept characteristic description of the concept. In this paper, the depth information is obtained by using the sememe tree structure, then the sememe similarity is computed by taking into account the hierarchical path of the sememe. Computing the similarity between two concept characteristic descriptions is based on characteristic type mapping. Word similarity is computed based on the sememe similarity, the concept characteristic descriptions similarity and the antonym information of sememe. Experimental results show that the calculating results of word similarity by this method are more in line with subjective cognition of the people.

【Key words】 word similarity; HowNet 2008; sememe; depth information quantity; path; characteristic description
DOI: 10.3969/j.issn.1000-3428.2015.09.040

1 概述

在自然语言处理领域,词语相似度计算被广泛地应用于信息检索、机器翻译、自动问答、词义消歧等方面,是一个具有基础研究性质的课题。例如:在信息检索中,词语相似度可以帮助匹配用户查询和符合条件的文本,提高检索的准确率和召回率;在基于实例的机器翻译中,词语相似度可以衡量 2 个不同词语在文本中的可替换程度;在自动问答系统中,词语相似度可以用来表示用户问题和答案之间的符合程度;在词义消歧中,词语相似度可以用来判断歧义词的词义。文献[1]认为 2 个词语的相似度是它们在不同的上下文中可以互相替换且不改变文本的句法语义结构的程度。简而言之,如果 2 个词语可替换的程度越高,它们的相似度就越大。词语的相似度和其语义的联系最密切,所以词语的相似度一般也指词语的语义相似度。词语的相似度一般用[0,1]区间的一个实数来表示。

基金项目: 国家自然科学基金资助项目(61363014,71171148);江西省自然科学基金资助项目(20151BAB207016)。
作者简介: 魏 韡(1983-),男,讲师、博士研究生,主研方向:自然语言处理,人工智能;向 阳,教授、博士生导师。
收稿日期: 2014-08-04 **修回日期:** 2014-10-13 **E-mail:** weiweihzkd@163.com

目前,词语相似度的计算方法大体上可分为2类,即基于大规模语料库统计的方法和基于本体或词典的方法。基于语料库统计的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,同时也容易受到数据稀疏和数据噪声的干扰。基于本体或词典的方法比较直观,易于计算,但需要有完备的本体或词典。

虽然2类方法各有千秋,但是由领域专家构建的本体或词典更具有权威性和完备性,因此,基于本体或词典的方法计算词语相似度得到的结果也更合理些。由于目前大多数基于《知网》的词语相似度计算使用的是旧版的《知网》,而2008版《知网》与旧版有较大的改动,比旧版更丰富和更完备。本文基于2008版《知网》提出新的词语相似度计算方法,将词语相似度分为词语概念的主类义原相似度和词语概念的概念特征描述相似度两部分。通过基于义原树的义原深度信息量及路径的混合方法计算主类义原相似度,采用层次特征类型匹配来计算特征描述相似度。

2 2008版《知网》简介

《知网》(HowNet)[2]是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。知网中的概念指的是词语的语义,一个词语如果有多个语义,也就有多个概念。概念是用义原进行描述的,义原是不可分割的最小意义单位,唯一且确定。义原之间的关系主要有:上下位关系,同义关系,反义关系,对义关系,属性-宿主关系,部件-整体关系,材料-成品关系,事件-角色关系。其中,主要的关系是上下位关系、反义关系、对义关系。2008版《知网》大概有2 000多个义原,大致可分为事件(Event)、实体(Entity)、属性(Attribute)、属性值(Attribute Value)、次要特征(Secondary Feature)等几个特征类别,每个特征类别的义原构成一个树状的层次体系,可称为义原树。

2008版《知网》中概念的描述架构和以前版本有很大不同,概念的定义由主类义原及特征描述两部分组成。主类义原是描述概念的最基本的语义,而特征描述是用特征角色和特征标注来详细定义概念,特征描述可以有多重嵌套。例如,词语“医院”在2008版《知网》中的定义如下:

NO. = 155586
W_C = 医院
G_C = noun [yi1 yvan4]
S_C =
E_C =
W_E = hospital
G_E = noun
S_E =
E_E =
DEF = {InstitutePlace|场所:domain = {medical|医},{doctor|医治:content = {disease|疾病},location = {~}}}

其中,NO.为概念编号;W_C,G_C,E_C分别是汉语的词语、词性和例子;W_E,G_E,E_E分别是对应的英语词语、词性和例子;DEF是《知网》对于该概念的定义。在DEF的描述中,场所是主类义原,场所冒号后面就是由特征角色和特征标注组成的特征描述,其中特征标注也是义原。这个特征描述对场所作了详细的说明,其中包括了2层嵌套。

3 词语相似度计算方法

词语相似度可以体现为描述词语概念之间的相似度,用公式表示为:

$$sim(W_1,W_2) = \max sim(S_{1i},S_{2j}) \tag{1}$$

词语W_1,W_2分别有n和m个概念;S_{1i}为W_1的第i个概念,S_{1j}为W_2的第j个概念,2个词语的相似度取W_1和W_2的各个概念相似度的最大值。其中式(1)在计算中结合概念的词性,词性相同的概念分类组合,减少不同词性的概念组合的计算量。本文提出了一种计算词语相似度的新方法,该方法可以在主义原相似度计算和特征描述相似度计算基础上得到概念相似度。

3.1 主义原相似度计算

主义原确定了概念最主要的语义,主义原相似度的计算对概念相似度计算有重大影响。主义原相似度的计算一般是依靠义原树的树形层次体系来计算。其中,一类方法是依据义原在树形层次体系中的距离来计算,2个义原的距离越近,则它们的相似度也越大;另一类方法是利用义原的信息量来计算,如果2个义原的公共信息量越大,则它们的相似度也越大。文献[1]给出的公式是:

$$sim(P_1,P_2) = \frac{\alpha}{\alpha + dis(P_1,P_2)} \tag{2}$$

其中,P_1和P_2表示2个义原;$dis(P_1,P_2)$表示2个义原在义原树层次体系中的路径长度;α是一个可调节的参数。文献[4]参考了文献[3]根据义原的层次深度计算相似度的思路,考虑了义原所在层次的影响,提出了修改后的公式:

$$sim(P_1,P_2) = \frac{\alpha \times \min(dep(P_1),dep(P_2))}{dis(P_1,P_2) + \alpha \times \min(dep(P_1),dep(P_2))} \tag{3}$$

其中,$dep(P_1)$和$dep(P_2)$分别表示义原P_1和P_2在义原树层次体系中的层次深度,根节点的层次深度为1。

文献[4]在借鉴文献[5]中利用WordNet计算

英文词语相似度的公式,提出了基于义原信息量来计算义原相似度的公式:

$$sim(P_1,P_2) = \frac{2 \times \log f(LCN)}{\log f(P_1) + \log f(P_2)} \quad (4)$$

其中,LCN 表示义原 P_1 和 P_2 在义原树中的最近公共父节点;$f(P)$ 表示该节点的子节点个数(包括自己)与树中的所有节点个数的比值。文献[6]综合基于义原在义原树中的距离以及最近公共父节点提出了如下公式:

$$sim(P_1,P_2) = \frac{2 \times \sum_{i=1}^{n} \frac{1}{(\alpha + i)}}{\sum_{j=1}^{m} \frac{1}{(\alpha + j)} + \sum_{k=1}^{h} \frac{1}{(\alpha + h)}} \quad (5)$$

其中,α 是一个可调节的参数;m,h,n 分别表示义原 P_1,P_2 以及 P_1 和 P_2 的最近公共父节点的层次数。

以上方法只简单考虑义原之间的距离以及所在义原树的深度,或者义原之间的公共信息量,并未综合考虑影响义原相似度的各种因素。因此,综合考虑义原所代表的信息量、所在义原树的深度及结构特征,提出基于义原在义原树的深度信息量及路径的混合方法来计算义原相似度。首先考虑义原在义原树的深度,定义义原 p 在义原树中的深度信息量 $IC(p)$,$IC(p)$ 的计算公式如下:

$$IC(p) = \left(1 - \frac{\log(num\_chi(p) + 1)}{\log(num(T))}\right) \times \left(1 + \log \frac{depth(p)}{\max depth(T)}\right) \quad (6)$$

其中,$num\_chi(P)$ 表示义原 P 的子节点个数;$num(T)$ 表示义原 P 所在义原树的总节点个数;$depth(P)$ 表示义原 P 在义原树的深度;$maxdepth(T)$ 表示义原树的最大深度。当义原的子孙节点个数越多,义原的深度越小,该义原的深度信息量越小,即该义原越抽象,包含的语义信息也越少。本文基于深度信息量及路径的混合方法所涉及的相关术语如下:

定义 1(路径)　$T = <P,E>$ 是一个有向树,设根节点 P_0 和 P 之间的路径 $V = (P_0,P_1,\cdots,P_n)$,其中,$P_n = P$,P_i 是 P_{i+1} $(0 \leqslant i \leqslant n-1)$ 的直接祖先,即 P_i 和 P_{i+1} 存在有向边连接。

定义 2(路径的交)　设有向树中节点 P_1 和 P_2 的路径分别是 V_1 和 V_2,则路径 V_1 和 V_2 的交记为 $V_{1\cap2}$,$V_{1\cap2}$ 包含的所有节点同时出现在路径 V_1 和 V_2 中。

定义 3(路径的并)　设有向树中节点 P_1 和 P_2 的路径分别是 V_1 和 V_2,则路径 V_1 和 V_2 的交记为 $V_{1\cup2}$,$V_{1\cup2}$ 由在路径 V_1 和 V_2 中全部节点组成。

若 2 个义原不在同一棵义原树上,则相似度取一个极小常数 0.001,若 2 个义原在同一棵义原树上,计算 2 个义原相似度的算法步骤如下:

(1) 分别计算出根节点到 2 个义原节点的路径。

(2) 分别计算出路径的交和路径的并。

(3) 计算出 $V_{1\cap2}$ 中各节点的深度信息量之和:

$$IC(V_{1\cap2}) = \sum_{i \in 1\cap2} IC(i) \quad (7)$$

(4) 计算出 $V_{1\cup2}$ 中各节点的深度信息量之和:

$$IC(V_{1\cup2}) = \sum_{j \in 1\cup2} IC(j) \quad (8)$$

(5) 计算出 $V_{1\cap2}$ 中各节点的深度信息量之和 $IC(V_{1\cap2})$ 与 $V_{1\cup2}$ 中各节点的深度信息量之和 $IC(V_{1\cup2})$ 的比值作为节点 P_1 和 P_2 代表的义原之间的相似度 $sim(P_1,P_2)$:

$$sim(P_1,P_2) = \frac{IC(V_{1\cap2})}{IC(V_{1\cup2})} \quad (9)$$

为了比较本文方法与其他方法的优劣,选取了 3 组义原对(A:“牲畜”和“禽”,B:“动物”和“植物”,C:“动物”和“禽”,其所在的义原树示意图如图 1 所示。分别使用式(2)~式(5)和本文方法来计算其相似度,在式(2)和式(3)中,参数 α 的取值均为1.6,式(5)中参数 α 的取值是 4,这些参数的取值均与对应文献中的一致,具体结果如表 1 所示。

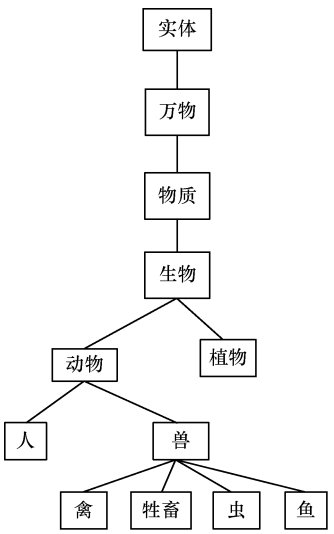

图 1　实体义原树的部分示意图

表 1　不同方法计算的义原相似度

方法	义原对 A	义原对 B	义原对 C
式(2)方法	0.444	0.444	0.444
式(3)方法	0.848	0.800	0.800
式(4)方法	0.643	0.723	0.720
式(5)方法	0.852	0.851	0.887
本文方法	0.424	0.285	0.354

虽然义原相似度的评价没有精确的数值来参考,但是可以依据义原的位置和人工判断来对义原相似度的大小排序。根据义原在义原树中的位置和人工直觉判断,3 组义原对的相似度从大到小依次为:A,C,B。在表 1 中的结果可以发现,式(2)计算

3 组义原对的相似度的值都相等,这是因为式(2)计算义原相似度时只考虑义原之间的路径长度,没有考虑义原深度差异的影响,而这 3 组义原对中义原之间的路径长度都为 2,由此看出式(2)的计算方法不太合理。式(3)计算 3 组义原对的相似度时比式(2)稍微合理一些,其中义原对 A 的相似度和义原对 B 及义原对 C 的相似度不相等,但是义原对 B 和义原对 C 的相似度仍然相等而没有区别,这是因为式(3)只考虑了义原的最小深度,而义原对 B 和义原对 C 的义原最小深度都是 5。式(4)计算 3 组义原对的相似度均不相等,但是计算出的相似度结果按大小排序与人工判断的并不一致,这表明只利用义原信息量计算义原相似度并不合理。式(5)计算 3 组义原对的相似度虽然有合理的部分,其中义原对 C 的相似度比义原对 B 的相似度大,符合人工判断,但是两者十分接近,区分不明显,同时义原对 A 的相似度不符合人工判断,显示出即使考虑了公共祖先和义原深度仍然不能较精确地刻画义原相似度。而本文方法计算 3 组义原对的相似度的值虽然偏小,但是大小排序完全符合人工判断,而且三者之间的相似度有一定的差距,区分较明显,表明本文方法比其他方法计算得到的结果更具合理性。

3.2 特征描述相似度计算

特征描述是概念在主义原的基础上进行详细说明,可以分为有动态角色和无动态角色 2 种,同时可以有多层嵌套。如例子“医院”在《知网》中的定义所示:对主义原“场所”的特征描述有 2 层嵌套,第 1 层次有“医”和“医治”,其中特征标注“医”前面还有动态角色“domain”,而“医治”则没有动态角色。第 2 层次有“疾病”和“场所”,其中,“疾病”的动态角色是“content”,场所的动态角色是“location”。计算 2 个特征描述相似度算法步骤如下:

(1)分别对 2 个特征描述层次分组。

(2)对同一层次的特征描述配对,计算其特征标注相似度。其中有相同动态角色的特征标注和没有动态角色的特征标注分别组成集合对。以没有动态角色的特征标注为例,设集合对是 T_1 和 T_2,T_1 中包括的特征标注为 $t_i(i=1,2,\cdots,m)$,T_2 中包括的特征标注为 $t_j(j=1,2,\cdots,n)$。首先利用本文的基于深度信息量与路径的混合方法计算 T_1 和 T_2 中的特征标注相似度 $sim(t_i,t_j)$得到特征标注相似度集合 S。取出特征标注相似度集合 S 中最大值 $sim(t_u,t_v)=\max\{sim(t_i,t_j)\}$ 加入集合 R,且在 T_1 中删除 t_u,在 T_2 中删除 t_v,并在 S 中删除涉及 t_u 和 t_v 的特征标注相似度值。重复以上步骤,直到 T_1 或 T_2 为空。将集合 R 中的特征标注相似度的平均值作为配对特征标注相似度 $sim(T_1,T_2)$,具体计算公式如下:

$$sim(T_1,T_2)=\frac{\sum sim(t_u,t_v)}{\max(m,n)} \tag{10}$$

如果存在不同动态角色的特征标注,则集合对 T_1 和 T_2 两者其中之一为空集,此时集合 R 中的特征标注相似度值只有一个值为 0,配对特征标注相似度 $sim(T_1,T_2)$也为 0。

(3)计算出每个层次的特征描述的相似度,即每个层次配对的特征标注相似度的平均值。设层次的特征描述的配对集合数为 s,每个配对集合得到的集合 R 中的特征描述相似度值的个数为 $t_h(h=1,2,\cdots,s)$,则层次的特征描述相似度 $sim(C_1,C_2)$可由以下公式计算得到:

$$sim(C_1,C_2)=\sum_{h=1}^{s}\frac{t_h}{\sum t_h}sim(T_{1h},T_{2h}) \tag{11}$$

(4)将每个层次的特征描述相似度加权得到总的特征描述相似度 $sim(D_1,D_2)$,公式为:

$$sim(D_1,D_2)=\sum_{k=1}^{n}\lambda_k sim(C_{1k},C_{2k}) \tag{12}$$

其中,D_1 和 D_2 分别表示 2 个概念的特征描述;C_{1k}和 C_{2k}表示对应层次的特征描述;λ_k 表示加权系数,且 $\sum_{k=1}^{n}\lambda_k=1$,$\lambda_k$ 的取值与层次个数有关,由于上一层次的特征描述比下一层次的特征描述对概念相似度的影响更大,因此上一层次的特征描述的权重比下一层次的特征描述的权重更大,即 $\lambda_k>\lambda_{k-1}$。经过反复实验,最终设定当层数 n 的值是 1 时,$\lambda_1=1$,当层数 n 的值是 2 时,$\lambda_1=0.6$,$\lambda_2=0.4$;当层数 n 的值是 3 时,$\lambda_1=0.5$,$\lambda_2=0.3$,$\lambda_3=0.2$;当层数 n 的值是 4 时,$\lambda_1=0.4$,$\lambda_2=0.25$,$\lambda_3=0.2$,$\lambda_4=0.15$,当层次 n 的值大于 4 时,只计算前 4 个层次。

3.3 概念相似度计算

在计算得到 2 个概念的主义原相似度和特征描述相似度的基础上,2 个概念的相似度 $sim(S_1,S_2)$可由以下公式计算得到:

$$sim(S_1,S_2)=\theta\times(\beta\, sim(P_1,P_2)+(1-\beta)sim(D_1,D_2)) \tag{13}$$

其中,S_1 和 S_2 分别表示 2 个概念;β 表示主义原相似度在概念相似度中的权重,当 2 个概念都没有特征描述时,β 等于 1;一般情况下取[0,1]的某个实数,经过反复实验后,设定 β 为 0.6;θ 表示惩罚因子,一般情况下,θ 等于 1,当 2 个概念中的义原存在反义或对义的关系时,则 θ 取[0,1]的某个实数。由于当 2 个概念的义原存在反义或对义的关系时差异较大,经过反复实验后,设定 θ 为 0.1。

4 实验与结果分析

目前基于《知网》的中文词语相似度计算研究,

除了文献[4,6],文献[7]从信息论的角度出发,改进了义原间的相似度计算公式。文献[8]利用义原的其他关系来考虑到词语的极性对词语相似度的影响。文献[9]引入弱义原概念,排除了弱义原对词语相似度计算的干扰。文献[10]根据不同类型的义原个数来调整类型义原的计算权重。文献[11]提出了新的义原描述式权重分配方案。这些文献都是在文献[1]的基础上基于旧版《知网》作改进。为了验证本文方法的有效性,选取了文献[1]中的一部分数据作为实验词语,分别比较采用文献[1]方法、2008 版《知网》提供的软件包以及本文方法计算词语相似度。实验结果如表 2 所示。

表 2　词语相似度计算结果

词语 1	词语 2	文献[1]方法	软件包方法	本文方法
男人	父亲	1.000	0.696	0.840
男人	经理	0.630	0.498	0.601
男人	和尚	0.861	0.529	0.600
男人	收音机	0.112	0.019	0.014
男人	鲤鱼	0.209	0.035	0.160
男人	苹果	0.171	0.029	0.076
男人	工作	0.112	0.021	0.001
工人	教师	0.722	0.591	0.685
工人	农民	0.722	0.646	0.846
中国	美国	0.936	0.684	0.867
中国	联合国	0.136	0.017	0.051
跑	跳	0.444	0.119	0.401
美丽	贼眉鼠眼	0.815	0.444	0.029
美丽	优雅	0.788	0.024	0.041
高尚	卑鄙	0.788	0.024	0.019

从结果来看,2008 版《知网》提供的软件包和本文方法都比文献[1]方法更符合人的主观认识,其中的原因可能是因为 2008 版《知网》比以前版本的《知网》对词语的定义更精确,所以可以得到更好的结果。例如,“男人”和“父亲”2 个词在以前版本的《知网》的定义完全相同,所以导致文献[1]方法计算出来的结果相似度为 1。而 2008 版《知网》对“男人”和“父亲”2 个词的定义则有区别,所以 2008 版《知网》提供的软件包和本文方法计算出来的相似度没有为 1。但是更主要的原因是本文方法充分分析与利用了 2008 版《知网》对词语的更准确定义,考虑了多种影响词语相似度的因素。

本文方法和 2008 版《知网》提供的软件包相比,在计算不具褒贬性的中性词语时大部分结果都比较接近,同时在某些数据上得到的结果更好。例如,“跑”和“跳”2 个词的相似度的结果表明本文方法计算得到的 0.401 比 2008 版《知网》提供的软件包计算得到的 0.119 要更合理些。在计算具有褒贬性的词语时,本文方法能有效地显示出词语的褒贬性对词语相似度的影响:即褒义词与褒义词的词语相似度要比褒义词与贬义词的词语相似度大。例如:本文方法计算褒义词“美丽”和褒义词“优雅”的词语相似度要大于褒义词“美丽”和贬义词“贼眉鼠眼”的词语相似度。而 2008 版《知网》提供的软件包计算褒义词“美丽”和褒义词“优雅”之间的词语相似度却小于褒义词“美丽”和贬义词“贼眉鼠眼”的词语相似度。这是因为本文方法在计算词语相似度时考虑了义原之间的反义和对义的关系。不同的参数选择,会对词语相似度计算产生细微的影响,例如某些词语的概念特征描述比较详细,可以考虑增加特征相似度的权重,即将 β 的取值调低一些,可能使计算结果更准确一些。

5　结束语

由于《知网》具有丰富的语义知识,因此《知网》是中文词语相似度计算的理想平台。但是目前中文词语相似度计算大部分是基于旧版的《知网》,由于旧版的《知网》在某些方面不够完善,因此会影响到中文词语相似度计算的准确性。本文分析和利用 2008 版《知网》的词语概念的描述架构,从概念的主义原定义和概念的特征描述两方面综合计算得到词语的相似度。实验结果表明本文方法得到的词语相似度和人的主观认识更趋于一致,且部分实验结果优于 2008 版《知网》提供的软件包方法得到的词语相似度。

参考文献

[1] 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论文集. 台北,中国:[出版者不详],2002.

[2] 董振东,董　强. 知网[EB/OL]. (2011-03-18). http://www.keenage.com.

[3] 吴　健,吴朝晖,李　莹. 基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机学报,2005,28(4):595-602.

[4] 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.

[5] Lin Dekang. An Information-theoretic Definition of Similarity[C]//Proceedings of the 15th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann,1998:296-304.

[6] 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.

[7] 夏　天. 汉语词语语义相似度研究[J]. 计算机工程,2007,33(6):191-194.

[8] 江　敏,肖诗斌,王弘蔚,等. 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(5):84-89.

[9] 林　丽,薛　方,任仲晟. 一种改进的基于《知网》的词语相似度计算方法[J]. 计算机应用,2009,29(1):217-220.

[10] 王小林,王小义. 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.

[11] 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.

编辑　顾逸斐