

多特征结合的词语相似度计算模型

张培颖¹, 房龙云²

(1. 中国石油大学(华东) 计算机与通信工程学院, 山东 青岛 266580;
2. 哈尔滨工业大学深圳研究生院 计算机科学与技术学院, 广东 深圳 518055)

摘要: 词语相似度计算在基于实例的机器翻译、信息检索、自动问答系统等有着广泛的应用。词语相似度的计算一般都是在基于《知网》的义原的基础上, 通过计算概念之间的相似度来获取。文中在综合考虑义原距离、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。为了验证算法的合理性, 利用 Miller 和 Charles 文献给出的基准词作为测试集合, 将计算得到的词语相似度的值与专家值进行比较, 计算其皮尔逊相关系数, 计算结果达到了 0.852。实验结果表明多特征结合的词语相似度计算和专家评定的词语相似度计算非常吻合。

关键词: 词语相似度; 知网; 同义词词林; 语义距离

中图分类号: TP391.1

文献标识码: A

文章编号: 1673-629X(2014)12-0037-04

doi:10.3969/j.issn.1673-629X.2014.12.009

Word Similarity Computation Model of Multi-features Combination

ZHANG Pei-ying¹, FANG Long-yun²

(1. College of Computer & Communication Engineering, China University of Petroleum (East China),
Qingdao 266580, China;

2. School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate
School, Shenzhen 518055, China)

Abstract: Semantic similarity computing has been widely used in machine translation based on example, information retrieval and automatic question answering systems. Word similarity computation is generally based on the original in "HowNet", through calculating the degree of similarity between concepts to obtain. In this paper, in consideration of the original distance, depth, width, density and contact ratio, use the method with multi-features to compute word similarity. In order to verify the rationality of the algorithm, using the benchmark of words given by Miller and Charles literature as a test set, make a comparison between the word similarity computation values and expert value, calculating the Pearson correlation coefficient, the calculation results is 0.852. Experimental result show that the word similarity computation of multi-features combination is identical with expert estimation.

Key words: word similarity; HowNet; Tongyici Cilin; semantic distance

0 引言

词语相似度计算在中文信息处理领域有着重要的应用。词语相似度研究的是用什么样的方法计算两个词语对之间的语义相似度, 它的主要应用有以下几点:

1) 在基于实例的机器翻译中, 词语相似度的计算有着重要的作用。例如要翻译“张三写的论文”这个短语, 通过语料库检索得到译例:

(1) 李四写的论文/ the paper written by Li Si.

(2) 去年写的论文/ the paper written last year.

通过词语相似度计算发现, “张三”和“李四”都是

具体的人名, 语义上非常相似, 而“去年”是具体的时间名词, 和“张三”相似度较低, 因此选择“李四写的论文”这个实例进行类比翻译, 就可以得到正确的译文: The paper written by Zhang San.

2) 词语相似度计算在信息检索中, 可以反映文本与用户查询在语义上的符合程度。

3) 在中文自动问答系统中, 词语相似度的计算主要体现在计算问句和语料库中的问句之间的相似度, 从而自动给出问句的答案。

4) 在文本分类领域中, 词语相似度可以用来计算

文本与给定的分类体系中某类别的语义相关程度。

词语相似度计算结果的准确性直接影响上面的几类应用。词语相似度计算方法大体可以分为两类^[1-4]：

(1) 基于统计的相似度方法。

这种方法利用大规模的语料库进行统计,主要是利用词语所处的上下文信息,将其信息分布概率作为一种度量来计算词语之间的相似度。但由于大规模语料库一般都是针对某个领域的,而且容易受到里面噪音和稀疏数据的干扰,所以即使这种方法有时能够取得较高的准确度,但并不常用。

(2) 基于本体知识的相似度方法。

这种方法主要是利用前人通过对词语间概念的层次结构关系来定义的词汇树来计算词语之间的相似度。目前英文比较著名的本体知识库是 WordNet,中文比较著名的本体知识库是《HowNet》和《同义词词林》。基于本体的词语相似度计算方法虽然有效,但词典毕竟是通过人构建的,存在着或多或少的主观因素,加之中文词语博大精深,两个本体库都没有完全涵盖中文所有词语,只是包含了大部分常见词语。因此在一定程度上也会影响计算的准确度。

在汉语词语相似度计算研究方面,文献[1]提出了一种基于《知网》的词汇语义相似度计算方法。该方法在计算两个概念的语义表达式之间的相似度时,采用了“整体的相似度等于部分相似度加权平均”的做法。在计算两个义原之间的相似度时,利用义原之间的上下位关系得到它们之间的语义距离并进行转换的方法。该算法充分利用了《知网》中对每个概念进行描述的丰富的语义信息,得到的语义相似度与人的直觉比较符合。

该算法主要考虑的是义原树之间的最短路径作为衡量两个义原之间的语义距离的依据,然后把义原距离转换为义原相似度。然而,义原之间的相似度不仅与义原树中两个节点之间的最短路径有关,还和义原节点所在区域的密度、深度有关。文中综合考虑义原之间最短路径、义原所在区域的深度和密度信息,提出了一种多特征结合的词语相似度计算方法。

1 多特征结合的词语相似度计算

1.1 语义深度相似度

直觉上,义原之间的相似度值应该随着义原在义原层次树中的深度不同而不同。在义原概念层次树中,路径长度相同的两个节点,如果位于概念层次的越低层,其语义距离较大。例如:“动物”和“植物”、“哺乳动物”和“爬行动物”,这两对概念间的路径长度都是2,但是前一对词语处于语义树的较高层,因此认为

其语义距离较大。

定义1:语义树根节点的深度为1,其他节点深度为其父节点深度加1。

定义2:假定两个义原节点A和B在义原层次树中的深度分别为DA和DB。如果两个义原节点相同,则义原深度相似度 $\text{Sim}_{\text{DEP}}(A, B) = 1$;如果节点B是根节点,则义原深度相似度 $\text{Sim}_{\text{DEP}}(A, B) = 1/DB$;如果节点A是根节点,则义原深度相似度 $\text{Sim}_{\text{DEP}}(A, B) = 1/DA$;如果节点B是节点A的祖先,则义原深度相似度 $\text{Sim}_{\text{DEP}}(A, B) = (DA-1)/(DB-1)$;如果节点A是节点B的祖先,则义原深度相似度 $\text{Sim}_{\text{DEP}}(A, B) = (DB-1)/(DA-1)$;如果两个节点没有相同的祖先,如果 $DA > DB$,则义原深度相似度 $\text{Sim}_{\text{DEP}}(A, B) = (DB-1)/(DA-1)$,否则 $\text{Sim}_{\text{DEP}}(A, B) = (DA-1)/(DB-1)$ 。

下面举例说明语义深度相似度计算方法。如图1所示, $\text{Sim}_{\text{DEP}}(\text{'实体'}, \text{'物质'}) = 1/3$; $\text{Sim}_{\text{DEP}}(\text{'生物'}, \text{'兽'}) = (4-1)/(6-1) = 0.6$; $\text{Sim}_{\text{DEP}}(\text{'鱼'}, \text{'兽'}) = 1$ 。

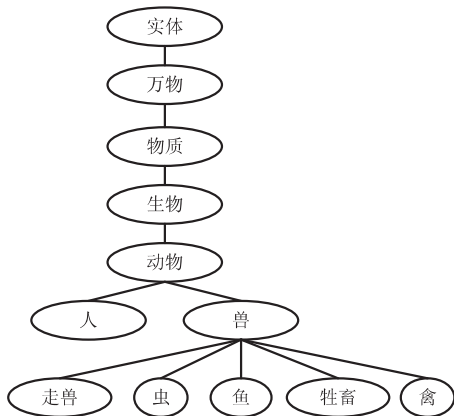


图1 义原层次树的结构示意图

1.2 语义宽度相似度

直觉上来讲,义原之间的相似度应该和义原在义原树中父节点的子女数有关。

例如:图2中节点4和7的父节点子女数为2,节点9和12的父节点子女数为3,两组节点的最短距离相同,深度因素也相同,如果采用文献[1]的算法,其相似度相同;但是义原父节点子女数越多,说明它继承

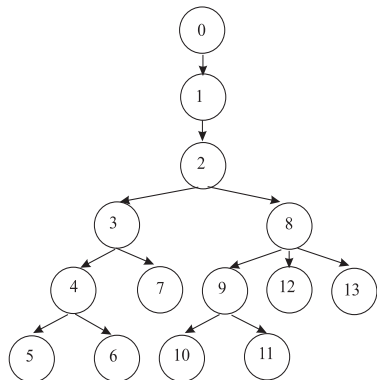


图2 知网义原层次树结构示意图

自父节点的属性越少,与其他节点的相似度越小。节点 4 和节点 7 的相似度应该比节点 9 和节点 12 的相似度要大。本节考虑义原父节点的宽度因素,来计算相似度。

定义 3:叶子节点的宽度为 1;其他节点的宽度定义为其节点含有的孩子节点数。

定义 4:如果两个节点相同,则其语义宽度相似度为 1;如果节点 A 和节点 B 具有祖先关系,则语义宽度相似度定义为节点 A 和节点 B 之间最短路径上所有节点宽度乘积的倒数;如果节点 A 和节点 B 不存在祖先关系,则节点 A 和节点 B 之间的语义宽度相似度为 A 和 B 父节点的相似度除以 A 和 B 父节点的宽度的乘积。

如图 1 所示,“生物”和“兽”之间的语义宽度相似度为 $1/(1 * 2 * 5) = 0.1$ 。

1.3 语义密度相似度

词语之间的相似度应该和分类树型结构中的区域密度相关。对于分类树型结构中密度区域较大的概念,分类比较具体,相似度较大;对于分类树型结构中密度区域较小的概念,分类比较抽象,相似度较小。例如:“动物”和“植物”、“水果”和“蔬菜”的路径长度都为 2,但“水果”和“蔬菜”所处的区域密度更大,因此相似度应该更大。

定义 5:义原节点的密度定义为以当前节点为根的子树的节点个数与以当前节点父节点为根的子树的节点个数的比值。

如图 1 所示,义原“生物”和“兽”之间的密度相似度为 $6/9$;义原“兽”和“走兽”之间的密度相似度为 $1/6$ 。

1.4 语义重合度相似度

从语义重合度角度来看,两个义原之间的相似度应该和它们在义原树中的重合程度相关。如图 2 所示,节点 3 和节点 5 与节点 3 和节点 8 的最短路径相同,如果采用文献[1]的算法,两者的相似度应该相同;但是在义原树中重合度较大的节点对之间的相似度应该比义原树中重合度较小的节点对之间的相似度要大。本节主要考虑义原节点对在义原树中的语义重合度相似度。

定义 6:记节点 A 到根节点的最短路径长度为 LA ,节点 B 到根节点的最短路径长度为 LB ,节点 A 到根节点的最短路径与节点 B 到根节点的最短路径重合部分的长度记为 C ;则节点 A 与 B 之间的语义重合度相似度为: $\text{Sim}_{\text{OVERLAP}} = 2 * C / (LA + LB)$ 。

基于这种算法,在图 2 中,节点 3 和节点 5 之间的相似度为: $\text{Sim}_{\text{OVERLAP}} = 2 * 4 / (4 + 6) = 0.8$;节点 3 和节点 8 之间的相似度为: $\text{Sim}_{\text{OVERLAP}} = 2 * 3 / (4 + 4) = 0.75$ 。

1.5 多特征结合的义原相似度

通过上面的分析可以知道,仅仅考虑义原对之间的最短路径计算义原之间的相似度是不合理的。通过上面讨论的影响义原之间相似度的各种因素,这里提出了一种多特征结合的义原相似度计算方法^[5-7]。该算法综合考虑多种因素的影响,最终的义原相似度计算公式如下:

$$\text{Sim}(A, B) = \lambda_1 \times \text{Sim}_{\text{DEP}}(A, B) + \lambda_2 \times \text{Sim}_{\text{WIDTH}}(A, B) + \lambda_3 \times \text{Sim}_{\text{DENSITY}}(A, B) + \lambda_4 \times \text{Sim}_{\text{OVERLAP}}(A, B)$$

其中, $\lambda_i (1 \leq i \leq 4)$ 是调整参数,并且满足 $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$ 和 $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.25$ 。

1.6 词语相似度计算方法

对于两个词语 W_1 和 W_2 ,如果 W_1 有 n 个义项(概念): $S_{11}, S_{12}, \dots, S_{1n}$; W_2 有 m 个义项(概念): $S_{21}, S_{22}, \dots, S_{2m}$ 。则 W_1 和 W_2 的相似度是各个概念的相似度的最大值。

$$\text{Sim}(W_1, W_2) = \max_{i=1,2,\dots,n, j=1,2,\dots,m} \text{Sim}(S_{1i}, S_{2j})$$

概念之间的相似度可以通过义原之间的相似度计算得到,根据文献[1]的算法,主要分为四部分:

(1)第一基本义原。直接计算两个义原之间的相似度,记为 $\text{Sim}_1(S_1, S_2)$ 。

(2)其他基本义原。可以看作是一个集合,通过建立两个集合中元素的对应关系进而计算两个集合的相似度,记为 $\text{Sim}_2(S_1, S_2)$ 。

(3)关系义原。关系义原是包含若干个“属性:值”对的集合。通过建立两个特征结构的特征之间的一一对应关系从而转化为计算相同“属性”对应的“值”的相似度,记为 $\text{Sim}_3(S_1, S_2)$ 。

(4)关系符号描述。其值为一个特征结构,转换为两个特征结构的相似度计算问题,记为 $\text{Sim}_4(S_1, S_2)$ 。

概念之间的相似度可以通过义原之间的相似度计算得出。计算公式如下:

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

其中, $\beta_i (1 \leq i \leq 4)$ 是调节参数并且满足等式: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ 和 $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ 。义原之间的相似度采用 1.5 节中的公式计算。

2 实验结果与分析

2.1 实验设计

Rubenstein 和 Goodenough^[8]建立了 65 对名词作为测试词语相似度的基准词语,分别邀请 51 位专家人工评定每对词语之间的相似度值,相似度值范围从 0 到 4。Miller 和 Charles^[9]采用同样的方法进一步对基准

词语进行筛选,挑选了 30 对名词进行测量,这 30 对名词之间的相似度值高、中、低分布均匀。

2.2 实验结果与分析

为了验证算法的合理性和有效性,文中采用 Miller 和 Charles 给出的一组词语的相似度为基准,用改进后的算法计算同一组词语的相似度,然后对这组标准词的两组相似度计算皮尔逊相关系数来衡量算法改进后的优劣程度。

由于中英文差异,翻译了 10 对词语进行相似度计算。计算结果如表 1 所示。

表 1 词语相似度值和相关系数

词语	Miller's	深度	宽度	密度	重合度
汽车、机动车	3.920	0.870	0.870	0.870	0.870
男孩、少年	3.760	1.000	0.750	0.750	0.875
工具、农具	2.950	0.917	0.413	0.398	0.841
兄弟、和尚	2.820	0.861	0.722	0.722	0.819
少年、兄弟	1.660	0.806	0.500	0.500	0.653
旅程、轿车	1.160	3.556	0.000	0.000 1	0.217
修士、圣人	1.100	0.875	0.500	0.500	0.688
公墓、森林	0.950	0.500	0.000 1	0.000 1	0.286
海岸、丘陵	0.870	0.167	0.000	0.000	0.100
森林、墓地	0.840	0.417	0.000	0.000	0.231
相关度	1.000	0.696	0.753	0.857	0.852

通过上面的结果,可以看到单一利用某个影响义原的因素进行相似度计算,有时结果不是令人很满意。下面综合利用各因素特征进行相似度计算,结果如表 2 所示。

表 2 各种方法的相关度系数

计算方法	皮尔逊相关系数
基于语义距离的方法	0.856 106
基于语义深度的方法	0.694 618
基于语义宽度的方法	0.753 767
基于语义密度的方法	0.857 034
基于语义重合度的方法	0.852 639
综合各项因素的多特征方法	0.852 874

如果皮尔逊相关系数在 0.8 以上,则表明两组值有很强的线性相关性。单一因素虽然有些也取得了较好的相关系数,但是多特征结合的词语相似度计算方法能够在大多数词语相似度计算中取得较为稳定的结果。

表 2 显示用多特征结合的词语相似度算法计算的相似度的值与采用的标准值之间的线性相关性较好,所以该算法具有良好的合理性和有效性。

3 结束语

文中基于《知网》探讨了多特征结合的词语相似度计算方法。针对文献[1]中算法只考虑义原在义原层次树中的最短路径的不足,综合考虑多方面影响义原相似度的因素,利用义原的深度、密度、宽度和重合度因素,提出了一种多特征结合的词语相似度计算方法。实验结果表明,该算法在进行词语相似度计算的结果与人工评定的结果非常接近,相关度系数达到了 0.852。

由于中文词语博大精深,知网中不可能涵盖所有的词语。有些在知网中不存在的词语,在进行相似度计算的时候会产生较大的偏差。下一步的研究准备借助其他本体库,例如《同义词词林》等,作为辅助,使得词语相似度的计算更加准确、合理^[10-14]。

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学研讨会论文集.台北:出版者不详,2002:59-76.

[2] 刘青磊,顾小丰.基于《知网》的词语相似度算法研究[J].中文信息学报,2010,24(6):31-36.

[3] 安建成,武俊丽.基于语义树的概念语义相似度计算方法研究[J].微电子学与计算机,2011,28(1):138-141.

[4] 冉 婕,孙 瑜,漆丽娟.基于本体的概念相似度计算及其应用[J].微型机与应用,2010(11):14-16.

[5] 张玉娟.基于《知网》的句子相似度计算的研究[D].北京:中国地质大学,2006.

[6] 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.

[7] 徐 瑛.一种综合加权的词语语义相似度计算研究[D].青岛:青岛理工大学,2011.

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

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

[10] Liu Hongzhe,Bao Hong,Xu De. Concept vector for similarity measurement based on hierarchical domain structure [J]. Journal of Computing and Informatics,2011,30:1001-1021.

[11] 刘宏哲,须 德.基于本体的语义相似度和相关度计算研究综述[J].计算机科学,2012,39(2):8-13.

[12] Liu Hongzhe,Bao Hong,Xu De. Concept vector for semantic similarity and relatedness based on WordNet structure [J]. Journal of Systems and Software,2012,85(2):370-381.

[13] 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.

[14] 郑 诚,刘娇丽,项 珑.基于 VSM 和 LDA 模型的 FAQ 问答系统[J].计算机技术与发展,2014,24(1):133-135.

多特征结合的词语相似度计算模型

作者：[张培颖](#)，[房龙云](#)，[ZHANG Pei-ying](#)，[FANG Long-yun](#)

作者单位：[张培颖, ZHANG Pei-ying\(中国石油大学 华东 计算机与通信工程学院, 山东 青岛, 266580\)](#)
[， 房龙云, FANG Long-yun\(哈尔滨工业大学深圳研究生院 计算机科学与技术学院, 广东 深圳, 518055\)](#)

刊名：[计算机技术与发展](#)[ISTIC](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(12)

引用本文格式：[张培颖](#). [房龙云](#). [ZHANG Pei-ying](#). [FANG Long-yun](#) [多特征结合的词语相似度计算模型](#)[期刊论文]-[计算机技术与发展](#) 2014(12)