基于信息熵的新的词语相似度算法研究

王小林,陆骆勇,邰伟鹏
(安徽工业大学 计算机科学与技术学院,安徽 马鞍山 243002)

摘 要:针对词语相似度计算中结果合理性的问题,文中基于对“知网”中词语、义项和义原三个层次概念的研究,提出一种结合信息论研究中熵的概念的新的词语相似度方法。首先是引入词表相似度计算对词语集进行合理选取,再根据义原信息熵对各义原进行权重上的平衡,抑制一些常见义原在词语的义原集中比重过大而导致计算结果与真实情况相比出现明显误差的情况。实验结果表明,与传统方法相比,文中方法在实验并未出现 1.000 这样过于绝对的结果,提高了结果的合理性;并且实验词语集而非两词语之间,说明比较的效率也得到了提高。

关键词:词语相似度;知网;义原;信息熵;词表相似度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2015)09-0119-04
doi:10.3969/j.issn.1673-629X.2015.09.026

Research of a New Algorithm of Words Similarity Based on Information Entropy

WANG Xiao-lin, LU Luo-yong, TAI Wei-peng
(School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: The words similarity computation is widely used in the area of natural language processing. In this paper, based on the research of words, concepts and sememe in HowNet, a new algorithm of word similarity based on information entropy is proposed. Firstly, similarity of words surface is led in this paper for selecting words from words set reasonably. Secondly, weight of each sememe would be balanced on the basis of information entropy to inhibition that common sememe would be much more than others in the sememe set what would result in obvious error comparing with physical truth. Experimental results show that compared with traditional methods, the unreasonable result like 1.000 is no-show, which means that the result is rational. In addition, this experiment is based on words set instead of two words, which means that the method is more efficient.

Key words: word similarity; HowNet; sememe; information entropy; similarity of words surface

0 引 言

词语相似度计算是自然语言处理领域一个非常重要的部分,在问句分类、智能翻译、语义排歧、信息检索以及文本相似度计算等方面有着非常广泛的应用。刘群等[1]认为词语的相似度可以通过词语间的语义距离来反映,词语之间的语义距离越大,相似度就越低;反之,相似度就越大。而语义距离可以根据义原树状结构中节点之间的距离来表示。Lin Dekang[2]则从事物的个性与共性的角度出发,认为词语相似度是指在不同的上下文中,词语相互替换后而不会改变原来文段中语法语义结构的程度[3]。李峰等[4]在文献[1]的基础上引入事物信息的概念,通过义原信息对义项相似度的的影响研究词语相似度。

“知网”[5]作为以上研究方法的基础,是由我国著名的机器翻译专家董振东先生创建的一个知识系统。它含有丰富的词汇语义知识和世界知识,内部结构相当复杂。它是以包括汉语和英语在内的词语概念为描述对象,揭示概念之间以及概念属性之间的关系的常识知识库。“知网”中的“义项”和“义原”两个概念是研究语义相似度的基础;“义项”是通过“知识表示语言”来描述词汇语义的,即每一个词汇所包含的几种不同意思,而这种“知识表示语言”称之为“义原”。“义原”是描述一个“义项”的最小意义单位。“义原”相互间又存在着复杂的关系,有上下位关系、同义反义关系、对义关系、部分与整体关系、成品与材料关系、主

收稿日期:2014-09-01 **修回日期**:2014-12-03 **网络出版时间**:2015-08-26
基金项目:安徽省高校自然科学研究重点项目(KJ2013Z023,KJ2013A058);安徽省振兴计划资助项目(2013ZDJY073)
作者简介:王小林(1964-),男,教授,硕士,研究方向为人工智能、中文信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1535.008.html

体与属性关系、事件与角色关系。其中在描述过程中最主要的是上下位关系,根据义原的上下位关系,所有的基本义原可以形成一个义原层次体系,如图 1 所示[6]。这个体系结构是进行词语相似度计算的基础。

```
—event|事件
  —static|静态
    —relation|关系
      |—isa|是非关系
      |  |—be|是
      |  |  |—become|成为
      |  |  |—mean|指代
      |  |——BeNot|非
      |—possession|领属关系
      |  |—own|有
      |  |  |—obtain|得到
      |  |  —receive|收受
      |  |—Being To|属于
```

图 1 义原层次体系

1 传统的词语相似度计算方法

1.1 义原相似度计算

在知网的体系中,利用“概念”对词语进行描述和定义,而“知网”并不是简单地将全部“概念”归结到一个树状结构层次体系中,而是要用义原对“概念”进行进一步的描述,义原相似度的计算是计算“概念”相似度的基础。在根据义原上下位关系构成的树形结构的义原层次体系中,研究人员通过树形结构中的节点距离计算义原相似度。国内外的诸多学者也结合大量研究提出了一些经典的义原相似度的计算公式。

刘群等提出的公式:

$$\mathrm{Sim}(P_1,P_2)=\frac{\alpha}{\alpha+\mathrm{dist}(P_1,P_2)} \tag{1}$$

式(1)是基于树形结构中两个节点的路径长度并结合义原的上下位关系提出的。其中,P_1,P_2 表示两个义原;$\mathrm{dist}(P_1,P_2)$ 表示 P_1,P_2 在义原层次体系中的路径长度;α 表示相似度为 0.5 时的路径长度参数,一般取值 1.6。该公式虽计算量小,但结果不够精确。

而李峰等认为除了根据节点之间的路径长度计算义原相似度外,还可以根据两个节点所含有的公有信息大小计算义原的相似度[3],并提出如下公式:

$$\mathrm{Sim}(P_1,P_2)=\frac{2*\log x(P_0)}{\log x(P_1)+\log x(P_2)} \tag{2}$$

其中,$x(P)$ 表示该节点的子节点个数与树形中所有节点个数的比值;P_0 表示距离这两个节点最近的共同的上层义原节点。式(2)在式(1)的基础上引入公有信息的概念使计算结果更加准确。此外,在实际计算中节点深度对相似度计算也有影响,吴健等[7]在此基础上引入了层次深度的概念。认为两个词语随着它们所处层次总和的增加使得分类更加细致,相似度就越高。反之,随着层次差的增加相似度会降低。

1.2 义项概念相似度计算

刘群等将知网对词语的概念即义项描述进行了细化。在概念描述中,第一描述是一个基本义原描述,这是对这一概念最重要的一个描述,描述了概念最基本的特征。此外,还包括其他基本义原描述、关系义原描述和关系符号描述。在计算得到这四类义原相似度的基础上,结合式(3)得到词语的概念相似度:

$$\mathrm{Sim}(C_1,C_2)=\sum_{i=1}^{4}\beta_i\,\mathrm{Sim}_i(P_1,P_2) \tag{3}$$

其中,C_1,C_2 表示两个概念义项;$\mathrm{Sim}_i(P_1,P_2)$ 表示四个义原组的相似度;$\beta_i(1\leqslant i\leqslant 4)$ 为各个义原组相似度的权值,并且为一个可调节的参数,且有 $\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1>\beta_2>\beta_3>\beta_4$。

1.3 词语相似度计算

在知网中,词语是通过概念义项来描述的,一个词语可能有几个概念义项。那么词语相似度的计算可通过计算这些义项相似度实现。假设有两个词语 W_1 和 W_2,其中 W_1 有 m 个概念,W_2 有 n 个概念,表示如下:

$$W_1=\{C_{11},C_{12},\cdots,C_{1m}\},W_2=\{C_{21},C_{22},\cdots,C_{2n}\}$$

刘群认为将 W_1 和 W_2 中所有概念两两组合计算概念相似度,这些计算结果中的最大值即为 W_1 和 W_2 的词语相似度,公式如下:

$$\mathrm{Sim}(W_1,W_2)=\max_{i=1,2,\cdots,m,j=1,2,\cdots,n}(\mathrm{Sim}(C_{1i},C_{2j})) \tag{4}$$

2 基于词语表层概念与信息熵的词语相似度改进算法

然而上述方法在对词语进行相似度计算时,需要对词语的所有概念义项进行组合,计算各组合的相似度,取最大值作为词语的相似度。文中认为此方法存在一些问题,例如对表 1 中的“篮球”和“运动”两词进行相似度计算时,“篮球”有两个概念义项,“运动”有三个概念义项,如果按照式(4)的方法进行相似度的计算就需要进行 6 次组合。并且得到的结果是“篮球”与“运动”的词语相似度为 1,此结果明显是错误的,说明上述算法的计算准确度有待改善。

表 1 “篮球”与“运动”词语相似度计算

篮球	运动	各组合相似度	结果
N 1 运动器材	N 1 事情 2 变空间位置	0.111	--
N 1 运动器材	N 1 事情 2 锻炼 2 政	0.111	--
N 1 运动器材	N 1 事情 2 活动 2 体育	0.111	--
N 1 事情 2 锻炼 2 体育	N 1 事情 2 变空间位置	0.591	--
N 1 事情 2 锻炼 2 体育	N 1 事情 2 锻炼 2 体育	1.000	1.000
N 1 事情 2 锻炼 2 体育	N 1 事情 2 活动 2 政	0.635	--

此外,在实际应用查询中不可能只是几个词语之间比较相似度,待查询的数据量往往是很大的。例如

文本相似度的比较就是在比较大量词语的相似度基础上结合诸如上下文关系等其他信息条件而进行的计算[8],可见计算词语相似度是其中一个很关键的步骤。而基于分词后会产生大规模数据量的情况,此时如果再利用式(4)的方法,效率和准确率都会变得很低。

在此基础上,文中提出的改进算法能够高效快速地对较大规模的词语集中的词语进行相似度计算。

改进算法如下:

(1)处理过程中,首先将词语表示成概念义项的集合: $W = \{C_1, C_2, \cdots, C_m\}$ 。其中, C_i 表示词语所包含的义项。

(2)引入词表相似度的概念,词语 W_1 和 W_2 的词表相似度为:

$$\mathrm{Sim}_s(W_1, W_2) = 2 * \Gamma(W_1 I W_2) / (\mathrm{Len}(W_1) + \mathrm{Len}(W_2)) \quad (5)$$

其中, Γ 运算表示求义原集合中的元素个数; I 表示集合之间求交集运算; Len 表示词语的长度。

两个词语的词表相似度越大,则待检测词语与大规模词语集中相同的义原就越多,从而对后续大规模词语集所要做的改动也就越少。这也说明了引入词表相似度的概念符合总体上降低时间复杂度的要求,并且有利于生成质量比较高的检测结果。

(3)在词表相似度计算的基础上引入熵的概念[9],义原所包含信息量的熵值:

$$H(P) = \lg(M/m) \quad (6)$$

其中, P 表示义原; M 表示词语集中的词语总数; m 表示其中出现过义原 P 的词语数。如果 $H(P)$ 的值越大,说明义原 P 在义原集合中出现的频率就越低,则改值对于词语区分的作用也就越大。在此基础上可以通过式(7)计算词语之间的信息熵相似度:

$$\mathrm{Sim}_H = \sum H(P_i) \quad (7)$$

其中, P_i 属于集合 $\{W_1 I W_2\}$ 。若两个词语的信息熵相似度越大,则从概率角度讲,待检测词语与词语实例在词义上更相似。此外通过引入信息熵的概念对一些比较常见的义原诸如{<事情>,<数量值>,<物质>,<属性值>,……}起到抑制的作用。

在进行相似度检测时,首先根据式(5)进行词表相似度的计算,先从大规模词语集中选出 m 个词语,再根据式(7)信息熵相似度的计算方法,从 m 个词语中选出 k 个词语,文中实验设定 $m = 200$, $k = 20$。在此过程中并不是通过信息熵的大小来选取词语集中的实例的,因为若在整个词语集中直接利用信息熵的大小来筛选,会导致筛选结果中非常见的义项所占比重过大。如在实验中义原<政>的信息熵是常见义原<事情>的 13.2 倍,这样会导致相似度检测结果的准确率大大降低。

(4)最后词语相似度的计算公式如下:

$$\mathrm{Sim}(W_1, W_2) = \alpha * \mathrm{Sim}_s(W_1, W_2) + \beta * \mathrm{Sim}_H \quad (8)$$

其中, α, β 为权值系数,分别表示词表相似度和信息熵相似度的权值。并且 $\alpha + \beta = 1$,本实验中取 $\alpha = 0.55$, $\beta = 0.45$。在实际实验过程中,由于信息熵的相似度是基于词表相似度计算出来的,所以设置参数时 α 应该稍大于 β[10-11]。

3　实验数据与分析

现代汉语中常用的词语个数在 56 000 个左右,由于实验说明和算法比较的需要,从文献[1]和文献[10]部分数据中随机选取其中的 200 个作为测试集,然后根据式(7)和式(8)利用文中所提方法对比刘群等提出的方法,得出如表 2 所示的实验结果。

表 2　改进的词语相似度计算实验数据

词语 1	词语 2	文献[1]	文献[10]	文中
发明	创造	0.615	0.615	0.743
美丽	丑陋	0.815	0.758	0.435
跑	跳	0.444	0.518	0.553
篮球	运动	1.000	0.406	0.552
男人	父亲	1.000	0.953	0.673
中国	美国	1.000	0.943	0.865
计算机	电脑	1.000	1.000	0.826
粉色	红色	0.074	0.782	0.833
问题	困难	0.768	0.768	0.745
分析	研究	0.444	0.493	0.762

实验结果分析:

(1)与第三列结果相比,第四列的结果更加符合正常理解的词语相似度比较结果[12]。因为文献[1]中只是根据义原在树状结构中各自之间位置关系计算,所得到的结果不是很准确;而文献[10]是在文献[1]的基础上引入了词语词性的概念,考虑到相同词性对词语相似度计算所起到的作用,使得结果更加合理。

(2)与第三列和第四列的实验结果相比,文中方法所得到的第五列数据结果的变化更加平滑,更加接近平常生活中的理解,没有出现过高或者过低的检测结果,说明文中所提方法的稳定性较好。

(3)第二行"美丽"与"丑陋"的比较结果中,无论是文献[1]和文献[10]所得到的结果相似度都挺高,显然是不合理的。在《知网》中"美丽"与"丑陋"的描述为:

美丽:{aValue|属性值,prettiness|美丑,beautiful|美,desired|良}

丑陋:{aValue|属性值,prettiness|美丑,ugly|丑,undesired|莠}

文献[1]中的方法对基本义原采用的是最大相似度求均值的方法,没有考虑到义原之间的上下位关系;文中所用方法抑制了<属性值>这一义原的相似度计算中所占的权重,降低了这两个词语的相似度,使得结果更加贴近人们正常理解的结果。

(4)第四行中,文中方法改变了“篮球”和“运动”两个词语在文献[1]中相似度为 1.000 的不合理结果,根据词表相似度和词语信息熵的计算使得两者相似度降低了很多,趋于合理。

(5)与传统的比较方法相比,文中方法无需对词语每个义项之间的相似度进行计算,而是通过对义原集合进行混乱度处理删减后进行计算,从而使计算效率得到一定的提高。

(6)第七行“计算机”与“电脑”在前两种方法所得到的实验结果中,相似度都是 1.000,表示这两个词语等价,这是符合日常概念的。然而在文中方法中,由于两者词语长度的不同导致在计算词表相似度时,两个词语之间的相似度降低,从而影响了最终结果的准确性,说明文中方法还有加以改进的地方。文中所提方法只是单纯在词语集中计算词语的相似度,比较适用于检测点信息或者简单句之间的关系。如果要将该方法应用到文本等块信息之间的比较的话,需要将相同词语在不同的语境当中所表达的不同含义考虑进来,这还有待于做进一步的研究。

4 结束语

词语相似度的计算在很多领域都有着重要的作用,尽管许多学者已经做了大量的工作,但应用到具体生活中时,还是存在许多有待改进的地方,所以词语相似度部分的研究仍然是自然语言领域研究的重点[13]。文中在多位学者所做研究的基础上,提出了改进的词语相似度计算方法。在传统词语相似度计算方法的基础上引入词表相似度的概念,提高了相似度检测的效率;并且在义原层面上引入信息熵的概念,利用信息熵降低常见的基本义原在比较过称中较高的权重,使检测结果趋于合理。而将文中的词语相似度计算方法结合词语前后关系进行语句相似度计算[14]将是下一步的研究重点。

参考文献:

[1] 刘 群,李素建. 基于《知网》的词语相似度计算[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] Lin Dekang. An information-theoretic definition of similarity semantic distance in WordNet[C]//Proceedings of the fifteenth international conference on machine learning. [s. l.]:[s. n.],1998.

[3] Rua L F,Jacobs P S. Creating segmented databases from free text of retrieval[C]//Proc of ACM SIGIR. [s. l.]:[s. n.],1991:337-346.

[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] 于江生,俞士汶. 中文概念词典的结构[J]. 中文信息学报,2002,16(4):12-20.

[6] 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.

[7] 吴 健,吴朝晖,李 莹,等. 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.

[8] 王家琴,李仁发,李仲生,等. 基于多层信息的本体概念相似度计算的研究[EB/OL]. 2009-06-10. http://www.paper.edu.cn.

[9] 王 成,吕学强,王弘蔚,等. 基于信息熵与词语活跃度的领域词抽取[J]. 北京信息科技大学学报:自然科学版,2011,26(5):49-52.

[10] Witten I H,Paynteer G W,Frank E,et al. KEA:practical automatic keyphrase extraction[C]//Proc of the 4th ACM conference on digital libraries. California,USA:ACM Press,1999:254-256.

[11] 鲁 松,李晓黎,白 硕,等. 文档中词语权重计算方法的改进[J]. 中文信息学报,2000,14(6):8-13.

[12] 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.

[13] Salton G,Buckley C. Term-weighting approaches in automatic retrieval[J]. Information Processing Management,1988,24(5):513-523.

[14] 吕学强,任飞亮,黄志丹,等. 句子相似模型和最相似句子查找算法[J]. 东北大学学报:自然科学版,2003,24(6):531-534.

基于信息熵的新的词语相似度算法研究

作者： 王小林， 陆骆勇， 邰伟鹏， WANG Xiao-lin， LU Luo-yong， TAI Wei-peng
作者单位： 安徽工业大学 计算机科学与技术学院,安徽 马鞍山,243002
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(9)