基于 HowNet 的词语相关度计算模型*

曾淑琴,吴扬扬
(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘 要:提出了词语相关度模型,作为在数据空间中发现数据源内容关联的一个基础。本模型基于 HowNet,可以计算同种词性以及不同词性之间的相关度,融合了词语的相似度、关联度和实例因素,综合获得词语的内在相关性。通过对比实验发现,本模型所计算的词语相关度值更加符合人们主观上对词语相关性的认识。

关键词:数据空间;HowNet;词语相关度
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1674-7720(2012)08-0077-04

The model of words relation computing based on the HowNet

Zeng Shuqin, Wu Yangyang
(School of Computer Science and Technology, Huaqiao University of China, Xiamen 361021, China)

Abstract: It is the basic of creating index, browsing, searching, querying and other services. The current search mostly is in the premise of having been obtained the relationship between the data. But it often is limited. The model of words relation proposed by this paper is the basic research of discovering the association of data source in the dataspace. The model of word relation is in view of HowNet, could compute the same and the different property of the words. It merges the word similarity, word association, the example factor and so on. Comprehensive these factors and receive the inner relation of words. Via the contract experiments, we find the model of words relation proposed by the paper is closed to and fit the cognition of words relation which people realize subjectively.

Key words: dataspace; HowNet; words relation

语义相关度的研究是自然语义处理 NLP(Natural Language Processing)的基础,广泛用于语义消歧、信息检索、文本分类、文本聚类等领域。本文将其作为数据空间[1]研究课题的基础性内容来研究,旨在从内容上发现数据空间中的数据源之间的关联。

关于语义相关度的研究在国外较多,目前的方法一般分为两类[2]:一种是统计方法,另一种是基于语义词典方法。Jiang 和 Conrath 利用 Wordnet 图的上位关系,通过合并概念 c1 和 c2 的信息内容以及最小的共同类属者,综合基于边以及结点的技术,再用语料库统计作为辅助因素进行矫正[2];Banerjee 和 Pedersen 在 Wordnet 的英文语境下,将单词的解释中重叠的单词数量的平方,及含有上下文等关系类型的词语的单词重叠的数量的平方之和,共同作为最后词语相关度的值[2]。

国内在语义相关方面的研究还较欠缺,且大多数选择英文环境,主要基于 HowNet、词林、维基百科等知识库[3-5]。参考文献[3]根据知网中的特征文件下位义原和上位义原拥有的属性以及纵向语义联系和实例信息计算词语的相关度。参考文献[4]通过挖掘直接或间接的关系而提出的新的语义相关度计算模型,适用于类似知网的知识体系。总结基于语义词典度量语义相关度所考虑的因素,即最短路径长度、局部网络密度、结点在层次中的深度、连接的类型、概念结点的信息含量以及概念的释义,将上述 6 个因素归为三大类:结构特点、信息量和概念释义。

本文在综合了参考文献[3]中所提到的基本义原相似度和关联度以及其他相关研究的基础上定义了一个词语相关度算法模型,实现计算同种词性、不同词性词

* 基金项目:福建省科技计划重点项目(2008I0021)

语之间的相关度。

1 知网

中国人民大学的董振东教授等人编写的《知网》以汉语和英语的词语所代表的概念为描述对象，包含丰富词汇，反映概念的共性和个性，是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。

知网中的语义通过义原描述，共有 1 618 个义原被分成 10 大类，每一类都是由一个树结构来存储，而不同类之间的义原构成一个网状结构，它们通过解释义原关联起来。知网中的词语关系类型[6]如表 1 所示。

2 语义相关度模型

2.1 语义相关概念

定义 1 语义相似度是指两个词在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度[7]。

定义 2 词语关联度是指词语在概念解释上所存在的语义关系的程度。

定义 3 词语相关度是指词语间含有表 1 中的关系类型或存在词语隐含传递等相互关联的特性，即两个词语相互关联的程度从侧面反映了两个词语在同一个语境中共现的可能性，其影响因素有词语的相似性以及关联性等。

鉴于目前国内还没有对相关度判断的标准和类似的专门人工判断的词集，本实验中对相关度的判断主要从两个方面来界定：一是依据上文的定义；二是通过对比参考文献[3]中相关度的实验结果，改进其中一些明显不合理的实验结果来确认本方法的改进性。

2.2 建立词语语义相关度模型

通过对知网结构的分析，根据如下几个因素计算语义相关度：

(1)词语的相似度

知网中的词语通过一个记录来表示，其中有一项语义表达式 DEF 对该词语进行描述，语义表达式由概念和义原组成。知网中义原有 3 个类别，另有一些关系符号对概念的语义进行描述的义原，因此，可以将义原分为基本义原、其他义原、关系义原以及关系符号义原。词语的相似度可以通过这 4 种义原类型求得。

采用下列方法计算两个词语之间的相似度：将两个词语的语义表达式中的义原抽取出来，计算对应义原类型的相似度。如果某一义原类型的对应项为空，则将任何义原（或具体词）与空值的相似度定义为一个比较小的常数；如果某一义原类型包含多个义原，则将各个义原的相似度加权平均作为该类型义原的相似度[7]。

第一基本义原即主要特征义原，两个词语的这一部分的相似度采用式(1)计算：

$$sim_1(p_1,p_2)=\alpha/(1+2^{dis}) \quad (1)$$

参考文献[7]中提到的第一基本义原直接用 path 的倒数计算，不够逼近相关度的实际曲线。本文的思想来源于 BP 神经网络的 S 型函数，该函数所划分的区域是一个非线性的超平面组成的区域，是比较柔和、光滑的任意界面，因而它的分类比线性划分精确、合理，且容错性较好，取值范围在[0，1]之间，其图像更加逼近相关度的实际曲线，故而将其作为第一基本义原的表达式。

其他义原即语义表达式中除第一基本义原以外的所有其他义原(或具体词)，其值是一个特征结构：$sim_2(p_1,p_2)$[6]。

关系义原即对应于所有关系义原描述式，其值是一个特征结构，记为：$sim_3(p_1,p_2)$。

关系符号义原即对应于关系符号描述式，其值是一

表 1 知网中的关系类型

类型	含义
,	多个属性之间，表示“和”的关系
#	表示“与其相关”
%	表示“是其部分”
$	表示“可以被该‘V’处置，或是该‘V’的受事、对象、领有物或者内容”
*	表示“会‘V’或主要用于‘V’，即施事或工具”
+	对 V 类，表示“它所标记的角色是一种隐性的，几乎在实际语言中不会出现”
&	表示“指向”
~	表示“多半是”，“多半有”，“很可能的”
@	表示可以做“V”的空间或时间
?	表示可以是“N”的材料，如对于布匹，标以“?衣服”表示布匹可以是“衣服”的材料
{}	(1)对于 V 类，置于[ ]中的是该类 V 所有的“必备角色”。如对于“购买”类，一旦发生，必然会在实际上有如下角色参与：施事、占有物、来源、工具。尽管在多数情况下，一个句子并不把全部的角色都交代出来；(2)表示动态角色，如介词的定义
()	置于其中的应该是一个词表记，例如：(China|中国)
^	表示“不存在”，“没有”，或“不能”
!	表示某一属性为一种敏感属性，例如：“味道”对于“食物”，“高度”对于“山脉”等
[]	标识概念的共性属性

个特征结构，记为：$sim_4(p_1,p_2)$。

于是，两个概念(义项)语义表达式的整体相似度为[6]：

$$C\_sim(c_1,c_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}sim_j(p_1,p_2) \tag{2}$$

其中 $\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_i$ 的值依次递减，反映了这 4 类义原对整体的相似度所起到的作用是依次递减的。

词语的相似度：$sim(w_1,w_2)=\max C\_sim(c_i,c_j)$

(2)词语的关联度计算

知网的每类义原都用一个树结构来存储，形成上下文的层次结构，而每个义原和不在同一个义原树中的义原彼此也可能存在关系，这样就表现出义原之间的横向联系，也就是关联关系，从而使整个义原体系形成一个网状结构。

本文通过对 HowNet 层次网络结构的分析，找到义原和解释义原之间的重叠部分，从而获取词语关联度的计算模型。

①义原之间的关联度

义原 p_1 和 p_2 的关联度可以表示为：

$$P\_ass(p_1,p_2)=\max\{\sum_{p_{1i}\in\exp(p_1)}q_i\times\frac{dis(p_{1i},p_2)}{n}+\sum_{p_{2j}\in\exp(p_2)}q_j\times\frac{dis(p_1,p_{2j})}{m}\} \tag{3}$$

其中 q_i、q_j 是常数，p_{1i} 表示 p_1 的第 i 个解释义原，p_{2j} 表示 p_2 第 j 个解释义原，$dis(p_{1i},p_2)$ 是分别求 p_2 与 p_1 的解释义原的基本义原的相似度之和，$dis(p_1,p_{2j})$ 是分别求 p_1 与 p_2 的解释义原的基本义原的相似度之和，n 和 m 分别是 p_1 和 p_2 解释义原的个数。

②义原之间的相关度

义原的相关度由义原的相似度及其关联度共同决定，表示为：

$$P\_rel(p_1,p_2)=s_1\times sim(p_1,p_2)+s_2\times P\_ass(p_1,p_2) \tag{4}$$

其中，s_1 与 s_2 为动态分配权值，其和为 1。

③义项(概念)之间的关联度

每个词语可能有几个义项，而义项是通过义原来描述的，故而义项的关联度要从义原的相关度上来计算，而词语的关联度则是从义项的关联度上来计算。

$$C\_ass(c_1,c_2)=\max P\_rel(p_i,p_j)$$

式中，p_i、p_j 分别是 c_1、c_2 中的解释义原。其中 $i\leqslant size(c_1)$，$j\leqslant size(c_2)$。

④词语之间的关联度

每个词语可能有几个义项，故而可以将词语之间的关联度表示为：

$$ass(w_1,w_2)=\max C\_ass(c_i,c_j) \tag{5}$$

式中，c_i，c_j 分别表示 w_1 的第 i 个概念和 w_2 的第 j 个概念。

(3)实例因素

实例因素模型即义项的实例单词的集合，实例因素对相关度的影响[3]：

$$E\_H(e_1,e_2)=\max H(p_{ei},p_j)(1\leqslant i,j\leqslant 2) \tag{6}$$

其中，p_{ei} 为第 i 个义项的实例单词集合的任意一个词的义项，用 p_i 的实例中词的义项与 p_j 计算相似度，取最大值。

(4)词语的相关度计算

词语的相关度就是将语义相似度与相关度结合起来，同时考虑实例因素，共同构成词语的相关度：

$$W\_rel(w_1,w_2)=\max\{q_1\times sim(w_1,w_2)+q_2\times ass(w_1,w_2)+q_3\times E\_H(w_1,w_2)\} \tag{7}$$

式中 $q_1+q_2+q_3=1$，若式(7)的第 3 项值为 0，这时应把 q_3 的值按比例分配给 q_1、q_2。

2.3 实验结果与讨论

本实验的数据来源于知网的数据文件，实验中所设置的计算相似度的参数与参考文献[3]和参考文献[7]中是一致的，所以存在可比性。此外，其他一些参数是随程序自动调整使得结果达到最佳效果。

关于第一义原的改进，通过与参考文献[7]的实验进行对比，结果如表 2 所示。

表 2 词语相似度实验结果

词语 1	词语 2	相似度(参考文献[7])	相似度(本文方法)
男人	女人	0.833	0.693 33
男人	父亲	1.0	0.408 88
男人	和尚	0.833	0.541 79
男人	经理	0.657	0.352 54
男人	高兴	0.013	0.065 39
男人	收音机	0.164	0.410 45
中国	联合国	0.136	0.231 68
医生	患者	0.574	0.472 36
美丽	丑陋	0.722	0.559 99
中国	美国	0.94	0.551 11
跑	跳	0.01	0.214 85
发明	创造	0.615	0.752 00
珍宝	宝石	0.13	0.309 72
粉红	深红	0.074	0.053 92
出兵	出征	0.105	0.128 71
青山	苍山	0.6	0.352 54
香蕉	苹果	1.0	0.525 79

从表 2 可知，“中国”和“美国”在参考文献[7]中的相似度特别高。主要是它用其距离的倒数作为其第一义原，会出现分类不明确的情况，本文采用的 S 型激活函数所划分的区域，分类比线性划分精确合理，所计算值也更合理。“男人”和“父亲”的相似度为 1，“香蕉”和“苹果”也为 1，显然太过粗糙，这种划分分类的方法确实存在着许多缺陷，且算出的值在客观事实之外，本文通过修改第一义原的定义和计算，所得出的相似度分别为 0.408 88 和 0.525 797，相比而言更合理。

上述实验都是同种词性的相似度，而相似只是相关的一个方面，故而进行下面实验，进一步量化同种词性和不同词性之间的相关度，通过对比参考文献[3]的结果

表 3 词语相关度计算的实验结果

词语对	词语 1	词语 2	相似度(参考文献[7])	相关度(参考文献[3])	相似度(本文方法)	相关度(本文方法)
1	削	苹果	0.074 074	0.103 927	0.069 565	0.065 391
2	削	刀	0.074 074	0.200 000	0.069 565	0.058 880
3	削	皮	0.117 647	0.082 353	0.101 333	0.096 533
4	苹果	刀	0.285 714	0.255 671	0.358 400	0.354 212
5	苹果	皮	0.285 714	0.394 121	0.358 400	0.354 212
6	面包	苹果	0.186 047	0.695 652	0.400 000	0.384 695
7	面包	巧克力	1.000 000	1.000 000	0.533 333	0.508 290
8	面包	报纸	0.242 421	0.272 000	0.266 666	0.261 472
9	吃	面包	0.074 074	0.200 000	0.069 565	0.309 565
10	吃	报纸	0.074 074	0.099 129	0.069 565	0.224 347

进行说明。结果如表 3 所示。

由表 3 可以看出,用参考文献[7]所述方法算出的相似度比较粗糙,例如面包和报纸的相似度比面包和苹果的相似度还要高,这显然不太合理,在义原树中,仅仅考虑语义距离,确实“面包”和“报纸”的距离更近,分析发现,这是因为没有考虑义原关联度原因导致的,而本文计算出来的结果对比参考文献[7]和参考文献[3],结果更合理些。

在参考文献[3]的结果中,“面包”和“巧克力”的相关度为 1,这显然与事实不符,通常认为相关度为 1 是完全相关,趋于同一个事物,虽然这两个词语同属于“食品”范畴,关联度方面确实很大,可是相似度方面却相差甚远,因此其相关度值不可能为 1。此外,对事物的看法倾向于一个动宾方式,“削”和“皮”与“削”和“刀”,后者的搭配中表明用“刀”进行“削”,但是也存在用别的东西来“削”,而“削皮”这个搭配在人的直观认知中应该更加相关,故而“削”和“皮”的相关度应该更甚于“削”和“刀”,在本文方法中前者为 0.096 533,后者为 0.058 880,也符合习惯使用上对相关度的主观判断。另外经分析可以看出,本文方法计算出来的数值都会偏小一些,且不会出现极端值问题,比较平稳,从整体上改进了参考文献[3]中的实验结果。

实验所存在的不足是结果对比不够明显,只是改进了偏差比较大的结果,其原因有两方面,一是对于相关度的度量确实是一个比较主观的做法,且目前没有基于统计的相关度的判断标准,因此很难从微观上细小地区分方法的优劣;其次,知网本身有待进一步完善和补充外,通过义原的相似度(相对稀疏的层次结构)来反映大量词语之间的相似度(相对密集)的方法本身是否存在一定的上限还需要进一步深入研究,且许多词语的编撰的定义项存在着一些不完整的方面。

本实验通过自适应的参数来进行调整,没有固定权值,考虑到的是动词间、名词间以及名词之间和动词间,其所侧重的因素不同,如名词之间的相关度计算,相似度占的比重更大,而在动词和名词间,相似度比重应该较小,关联度应占更大的比重,这样才更加合理,因此,自动调整好各参数,偏向各自比较侧重的因素,以便获得更好的效果。

词语的语义相关度研究在国内并不多,本文以知网为知识库,在参考文献[3]的基础上改进算法模型,以此提出的相关度模型所得出的结果比较符合人类主观上对相关度的认识。

今后的工作主要是将此词语相关度模型应用到数据空间中数据源内容关联性的发现机制中去,提出一个基于语义模式匹配的相关性匹配策略,以本文中的词语相关度模型为依托,从而发现数据空间内部的各种数据源的联系性。

参考文献

[1] 李玉坤,孟小峰,张相於.数据空间技术研究[J].软件学报,2008,19(8):2018-2031.

[2] Hua Yu,Jiang Hong,Zhu Yifeng,et al.Smart Store:a new metadata organization paradigm with metadata semantic-awareness for next-generation file systems[C].University of Nebraska-Liclon,Computer Science and Engineering,2008.

[3] 许云,樊孝忠,张锋.基于知网的语义相关度计算[J].北京理工大学学报,2005,25(5):411-414.

[4] 王红玲,吕强,徐瑞.一种基于知网的中文语义相关度计算模型[C].苏州:第三届全国信息检索与内容安全学术会议,2007.

[5] 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):101-107.

[6] 李素建.基于语义计算的语句相关度研究[J].计算机工程与应用,2002,38(7):75-76.

[7] 刘群,李素健.基于《知网》的词汇语义相似度计算[C].台北:第三届汉语词汇语义学研讨会,2002.

(收稿日期:2011-12-27)

作者简介:

曾淑琴,女,1987年生,硕士研究生,主要研究方向:数据库应用技术。

吴扬扬,女,1957年生,教授,硕士生导师,主要研究方向:数据库技术和数据挖掘。