一种高稳定性词汇共现模型

乔亚男，齐勇，侯迪
（西安交通大学计算机科学与技术系，710049，西安）

摘要：针对传统词汇共现模型存在的缺乏理论基础和稳定性欠佳等问题，提出了一种基于项场的高稳定性词汇共现模型.借鉴经典物理学中场的概念给出了项场的定义，其中项是语言的基本单位，是概念的抽象描述，而项场则是项在文档中的影响范围.在此基础上，引入量子场论将项与项的相关度类比为项场的叠加，由此给出了项与项之间距离和相关度的函数关系，并用其建立了词汇共现模型.实验结果证明，在小距离的情况下，所提模型中项的相关度大体呈常数，具有一定的窗口内稳定性，而同范畴的项对相关度振幅只有对照模型中最小振幅的 26%，表明它具有较好的数据集稳定性.

关键词：项场；词汇共现；窗口内稳定性；数据集稳定性
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2009)06-0024-04

A Highly Stable Term Co-Occurrence Model

QIAO Yanan, QI Yong, HOU Di
(Department of Computer Science and Technology, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: To address the issues that traditional term co-occurrence models are lack of theoretical basis and poor stabile, a highly stable term co-occurrence model based on term field is proposed. The model uses the concept of field in classical physics for reference to define the term field (terms are the basic units of language, which describe the abstract concepts, and the term field is the area affected by a term in document). Based on the definition, the model regards correlation as a superposition of term fields, and gets the functional relations of terms correlation and the distances between terms. Experimental results show that the terms correlation in this model is almost a constant for small distances and stable enough in window. While the correlation amplitude of the terms in same category is only 26% of the best result obtained with other models, which means the model is stable enough in dataset.

Keywords: term field; term co-occurrence; in-window stability; in-dataset stability

词汇共现模型(term co-occurrence model)可以对词与词之间的相关性进行量化比较，因此被广泛应用于信息检索、文本聚类等研究领域[1-3].

最早提出的词汇共现模型被称为常数模型[4]，它简单地认为同一个共现窗口内的 2 个词无论位置如何，其相关度都是恒定的，该模型无疑将问题过于简化.近年来，在常数模型的基础上提出了多种改进模型，包括多项式递减模型[5]、指数递减模型[6]、吸引和排斥模型(指数模型的变形)[7]等.这些改进模型都不同程度地提高了 2 个词相关度计算的精确度，但都存在着 3 个主要的问题：①对于大部分改进模型来说，距离和相关度之间的函数关系都是作者的经验推测，并没有太多的相关理论基础；②它们是基于递减模型的，即 2 个词的距离越大，相关度越小，这种相关度缺乏必要的稳定性，而在语言学研究中发现，当 2 个词距离比较小的时候，相关度与距离

收稿日期：2008-09-05. 作者简介：乔亚男(1979—)，男，博士生；齐勇(联系人)，男，教授，博士生导师. 基金项目：国家高技术研究发展计划资助项目(2006AA01Z101)；教育部高等学校博士学科点专项科研基金资助项目(20060698018).

的变化基本无关;③改进模型即使针对同一个数据集,同一范畴词对的相关度值也很容易受到具体选词的影响.

针对上述问题,通过借鉴经典物理学中场的概念和量子场论,本文提出了一种基于项场的高稳定性词汇共现模型,即利用基于项与项之间项场叠加构造出的相关度函数来描述词汇之间的关系.

1　项　场

项(term)是词汇共现模型研究中最基础的概念.文本主要是用它所含有的基本语言单位(字、词、词组或短语等)来表示内容特征,这些基本的语言单位统称为项.由此可见,项和一般意义上的词(word)既有联系又有区别,项偏重于描述抽象的概念,是跨语言的,一个项在一种语言中可能是一个词,而在另外一种语言中可能是一个词组甚至是一个短语.所以,词汇共现、词汇共现模型称为项共现、项共现模型更为准确.

在实际应用中,词语的上下文是词语的语言环境,所以核心词语上下文有效范围的确定可以转化为求解词语上下文中各个位置对核心词语信息量贡献大小的问题.文献[5]将核心词及其上下文形式化为一个符号信息系统,信源的先验不确定性(熵)就是核心词的统计不确定性,信宿的后验不确定性就是在已知一位置上下文情况下的不确定性,二者之差即为相对于已知一上下文位置的情况下条件熵的信息增益,以此可确定各位置的信息量.文献[5]还对中英文 2 种语言进行了实验,对计算结果进行了 3 次曲线拟合,最后得到了中英文上下文位置信息量函数.该函数也可以理解为:文档中的每一个项都有一个影响范围,影响的强度遵循上下文位置信息量函数.这个影响范围的概念和电荷产生的电场或磁荷产生的磁场在某种程度上十分类似,我们称它为项场.

定义 1　项在文档中的影响范围称为项场.文档中的每一个项都有自己的项场,同类项的项场互相叠加,非同类项的项场互不干扰.

任何一个项的项场都对其他的项产生影响.当考虑一个项的项场时,相对其他项,这个项称为源项,而被作用的项称为汇项.类似于磁场、电场,项场也有描述场强分布状态的函数,即项场分布函数.

定义 2　描述项场相对场强分布状态的函数称为项场分布函数.对于源项 t_s 和汇项 t_c,项场分布函数为

$$E_s = F(D_{sc}, P_s) \quad 0 \leqslant E_s \leqslant 1$$

式中:P_s 为 t_s 项场的最大作用范围,与 t_s 的词性和语义相关;D_{sc} 是 t_s 与 t_c 之间的距离(间隔的单词数).将文献[5]中上下文位置信息量函数经过适当的变形和归一化(主要为了把相关度限制在 0 和 1 之间),就可以作为项场分布函数.

根据项场分布函数的定义域(其场强值有意义的区域),项场有一个有限的影响范围,称为项场的作用域.

本文中的论述及实验以文献[5]提到的上下文位置信息量函数为基础.

2　基于项场的词汇共现模型

项之间的距离与相关度的函数关系是词汇共现模型研究的根本问题,项对间关系准确量化的关键就在于相关度函数的确定.

在基于项场的词汇共现模型中,项对相关度的定义借鉴了量子场论中有关场与场相互作用的一些理论.根据量子场论,场是物质存在的一种基本形式,这种形式的主要特征在于场是弥散于全空间的.目前,人类所知的自然界有 4 种基本的相互作用:强作用、电磁作用、弱作用、引力作用,这 4 种基本作用都可以归结为相应的场对场的作用.将量子场论的一些概念与词汇共现模型相类比可以发现很多相似之处,因为量子是一个不可分割的基本个体,量子与量子之间的作用是离散且非连续的,而文本中的项之间的作用也是离散的,所以项可以类比为一个量子.项是组成文本的基本单位,因此项之间的距离必然是一个整数.

每一个项的项场理论上都是弥散于整个文本的,当项场的作用范围超出其最大作用范围时,可以视它为 0.

本文将量子场论中基本作用力是场对场作用的概念引入到了项场之中,得到 a 项与 b 项之间的相互作用即为 a 项场与 b 项场之间的相互作用,并将这种相互作用定义为项对的相关度.

定义 3　对于 2 个项组成的项对,这 2 个项的项场作用域重叠部分(即相互作用部分)的平均项场强度被定义为 2 个项的相关度.也就是说,对于项对 $[a,b]$,这 2 个项的项场强度分布函数分别为 $F_a(x)$ 和 $F_b(x)$,如果 d 为 2 个项场作用域的重叠部分,x 为 d 中的一个特定位置,$|d|$ 为 d 的长度,则项对 $[a,b]$ 的相关度为

$$r([a,b]) = \frac{1}{|2d|}\int_{x\in d}(F_a(x)+F_b(x))\mathrm{d}x \tag{1}$$

在文献[5]上下文位置信息量函数中,中文的定义域为[-13,-1]∪[1,14],英文的为[-18,-1]∪[1,18],而在核心词位置(0位置)没有定义,此时要计算项对的相关度显然需要补充这个位置的定义,也就是定义一个项的项场在其自身处的场强.本文采用类似曲线拟合的方法,将项场强度分布函数$F(x)$在$x=0$时的函数值定义为该项的自身场强.

如图1所示,考虑"horrible"(a项)和"dog"(b项)2个项,设项场的作用域宽度为4个单词,则a项的项场和b项的项场的重叠部分为自"my"至"yesterday"宽度为6个单词的区域.为了简化问题,设项场强度$F_a(x)$和$F_b(x)$是线性递减的,即项场分布函数$F(x)=1-\frac{x}{k+1}$,其中k为项场作用域的半径,则a项和b项的相关度为

$$r([a,b])=\frac{1}{12}\sum_{x=\text{"my"}}^{\text{"yesterday"}}(F_a(x)+F_b(x))=$$

$$\frac{1}{12}((0.8+0.2)+(1.0+0.4)+$$

$$(0.8+0.6)+(0.6+0.8)+$$

$$(0.4+1.0)+(0.2+0.8))\approx 0.633\ 3$$

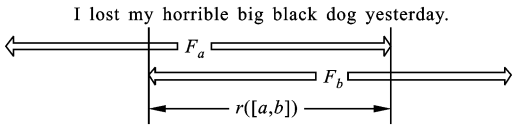

图1 项场模型中项对相关度的定义

3 实验和结果分析

利用文献[5]中上下文位置信息量的计算结果,根据项对相关度定义,计算出不同距离下项对的相关度,并与传统的常数模型、线性模型和指数模型做比较,结果如图2所示(对相关度进行了适当的归一化处理,以便于比较).

从图2可以清楚地看到,在项距(距离)较小的情况下(约为1~17个单词),项对的相关度整体变化是很小的,大体呈常数(0.85~0.98),甚至最大值并不出现在距离最小的地方(约在9的位置),此后随着项距的增加,相关度急剧下降,在项对的叠加项场的作用域边界降至最小值.

实验结果在理论上解释了概述中提及的情况.在项距较小的情况下,项对的相关度应当与项距基本无关,具有较好的稳定性.项距一旦跨越特定的临界值(约为17),相关度与项距呈线性关系.我们发现:特定临界值与英文的平均句子长度比较吻合;在同一个句子中,项对的相关度基本一致;在不同句子中,项对的相关度随项距增加而减小.这个结果相对于传统模型,更符合常理.

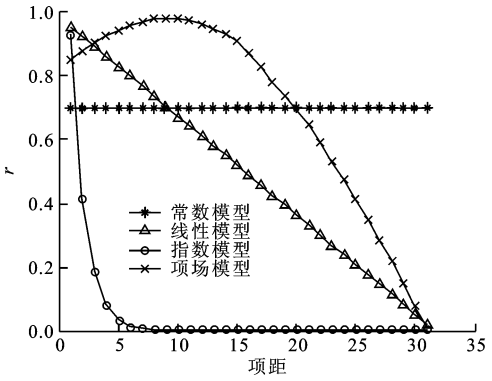

图2 项对相关度和项距的关系

此外,从语言学的角度来看,相关度的最大值并不出现在项距最小的地方,即使对语料进行统计分析之前剔除了最常见的停用词,与核心词距离最小的词语也多半是研究意义不大的连词、副词、助词等,真正联系紧密的共现词一般距核心词不太近也不太远.

项场模型的这些特性也有利于提高同范畴项对相关度计算的稳定性,在对实际数据集进行的另一组实验中也证明了这一点.在本实验中,数据集是本实验室编纂的《人民日报》英文版数据集,共有71 158篇文章,纯文本大小为328 MB,涉及1999年~2007年刊登的大部分有关时事、经济和科技方面的新闻报道.我们还挑选了《人民日报》中经常出现的4个城市名(Beijing,Shanghai,Nanjing,Guangzhou)穷举构造了6组项对作为测试用例(见表1).对这6组项对分别进行计算,得出基于4种词汇共现模型的平均相关度,如图3所示.

表1 项对实例

实例序号	项对	
	项a	项b
1	Beijing	Shanghai
2	Beijing	Nanjing
3	Nanjing	Shanghai
4	Shanghai	Guangzhou
5	Beijing	Guangzhou
6	Nanjing	Guangzhou

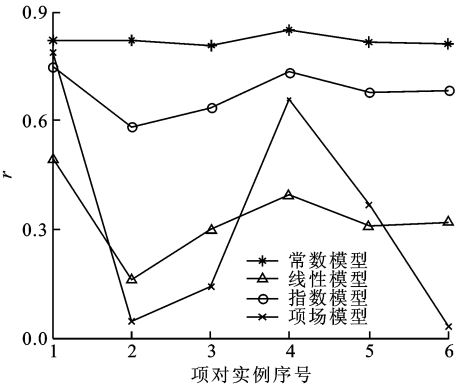

图 3　不同词汇共现模型对于同范畴项对相关度计算的稳定性比较

挑选的这 4 个项均为中国的大城市名，由常理判断穷举构造的 6 组项对的相关度不应该相差太大. 从图 3 中可以明显看出，常数模型和指数模型计算出的相关度稳定性很差，线性模型稳定性略好，但相关度的振幅(0.169 6)仍然是项场模型相关度振幅(0.043 3)的 4 倍.

4　结　论

基于项场的词汇共现模型将项与项的相关度视为项场的叠加，对模型中项距和相关度的函数关系在理论上进行了说明，解释了在项距比较小的情况下项对相关度基本不变的事实. 实验所得出的满足“小距离”特性的临界值与实际的平均句子长度比较吻合，证明了基于项场的词汇共现模型的合理性. 在同范畴项对相关度稳定性对比实验中，项场模型的优势也比较明显，与传统模型相比，更适用于在大规模文本数据集中测算词汇相关度.

在对精确度要求不高的场合下，传统的共现窗口模型的计算效率比较高，而本文的与“小距离”临界值有关的结论也适用于确定词汇共现窗口的大小. 文献[5]得出的上下文有效范围为[－16，＋13](英文)，本文得到的结果为[－17，＋17]，这个结果的合理性尚需理论和实验进一步证实.

参考文献：

[1] BEIGBEDER M，MERCIER A. An information retrieval model using the fuzzy proximity degree of term occurrences [C]// Proceedings of the 2005 ACM Symposium on Applied Computing. New York，USA：ACM，2005：1018-1022.

[2] PETKOVA D，CROFT W B. Proximity-based document representation for named entity retrieval[C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York，USA：ACM，2007：731-740.

[3] RASOLOFO Y，SAVOY J. Term proximity scoring for keyword-based retrieval systems [C]// Proceedings of 25th European Conference on IR Research. Berlin，Germany：Springer，2003：207-218.

[4] YAROWSKY A D. One sense per collocation[C]// Proceedings of the ARPA Human Language Technology Workshop. Stroudsburg，PA，USA：Association for Computational Linguistics，1993：266-271.

[5] 鲁松，白硕. 自然语言处理中词语上下文有效范围的定量描述 [J]. 计算机学报，2001，24(7)：742-747.
LU Song，BAI Shuo. Quantitative analysis of context field in natural language processing [J]. Chinese Journal of Computers，2001，24(7)：742-747.

[6] GAO Jianfeng，ZHOU Ming，NIE Jianyun，et al. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations [C]// Proceedings of the 25th Annual International ACM SIGIR. New York，USA：ACM，2002：183-190

[7] 郭锋，李绍滋，周昌乐，等. 基于词汇吸引与排斥模型的共现词提取 [J]. 中文信息学报，2004，18(6)：16-22.
GUO Feng，LI Shaozi，ZHOU Changle，et al. Co-occurrence word retrieval based on the lexical attraction and repulsion model [J]. Journal of Chinese Information Processing，2004，18(6)：16-22.

（编辑　苗凌）