Confirmation Number: 70

Submission Passcode: 70X-D3H4J2H6A6

基于偏向相似性的自然语言关联和聚类研究*

陈振宁[1]，陈振宇[2]

(1.浙江大学，浙江省杭州市，310058；2.复旦大学，上海市，200433)

摘　要：聚类按关联进行分类，关联和聚类分析的基础是相似性计算。通常相似性是指绝对相似性，具有对称性。但自然语言研究中发现大部分规律都是偏向的，具有不对称性，需要用偏向的思路来考察不对称的关联和聚类策略：以类似条件概率的概率蕴涵指标来描写特征间的不对称关联，并在此基础上定义优势关系、紧密关系、控制中心、中途岛等关联特性；基于偏向相似性的聚类策略，从而能更好地处理语言本体研究中的“假性孤立点”、数据稀疏问题和家族象似性类型的聚类。

关键词：不对称性，条件概率，关联，聚类

中图分类号：TP391　　文献标识码：A

Cluster and Association Analysis of Natural Languages Based on Inclined Similarity Measures

Chen Zhenning[1], Chen Zhenyu[2]

(1. Zhejiang University, Hangzhou, Zhejiang, 310058, China; 2. Fudan University, Shanghai, 200433, China)

Abstract: Cluster analysis is the task of grouping a set of objects by associations of these objects. The diameters of cluster and association analysis is similarity measures. The absolute similarity which is symmetry property are commonly used in analysis. But most rules found in natural languages is inclined and have asymmetrical performs. We will give some asymmetrical associations: There is a parameter, Probability Entailment, resembles the conditional probability to represent the asymmetrical associations among features, as well as Domination Relation, Tight Relation, Control Center, Midway island. A strategy for cluster based on inclined similarity measures will be put forward in this article for dealing with problems in noumenal research on languages: False isolated points, Data Sparsity and Family iconicity.

Keyword: Asymmetry, conditional probability, association, Cluster

1 引言

聚类分析是在无监督情况下将对象按一组特征进行分类的统计方法：按研究问题要求确定特征和对象“相似性”或“距离”的计算方法，并根据计算得到的距离或相似性，按一定策略聚类。[1][2][3]

针对自然语言系统的聚类研究，目前为止主要的应用领域是对自然语篇的聚类，聚类后的类型主要和语篇的题材、话题、语体、风格有关[2]。从语言学的角度来看，就是对语篇级别的“语用、修辞”领域的研究。而在句法语义这些语言本体研究的“重镇”，相关研究还

* **收稿日期：**　　**定稿日期：**

基金项目：教育部人文社会科学规划基金项目“现代汉语句法与语义计算研究”（13YJA740005）

作者简介：陈振宁（1977——），女，博士研究生，主要研究方向为计算语言学；陈振宇（1968——），男，副教授，主要研究方向为汉语句法语义。

非常有限。

在语言本体尤其是句法语义研究中，研究者常常遇到这样一个问题：**研究的特征和对象表现出很强的“偏向”，偏向的规律也往往是语言本体研究的重点。**但是现有相似性和距离计算有一个基本的前提条件：相似性和距离都是绝对的，互成反比：绝对距离越小越相似，越大越不相似并且都是对称双向的[1]，如公式 1。

$$c_{ij} = \frac{k}{d_{ij}} \quad k\text{为根据实际问题确认的系数} \qquad \text{公式 1}$$

其中c_{ij}为任意 i，j 间的相似性，d_{ij}为任意 i、j 间的距离[1]，且：

A. $d_{ij} = d_{ji}$ ；

B. $c_{ij} = c_{ji}$ ；

因此，本文将针对句法语义研究的实际需要，设计偏向的相似性指标，挖掘特征间的偏向性规律，并用偏向的策略进行聚类。

2 偏向相似性/概率蕴涵、偏向聚类策略

2.1 偏向相似性和概率蕴涵

语言类型学研究中基本语序和性词缀语序的共性几乎都是偏向的，例如[5]：

共性：如果性标记采取前缀，那么基本语序就是 VO；且如果基本语序为 OV，那么性标记采取后缀。

也就是说：前缀性标记对 VO 语序的选择或者主要语序 OV 对后缀性标记的选择都是偏向的，反之不然。这一共性描写基于如下调查表 1[6]：

表 1 基本语序和性标记语序的共现频次表

	性前缀	性后缀
VO	20	30
OV	0	50

某一特征能在越多的语言样本中贡献，说明人们的心理上越容易认识到它们是“相似的”，那么我们可以简单地把特征两两直接共现的频次视作绝对相似性，那么这一调查可以绘制成一个语图 1：

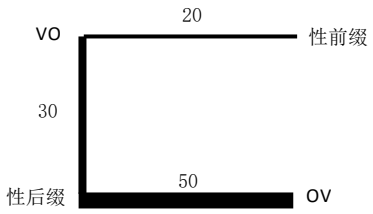


图 1 基本语序和性标记图

从对称的绝对距离/相似性来看，如果用均值（20+30+50）/4=25 作为聚类的阈值，就分为两类：{VO，OV，性后缀}，{性前缀}，性前缀是“孤立点”。

在这种对称的绝对距离/相似中隐含了不对称的“相对远近/距离/相似”：

VO：相对“性后缀”最近，“性前缀”较远，“OV”完全无关；

OV：相对“性后缀”最近，和其他无关；

性前缀：相对“VO”最近，和其他无关；

性后缀：相对“OV”最近，“VO”较远，“性前缀”无关。

这种“相对远近/距离/相似”是一致不对称的偏向相似。我们可设置**偏向相似性指标 P** 来量化它：

$$P_{ij} = \frac{c_{ij}}{\sum_{a=1}^{n} c_{ia}}, \quad n\text{为}i\text{连接的所有点的总数} \qquad \text{公式 2}$$

可以看出，**这一指标在本质上和条件概率是等价的**：i 确立的条件下，i 和 j 共现的概率。条件概率本身即有 P(i/j)≠P(j/i)的性质，正是偏向的。在特殊的情况下，这种偏向性等同于逻辑中的“蕴涵（→）”关系。因此本文将这一指标称之为**“概率蕴涵”**。

表 2 基本语序和性标记语序的概率蕴涵

性前缀/VO	20/50=0.4	VO/性前缀	20/20=1
性后缀/VO	30/50=0.6	VO/性后缀	30/80=0.375
性前缀/OV	0	OV/性前缀	0
性后缀/OV	50/50=1	OV/性后缀	50/80=0.625

如表 2，有两个最强的偏向相似（**逻辑蕴涵**）：性前缀对 VO、OV 对性后缀的选择性，这两个共性已被前人归纳为“蕴涵共性”。

同时还有两个**优势**的偏向相似未被归纳：VO 对性后缀的选择性、性后缀对 OV 的选择性。这里两个优势偏向相似看来似乎有一定矛盾，但在“谁先出现从而影响谁的出现”这种历时考察中是有意义的。

2.2 偏向聚类策略

按偏向相似性，我们可以看到，“性后缀”和“OV”的偏向相似“凑巧”都以对方为“相对最近”的目标，因此它们自然可以合为一类{OV，性后缀}，VO 则“偏向”选择了“性后缀”，被合并进来{VO，OV，性后缀}，同理，“性前缀”虽然绝对距离更远一点，但它并不是“完全被孤立没有联系的点”，它一样偏向选择了“VO”，再次被拉进来{性前缀，VO，OV，性后缀}。

这就是**偏向聚类策略：每个对象寻找自己的“相对最近”，能找到（即没有出现“绝对相似性直接为零”或“绝对距离为‘不能相通’的无穷”）就有其类的归属**。

这种不对称的偏向聚类策略建立在这样的观察上：1、除非真正得出某种“距离/相似性为零”，就没有真正意义上的“孤立点”，绝对距离远的点总是在自己的范围内搜索相对自己最近的对象进行“联系”，是一种**“伪孤立点”**；2、在两个以上对象共存的系统（或考察范围）里，两个对象之间纯粹的“两两联系”总是会受到其他对象的影响，绝对距离的对称性一旦放入多个对象互相有所联系的系统中，就肯定会衍生出**相对不对称**的问题，各点的地位就会变化。

“性前缀”正是一个“伪孤立点”，事实上性前缀对 VO 有强烈的依附性，已经被归纳为重要的语言学蕴涵共性，**但按照绝对聚类策略，“性前缀“被直接”踢“了出去成为孤立点，根本没有规律可循，这显然是不符合语言学研究需要的**

另外，似乎出现了一个问题：**最后只有一个类！**

但是，反过来这似乎反而是**绝对聚类策略的问题：不论数据实际情况到底如何，已预先假设至少分成 2 个以上的类，这其实并不符合“无监督”分类的原则。**

事实上，伪孤立点“性前缀”用偏向性指标概率蕴涵来衡量的话，对 VO 的“依赖性”极强，也就是说这个绝对距离“远”的点恰恰通过 VO 对整个类的依附性非常强烈。似乎荒谬。但在现实世界中其实未必不合理。主要有两个原因：

①“数据稀疏”问题：自然语言是一个数据稀疏的系统[2]，这意味着一些规律在语篇中显现的数据有限，从绝对性来看这些规律是不强的，相对来看仍旧是很强的规律。

②人类认知中“家族像似性”造成的类型划分：各成员间的共同特征可以差别很大甚至完全没有一个相同的共性，如语词 Game 所指的对象，全部合在一起会几乎没有一个能说得清的共性。它们就是通过这样一种“拍皮球相对和荡秋千接近、打球相对和拍皮球接近、职业球类竞赛相对和打球接近……”的关系“串联”起来的。[4]

3 语言学实例研究

3.1 一系列蕴涵共性数据的再分析

蕴涵共性研究中语言学家还调查了一系列基本语序和语法标记的共现情况，如表 3[7]：

表 3 基本语序和多种前后缀形式标记的概率蕴涵

CC-map	概率蕴涵			
VO 17 前缀 7 21 后缀 55 OV **图 2 基本语序和否定标记图**	P（前|VO）	**0.708**	P（VO|前）	0.447
	P（后|VO）	0.292	P（VO|后）	0.113
	P（前|OV）	0.276	P（OV|前）	0.553
	P（后|OV）	**0.724**	P（OV|后）	**0.887**
VO 19 前缀 25 35 后缀 21 OV 图 3 基本语序和所有者一致性标记图	P（前|VO）	0.432	P（VO|前）	0.352
	P（后|VO）	0.568	P（VO|后）	0.543
	P（前|OV）	0.625	P（OV|前）	0.648
	P（后|OV）	0.375	P（OV|后）	0.457
VO 26 前缀 15 19 后缀 40 OV 图 4 基本语序和主语一致性标记图	P（前|VO）	0.634	P（VO|前）	0.578
	P（后|VO）	0.366	P（VO|后）	0.273
	P（前|OV）	0.322	P（OV|前）	0.422
	P（后|OV）	0.678	P（OV|后）	0.727
VO 10 前缀 32 32 后缀 26 OV 图 5 基本语序和宾语一致性标记图	P（前|VO）	0.238	P（VO|前）	0.238
	P（后|VO）	0.762	P（VO|后）	0.552
	P（前|OV）	0.552	P（OV|前）	0.762
	P（后|OV）	0.448	P（OV|后）	0.448

表 3 中绝对相似高的规律只有：“否定后缀-OV”。因为从偏向概率来看，它们正好彼此“偏向选择”对方。我们可以将其看作一种“**紧密关系**”。

但从偏向选择来看，还有：“否定前缀-VO”、“主语一致后缀-OV”、“宾语一致后缀-VO”和“宾语一致前缀-OV”四个“**优势关系**”。

如果用绝对相似的聚类策略来分类，这些重要共性中的成分往往会被“孤立”出去，如按绝对共性对基本语序和否定标记分类，VO 就会成为孤立点难以参与任何规律。

3.2 从成都话语气词“哇”考察“传疑/传信”连续统

疑问语气研究中“传疑/传信”是一个连续统：1）真性疑问，说话人语用目的是“对疑问内容不确定并要求对方给出答案”；2）反问，说话人“无疑而问不要求对方给出答案”；3）介于真性问和反问之间的“猜测、求证、求认同”等，对内容的确定性较弱（猜测、求证），可能要求答案也可能不要求答案。这样，我们可以从两个维度来考察汉语的“疑问”范畴的句子或标记：

确定性：不确定、弱确定、确定；

求答性：求答、弱求答、不求答。

乍一看这两个维度是一一对应的完全可以合为一个。但在真实语料中调查会发现实际上是有参差的。在确定和弱确定的情况下，对“求答”都有多种选择。表 4 是成都话常用于疑问形式的语气词“哇”的部分研究数据[1]：

和语气词“哇”有关的疑问形式中，真正双向概率都大的紧密关系是“不确定-求答”。弱求答、不求答对确定的偏向选择性虽然强，确定却可以在两者间游弋。

也就是说，我们在不确定的时候，当然要求给出答案，反之亦然。但是在确定的时候，还是有较大可能希望对方回话，因为交际的顺利进行为往往依赖于“互动”，说话人讲的确定情况也需要得到听话人的反应（认同或反驳）才能更好地将交际进行下去。

[1] 以五部成都话方言小说（约 49 万字）构成语料库（成都国家开放大学副教授杜克华负责建立），进行全文搜索，共检得“哇”字句 144 句，由陈振宇提供。

表 4 成都话“哇”字句中确定性和求答性各特征的概率蕴涵

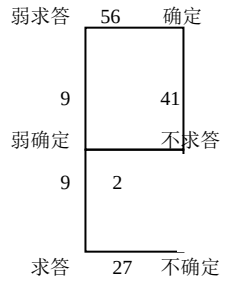
CC-map	概率蕴涵			
	P（确定|求答）	0	P（求答|确定）	0
	P（确定|弱求答）	0.862	P（弱求答|确定）	0.577
	P（确定|不求答）	0.953	P（不求答|确定）	0.423
	P（弱确定|求答）	0.25	P（求答|弱确定）	0.45
	P（弱确定|弱求答）	0.138	P（弱求答|弱确定）	0.45
	P（弱确定|不求答）	0.047	P（不求答|弱确定）	0.1
	P（不确定|求答）	0.75	P（求答|不确定）	1
	P（不确定|弱求答）	0	P（弱求答|不确定）	0
	P（不确定|不求答）	0	P（不求答|不确定）	0

图 1 确定性和求答性特征图

按偏向策略给确定性和求答性 6 个特征聚类：

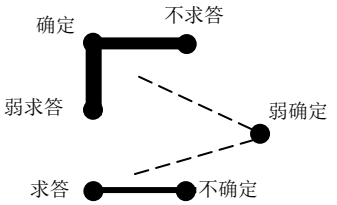


图 7 确定性和求答性特征聚类图

如果歧义情况单独“拎”出来的话，偏向聚类可分 3 类：

{不求答，确定，弱求答}、{求答，不确定}、{弱确定}。事实上，如果我们用绝对聚类，并简单地以均值为阈值，低于阈值（(56+41+9+9+2+27)/6=24）就分开，也可以得到同样的 3 类。

但是，从偏向策略来看，可以看出“弱确定”的“相对相似性”有歧义，即它即可能和“弱求答”最相似也可能和“求答”最相似。这里的“弱确定”不是简单的“孤立点”，恰恰是一个沟通两大类之间的比较“脆弱”的“**中途岛**”，在系统里其实其中重要的作用。

另外，配合表 4 的特征概率蕴涵，我们还可以注意到：“不求答”和“弱求答”都依附于“确定”，亦即“确定”被多个点依附，具有“**控制中心**”的位置。

3.3 汉语体标记的聚类

考察汉语常用动词和 9 个体标记搭配得到的频次数据[2]。从这些特征在动词中的同现情况，我们将其绘制为 CC-map 语图(图 8)，其算法为“赢多输少”[9]，各边的权重就反映了特征两两之间的绝对相似值。

从绝对相似性考虑，以均值 207.56 为阈值，那么汉语体标记可以分成类：{过,正在/在,着,起来,重叠,了 1,了 2}、{下去}、{了 3}。

偏向聚类策略中“下去”和“了 3”各自按相对“最近”建立关联，“下去”连接上“过”、“了 3”连接上“了 1”，**成为一个类**。如图 9。

尽管偏向策略聚类后所有的点只有一个类，但从图 8 我们可以看出。所有体标记“最主流”的关联基本上都“汇聚”于一个标记“过”，“过”是**控制中心**。这是“控制关系”非常严明的星型网络，是一个有较明确中心的“原型”类型。反观调查数据，汉语的常用动词

[2] 考察 2000 个汉语常用动词[8]和体标记的搭配，如一个动词能搭配多个体标记，则视为多个体标记的一种共现情况。是否能够搭配的判定标准为整理者母语语感，并参考“北京大学 CCL 语料库检索系统（网络版）”的检索情况。除去完全不能搭配体标记的 106 个动词，整理后共得到 9 个体标记共现频次表 163 行，由陈振宇提供。另外，“重叠”是指动词重叠式，这是动词和时间信息有关的一种屈折形态变化。

中，只要可以和多个体标记搭配的[3]，绝大部分都可以搭配“过”。即：“过”包含的某种和时间有关的信息，能够适应绝大多数常用动词，并且和其他体标记相容或对比。

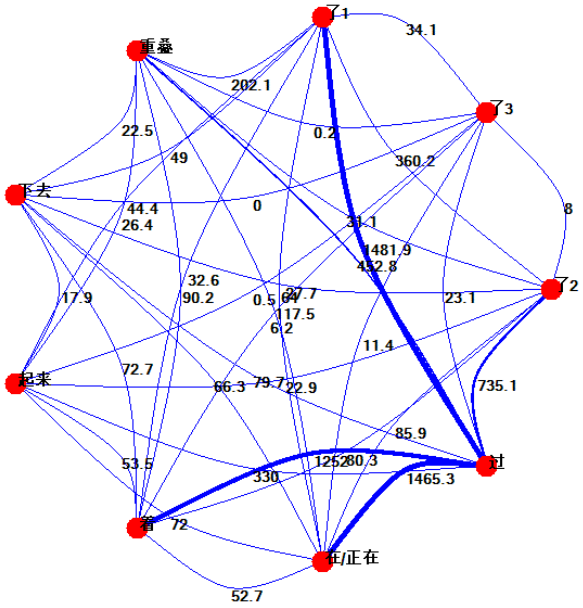


图 8 汉语体标记地图

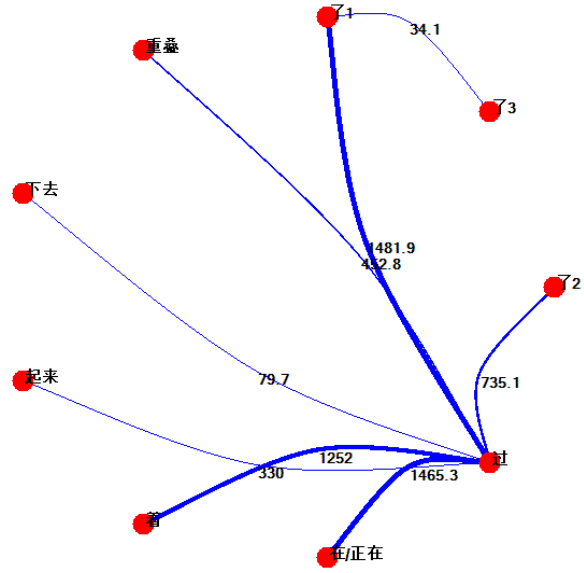


图 9 汉语体标记偏向相似性聚类图

从时间信息相关标记来说，世界语言最常见的时间相关标记区分是“过去/非过去（现在、未来等）”的区分。汉语体标记虽然不是时标记，但也包含可推导出时标记的信息。由此可见，“过”这个经历体标记是汉语体标记中和过去时联系最紧密的，因此成为体标记类的中心。可以对比的是“了 1”，一度也被认为是和过去时联系紧密的一个体标记。但事实上，“了 1”是一个实现体标记，和现在时的关系可能更大[10]。

再看绝对聚类策略中成为鼓励点的“了 3”。语言本体研究中，汉语“助词‘了’”的问

[3] 汉语有一大类动词（“强静态”动词如“是、属于”）基本不和任何体标记搭配，还有极少数动词只能搭配一个体标记。这些动词暂时未纳入考虑范围内。因为如果要考虑的话，就要加入对“无某一特征”这种否定性特征的研究。

题一直是体标记中最复杂的。这个助词在不同场合的功能表现差异较大，一般认为至少可以分为“了 1、了 2”：“了 1”位于句中主要谓词之后，且后面还有宾语或时量动量等成分，是比较纯粹的表示句子时间信息的“体标记”；“了 2”位于句子末尾，即 VO 或“V+时量/动量”之后，一般认为除了能表示时间相关信息外，有较强的语气作用，甚至可能主要是个语气词而不是体标记[10]。

但也有认为汉语的“了”应该三分，即“了 1、了 2、了 3”。这一分化主要依据如下：“了 1”和“了 2”之间明显有一种“中间状态”，即主要谓词本来就是不带宾语的一价谓词（一价动词或形容词），同时也没有时量/动量短语，这样的“了”可视作直接位于主要谓词后面，同时也可视作位于句子末尾，其功能似乎在体标记和语气词之间更加模糊[11]。

由此可见，“了 3”从理论划分上就不可能是“孤立点”，之所以绝对聚类策略会出现这种“误会”，是因为：我们考察的主要是常用动词，那么“了 3”的总体数据量自然偏小，因为一价谓词大多是形容词。同时，具备“了 1、了 2、了 3”搭配能力的一价动词往往是那种可以在语言实际使用中“变价”的动词，总体数量上肯定比较小。但恰恰是在这样的环境里我们才能真的考察“了”三分的关系之所在。而统计数据显示“了 3”和“了 1”关系远比“了 2”紧密，能帮助我们考察“了”功能的语法化关系。即：

如果“了 3”确实已经分化出来[4]，很可能来自“了 1”特殊用法的逐渐语法化。如果“了 3”并未分化，那么它应归于“了 1”而非“了 2”。

还有一个绝对相似性考察下容易成为“孤立点”的“下去”，它也是一个使用环境相对受限的体标记。“下去”使用基本条件是：一个活动的进行过程可以被打断，然后从这个断点在继续“下去”。通常事件情状更容易让人注意到的是起始和结束两个“端点”，进行过程中被打断更特殊一些[10]。因此，“下去”查到的绝对数据也很可能是有限的，会导致其绝对相似性偏小。

表 5　汉语体标记的概率蕴涵表

→		概率	→		概率	→		概率	→		概率
了 1	了 2	0.16	了 2	了 1	0.27	了 3	起来	0.01	起来	了 3	0.00
了 1	了 3	0.02	了 3	了 1	0.41	了 3	下去	0.00	下去	了 3	0.00
了 1	在/正在	0.03	在/正在	了 1	0.03	了 3	重叠	0.00	重叠	了 3	0.00
了 1	着	0.01	着	了 1	0.02	在/正在	着	0.03	着	在/正在	0.03
了 1	过	0.65	过	了 1	0.25	在/正在	过	0.76	过	在/正在	0.25
了 1	起来	0.02	起来	了 1	0.08	在/正在	起来	0.04	起来	在/正在	0.13
了 1	下去	0.02	下去	了 1	0.15	在/正在	下去	0.03	下去	在/正在	0.20
了 1	重叠	0.09	重叠	了 1	0.21	在/正在	重叠	0.06	重叠	在/正在	0.12
了 2	了 3	0.01	了 3	了 2	0.10	着	过	0.76	过	着	0.22
了 2	在/正在	0.06	在/正在	了 2	0.04	着	起来	0.03	起来	着	0.09
了 2	着	0.06	着	了 2	0.05	着	下去	0.04	下去	着	0.22
了 2	过	0.54	过	了 2	0.13	着	重叠	0.05	重叠	着	0.10
了 2	起来	0.02	起来	了 2	0.04	过	起来	0.06	起来	过	0.58
了 2	下去	0.02	下去	了 2	0.08	过	下去	0.01	下去	过	0.24
了 2	重叠	0.02	重叠	了 2	0.03	过	重叠	0.08	重叠	过	0.48
了 3	在/正在	0.14	在/正在	了 3	0.01	起来	下去	0.03	下去	起来	0.05
了 3	着	0.07	着	了 3	0.00	起来	重叠	0.05	重叠	起来	0.03
了 3	过	0.28	过	了 3	0.00	下去	重叠	0.07	重叠	下去	0.02

[4] 究竟是二分还是三分，其实和汉语不同方言“演化”的进程有关。

从概率蕴涵表我们可以看到：

没有一个真正的“紧密关系”，这说明汉语体标记的功能和来源还比较分散，演化线索很可能颇为复杂。

尽管在聚类上几乎所有体标记（“了 3”除外）汇聚于“过”是很明显的，但“过”作为一个“控制中心”的地位还不是最强：“正在/在、着”这两个动态/静态持续体标记对“过”的蕴涵概率超过 0.75，非常强；“了 1、了 2、起来”等对“过”的蕴涵概率只略超过 0.5，较强；“下去、重叠式”对“过”依赖性。这说明这几个标记之间的关联性也不弱，意味着排除“过”以后它们之间的关联还可以继续进行分类。这为下一步做层次聚类研究提供了入手点[5]。

4 更多偏向性指标

从上面的研究我们看到，概率蕴涵这个指标可以描写偏向的相似性。在概率蕴涵的基础上，我们还可以设置更多的指标来描写特征中“重要的点”。这些点其实也可以反映在聚类图上。

首先，紧密关系。

如果两个点彼此都能大概率蕴涵，那么它们肯定是一个“排外的紧密”关系，同时在这个关系里，尽管轻重多少有差异，但差异不够显著，形成一个相对对等的排他性类。

我们暂且设置“大概率”的含义是 P≥0.75。

定义 1 如果有点 i 和点 j 的概率蕴涵如下：

$$P(i/j)\geq 0.75 \quad \text{且} \quad P(j/i)\geq 0.75$$

那么，i 和 j 是一个紧密关系。反映在语言地图中就是一对独立成类的点。

其次，控制中心。

逻辑中蕴涵“P→Q”的集合论本质是 P 为小集合 Q 为大集合：$P \subseteq Q$ 。也就意味着 Q“控制”了 P。概率蕴涵里面，这种蕴涵关系变成概率上的大小。如果某个点被别的点大概率蕴涵，那么这个点就对别的点具备了控制关系。如果“别的点”是多个点，那么无疑这个点就成为一个“控制中心”。

定义 2 如果有点 i 和另一组点 $\{j_1, j_2 \cdots j_n\}$ (n≥2)，它们之间有这样的概率蕴涵关系：

$$P(j_k/i)\geq 0.5,\ k=1,2,\ldots,n$$

那么，点 i 是一个“控制中心”。

注意到，这里我们可以很容易证明，按相对相似性的聚类策略，任何“控制中心”一定会将它控制下的点汇聚为一个不能分离的类。

第三，中途岛。

有的点“对外”的概率蕴涵具有歧义性，从它出发来看无法确定它可以属于哪个类，则它属于多类。同时“别的点”对它的概率蕴涵不强，这使得它们无法成为一个控制中心。这样它就成为了一个类似“中途小岛”的存在，也就是连接两个或两个以上类的“脆弱桥梁”。节 3．2 确定性和求答性聚类中的“弱确定”就是这样一个点。

定义 3 如果有点 i 和一组点{j1,j2…jn}(n≥2)，它们之间具有这样的概率蕴涵性质：

{j1,j2…jn}是所有和 i 有直接关联的点；

存在 k1=1,2…n,k2=1,2…n,能满足 $P(j_{k1}/i)=P(j_{k2}/i)$；

对任意 k=1,2…n，对任意 j_k 直接关联的点 a，能满足 $P(j_k/i)<\max(P(a/j_k))$, max 为最大值；

那么，点 i 是一个联系两个类的中途岛。可以想见，中途岛点在研究语言演化时往往有重要作用。

5 小结

[5] 从专家直觉来看，语言本体研究中涉及的聚类很可能大多数都是层次性的。层次聚类涉及的问题更复杂，篇幅有限，相关问题另行撰文。

语言学本体研究中最常见的规律是“偏向”，很少是真正“双向”的。即使“双向”都较强的规律，也几乎不是等同的，多多少少有强弱之别。因此，依据偏向的相对相似性/距离来考察关联和聚类更适合语言学研究，具体来说有助于：

发现数据较少的“相对”强规律；

更复合人类类型划分的多种认知策略，即不仅仅是原型策略，还有家族象似性策略；

多种偏向指标挖掘类中和类间多种性质的关联，得到更精确的语言系统描写。

本文数据运算已经编制为程序，可在线应用详见网站“永新语言学（http://www.newlinguistics.org/）”。

参考文献

[1]方开泰,潘恩沛.聚类分析[M].北京:地质出版社,1982.

[2]龚静.中文文本聚类分析[M].北京:中国传媒大学出版社,2012.

[3]白雪.聚类分析中的相似性度量及其应用研究[D].北京交通大学博士论文,2012.

[4]Wittgenstein, Ludwig. Philosophical Investigations, translated By Anscobe, 2ndedn[M]. Oxford: Blackwell, 1958.

[5]Greenberg, Joseph H. Greenberg, Joseph. Universals of Language[M]. London: MITPress,1963:73-113.

[6]Hawkins J A, Gilligan G. Prefixing and suffixing universals in relation to basic word order[J]. Lingua, 1988, 74(2-3): 219-259.

[7] Dryer, Matthew S. Position of Negative Morpheme With Respect to Subject, Object, and Verb [DB/OL]. 2016. In The world atlas of language structures online (http://wals.info/feature/114).

[8]孟琮,郑怀德,孟庆海,蔡文兰.现代汉语动词用法词典[M].北京:商务印书馆,1999.

[9]陈振宁,陈振宇.用语图分析揭示语言系统中的隐性规律——赢家通吃和赢多输少[J].中文信息学,2015(5):20-31.

[10]马希文.关于动词“了”的弱化形式 lou[A].朱德熙.中国语言学报(第一期)[C].北京:商务印书馆,1983:1-14.

[11]戴耀晶.现代汉语时体系统研究[M].杭州:浙江教育出版社,1997.

作者联系方式：陈振宁，浙江大学西溪校区北园 16 幢 206 室，310058，手机:18986290913，Email:706867589@qq.com；陈振宇，上海市邯郸路 220 号复旦大学中国语言文学系，200941，手机:13585693648，Email: chenzhenyu@fudan.edu.cn。

陈振宁照片

陈振宇照片

稿件修改说明：根据评审意见，稿件修改具体情况如下表。

正文	it seems not enough to just use one word as an example.	增加了多个语言学研究中的实例。
	the advantage of the proposed	在实例中指出基于绝对相似的对称式聚类的问题

	definition is not presented very clearly	
	有内容缺失现象：例如定义 2 缺失蕴涵关系公式。	已添加。
	论文篇幅量大，应适当精简论文内容	将原文和语言学无关的实例讲解全部删除，在增加语言学实例前提下融合到语言学实例中做理论分析，修改篇章结构，从 10000 字精简到 8000 字规模
	论文“考察汉语常用动词和 9 个体标记搭配得到的频次数据”，对原始数据来源、规模均未作介绍，应补充。	已补充。