Confirmation Number: 55

Submission Passcode: 55X-A3H9F8D4P6

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算法

陈振宁[1]，陈振宇[2]

（1.浙江大学，浙江杭州，310058；2.复旦大学，上海，200433）

摘　要：本文用“图”这一数学工具，通过定量分析来揭示语言系统中的隐性规律。设计了赢家通吃和赢多输少两种生成算法，将理想算法“步步竞争、择优而行”的博弈论思路贯彻到非理想状态。两种新算法都较前人有更好的概括能力。赢多输少算法更兼顾了充分概括和适度概括均衡。生成语图后设计着重准确率的最小简图和着重覆盖率的最大简图归纳算法，挖掘控制的主流规则、分析语言系统的语言学规律。在最小简图基础上提出控制度公式以评价语言系统。

关键词：隐性规律，图论，博弈论，规则挖掘

中图分类号：H030　　**文献标识码**：A

Revealing Covert Laws in Language Systems through Graphs

——Winner-Get-All & Winner-More-Loser-Less

Chen Zhenning, Chen Zhenyu

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

Abstract: We tried to reveal convert laws with quantitative analysis through graphs and designed two generating algorithms of language graphs: Winner-get-all and Winner-more-loser-less, which extrapolated game theory used by idea-algorithm to none-perfect state. Given consideration to full generation, our algorithms are more powerful than algorithms made before. Furthermore, we created a balance between full and modest generation in the Winner-more-loser-less algorithm. There are two kinds of inductive algorithms to mine mainstream rules and analyze linguistic laws: Min-Subgraphs for accuracy, as well as Max-Subgraphs for coverage. A formula for control degree based on min-subgraphs was put forward to evaluate language systems.

Keywords: Covert Laws; Graph Theory; Game Theory; Rules Mining

1 引论：

类型学在跨语言比较研究中引入语义地图（Semantic Maps）理论，它以基元（即所调查的语言项目）为“点（node）”，根据这些项基元的共现“关系（relationship）”连接成“边(edge)”，生成一个“图（graph）”。然后用这一地图去挖掘各项目间的规律。[1][2][3]这种地图其实就是“图论（graph theory）”研究的内容[4]；这种关系，是研究在“交际”中形成的“隐性控制”关系。

另外，“语义地图”并不限于狭义的语义。“任何形式、语义甚至语用项目，只要对象个体间具备某种联系，或者说相关性”，就都可以其研究[1][3]。因此，本文扩展这一术语为“语图”（Graphs of Languages）。

* **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）

1.1 交际—控制理论

交际—控制是一个社会学概念。“社会”（society）是人的集合，但仅仅把人弄到一起还不够，其中必须有一套内在控制机制，令人群成为有类别有等差、一体运行的集体。粗略地讲，“社会=成员集合+控制机制”。

存在着两种控制模式：[5]

1）显性控制：成员产生明确的关于某种运行规则的认识，这一规则“外化”于社会，有明确标记，相对独立、静止。

2）隐性控制：未曾事先规定任何规则的社会在其自身的运行中，会自发地形成运行机制，但它仅仅是现在的、当下的、自动地形成着。

隐性控制难以认识与把握，具有模糊性、即时性、变化性等特征，即难以识别，又可能产生过强的识别。“**错误理解**”（mis-understanding）和“**过度理解**”（over-understanding）都是对事实的非真实的反映。所以，隐性控制最需要**定量**的分析。

“交际—控制理论”把隐性控制机制看成是一系列交际过程中呈现的即时的事实，试图通过对论域中的交际活动的定量分析，来构建隐性控制机制的轮廓。

在交际的过程中，各个成员（称为“基元”）参与交际的程度并不一样，其中有的参与程度高，从而成为“控制中心”，并形成一定的“控制路径”，对整个系统起着主要的甚至是决定性的作用 [5][6]。如一个俱乐部有 A、B、C 三个成员，假定他们有两种共现情况，分别如表 1、表 2 所示（其中“+”号表示共现关系，“yes 数”表示共现成员的数量）：

表 1　三基元的理想控制关系

A	B	C	yes 数
+	+		2
+		+	2
+	+	+	3

表 2　三基元的理想非控制关系

A	B	C	yes 数
+	+		2
+		+	2
	+	+	2
+	+	+	3

根据 Haspelmath、Haan,Ferdinand 提出的理想状态下的经典绘制算法 [3][7]（简称理想算法），我们只考虑这些基元之间共现关系的“有无”：无共现的点不能连接；有共现的点则加以连接；3 以上多元共现要核对“两两排他共现”以避免出现“圈”（cycle）。因此从表 1、表 2 分别生成两个关系图：

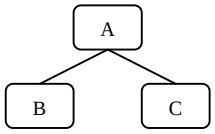

图 1 三基元的理想控制关系

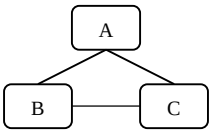

图 2 三基元的理想无控制关系

表 1、表 2 都有 A、B、C 三者共现。但在表 1 中，B、C 之间并无两两排他共现，所以 B 与 C 之间没有直接关系，是经由 A 作为“中间人”才能沟通，因此在图 1 中有以下隐性规律：A 点作为辐射中心，B、C 只能和 A 点直接交际，A 可视为星（star）图的中心，是典型的隐性控制中心：B 与 C 共现时，一定是 A 沟通的，因此 A 一定出现。

在表 2 中，A、B、C 两两之间都有直接关系，是图论中的“完全图”（completed graph），因为任意两点之间都有边，所以整幅图的关系“均匀划一”，所有成员“人人平等”，没有任何控制关系，也称为“空地图”。

下面是基于上述原理构建的不定代词（indefinite pronoun）各个功能项之间的关系地图 [7]，可以看到，其中有很好的控制关系，如（2）控制（1），（6）控制（7），（8）控制（9）；但功能项（3）、（4）、（5）之间是完全图，（4）、（5）、（8）、（6）是“圈”（cycle），都无法找到隐性控制者。

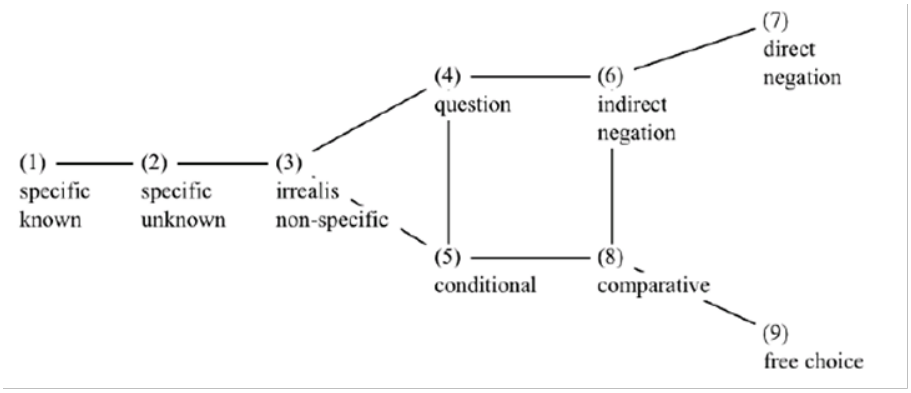

图 3　世界语言中“不定代词”各功能项之间的关系

注意，排除调查数据有误，确实可能有无法去除的圈，说明这一系统局部仍处于“自由竞争”状态，本身不具有稳定的隐性规律 [5]。

1.2 非理想系统已有分析算法：完全加权

现实中的真实数据并不总是理想的，大多数情况下，基元之间的共现关系并不是绝对的“有无”，而是以不同频次体现出来的相对“多少”倾向。Cysouw 研究跨语言人称语义时就遇到这个问题。他将人称语义分解为 8 个基元，调查了这些基元的跨语言共现，如表 3：[7]

表 3　人称 8 个基元的共现情况表

频次	1	2	3	12	123	13	23	33	yes 数	频次	1	2	3	12	123	13	23	33	yes 数
125			+					+	2	3	+	+	+						3
97				+	+				2	2				+	+		+		3
84		+					+		2	1					+	+	+		3
29	+					+			2	1				+	+			+	3
17							+	+	2	1	+			+	+				3
10	+		+						2	35	+			+	+	+			4
7		+	+						2	18				+	+	+	+		4
3					+	+			2	11				+	+	+		+	4
3	+	+							2	6		+	+				+	+	4
2				+		+			2	5		+		+	+	+			4
2						+	+		2	4		+		+	+		+		4
2			+				+		2	1			+	+	+			+	4
1						+		+	2	5				+	+	+	+	+	5
1	+			+					2	2	+			+	+	+	+		5
1	+						+		2	1	+	+		+	+	+	+		6
100				+	+	+			3	1		+		+	+	+	+	+	6
5			+			+		+	3	1	+	+		+	+	+	+	+	7
4		+				+	+		3										

人称 8 基元含义：1 第一人称；2 第二人称；3 第三人称；12、123、13 第一人称复数；23 第二人称复数；33 第三人称复数

人称基元的共现很复杂，光看“有无”基本判断不出什么来，但不同共现间的频次差异很大。Cysouw 提出了基于共现频次高低的加权算法 [8]：n 个基元共现，认为它们两两之间**全部**存在同一关系，于是直接**两两全部连接**起来形成 n*(n-1)/2 条边，所有 n*(n-1)/2 条边都直

接加上共现频次 f 作为权重，如表 4。

表 4　完全加权生成的人称语图权重矩阵

	2	3	12	123	13	23	33
1	8	13	41	40	68	5	1
2		16	12	12	4	101	8
3			1	1	5	8	137
12				286	181	34	20
123					184	35	20
13						35	24
23							30

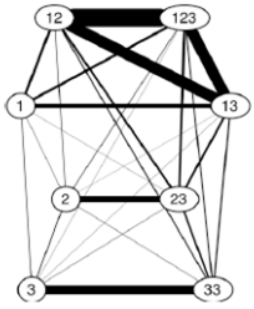

图 4　完全加权生成的人称全图

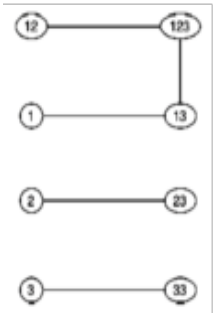

图 5　简图 1

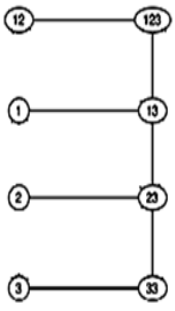

图 6　简图 2

这样，每个共现记录局部是**完全图**，表 4 中所有边的权重都大于 0，所以本文简称其为**完全加权**算法。显然，完全加权生成的语图包含大量的圈。Cysouw 按主观判断删略一定的“粗边”得到揭示跨语言人称语义蕴含规律的简图。全图如图 4，因为主观取舍的不确定性，简图有多幅，如得图 5、图 6：

完全加权不兼容理想算法，图 4 控制能力差，无法很好地归纳规律。这就产生了一系列问题，例如：

表 5：3、13、23 的两两排他共现

3	**13**	**33**	**两两排他共现频次**	**理想算法处理结果**	**完全加权处理结果**
+	+		**0**	**3-33-13**	**3-13-33-3**
+		+	132		
	+	+	19		

问题 1：基元 3、13、33 有共现且频次为 5，这 3 个基元的两两排他共现见表 5：

这个局部共现明明是理想状态，应生成控制链（chain）3-33-13，却被完全加权处理为圈 3-13-33-3，其中本应权重为 0 的 3-13 在表 3 中有权重 5。

问题 2：基元 12-13 完全加权后累计权重 181，是图 4 中第三“粗”的，两幅简图都删掉了它。边 1-13 权重 68，相对较“细”，却在简图中保留。这样做不是基于算法而是基于研究者的直觉，其主观性很难操作。

国外其他学者的研究也多以局部完全加权为基础 [9]。在国内，郭锐提出的完全关联度算法 [2]大体上也是一种完全图，所以未能避免有关的问题。

1.3 本文的研究目标与技术路线

本研究致力于解决在非理想状况下系统的隐性控制规律分析问题。我们认为：

1、加权算法引入共现频次来处理非理想数据是合理的。这一点上我们的技术路线与它相同：**定量分析**，按频次为每条边逐步累计加权 [5]，按权重之和确定倾向性，得到一语言系统的控制规律“主流”（mainstream）。

2、我们不同意**完全**加权，这种在每个局部生成完全图的做法反而违背了定量分析的要求，不符合图论与隐性控制的基本原理，和理想算法在数学方法上相悖，最终概括力度太弱。我们的技术路线修订为：每一步计算累计都**综合其他记录提供的竞争参数，按竞争参数定量分析，设计博弈论（Game Theory）的优先决策算法，对“赢家”和“输家”边给予不同的加权策略**。

另外，隐性控制的探索还要注意两点：充分概括，建立具有充分概括力的算法，把各基元、各边之间的不平等关系充分地体现出来；适度概括：过强的概括力可能会把较小的差异“放大”为显著的区别，“过犹不及”，需加以压制。

就已有的研究看，**尚未能找到充分概括的算法**是主要矛盾，但也不能忽视次要矛盾，在**找到充分概括的道路后应关注适度概括**。。

2 我们的方案

2.1 赢家通吃算法 [6]

赢家通吃将理想算法的基本原则扩展到非理想状态：

1、对每个 n≥3 的多元共现，提供竞争参数“**两两独立共现频次**”，按参数大小竞争。先计算局部共现中所有“两两对子”的两两独立共现频次，再按从大到小顺序排列这些两两对子，选取频次大的 n-1 个对子为“赢家”，剩下的对子都是“输家”；

2、 “**优胜劣汰**”博弈策略：赢家获得全部加权，输家无加权。

其中，**两两独立共现包括：1、两两排他共现；2、不同多点共现中出现的两点单独共现。**

以表 3 中人称 12、123、13 三者共现 100 次的记录为例。12、123、13 能形成最多 3 个对子：12-123、123-13 和 12-13。这 3 个对子的“两两独立共现频次”计算如表 6：

表 6　12、123、13 的两两独立共现频次

12	123	13	两两排他共现频次	不同多点共现中的两两单独共现频次	两两独立共现频次
+	+		97	9	106
	+	+	3	1	4
+		+	2	0	2

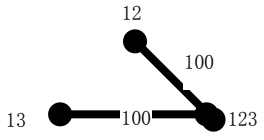

图 7　赢家通吃生成的 12、123、13 局部语图

然后，保留 n-1=2 个“赢家”：两两独立共现频次相对大的 12-123、123-13，全部加权 100；剩下 12-13 是输家，无加权。如图 7。

对人称语义应用赢家通吃算法，可得到权重矩阵如表 7，语图如图 8。因为兼容理想算法，理想状态下明确可以删除的边权重都为 0。

最后，赢家通吃算法设计了简单的归纳算法：严格按权重阈值“删细留粗”。如果设置阈值为 35，得到完全加权主观简化的简图 5。设置阈值为 30，得到完全加权的简图 6。

表 7“赢家通吃”生成的人称语图权重矩阵

	2	3	12	123	13	23	33
1	3	13	2	0	68	1	0
2		10	5	5	0	101	0
3			0	0	0	2	137
12				285	2	0	1
123					183	25	2
13						35	17
23							30

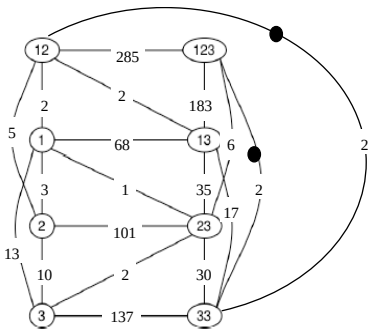

图 8：赢家通吃生成的人称全图

注意，赢家算法中每个局部的赢家选择 n-1 个，遵循的是图论的定理及其推论 [4]：

定理 1：n 个顶点的连通图是一颗树，当且仅当它有 n-1 条边。

推论 1：每个连通图均包含一棵支撑树。

由此，不考虑全图本身可能是“森林”、图中有“歧义”、“可恢复边”[1]等特殊情况，选取竞争参数相对最大的 n-1 条边，是为了归纳局部语图的“最大支撑子树（max spanning

[1] 森林：由几棵彼此不连通的树构成的图[4]。歧义和可恢复边的数学定义见节 3。

subtree）”。

这意味着赢家通吃算法局部**最大限度地加强概括力度**，反过来说可能造成**概括过度**：赢家与输家差别不大时，完全不赋予输家权重可能太过分了。

2.2 赢多输少算法

赢多输少对赢家通吃可能出现的过度概括进行了均衡：博弈采取“**优多劣少**”，按照两两共现频次的“多少”倾向程度，对赢家输家按比例加权。这样也能在加权策略上**更彻底地贯彻定量分析**方法。

分配比例理论上应按“连接所有基元的路”来分配，但这样算法复杂度高达 O(n!)[2]。为降低算法复杂度，本文采用一个近似的比例分配算法：前面 n-2 个赢家都直接 100%加权；最后一个（第 n-1 个）赢家和所有输家一起按比例分配加权。这样连接所有基元的路最多可能有 n-1 条，算法复杂度降为 O(n)。

如对前述 12、123、13 局部共现运用赢多数少算法生成，结果如表 8 和图 9。

表 8　赢多输少按 12、123、13 的两两独立共现频次比例加权

两两对子	两两独立共现频次和	**各路加权**
12-123	106	100
13-123	4	100*(4/(4+2))≈66.7
12-13	2	100*(2/(4+2))≈33.3

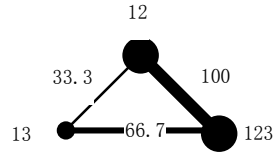

图 9　“赢多输少”算法的例 3 结果

赢多输少算法的概括能力趋向于“均衡”，各边粗细差异图 9 比图 7 小：

1、输家 12-13“彻底失败”，多少能分配到一些权重，劣势不那么明显；

2、处于赢家末位的“小赢家”13-123 的竞争参数并不比输家高多少，分到的权重被“压低”，优势不那么明显。

赢多输少算法生成人称语图的权重矩阵如表 9，全图如图 10。

表 9　“赢多输少”算法生成的人称语图权重矩阵

	2	3	12	123	13	23	33
1	4.1	12	4.6	0	65.7	1.1	0
2		11.6	2.8	2.8	1.02	101	0.3
3			0	0	0	2.2	137
12				285	35.3	7.3	2.6
123					1497	14.4	3.7
13						15.0	14.9
23							27.8

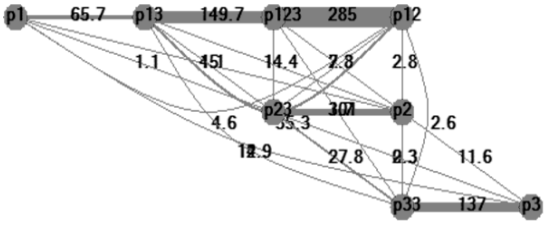

图 10　“赢多输少”生成的人称全图

赢多输少权重为 0 的边比赢家通吃少，如边 2-33，表 7 中权重 0，表 9 中权重 0.33。这是因为 2-33 出现过的局部其实不是理想状态，但因为 2-33 的两两独立共现频次很低，次次当输家。在赢家通吃算法输家无法加权，被“伪装”成了理想状态下的断路。赢多输少算法对输家多少有权重，剥除了 2-33 的“伪装”。

---

[2] “n 个基元共现于同一语言形式”的数学定义：n 个基元共现于同一语言形式，是指基元间至少存在一条能够连接所有基元的非圈最长“路（path）”。并有推论：

推论：每个局部最多可能有 n!/2 条连接所有基元的路。

于是，比例以“路”为单位来分配，就要计算 n!/2 条路，算法复杂度为 O(n!)，以阶乘增长。

3 归纳算法和非理想系统的评价

本文前述讨论的算法都是语图的“生成”算法。在理想状态或数据很少的时候，研究者很容易看出一个图的性质：典型的控制？典型的无控制？还是居于其间的状态？

但数据量较大的非理想数据复杂性高，使得任何生成算法得到的语图都还是太过复杂，无法仅仅看一看就靠主观评判全图性质，因此：1、需要用可操作的算法进行简化，但现有简化或者太主观（等于没有算法）、或者太简单（阙值简化）、或者只对基元分类根本没有控制关系（MDS 算法等[1][2][3]）；2、也需要提供评估参数，迄今为止尚未看到有研究者提出这一问题。

为此本文设计了两种归纳算法做规律“挖掘（mining）”。根据挖掘出的规律，进一步评估不同生成算法的合理性，同时提出对非理想系统隐性规则“强弱”的评价参数。

3.1 最小简图和控制度

主要思想：**找到每个基元“关联性最强”的关系**。

操作流程：从任意基元出发，检查基元 P 关联的所有边，**保留且只保留权重最大的一条边**；以此类推，直到遍历所有基元。

这一算法保留最少的边、同时保证保留下来的边权重最大，因此挖掘的是“**主流**中**最简约控制**规律”。因为最简约，所以能最大程度上保证规则的**准确率**。

在最简约的最小简图基础上，我们引入“控制度”这一概念，其计算公式是：

公式 2　控制度=(∑最小简图权重-∑最小简图歧义边权重) / ∑全图权重

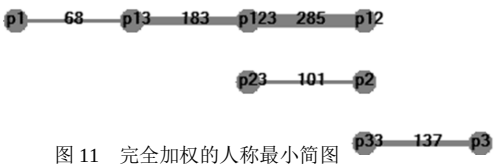

图 11　完全加权的人称最小简图

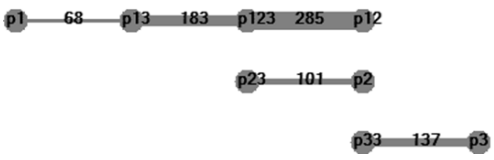

图 12　赢家通吃的人称最小简图

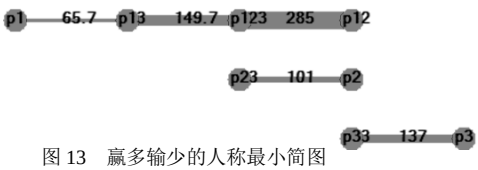

图 13　赢多输少的人称最小简图

其中“歧义”定义为：点 P 有歧义边，指和 P 关联的边中，权重相等的边数 m 大于等于 2，这 m 条边则是“关于点 P 有歧义的边”。

简化如果遇到点 P 有“权重最大的 m 条歧义边”，就无法确定点 P 到底通过谁主要和其他边相连，因此 m 条边都不可删除，留在简图内形成无法简化的子圈。无法简化的子圈无法预测控制路径，对控制度无贡献，因此需要减去。

归纳算法可以独立应用，我们对前文各算法生成的人称语图应用最小简图归纳算法，得到图 11、图 12、图 13。再根据最小简图计算各算法控制度如表 10。

各算法的最小简图拓扑结构一致，可见最小简图因为“最简约”准确率确实可观。

各算法的最小简图还和前文 Cysouw 凭主观简化得到的简图 5 拓扑一致，可见“语言学家的直觉”确实是有数学规律可循的。

最小简图所揭示的规律比 MDS 等基元分类法更全面：

1、可以确定分类：人称 8 基元分成 3 类，第一人称（1、13、123、12）、第二人称（2-23）和第三人称（3-33）；

2、可以确定最主流的控制路径：第一人称内部控制路径为 1-13-123-12；“我”与“我们”间的主要控制中心是排斥听者 13；“我们”中包含三方的 123 居于主要控制中心位置，各排斥了某一方的 12 和 13 之间语义关系疏远；

3、第二人称、第三人称内部只包含两个基元，谈不上控制路径，只表示各自的单复数之间关系最紧密。

表 10　跨语言人称系统的各算法控制度

算法	生成人称语图的控制度	规律性	专家评价
完全加权	57.93%	弱	强
赢家通吃	83.5%	强	
赢多输少	81.86%	强	

尽管最小简图拓扑结构一致，权重差异去很大，各算法所得控制度颇为不同。

完全加权所得控制度颇低，近 58%的控制度意味人称系统很“松散”，“最主流”的一、二、三人称之间混淆得很厉害，但研究者直觉上对“人称三分”的规律性评价是较强的[8]，这就产生了矛盾。

赢家二算法算出人称系统控制度高达 80%以上，略有差异而在一个数量级中，更加合理。

3. 2 最大简图

节 2.1 论及赢家通吃算法在局部生成“最大支撑子树”，这正是一种归纳算法：删除语图中任意圈里权重相对最小的边，从而把语图中每个圈都“打破”，最后必然得到语图权重最大的支撑子树。

所谓“最大”支撑子树，主要是：1、保留的边权重相对最大；2、子树支撑全图，**最大限度连通所有基元**，挖掘的是“**覆盖率最大**的主流控制规律”，因此称之为“最大简图”。

各算法的人称语图可生成最大简图如图 14、图 15、图 16。

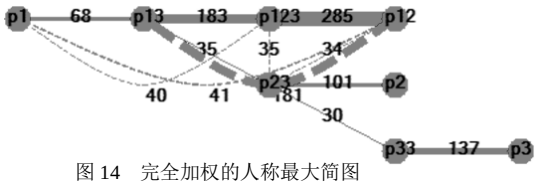

图 14　完全加权的人称最大简图

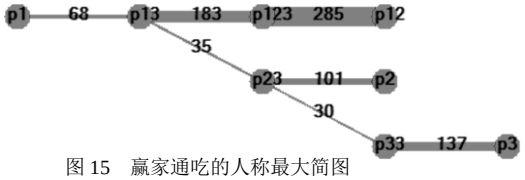

图 15　赢家通吃的人称最大简图

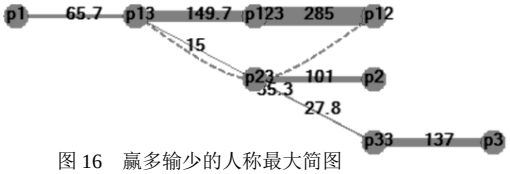

图 16　赢多输少的人称最大简图

所有最大简图在“主流”上依旧是拓扑一致的，且与 Cysouw 主观删减得出的简图 6 一致。可见这一算法的准确率还是很高，同时语言学家的直觉有数学规律可循。

但是，图中有 3 条不同的虚线边。

虚线边的权重比最大简图中的“最细边”高，这意味着有些权重可以跻身“主流”之列的关系十分“纠结”，很难概括其明晰的控制路径，由此而成圈是“可保留的圈”，相应的虚线边本文称之为“可恢复边”。如果硬要删除不免过度概括。

问题是语图的生成算法不同，可恢复边的情况就不同。不同算法得到人称语图可恢复边共计 3 条：1-123、1-12、13-12。

1、1-123：“我（1）”和典型的“我们（123）”完全没有两两独立共现，恰恰是理想的没有关联的基元。完全加权不兼容理想算法，因其在 1、123、13 三点共现中出现过，每次都给 1-123 完全加权，最终其权重较高可恢复，是不合适的。赢家二算法在 1、123、13 中都只连接 1-13，保持权重为 0。

2、1-12：“我（1）”和“咱们（12）”的两两独立共现频次为 1，是一个“非主流”规律。完全加权因其在 1、12、123、13 四点的多点场合里共现过，局部完全图累计较高权重，把“非主流”推成了“主流”，也不大合适。赢家通吃算法把输家 1-12 断开，赢多输少则保持其为非主流。

3、12-13：两个不太典型的“我们”间两两独立共现频次为 2，低。但是，它们主要在包含 12、123、13 三点的多点场合共现，“我们”集成 12、123、13 是极其主流的现象，有关共现频次数百，远超其他所有共现。因此，12-13“瘦死的骆驼比马大”，获得较高权重。

这确实是非常特殊的情况，赢多输少也能“独独”挑选出来。而赢家通吃算法因其生成时先行局部最大概括，所有输家都被直接“杀掉”，不免出现概括过度的“误杀”。

4 案例分析

4.1 汉语常用动词和时间标记的搭配

郭锐调查的汉语常用动词和时间标记搭配如表 11[10]：

表 11　汉语动词与时间标记的搭配

动词数量	了 I	了 F	时量 I	时量 F	着	在/正在	过	O		动词数量	了 I	了 F	时量 I	时量 F	着	在/正在	过
111								+		5		+			+		+
13	+		+							4		+			+	+	+
63	+		+				+			7	+				+		+
72	+		+		+		+			5	+				+		
247	+	+	+		+		+			14					+		
643	+	+	+		+	+	+			9					+		+
465	+	+	+	+	+	+	+			3		+				+	+
32		+		+	+	+	+			12							+
24		+		+		+	+			1	+		+		+	+	
220		+		+			+			3		+			+		
52	+	+	+			+	+			2	+				+		
21	+	+	+	+		+	+			1	+	+					+
1	+	+	+	+			+			3	+	+			+	+	+
29	+	+	+				+			2	+						
11		+		+						1			+				+
17		+								1		+		+	+		+
1	+	+	+				+			1						+	
1		+	+	+		+	+			2		+		+			
17	+						+			1			+		+		+
50		+					+										

“了 I”指动词可加“了”表示事件的开始，“了 F”表示事件的完结；“时量 I”指动词加时量成分表示事件持续的时量，“时量 F”表示事件完始后的时量。

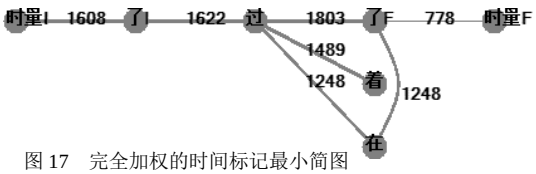

图 17　完全加权的时间标记最小简图

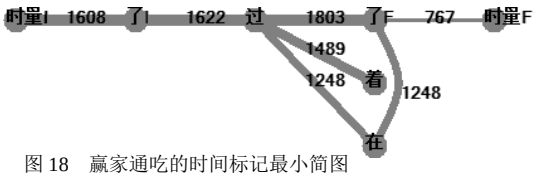

图 18　赢家通吃的时间标记最小简图

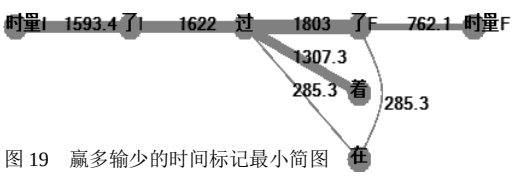

图 19　赢多输少的时间标记最小简图

暂不考虑第一行不能和所有时间标记搭配的动词，整理其他行数据，各算法最小简图如图 17、图 18、图 19，最大简图如图 20、图 21、图 22。

所有最小简图拓扑一致，“最主流”的规律准确率高：

1、汉语的时间标记统为一类。

2、不论歧义，基本上是**以“过”为控制中心的星图**。

语言学解释：“过”的语义模型包含事件“开始、持续、结束”阶段整体，因此分别控制表开始的“了 I”、结束“了 F（结束）”、持续“着、在”。

3、“时量 I、时量 F”分别只与“了 I、了

F”关联，符合其语言学定义。

4、“在”有歧义。

“在、着”都表示持续阶段，其中“在”是动态持续，“着”是静态持续。那么，我们是否可以考虑：动态和静态的差异在于，动态更倾向于结束，而静态的结束点相对“遥遥无期”？

各算法最大简图的主流是一致的。“可复活边”差异很大。

赢家二算法没有可复活边，最大简图和最小简图合一了。可见在这两种算法中，主流控制规律是很明晰的。

完全加权算法却大大不同，它的可复活边极多，各种关联纠结在一起。似乎“汉语时间标记关联混乱，几乎难以确定规律”，但这正是完全加权违背了理想算法所造成的“误会”。

例如“了 I、了 F”，它们的语言学定义就是分化“了”的两种情况，不可能出现大量纠缠不清的关联。但完全加权后边“了 I-了 F”的权重高达 1463，显然不合理。

计算时间标记系统各算法的控制度，如表 12：

图 20　完全加权的时间标记最大简图

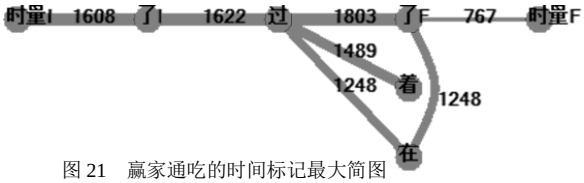

图 21　赢家通吃的时间标记最大简图

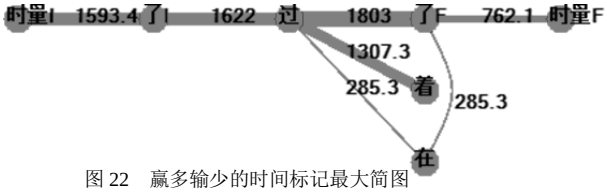

图 22　赢多输少的时间标记最大简图

表 12　时间标记系统的各算法控制度

算法	生成时间标记语图的控制度
完全加权	29.33%
赢家通吃	74.42%
赢多输少	82.38%

完全加权算法的控制度极低，这与其可复活边畸多的现象一致。赢家通吃和赢多输少的控制度相对极高，因为它们没有可复活边，主流控制规律明晰。

确实，汉语是显性规律很少的语言，汉语的“时间标记”没有彻底标记化，时间标记系统没有 100%控制度。

但是，研究者普遍称之为时间“标记”，将之归类为“虚词/功能词”，汉语时间标记即使没有完全标记化，其标记程度还是比较高的，赢家二算法明显比完全加权更符合“语言学家的直觉”。

值得注意的是时间标记系统里控制度最高的不是赢家通吃，而是赢多输少。

究其原因在于歧义：遇到歧义无法取舍，赢家通吃会直接给予所有歧义边都加权 100%，赢多输少则认为“m 个歧义=m 条机会相当比例均等的路”，因此给每条歧义边 1/m 加权。所以歧义边越多、越“重”的系统中，赢家通吃的歧义会比赢多输少“重”得多，按公式 2 反而减弱了控制度。

可见，赢家二算法的概括力度高低不可一概而论，有待深入研究。

4. 2 多个语言系统控制度参数研究

对于不同的系统，我们需比较它们的控制度。作为社会性系统，其隐性控制的程度会会有差异，呈现出一种动态的梯级，其中一端是最为严格的控制系统，其控制度为 1，即最小简图与全图完全一样，这种系统就可以直接显性化了；另一端则是完全没有隐性控制的自由状态的系统，控制度为零，即无法抽取出最小简图。

节 1.1 中的表 1 理想状态下控制度为 1，表 2 完全无控制则为 0，大部分系统则居于中间。我们对语言现象做了大量的实证研究，其控制度如表 13：

表 13　不同系统控制度参数举隅

系统	数据来源	赢家通吃的控制度	赢多输少的控制度	研究者对其规律性的评估
汉语介词“在、给、向、往、到”之间的关系	复旦中文系计算语言学小组研究	0.998	0.958	极强
汉语“名宾、动宾、小句宾、形宾”之间的关系	亢世勇 [11]	0.953	0.861	强
汉 AABB 式形容词作“状语、谓语、定语、补语、宾语”的关系	复旦中文系计算语言学小组研究	0.886	0.735	未评价
世界语言人称代词语义功能	本文表 4	0.835	0.819	强
汉语常用 AB 式动词的五类变体	复旦中文系计算语言学小组研究	0.823	0.753	未评价
汉语标记“了、着、过、在/正在、时量”之间的关系	本文表 10	0.744	0.824	强
我国境内 35 种语言中定语标记与定语类型的搭配关系	陆丙甫、屈正林 [12]	0.500	0.691	未评价
东南亚 54 种语言里“得”义语素的功能	吴福祥 [13]	0.565	0.748	弱

上述研究中：同一语言内部一般的系统控制度普遍高，跨语言的对比分析中则有高有低这可能是两个原因造成的：

1、同一语言内部共性普遍较强，跨语言间的共性偏弱；

2、同一语言内部数据调查容易些，数据多歧义易分化；跨语言调查困难，数据不足导致歧义畸多。

5 结论

语图是一种研究系统规律的工具：在多基元共现调查数据的基础上，通过算法生成一张语图，再从中归纳隐性规律。

以“理想数据”为出发点的理想算法遵循图论的原则，是一种极好的算法。但对现实中大量出现的“非理想数据”无能为力。而过去采用“完全加权”或与之本质相同的算法（如“完全关联度”等）来处理“非理想数据”，导致算法在每个局部没有概括力，生成的整个语图概括度偏弱。本文的研究即致力于解决这一问题，同时也注意到需要避免概括过度。

笔者提出的“赢家”二算法，试图在非理想数据中继续贯彻理想算法的策略：生成时步步按整体情况统计的“两两独立共现”频次计算各边的优先顺序，对 n 基元共现取 n-1 个边为赢家，其余为输家，“优胜劣汰”博弈优化，大大增加了概括力度。

其中，赢家通吃把共现频次只赋予赢家，在每个局部达到最大概括，从而使整个语图概括力度最大化，缺点是造成过度概括。赢多输少则更注意兼顾均衡，对赢家与输家按比例分配权重，在保证概括力度的同时防止出现过度概括。

本文还提出了前人尚未考虑到的问题，即对“非理想数据”如何评估其规律化的程度？为此引入了最小简图，并通过它与全图的权重比较计算出系统的控制度参数。

挖掘规律要兼顾准确率和覆盖率。最小简图的“准确率”最大，但“覆盖率”不足。为此，本文又构建了“最大简图”分析。

文中对语言学若干案例进行了研究，赢家二算法较之过去的算法更吻合系统的数据表现和语言学解释。赢多输少的归纳更适中，尤其在着重覆盖率的最大简图算法中所得简图更精

确。另外对若干语言系统的赢家二算法控制度比较，确实是参数取值越大，系统规律性越高。以上具体研究，今后还需要更多的检验和深入研究。

笔者为本文讨论的所有算法编制了程序，可在本文两位作者建设的网站“永新语言学（http://www.newlinguistcs.org）”输入数据自动计算权重控制度、绘制语图。据作者所知，本文研究至少在国内尚属首创，虽然具有填补空白的功效，但也难免会出现考虑尚不够周全之处。网站的目的既是为广大同行提供可资运用的技术手段，也是为了请研究者们提出批评意见。

参考文献

[1] 曹晋.语义地图理论及方法[J].语文研究,2012(2),P3-6.

[2] 郭锐.语义地图概念的最小关联原则和关联度[A].李小凡,张敏,郭锐.汉语多功能语法形式的语义地图研究[C].北京:商务印书馆,2015,P152-172.

[3] Haspelmath, Martin 2003 The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In Michael. Tomasello (ed.), The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, vol. 2, Mahwah, NJ: Erlbaum, 211-242.

[4] Reinhard Diestel,于青林,王涛译.图论（第四版）[M].北京:高等教育出版社,2013.

[5] 陈振宇,陈振宁.通过地图分析揭示语法学中的隐性规律——“加权最少边地图”[J].中国语文.2015(5).

[6] Nooy, Mrvar, Batagelj, 林枫译.蜘蛛：社会网络分析技术（第二版）[M].北京:世界图书出版公司,2012.

[7] Haspelmath, Martin 1997 Indefinite Pronouns. Oxford: Clarendon.

[8] Cysouw, Michael. Building Semantic Maps: the Case of Person Marking. In: Matti Miestamo & Bernhard Wälchli (eds)., New Challenges in typology: Broadening the horizons and redefining the foundations. Berlin: Mouton, 2007, P225-248.

[9] de Haan, Ferdinand. On Representing Semantic Maps. Ms. University of Arizona, 2004.

[10] 郭 锐（1993）汉语动词的过程结构，《中国语文》1993 第 6 期。

[11] 亢世勇（2004）《面向信息处理的现代汉语语法研究》，上海：上海辞书出版社。

[12] 陆丙甫、屈正林（2010）语义投射连续性假说：原理和引申———兼论定语标记的不同功能基础，《语言学论丛》（第四十二辑），北京：商务印书馆，112-128 页。

[13] 吴福祥（2009）从“得”义动词到补语标记 — 东南亚语言的一种语法化区域，《中国语文》第 3 期。

作者简介：陈振宁（1977——），女，博士研究生，主要研究领域为计算语言学。Email:706867589@qq.com；陈振宇（1968——），通讯作者，男，副教授，主要研究领域为汉语句法语义。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。

摘要	在摘要中应说明本文的主要贡献是提出两个新的算法。	摘要已改写，提出两个新的算法，并指出它们的主要贡献：都比过去的算法更有概括力度，同时赢多输少算法更好地平衡了充分概括和适度概括。
	摘要应突出对以往算法改进，以及运用改进的算法所得到的研究结果。	
正文	应该添加一个引言部分，说明本文的研究背景、研究目标、研究内容和主要贡献，并着重指出已有算法存在的局限性与本文所提出的算法的优势	加入引言，分 3 个部分，主要内容分别为：理论背景、已有“完全”算法分析、我们的研究目标。
	“基本理论”这一节过于冗长。	删节，并和理想算法的基本理论介绍合并为 1 节。
	不太确定的第一点：是否控制越高就意味着算法效果越好？	引言提出：虽然当前加强算法概括度是主要目标，但也不忽视避免概括过度； 我们的算法中赢多输少算法更兼顾适度概括； 增加第 3 节，提出最小简图、最大简图算法，从挖掘的主流控制规律的准确性、覆盖率等角度例证譬讲赢多输少算法兼顾适度概括后，其规则更“恰当”。
	不太确定的第二点：是否在一个语言现象上的效果好就意味处理其它现象时效果也好？作者或者应该处理更多的语言现象以证明算法的普遍有效性。	增加第 4 节。4.1 详解了汉语时间标记系统的语图分析，证明了本文算法的概括能力强且是适当的。因篇幅有限，应用本文算法做的其实一些研究无法一一详论，在 4.2 节做了一个关于控制度的简要概述。
		同时，因摘要、正文添加内容较多，原有标题也做了较大修改。
格式	英文摘要应该两端对齐	格式已修正
	表标题应位于表格上方	