《红楼梦》中社会权势关系的提取及网络构建

陈蕾[1]，胡亦旻[1]，艾苇[1]，胡俊峰[1,2]*

([1]北京大学，北京，100871

[2]计算语言学教育部重点实验室，北京，100871)

Email: {1100012154, 1300011764, aiwei, hujf}@pku.edu.cn

摘 要： 社会地位与权势的研究一直是社会语言学领域的一个热点话题。本文借助数据挖掘中的关系提取方案雪球算法（Snowball Algorithm），实现了红楼梦文本中候选的特征语言模式（pattern）和人物关系对之间的相互定位与赋权，对小说中频繁同现的人物对之间的社会等级关系进行挖掘，以此建立了能反映人物等级关系的有向加权人际关系网络。进一步应用最小树形图算法，生成了涵盖 192 个红楼梦主要人物的单向联通的树状社会关系图。通过这种方法生成的社会关系图不但能有效反映人际交往亲密度与社区影响力，同时还透视了人与人之间的社会等级差异。相较于单纯基于人际交往亲密程度的无向关系网络，能更加客观的表达出社会交往中人际关系网络的真实图景。

关键词：关系提取，权势关系，社会关系网络，最小树形图

中图分类号：TP391　　**文献标识码：**A

Extraction of Power Relationships in *The Story of Stone* Corpus and Construction of Social Network with Power Relationships

Lei Chen[1]，Yimin Hu[1]，Wei Ai[1]，and Junfeng Hu[1]*

([1]Peking University, Beijing, 100871，China)

Email: {1100012154, 1300011764, aiwei, hujf}@pku.edu.cn

Abstract: The study of social status has always been a hot spot in sociolinguistic research. In this study, we applied Snowball Algorithm and HITS Algorithm to discover the social relationships in the Chinese novel *The Story of the Stone*. By locating and weighting "Patterns" and "Tuples" iteratively, we built a relationship network with social class information. Finally, we generated a min-cost arborescence of the social relationships of 192 main characters of *The Story of the Stone* with Chu-Liu/Edmonds' algorithm. The generated social relationship not only reflects the intimacy and social influences, but also the hierarchical inequality of people. We regard it as a more objective and authentic reflection of social relationship network in class society.

Keywords: relationship extraction, power relationships, social network analysis, min-cost arborescence

1. 前言

社会语言学研究作为一门新兴学科，其主题围绕着语言和社会之间的相互作用展开，社会权势关系和不同社会阶层的语言使用是其中常见的研究方向之一。[1] 不同身份地位的人群所使用的语言有特异性，特殊的用语往往也会成为特定社会关系的语言标志。据此，如果收集人物间两两互动的语料，并提取出一些反映相对权势关系的特征词语，理论上就可以通过这些特征词语在群体中评估人物地位高低、并定位出具有权势差距的一对对个体。本文旨在通过文本提取信息，构建《红楼梦》一书中微型社会的权势网络。

权势是一种等级化、易于度量的单向社会关系。关于权势的社会语言学研究可以追溯到上世纪 60 年代，美国语言学家 William Labov 在 1966 年出版的 The Social Stratification of English in New York City 一书中报道了用“隐蔽式录音”的方法研究纽约市百货公司职员口语中对(r)音的着重程度和其社会地位之间的关系[2]，发现社会地位越高的人职员越倾向于将(r)音发出。1972 年，英国语言学家通过采集英国诺里奇市方言的语音资料，得出性别和潜在声望相关的音位和语音变素[3]。早期的社会心理学家也曾经尝试通过分析欧洲语言中权势与同等关系的代词的使用，揭示在历史进程中不同社会阶级之间的人际关系演变[4]，探讨了社会地位高的人自称和他称方式从明显与社会地位低的人用语方式分开，到逐渐也用权势低者的用语进行自称和他称的变化。社会语言学在中国发展起来后，国内相关研究也逐渐发展起来。2009 年胡美馨等通过分析前秦到晚清的文本，揭示女性身份认同的话语从强调男女差异（如在文学作品中“妳”和“你”的性别区分，暗示女性社会地位较低）逐渐过渡至男女“平等”（如逐渐趋向于“你”的统一化使用，代表女性社会地位趋于平等）的变化，探讨了女性社会地位的变迁[5]。2013 年李佳静等通过对杭州市“老板娘”一称呼语的调查，认为“老板娘”一用语包含上对下的社会权势关系，而这种用语的逐渐减少和废弃，也从另一方面反映出女性地位的提升[6]。传统的社会语言学研究方法能够以专业角度结合社会历史发展进程和语言元素的变化，然而往往也需要投入大量时间和人力进行采样。本研究中，我们采用了程序筛选结合人工监督过程，有效提高研究效率，同时更多从文本和数据本身入手，研究角度有别于前述“由假设推动的（hypotheses driven）”的研究。

近年来，随着计算科学的介入，基于文本的权势研究中出现了更多机器学习和统计模型的方法。大多数研究针对易于根据团队角色明确划分强弱势团体的情况。如 2012 年 Danescu-Niculescu-Mizil 等[7]于 World Wide Web Conference 发表文章，分别采集维基百科中管理员、管理员申请者、非管理员的网络讨论记录和美国最高法院的辩护记录，根据不同群体间互动时使用与对方相同语言模式的频率差异，分析“附和”（coordination）行为与权势的关系。同年，Gilbert[8]使用开源的 Enron 公司内部电子信件，根据职位建立权势层级结构，并据此提取不同权势阶级在词汇选用上的不同偏好。2014 年 Agarwal 等[9]使用同语料，说明交谈中被提到的次数越多，社会地位就越高的现象。以上研究与前文提到的传统社会语言学研究思路较为相似，都是在已知个体或群体的社会地位的基础上，寻找分布特点对应权势差异的语言因素，如词语、词性、语言习惯等。另外一些研究则采用逆向思维，通过少数已知权势关系，提取特征语素，再用这些特征语素建立分类器，进行未知权势关系的预测。如 2011 年 Bramsen[10]等发表的研究，同样利用 Enron 公司 Email 文本资料，将雇员间两两通邮的信件分为训练集和测试集，并通过在训练集中统计 N-gram 频率，筛选特征，借助支持向量机模型（Support Vector Machine）预测寄信者相对于收信人的地位差异。本文中，我们希望能够通过地位关系和语言特征之间的互证从而扩增已知信息（提取），这一点与前人研究相似。然而，我们同时也尝试探索结构信息，在《红楼梦》的虚拟社会体系中构建权势关系网络，一方面修正两人交互的偶然性偏差，获得人物之间社会地位关系的全局最优解；另一方面，清晰阐述小说的社会关系和权势结构。这一点由于应用文本的特殊性，则是在前述研究中鲜少出现的。

本文选用《红楼梦》作为研究语料主要基于以下三点考虑：首先，红楼梦中出场人物数量多、人物间阶级关系相对稳定且鲜明。其次，针对该语料的研究能够比较容易的通过人们对小说内容的理解进行验证与评测。最后，为该项研究今后在更加广泛的领域开展研究奠定可靠的基础。

2. 实验方法

2.0 实验背景介绍和方法概述

本实验采用已分词的红楼梦文本和包括了各人物所有称谓的红楼梦人名文本，在预处理阶段提取两个人名同现的语句（如“惜春 又 谢 了 王夫人”）。目标是从出现在人名之间的词语中提取模式，并用模式词语预测人物对间权势关系。由于小说文本容量有限，相当一部分人物对之间的交互频率不高，以前研究中普遍是基于统计的方法使用分类器系统，对于样本量小的情况不甚适用。在此处我们引入的雪球系统本质上采用了 HITS 算法，能够通过不断迭代，强化最具优势的特征，过滤掉一些偶发的干扰特征。在关系提取阶段，会尽量保留人物对之间双向的可能关系，最后通过生成有向图的单向连通最小支撑树的方案来削减偶然交互造成的异常值。

主要方法部分，本文先借鉴经典雪球系统，由权势人名对提取特征模式词语。后用同义词林扩充，经 HITS 系统筛选后，对得分低的词语进行去除，保留质量较高的特征模式词语。接下来对上述特征词在文中进行定位，并据此计算每一对存在交互的人物之间的权势值。最后，用最小树形图算法生成整个红楼梦社区中可定位人物组成的有向无环权势关系图。

2.1 经典雪球系统对研究有向关系的启发

1999 年哥伦比亚大学的 Agichtein 和 Gravano 等发表了一个用于关系提取的经典算法，命名为“雪球”（Snowball）系统[11]。继发表后，雪球系统及其各类变体多应用于开放系统中实体提取，如互联网中的问题发掘等。其基于“关系”（Relationships）的筛选机制，对本文研究小说文本这一封闭集合中社会关系结构具有深刻启发。研究者们观察到，在《红楼梦》中具有权势差的个体之间，普遍存在不少重复出现的“相处模式”，如当**权势高**的一方对**权势低**的一方常常有“命令”、“驱使”等行为[12]：

“原来**宝玉**心里有件私事，于头一日就吩咐**茗烟**…”

“**宝玉**便命**晴雯**来吩咐道…”

“**黛玉**不时遣**雪雁**来探消息 ...”

而**权势低**的一方对**权势高**的一方常常有“伴”、“从”等行为：

“**惜春**又谢了**王夫人**。”

“这里**紫鹃**扶着**黛玉**躺在床上 ...”

“这里**雪雁**正在屋里伴着**黛玉** ...”

这些在文本中反复出现的特征词汇和经典雪球系统中的“模式”非常相似，而具有权势差的一对人物可看做主体。因此，在最初的尝试中，本文作者们尝试了通过经典雪球系统进行实体与模式的迭代提取，后考虑到文学作品的修辞特点和人际关系的信息复杂性，在传统算法的思路基础上做出以下改进：

（1）改用单个词语取代词向量作为模式

（2）使用 HITS 算法对候选的语言模式和关系实体进行加权评估

（3） 考虑到封闭系统的特点减少迭代次数、就每一步扩展和提取采用不同的策略（如图 1）。

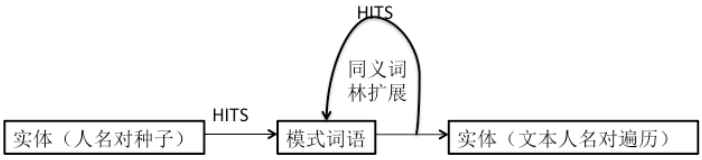


图 1. 改进后的实体和模式提取流程

（4）原始雪球系统中，辨别的是“存在目标关系的实体”和“不存在目标关系的实体”。而在我们的假设中，每对人物之间总是存在一定的权势差，只是实体之间社会地位相差的程度有异，因此我们根据模式对文中所有实体共现场景进行遍历，最后得出的并非一个区分系统，而是一个 N×N 打分矩阵，储存每一个人相对于其他所有人的权势分值。

（5）根据打分矩阵确定主要人物间权势关系，初步决定图中大多数边的方向

（6）引入有向图的最小生成树算法，以交互频率为边权，生成主要人物间社会权势关系的有向加权无环图。

2.2 用种子实体提取模式词语

首先，通过文本阅读和资料分析，我们列出 100 对存在明确地位差异的人物对作为种子实体，其中主要以“主-仆”（如“黛玉-紫鹃”、“宝玉-袭人”）、“长-幼”（如“贾母-凤姐”、“贾政-宝玉”）关系为主。按照上位者所处的位置顺序分为“上对下”和“下对上”两组种子包。

然后，提取原文中所有在种子之间出现的词语，统计其在不同种子之间出现的频率，并根据频率（经过词频修正）各筛选出前 100 个“上对下”和“下对上”的模式词语。

2.3 引入 HITS 算法进行权威度评估

HITS (Hyperlink-Induced Topic Search) 算法是 1999 年由康奈尔大学的 Jon Kleinberg 提出的一种基于“枢纽值（hubs）”和“权威值（authorities）”进行网页质量评价的算法思想[13]。本文引入此方法实现对实体和模式的质量控制：假设人物对主要具备“权威性”，模式词语主要具备“枢纽性”——即被具有高枢纽性的模式所命中的人物对，具有更为显著的地位差异；而存在于权势差更显著的人物之间的模式词语， 能更更有效地区分人物之间的地位差异。最终根据迭代至基本稳定的分值，将“上对下”和“下对上”的模式词语进行排序。

2.4 通过同义词林扩充模式词语范围

考虑到意义相近的词语在揭示权势关系的作用上有最大概率和原模式词语相同，我们运用《哈工大信息检索研究室同义词词林扩展版》对模式词语列表进行扩增。扩展后，分别得到“上对下”模式词语 1494 个和“下对上”模式词语 1214 个。然而，由于汉语词汇的一次多义现象，其中很多结果都可信度较低。因此，对各 1000 余个词语再次使用 HITS 算法评估其质量，将小于底限分数（0.0001）的结果去掉，并将“上对下”和“下对上”中都出现的重复词汇去掉，最终得到“上对下”模式词语 112 个，“下对上”模式词语 124 个，作为对 2.2 中所得词语的修正和扩充。

2.5 人物关系加权有向无环图的生成

将模式词语作为地位差距的标志，遍历文中所有人名对，对其交互频率和出现权势差异的次数进行统计，得出一个交互频率矩阵和双向的权势矩阵。以两个矩阵为数据基础，结合最小树形图算法，我们希望得到人物的关系的加权有向无环图：将两点之间交互频率的对数值赋值为两个点之间的交互边权，作为亲疏程度的衡量。亲疏程度在某种程度上反映了社会关系中子群落的信息，我们使用这种信息对一些偶然交互造成的误判进行校正。例如彩屏在权势矩阵中体现出比贾母更高的地位，而两人在文中仅有一次交互，数据可信性极低，故用对数计算剔除是合理的，同时对于交互次数多的两人，其边权值自然就大，体现出两者关系的紧密。

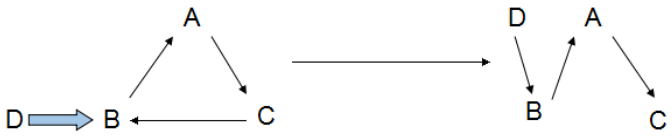
接下来，再根据权势矩阵，考察图中每对人物之间的权势方向，以明确上述带权图边的指向。首先计算出所有人名对的权势差的绝对值的平均值，将其作为筛选的阈值。当权势差高于阈值，保留权势更大的方向为最终无环图中两结点间方向，若小于等于阈值，则暂时保留结点间的双向关系，认为两个人之间的相对权势并不明显，但对于权势值较高的方向，增加 10%的边权，以保证在之后生成树的过程中实际存在的微弱地位优势不会被过强的交互频率所逆转。

在此图的基础上，运行最小树形图算法最终得到确定的方向。使用最小树形图的目的在于得到全局边权的最优的情况，并依此得到每个人名对确定的单一权势方向。具体来说，对于我们之前得到的有向带权的图，假设一个“权势至高者”作为根节点（本文中假设贾母在文中的地位最高），从根出发，选择其伸出的边权最大的边来扩展下一个点，并从下一个点重复这一扩展方法，直至所有的点连入图中，从而得到一个较优解。考察每一个点的入边，如果有比其值更大的未选边，就要考虑替换，由于图中不可成环，故有两种情况：

（1）如果待替换的边与原来的边共圈，替换不产生环，则直接替换。（图 2a）

（2）如果代替换边与其他边成环，先替换掉原边，再考虑打开所成的环。在从所有连到环上某点的未选边中选择与该点原入边边权差最小的替换环中边，若还有环则放弃，选差第二小的反复进行直至无环。（图 2b）

a.BC 边权比 AB 大，两者共圈可直接替换



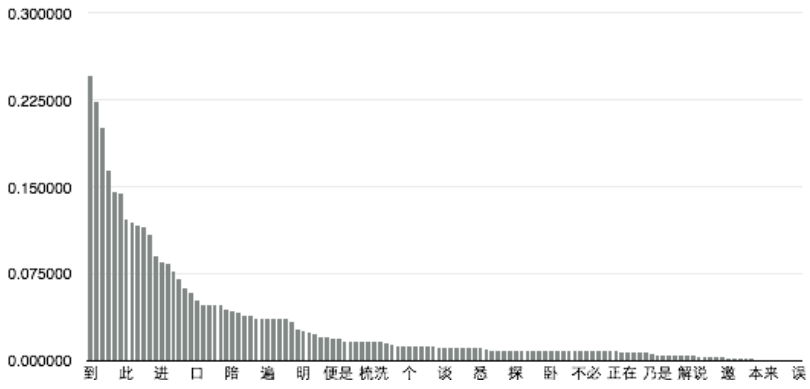
b. ABC 是替换后产生的环，DB 是符合要求的未选边，损失最小的边权来打开环

图 2 最小树形图算法思路图解

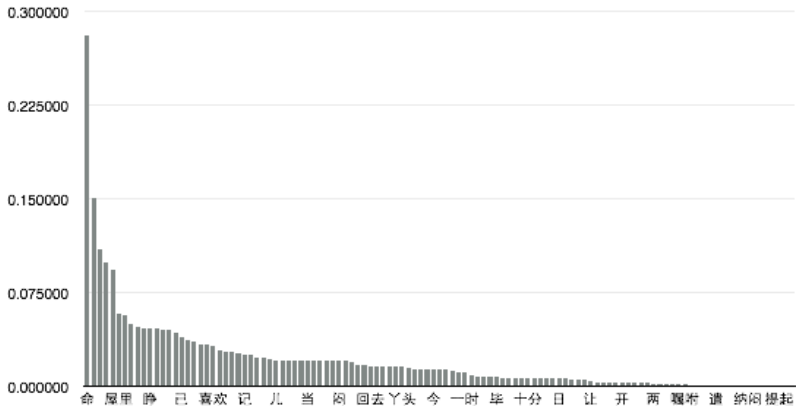
通过这样的算法，我们就成功的得到了边权和最大《红楼梦》人物关系有向无环图，即最小树形图 [14]

3. 实验结果

3.1 模式词汇提取



a.“下对上”关系模式词汇权重分布



b.“上对下”关系模式词汇权重分布

图 3 关系模式词汇权重分布示意图

列举“上对下”、“下对上”两种关系中最终权重较高的模式词汇（图 3/表 1、表 2），可看出，在“上对下”关系中，模式词汇之间权重差距更为明显；而“下对上”关系中，模式词汇的权重差异则较为缓和。根据得分最高的模式词汇，可推测其中社会地位相对较高的人对社会地位较低的人在“命”一词的使用上有很高的频率，且一旦这一语素出现于两个人之间，二者社会地位悬殊的事实就很容易被确定下来。而从“下对上”的关系词中，直观上应该更为显著的如“陪”、“扶”等词汇实际上得分却并不如“到”、“睡”一类从词义本身偏向中性的词汇那么高。推测出现这种差异的原因是，在《红楼梦》这一作品中对于地位高者的威严和权势的形象塑造着墨更重（致使相似命令式词汇出现频繁），而对于丫鬟和小辈这样的地位相对较低者，则

一来更少作为交际中的主动者（模式词汇描述的更多是回应和反应的行动），二来《红楼梦》中对他们的描写也更注重人物的独特个性（使得如同“命”一样千遍一律的词汇很少出现）。

从模式词汇的提取结果上看，我们并不能下结论说每一个词汇都能够独自代表一种关系，甚至其中也有可能出现一些由于主被动关系无法区分而混淆的结果。但是从另一方面说，在对文本进行深入研究之前，也无法根据对词义的直观理解来排除结果。因此我们选择在关系提取一步中验证这些模式词汇对权势关系的预测准确度，来判断模式词汇对文中社会地位差异场景的敏感性。

表 1.“上对下”关系模式词汇举例

位序	1	2	3	4	5	6	7	8	9
模式词	命	中	回来	带	和	屋里	不	想	犹
权重	0.2808	0.1504	0.1090	0.0988	0.0936	0.0574	0.0571	0.0493	0.0478
位序	10	11	12	13	14	15	16	17	18
模式词	房	睁	打发	哭	还	声	已	那边	打
权重	0.0466	0.0464	0.0462	0.0446	0.0446	0.0432	0.0393	0.0366	0.0359

表 2.“下对上”关系模式词汇举例

位序	1	2	3	4	5	6	7	8	9
模式词	到	睡	只	告诉	请	知	此	坐	答应
权重	0.2453	0.2231	0.2006	0.1635	0.1447	0.1434	0.1212	0.1185	0.1161
位序	10	11	12	13	14	15	16	17	18

模式次	接	怎么	进来	进	上	至	拿	扶	正
权重	0.1142	0.1077	0.0900	0.0843	0.0830	0.0770	0.0703	0.0620	0.0577

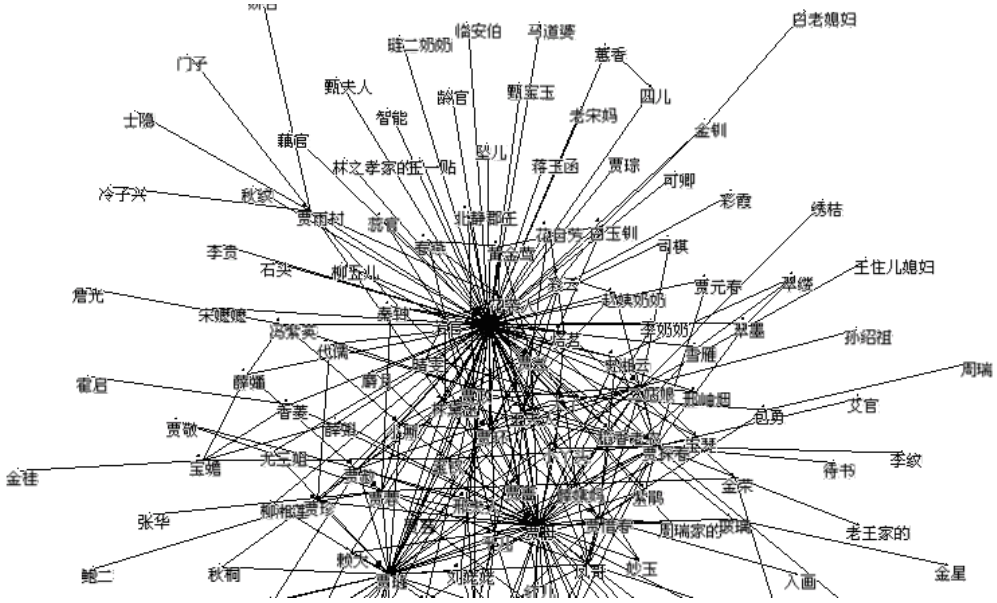
3.2《红楼梦》权势关系人物对的提取

对于主要的 192 个人物之间的社会地位差异，我们使用模式词语在其间出现的频率计算，对于每两个人之间出现双向有权边的情况，保留得分更高的一条，作为权势降低的方向。之后，用已知 158 对具有相对权势差异的人物对，进行准确度测试。具体地，对于二者能够通过一条边直接连通的人物对，观察连通方向是否与假设方向相同，若相同则记为“正确”，反之记为“不正确”；对于二者不能够通过一条边直接连通的人物对，在只能往权势降低方向行进的前提下，观察从假设中地位高的一方是否能够间接连通地位低的一方，以及地位低的一方是否能够连通地位高的一方。若前一种情况通畅而后一种情况无法到达，记为“正确”，反之记为“不正确”，若两种情况都可以连通，则记为“不定”。最终，我们得到 92 个正确结果，23 个不正确结果，以及 43 个不定结果。

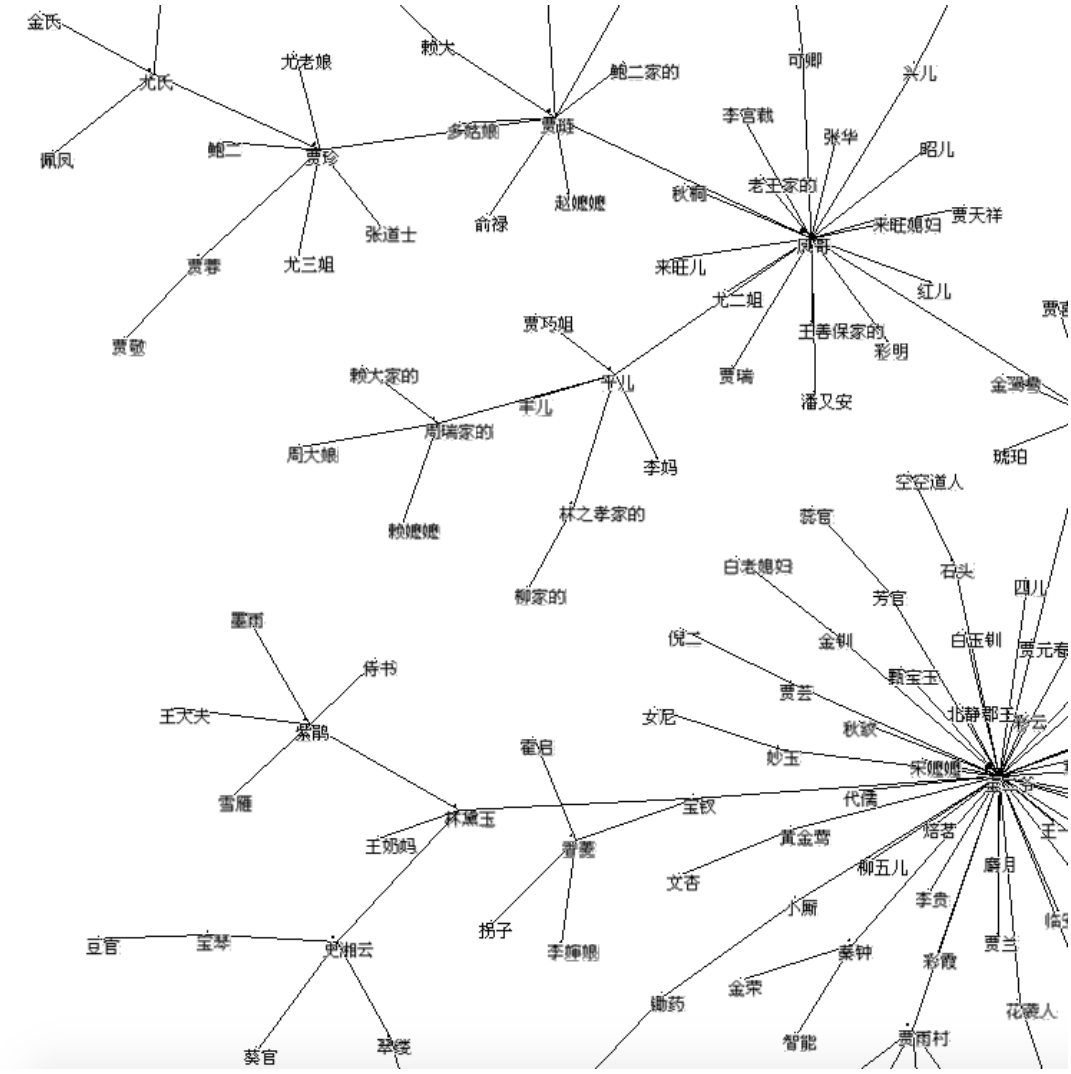
3.3《红楼梦》社会关系网络模型初探

原始的有向关系网络存在相当数量的环路，这反映了人际交往过程中地位关系的复杂性。因此存在 43 对人物之间权势方向无法确定。比如说湘云和岫烟、湘云和鸳鸯、凤姐和探春、凤姐和宝钗等。因此，直接观察该有向关系网络其人物间等级化关系和社区结构划分并不清晰（图 3a）。

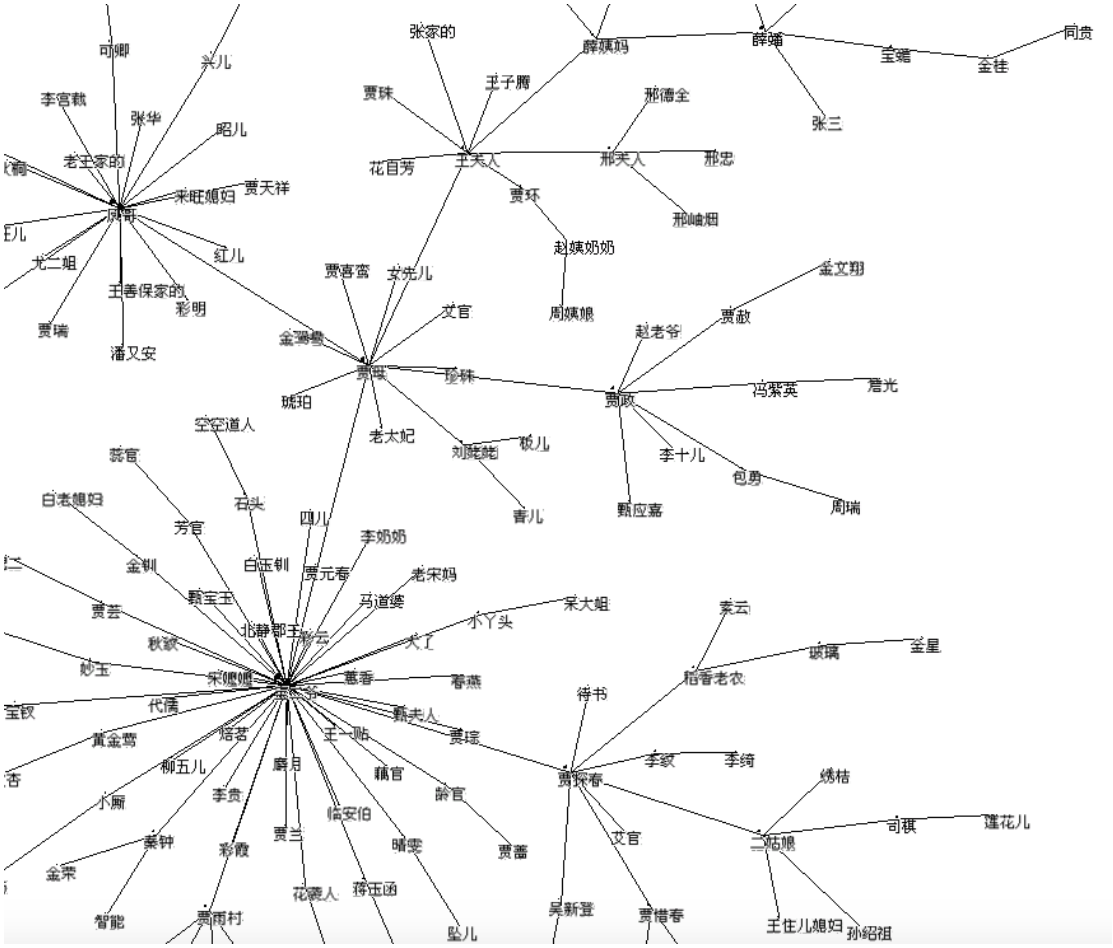
考虑到个人之间的关系在实际交往中可能会有偶然性，即跨越等级的表现（比如宝玉和晴雯之间常常出现僭越主仆关系的互动），但从社群整体来看等级关系则是相对稳定的。因此，我们利用最小树形图算法将有向关系网络中的次要的边去除，形成一个整体上拥有最强单向依赖关系的树，由此得到以数个主要人物为中心的多中心辐射状树形图（图 3b、c）。考虑到贾母在红楼梦中的地位，我们这里选取贾母作为树根，默认没有权势地位明显高于贾母的人。大多数（134 个）结点都只有一条关联的边（叶子结点），而只有少数结点（8 个）被多余 5 条边连接，成为每一簇小社群的中心，通常都是在《红楼梦》中社会地位较高的人物（表 3）。其中，贾宝玉的主角效应非常显著。其他人物社会关系也能在图中很好的体现出来。



a. 原始网络（局部）



b. 树形网络（左半部分）



c. 树形网络（右半部分）

图 4. 《红楼梦》192 个主要人物网络模型（未示权势方向）

表 3. 树形网络中出入边总数大于 5 的人物

位序	1	2	3	5	6	8
人物	贾宝玉	王熙凤	贾母，贾琏	贾探春	贾政，贾珍	王夫人
边数	48	18	11	8	7	6

可以预期，由于人物间交互信息繁杂，不怎么打交道的两人之间，容易在少数往来中偶然命中特征词汇，造成原始网络中一些误保留的边。通过生成最小树形图删除一些边后，这种情况有所改善，使得社会关系结构能够更好地体现出来，例如图 3 所示情况。在原图中，除去权重显著低于反向边权的边后，紫鹃相连的边共有 30 条，而雪雁所连的边有 12 条。在

最小树形图算法处理下，许多边由于交互频率过低而被消除，如雪雁和紫鹃与贾母、宝玉之间的边。然而，这并不代表我们放弃了对这些关系的判定，虽然没有被直接相连，我们依然可以从树形图中得到紫鹃、雪雁和贾母、宝玉等人之间的关系。从而很大程度上去除了冗杂交互信息，促进有向社交网络的可视化。事实上，这样的树状网络直观地反应出了人物的行政权势关系。最小树形图算法不仅删除了许多可疑的边，还删除了非直接隶属（联系不够紧密）关系的边，这样留下的边往往连接的是有直接上下级关系的两人，有利于我们对整个网络的权势脉络有更加清晰、正确的认识。同时，对于文本中没有直接产生交互关系的个体，只要在树上存在直接连通的通路，就可以预测其在《红楼梦》中的相对权势关系。举例来说，墨雨很少与其他人物有交集，但其处于紫鹃的下级，从而我们可以合理的推断，与其在同一路径且位于上层的贾宝玉对于墨雨有社会地位上的优势。也就是说，即使对于非直接隶属的关系，我们通过权势的可传递性以及树的特点，能够作出合理的推断。

图 5. 树形网络局部特写（宝玉——黛玉——紫鹃——雪雁）

当然，最小树形图也有其局限性，对于数据稀少的个体，可能由于对全局最优的需要而生成我们意料之外的边，比如原始关系数据极少的北静郡王和贾元春，就接入了宝玉的下方，并不太符合实际的关系。同样的，有些与他人交集较少的底层的丫鬟或奴仆，也可能作为个例接入并非其主人的父节点。根据观察，若不考虑一些个体由于数据不足而产生的问题，树形图整体上以很高的准确度反应《红楼梦》中的权力制约关系。

4. 结语

本研究尝试了在文本语料中提取人物社会阶层关系，建立了反映社会阶层关系的红楼梦人际关系网。实验表明，通过该有向关系网做出的最小树形图能较为准确地反映《红楼梦》中 192 个主要人物之间的社群结构，对多数人物对之间的社会地位差异的预测结果也比较可靠。

相对于以前的研究，此方法的特点有三个。其一，适用于文学作品一类的小文本，人物关系复杂，而交互信息有限的情况。其二，相较于以往的社区划分算法，在加入了权势依赖关系是单向且无环路的约束后，实现了整体权势结构的最优。能有效地消除个别人物角色之间偶然发生的阶层越位的互动带来的干扰，因此在社会地位的判定上更为精细。由于阶层关系并非可以单纯依据人物之间的两两互动来确定的，因此在本研究中我们没有使用常见的分类器的方案，一开始就尽可能的保留了人物之间所有的双向关系，然后再局部对比和全局考量过程中逐渐选择性删除边。最后达到了好的效果。其三，所得到的权势关系不再局限于有交互事件发生的个体之间，而是可以借助连接其他节点形成通路来间接比较。因此能有很好的预测性。在权势网络中的两个人物只要有通路，就能唯一判定相互之间的权势关系，而并不要求在文本中两个人有实际的互动。

在人际关系网络研究中加入等级关系更真实地还原了社会网络中人物之间的社会交往形态。可以认为本文的方法在研究社群划分、社会关系变迁和社会结构分析中都存在更大的应用潜力。

同时，本研究仍然存在一些局限性：

（1）可应用语料的有限性。如《红楼梦》这样出场人物众多、存在明确而复杂的人物关系、等级森严的社会制度的小说非常少。因此，在后续的探索中，我们考虑尝试在网络论坛的社区环境下考察此方法的有效性，并同时尝试寻找其他可用语料和应用场景。

（2）由于文学作品侧重于主角的描写，众多配角的出场多是围绕主角进行，而现实生活中，这样以一人为核心、其他人之间的关系都很疏离的情况是不太常见的。且由于最小生成树的算法特征，无法连入剧情主干的一些成独立“小圈子”的节点们在建树过程中被逐渐删去，边缘化群体之间的关系无法被观测到。因此，若考虑将本文方法应用于现实生活中的网络社区，尚且需要做更多的尝试和调整。

（3）本文以词语提取而非词包提取为主，并没有特别考虑被动式。分辨“上对下”和“下对上”关系主要依靠两个人物在文本中出现的顺序。当被动式一类可能造成词义反转的情况出现时，词语在两种关系方向中的权重都会降低（主动式和被动式的权重互相抵消）。然而这就导致本研究在模式词语的提取上始终比较保守。在未来大文本的工作中，可以考虑进一步使用词袋模型或更复杂的语言元素代替单独词组，将被动式等可能造成词义削弱或反转的因素纳入模型中。而在现有的小文本情况下使用词袋模型等可能会导致每个候选模式的频率都比较低。

5. 致谢

本研究得到了国家自然科学基金项目（M1321005）；国家自然科学基金项目（61472017）的支持。

北大信息科学与技术学院张梦楠、苗睿同学，北大地球空间学院李丰翔同学为研究工作提供了帮助和支持。

参 考 文 献

1. 赵蓉晖 编. 社会语言学[M]. 上海：上海外语教育出版社，2004

2. Labov W. The social stratification of English in New York city[M]. Cambridge University Press, 2006.

3. 祝畹瑾 编. 社会语言学译文集[M]. 北京：北京大学出版社，1985

4. 祝畹瑾 编. 社会语言学译文集[M]. 北京：北京大学出版社，1985

5. 胡美馨，吴宗杰. 从先秦与晚清文本看女性身份的话语变迁——一种谱系学的跨文化分析[J]. 中国社会语言学，2009,2(13): 141~151.

6. 李佳静，孙德平. 杭州市称呼语“老板娘”调查[J]. 中国社会语言学，2013,1(20): 27~37.

7. Danescu-Niculescu-Mizil C, Lee L, Pang B, et al. Echoes of power: Language effects and power differences in social interaction[C]//Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 699-708

8. Gilbert E. Phrases that signal workplace hierarchy[C]//Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, 2012: 1037-1046.

9. Agarwal A, Omuya A, Zhang J, et al. Enron Corporation: You're the Boss if People Get Mentioned to You[C]//Proceedings of the 2014 International Conference on Social Computing. ACM, 2014: 2.

10. Bramsen P, Escobar-Molano M, Patel A, et al. Extracting social power relationships from natural language[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 773-782.

11. Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[C]. Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000: 85-94.

12. 曹雪芹，高鹗. 红楼梦[M]. 北京：人民文学出版社，2000

13. Kleinberg J M. Hubs, authorities, and communities[J]. ACM Computing Surveys (CSUR), 1999, 31(4es): 5.

作者简介：

陈蕾（1993—），女，美国圣路易斯华盛顿大学博士生，主要研究领域为生物信息学和统计遗传学，Email:1100012154@pku.edu.cn；

 胡亦旻（1994—），男，北京大学本科生，主要研究领域为计算机科学与技术，Email: 1300011764@pku.edu.cn；

 艾苇（1990—），男，美国密歇根大学博士生，主要研究领域为数据挖掘与推荐系统，Email: aiwei@pku.edu.cn；

胡俊峰（1967—），男，通讯作者，北京大学副教授，主要研究领域为计算语言学， Email: hujf@pku.edu.cn