

# 《水浒传》人物关系网络分析

张紫芸 王文发 何进荣 马乐荣

(延安大学数学与计算机科学学院 陕西 延安 716000)

**摘要** 复杂网络作为复杂系统的一种抽象描述,可用于文学作品中人物关系网络的建模工具。以《水浒传》中的108将为研究对象,采用复杂网络分析方法对该作品中的人物关系进行构建与分析,小说中的人物对应于网络中的节点,人物之间的关系对应于网络中的边。人工搜索标记人物名称,以任意两个人物在同一章回的共现关系为前提,构建人物关系网络,并计算度、介数、集聚系数、稀疏性、相关性等网络指标。结果表明《水浒传》人物关系网络是异配网络,并且具有小世界特性和社区特性。

**关键词** 水浒传 复杂网络 人物关系 小世界网络

中图分类号 TP3

文献标志码 A

DOI:10.3969/j.issn.1000-386x.2024.05.049

## CHARACTER NETWORK ANALYSIS OF WATER MARGIN

Zhang Ziyun Wang Wenfa He Jinrong Ma Lerong

(School of Mathematics and Computer Science, Yan'an University, Yan'an 716000, Shaanxi, China)

**Abstract** Complex network, as an abstract description of complex system, can be used as a modeling tool for the relationship network of characters in literary works. Taking 108 characters in Water Margin as the research object, the complex network analysis method is used to construct and analyze the relationship between the characters in the work, in which the characters in the novel correspond to the nodes in the network, and the relationship between the characters corresponds to the edges in the network. We searched the names of the characters manually, and constructed the network based on the co-occurrence relation of any two characters in the same chapter. We calculated network indicators such as calculation, meso, agglomeration coefficient, sparsity, correlation, etc. The results show that the relationship network of characters in "Water Margin" is a heterogeneous network and has the characteristics of small world and community.

**Keywords** Water Margin Complex network Character relationship Small world network

## 0 引言

《水浒传》是我国四大名著之一的一部长篇英雄传奇小说,为中国古典长篇小说的代表作,第一部以农民起义为题材的小说。故事创作于元末明初,是作者将元、宋以来的故事、戏曲以及话本通过加工整理,然后以宋江所领导的农民起义为线索进行撰写,小说具有很高的文学成就,在中国和世界的文学史上都具有极其重要的影响。

阅读优秀的文学作品能提高一个人的文学素养,

因此对其进行全面研究和分析对我们挖掘文学作品中的隐藏信息、理清小说中人物之间的关系、读懂文学作品的故事情节、理解作品的创作背景以及了解当时的社会背景都具有非常重要的作用,是理解文学作品的关键点。在科技飞速发展的当下,我们可以使用互联网的技术,将文学作品的人物关系抽象表示成网络的结构形式,从而对小说整体进行分析,抛开主观分析,得到一个对故事的客观全面理解<sup>[1-2]</sup>。

在对文学作品的人物关系网络进行构建和分析的过程中,产生了许多基于自然语言处理、复杂网络技术、文本挖掘,基于数据挖掘中的关系提取方案以及基

于语料库的方法<sup>[3-6]</sup>,且都取得了较好的成果。

涂轶文<sup>[7]</sup>用基于图嵌入的关系网络的研究方法进行分析。林峰等<sup>[8]</sup>从社会网络的角度分析,以文中人物出现的位置作为人机距离量化指标,使用 Pajek 软件绘制网络社群图,然后对人物关系进行分析。唐毅等<sup>[9]</sup>利用文本挖掘建立人物关系矩阵,然后使用社会网络的方法对人物关系进行分析。李娇<sup>[10]</sup>利用共现词对完成人物关系图谱进行绘制,然后根据关联规则,用 Apriori 算法挖掘出人物间的关联规则。胡岚曦<sup>[11]</sup>提出一种基于人物行为分析发掘人物关系网络的方法,以特定领域行为分析为对象构建行为模型,获取该领域的核心人物,然后训练生成领域关系模型,最后用社会网络分析技术发掘人物领域关系网络。

目前对于《水浒传》人物关系网络的研究主要集中在文学作品的翻译、文章中茶和酒文化,以及对小说中单个人物的行为语言的剖析,还未对大量人物及人物之间的关系进行综合分析<sup>[12-14]</sup>。本文在之前学者的研究基础上采用复杂网络分析的方法,使用 Python 以及 NetworkX 包和可视化软件 Gephi 实现了《水浒传》人物关系网络的构建和分析。

## 1 人物关系网络的构建

### 1.1 数据集简介

本文实验部分数据集采用的是有 120 回本的《水浒全传》,名著中 108 名好汉是主要人物,其他人物在文章中出现次数较少,对小说内容走向并不起主要推动作用,因此本文选用 108 名好汉(即三十六天罡星和七十二地煞星)为主要研究对象,使用复杂网络的方法对其进行了分析。

### 1.2 人物关系抽取方法

小说中人物关系的构建规则通常有两种,第一种是基于人物对话,也就是在小说中若两个角色之间有对话,就认为他们之间有关系,即网络中存在连边。这种大多适用于以角色之间对话来推动小说情节发展的文章,例如剧本类小说。第二种是基于人物共现<sup>[11]</sup>,就是当两个角色出现在同一段落、一句话,或者一个故事情节中时,认为人物之间存在关系,构建人物关系网络。

《水浒传》叙述体式为章回体,即将小说全文分为若干章,称为“回”或者“节”,每回前用一句或者两句作为本回标题,高度概括本回故事内容。通常每一回都描述了一个较为完整的故事,回与回之间相对独立,

因此每一回中出现的人物之间都具有一定的关系。由于其特殊叙述体式,本文在构建人物关系的条件是认为出现的同一章的人物之间有边存在。

用复杂网络方法构建人物关系网络,首先搜索获得 108 单将的姓名文档,将《水浒传》分章节存储,然后使用 Python 以及 NetworkX 软件包获取人物之间的连边关系,若两个人物同时在不同章回出现,则其权重加一,构建无向加权的人物关系网络  $G = (V, E)$ ,其中  $V$  表示网络节点,分别是 108 单将,  $E$  表示边,代表人物之间存在关系,得到如表 1 所示的节点数为 108、连边数为 5 765 的《水浒传》加权人物关系网络。

表 1 《水浒传》108 位人物关系网络的网络信息

文学作品名	《水浒传》
方向	无向
节点数	108
边数	5 765
节点	名著人物
连边	人物关系
章节	120

### 1.3 数据分析目标

使用 NetworkX 软件包计算网络的拓扑结构统计性质,包括度、聚集系数、最短路径、网络直径、平均路径长度、介数、相关性、稀疏性。

1) 度。网络中节点的度指的是与节点直接相连的连边数,度反映到现实生活中就是你有多少个朋友,所以在水浒传的网络图中反映的就是一个人物有多少认识的人。

2) 平均度、度分布。平均度就是所有节点的度之和与网络总节点数之比。度分布定义为将网络中节点的度值从小到大排列,统计度值为  $k$  的节点占整个网络节点数的比例。《水浒传》的人物关系的度分布直方图如图 1 所示,横轴表示节点的度数,纵轴表示  $N_k$ (即度为  $k$  的节点的个数),《水浒传》人物节点度有三种,分别是 94、106、107,我们可以看出度为 107 的节点的度个数占了总节点较大比例。说明人物之间的联系比较紧密。平均度  $\langle k \rangle$  和度分布  $p(k)$  定义为:

$$\langle k \rangle = \frac{2L}{N} \text{ 或 } \langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i \quad (1)$$

$$p(k) = \frac{N_k}{N} \quad (2)$$

式中: $N$  表示网络中的节点数; $L$  表示网络中的连边数; $N_k$  表示度为  $k$  的节点的个数。

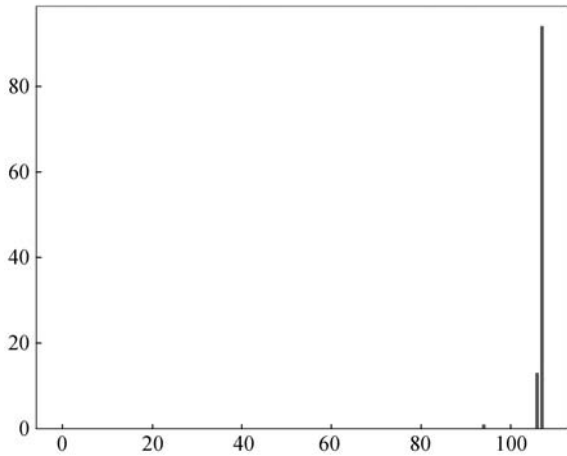


图1 《水浒传》人物关系网络的度分布

3) 网络直径、平均路径长度。路径指的是网络中节点之间存在连边,一条从节点  $i_0$  到  $i_n$  的长度为  $n$  的路径  $P$  经过  $n+1$  个节点和  $n$  条连边。最短路径指链接两个节点的边数最少的路径。网络直径指网络中任意两节点间距离的最大值。平均路径长度指任意两点间的最短距离的平均值。最短路径表示为:

$$\langle d \rangle = \frac{1}{L_{\max}} \sum_{i,j>i} d_{ij} \quad (3)$$

式中:  $d_{ij}$  表示节点  $i$  到节点  $j$  的距离。

4) 介数。介数分为点介数和边介数,其反映了相应的节点或边在整个网络当中的作用和影响力,是一个全局几何量。点介数表示在求任意一对节点之间最短路径,路径所经过的节点次数称为节点的点介数,边介数表示路径经过某条边的次数,对应的就是某条边的边介数。

5) 集聚系数。定义为节点  $i$  的  $K_i$  个邻居节点之间实际存在的变数  $E_i$  和总的可能边数之比。集聚系数的最大值为 1,其中全连通图的集聚系数为 1,一般图都是小于 1。集聚系数反映一个节点的邻接节点之间是否有连边,描述网络中节点的集聚状态,在社交网络中表示互相认识的两个人朋友圈子的重合程度,即两个人的朋友之间也是朋友的状态。节点集聚系数  $C_i$  和平均集聚系数  $\langle C \rangle$  定义为:

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \quad (4)$$

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i \quad (5)$$

式中:  $e_i$  表示当前节点的邻居节点之间存在连边的个数;  $k_i$  表示当前节点的度值。

6) 稀疏性。网络的稀疏性定义为网络中实际存在的边数与最大可能的边数之比,反映一个复杂网络是稀疏还是稠密的。网络稀疏性应满足的条件:  $L \ll L_{\max}$  或  $\langle k \rangle \ll N-1$ ,稀疏性定义为:

$$\frac{L}{L_{\max}} = \frac{2L}{N(N-1)} \quad (6)$$

7) 相关性。人物关系网络的相关性我们使用皮尔逊相关系数来进行判断,取出网络中的所有连边,计算每条连边两端点的度值,并将其按从大到小排序,得到度小序列和度大序列,最后计算他们的皮尔逊相关系数,若皮尔逊相关系数的数值大于 0,则称网络为同配网络;小于 0,则为异配网络。

## 2 实验

本文实验部分主要用到了三个技术,分别是复杂网络分析、NetworkX、Gephi。复杂网络简而言之就是呈现高度复杂性的网络,指的是具有自组织、自相似、吸引子、小世界、无标度中部分或者全部性质的网络,特性包括小世界、集群、幂律的度分布概念等。NetworkX 用户能简便地对网络进行创建、操作和学习,人物可视化部分用到了可视化工具 Gephi。

### 2.1 网络可视化

人物关系可视化是将人与人之间通过某种属性进行连接而形成的关系网络以图形化的方式表示出来,对于我们的数据集,我们使用 Python 编程语言得到的人物之间连边表格,以及 108 单将的姓名表格,利用网络可视化工具 Gephi 对其进行了人物关系网络可视化,网络可视化的结果如图 2 所示,由于网络的节点度值有三种,分别是 107、106、94,可以看到人物节点的度均较大,因此 108 单将之间关系有明显的聚集性,符合小说故事情节描述。



图2 《水浒传》人物关系可视化

### 2.2 网络指标计算

复杂网络的相关计算指标的概念如 1.3 节所示,《水浒传》人物关系网络的网络指标相关数值如表 2 所示。

表2 网络指标相关指数

最大度	107
最小度	94
平均度	106.759 2
网络直径	2
平均路径长度	1.002 2
最大点介数	0.000 02
最大边介数	0.020 9
平均集聚系数	0.998 0
相关性	-0.025 8
稀疏性	0.997 7

根据网络指标计算的实验结果显示,网络中的度值排名靠前的包括宋江、林冲、吴用、公孙胜、李逵、朱武、史进、秦明、花荣等人,其度值为107。说明他们在《水浒传》的人物关系网络中地位或影响力较高,且与其他人物有着密切的联系,度值最小为97,是彭圯,即他的地位和影响力低于其他人。对比《水浒传》的108位主要人物,其中除单廷、彭圯、王定六、郁保四、王英、焦挺、邹润、龚旺、丁得孙、鲍旭、皇甫端、魏定国、董平、张清外,其他人的点介数均较高,扮演着重要的角色,符合小说的实际情况。人物关系网络的平均集聚系数较高为0.998 0,说明网络中人物之间朋友圈子的重合度较高,即小说中人物之间的关系比较紧密,符合小说描述。结合表中网络的稀疏性值较高为0.997 7,即我们构建的是一个稀疏的无向加权网络。

在度相关网络中,若总体上度大的节点倾向于连接度大的节点,那么就称网络是度正相关的;如果总体上度大的节点倾向于连接度小的节点,称网络为负相关的,或称为是异配的。度相关系数  $r \in [-1, 1]$ ,  $|r|$  的大小反映了网络同配或异配的强弱程度,若  $r > 0$ ,则网络是同配的;  $r < 0$ ,则网络是异配的。《水浒传》人物关系网络的度相关性为  $-0.025 8$ ,即该网络是异配网络。

小世界网络又称为六度分割理论,是一类平均距离较短,集聚系数大的网络的总称,其最直接的特性就是在小世界网络上,信息传播的速度非常快。它起源于社会心理学家 Stanley 的一项实验,在《水浒传》的人物关系网络中最短路径长度为 1.002 2,集聚系数为 0.924 6,该网络具有小世界网络特性<sup>[15]</sup>。

### 2.3 模块化

模块化的度量值是由 Newman 等提出,是目前社团检测常用的一种衡量网络中社区稳定度的方法,社团检测就是要在一个图上发现社团结构,即把图中的节点进行聚类,构成一个个小小的社团。聚类就是用

个体的相似性作为距离的度量标准,节点间相互距离较小的归为一类,其反映了不同个体之间的差异性和共同性<sup>[16]</sup>。本文使用 NetworkX 的软件包对人物关系网络进行了模块划分,划分成为一个模块说明两个节点的网络拓扑结构相似,模块化结果如图3所示,网络被划分成了三类,分别是吴用、宋江、李逵、秦明、宋立等人,周莹、王通、王定六、孔亮、龚旺等人以及李立、白胜、宋清、宋万、孟康等人组成的模块。

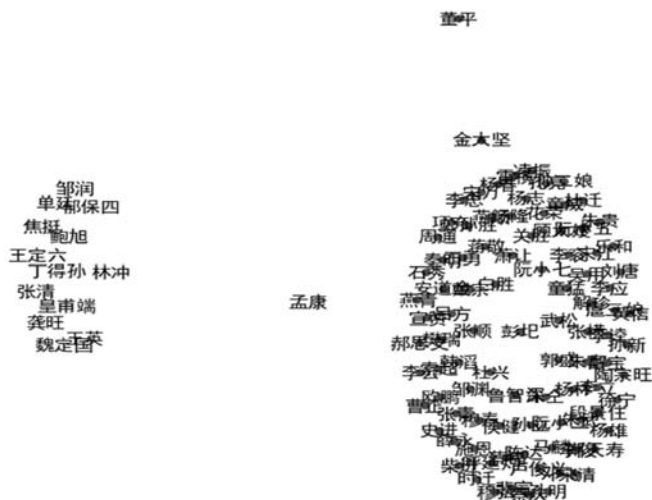


图3 《水浒传》人物关系网络模块化结果

### 2.4 链接预测

在复杂网络中链接预测指的是通过已知的网络结构信息,预测尚未有边存在的两个节点之间是否有连边的可能性,预测包括三方面的内容:一是实际存在但尚未连接的边;二是预测未来可能出现的连边;三是对错误连接的预测。链接预测解决的事节点间的相似性问题<sup>[17]</sup>。

在链接预测的实验中,本文在原始构建的网络基础上随机删除 20% 的边,同时确保留下的是一个连通网络,从而形成一个有随机边缺失的网络,从而在随机边缺失的《水浒传》的人物关系网络图上预测这些被删除的边。从以下这三个方面评估在缺失边的人物关系网络上的链接预测结果:

杰卡德系数(Jaccard's Coefficient):

$$\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \quad (7)$$

Adamic-Adar 指数(Adamic-Adar Score):

$$\sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|} \quad (8)$$

优先链接(Preferential Attachment):

$$\Gamma(u) \cdot \Gamma(v) \quad (9)$$

式中: $u, v$  分别代表不同节点,  $\Gamma(u)$  表示节点  $u$  的邻居节点,具体计算结果如表3所示(结果保留4位小数),AUC 是从整体上来衡量一个链路预测算法的精确度,

其最小值为 0.5,通常用一个 AUC 大于 0.5 的多少来衡量其比链路预测其比随机相似度算法好的程度<sup>[17]</sup>。通过表 3 可以看到这三个数值均大于 0.5,因此实验能够预测到那些随机删除或未来可能存在的连边,且比随机相似度算法好。

表 3 链接预测的曲线下面积(AUC)评分

数据集	杰卡德系数	Adamic-Adar 指数	优先链接
《水浒传》	0.524 1	0.521 9	0.518 8

### 3 结 语

阅读文学作品可以培养人们的文学素养,其一般都反映了当时的社会背景,因此理清人物之间的关系对文学作品的分析以及正确理解文章很重要,然而人们往往是通过阅读全文的方式,获得人物关系以及对文章的理解,但这样得到的结果有很大一部个人主观因素影响,不同的人可能对同一篇文章有不同的理解。本文使用复杂网络的方法,搜索获得人物名称文件,代码得到人物之间关系的连边文件,以任意两个人物在不同章回的共现次数作为权重,构建无向加权网络,对文学作品进行了客观分析,实现了《水浒传》的一百单八将人物关系网络的可视化,在实验部分计算了人物关系网络的度、网络直径、平均路径长度、介数、集聚系数,模块度以及链接预测分析了《水浒传》的人物关系网络,最后发现其具有小世界网络特性以及社区特性。

### 参 考 文 献

[1] 吉红宇. 基于复杂网络分析的人物关系挖掘[D]. 成都: 电子科技大学, 2017.

[2] 张荣杰. 感知与现实人际关系网络: 经典文学作品分析[D]. 漳州: 闽南师范大学, 2018.

[3] 陈蕾, 胡亦旻, 艾苇, 等. 红楼梦中社会权势关系的提取及网络构建[J]. 中文信息学报, 2015, 29(5): 185 - 203.

[4] 赵京胜, 张丽, 朱巧明, 等. 中文文学作品中的社会网络抽取与分析[J]. 中文信息学报, 2017, 31(2): 99 - 106, 116.

[5] 董晓烨, 柴静. 语料库辅助的文学作品主题分析[J]. 西安电子科技大学学报(社会科学版), 2018, 28(3): 106 - 111.

[6] 楼锴毅, 霸元婕, 李绍昂. 基于社交网络的小说聚类[J]. 软件工程, 2018, 21(10): 14 - 16.

[7] 涂轶文. 基于人物相似度的互联网络人物关系分析方法研究[D]. 成都: 电子科技大学, 2019.

[8] 林峰, 赵广平, 林娜, 等. 《红楼梦》文本的社会网络结构分析[J]. 石家庄铁道大学学报(社会科学版), 2018, 12(1): 58 - 63.

[9] 唐毅, 王硕, 胡桓. 《水浒传》人物关系网络的文本挖掘[J]. 社科纵横, 2018, 33(4): 117 - 120.

[10] 李娇. 基于共现与关联挖掘的人物关系图谱研究与实现[D]. 兰州: 西北民族大学, 2019.

[11] 胡岚曦. 一种基于行为分析的人物关系网络发掘方法[J]. 计算机应用与软件, 2009, 26(10): 256 - 258.

[12] 任东升, 马婷. 基于语料库的《水浒传》沙博理英译本意合句式研究[J]. 外语研究, 2015, 149(1): 64 - 70.

[13] 杜贵晨. 《水浒传》茶事考论[J]. 陕西理工学院学报(社会科学版), 2016, 3(4): 1 - 10.

[14] 李桂奎. 论《水浒传》“怒气”摹写之“乖错”情理[J]. 中原文化研究, 2020, 8(3): 92 - 100.

[15] 李泽荃, 杨墨, 刘嵘, 等. 复杂网络与机器学习融合的研究进展[J]. 计算机应用与软件, 2019, 36(4): 11 - 28, 62.

[16] 熊中敏, 朱春卫, 郭振辉, 等. 基于 OLAP 和聚类分析的关联规则挖掘方法[J]. 计算机应用与软件, 2018, 35(5): 58 - 61.

[17] 司帅宗. 社会网络中的链路预测及网络重构[D]. 沈阳: 东北大学, 2014.

### (上接第 331 页)

[3] Stillerman J, Fredian T, Klare K, et al. MDSplus data acquisition system[J]. Review of Scientific Instruments, 1997, 68(1): 939 - 942.

[4] Choi Y, Loo Y, LaCroix K. Cookies and sessions: A study of what they are, how they can be stolen and a discussion on security[J]. International Journal of Advanced Computer Science and Applications, 2019, 10(1): 32 - 36.

[5] 李慧琴. 基于动态令牌的网关服务访问认证的研究与实现[D]. 郑州: 河南大学, 2020.

[6] 黄伟民, 陈可新. 基于 Token 的物联网云平台系统身份认证机制研究[J]. 智库时代, 2018(42): 195 - 196.

[7] 汪昱. 应用程序认证机制安全研究[D]. 西安: 西安电子科技大学, 2018.

[8] Jones M, Hardt D. The OAuth 2.0 Authorization Framework: Bearer Token Usage[EB/OL]. (2020 - 03 - 07). <https://tools.ietf.org/html/rfc6750>.

[9] Jones M. JSON Web Token[EB/OL]. (2020 - 03 - 07). <https://tools.ietf.org/html/rfc7519#section-10.2.1>.

[10] 程治胜. 一种多因子策略混合 Token 签名算法[J]. 电子技术与软件工程, 2019(5): 149.

[11] 顾洁. 面向无线社区的身份认证及密钥管理技术研究[D]. 上海: 上海交通大学, 2012.

[12] 陈晓. 电力企业信息系统中统一身份认证与访问控制应用研究[D]. 北京: 华北电力大学, 2013.

[13] 李春燕. 云计算环境下基于角色的访问控制模型研究[D]. 天津: 天津大学, 2012.

[14] 何修宇. 微服务环境下访问控制技术的研究与应用[D]. 北京: 北京邮电大学, 2018.

[15] 樊昊鹏, 袁庆军, 王向宇, 等. 针对 AES-128 算法的密钥优势模板攻击[J]. 电子学报, 2020, 48(10): 2003 - 2008.

[16] Lopez D. Full-Stack web development with Jakarta EE and Vue.js[M]. Apress Berkeley, 2021.