

基于在线百科的大规模人物社会网络抽取与分析*

林泽斐 欧石燕

摘 要 在线百科词条中蕴含着海量的人物间关系信息,基于这些信息可以抽取大规模社会网络,为数字人文和社会计算研究提供数据支撑。本研究以百度百科为例,首次对面向中文在线百科的大规模社会网络抽取进行探索,提出一种新的人物社会网络抽取方法。该方法利用排序学习综合多种特征计算人物关系权重,通过估计人物生存时空来发现人物间的时空耦合关系。由此,从百度百科中抽取出一个带权重的跨时空人物社会网络和一个时空耦合的人物网络。这两个人物网络具有良好的小世界和无标度特性,并存在清晰的社区结构。最后,通过可视化分析展示了百科人物网络在数字人文研究中的应用模式和应用价值。图 8。表 6。参考文献 39。

关键词 社会网络抽取 社会网络分析 人物社会网络 在线百科 数字人文

分类号 G255.76

Large-Scale People Social Network Extraction and Analysis Based on Online Encyclopedia

LIN Zefei & OU Shiyen

ABSTRACT

Social Network Extraction (SNE) is an emerging research field which focuses on automatic extraction of hidden social networks from a wide variety of information sources. The articles of online encyclopedia contain massive information about persons as well as their interpersonal relationships, from which a people social network can be extracted and used for the research of digital humanities and social computing. The extracted people social network involves both real persons who may span thousands of years and virtual persons who may come from a large number of literary works. However, most of people social network extraction methods ignore the types and spatio-temporal characteristics of persons, and only consider text similarity or other related features to measure the degree of relevance between persons. This may result in restrictions on the accuracy and application field of the extracted people social networks.

This study explored the automatic extraction of a large-scale people social network from Chinese online encyclopedia for the first time by taking Baidu Encyclopedia as an example. It proposed a new method of

* 本文系国家社会科学基金重点项目“基于关联数据的学术文献内容语义发布及其应用研究”(编号:17ATQ001)的研究成果之一。(This article is an outcome of the key project “Research on Linked Data-based Semantic Publishing of Academic Articles and its Application” (No. 17ATQ001) supported by National Social Science Foundation of China.)

通信作者:欧石燕,Email:oushiyan@nju.edu.cn, ORCID:0000-0001-8617-6987 (Correspondence should be addressed to OU Shiyen, Email:oushiyan@nju.edu.cn, ORCID:0000-0001-8617-6987)

social network extraction, which distinguishes the types and spatio-temporal characteristics of extracted persons and more accurately measures the weight of interpersonal relationships based on multiple relevance features. This method contains three phrases—generating an initial people social network, computing the relationship strength between different persons and analyzing the spatio-temporal characteristics of persons. In the first phase, the articles on persons (hereinafter referred to as “person articles”) were identified from Baidu Encyclopedia, and then an initial undirected and unweighted people social network containing more than 0.54 million nodes and 2.22 million edges were generated based on the links between person articles. In the second phase, the strength of the relationships between persons in the initial network was calculated as a ranking task. It was solved with a supervised learning to rank (L2R) method to combine five similarity features for measuring the relevance degree between persons. Based on this method, the initial unweighted people network was then transformed to a weighted network in which person nodes are across time and space. In the third phase, the living time-space of each person in the people network was estimated. For a real person, his/her living time-space was estimated based on the years (including reign titles) occurring in the article on him/her, whereas for a virtual person, his/her living time-space was one or more works depicting him/her. In this way, a time-space coupling network, which contains about 0.45 million nodes and 1.70 million edges, was derived from the previous cross-time-space weighted people network.

The characteristics of the extracted two people social networks were investigated with social network analysis. The results showed that the two networks were both small-world and scale-free networks and have a clear community structure. Furthermore, three types of visual analysis were also performed on the two people networks: point analysis was used to detect related persons of a central person; chain analysis was used to discover the path between two persons (i.e. their direct or indirect relationships); and network analysis was used to reveal high-central persons and person communities in a specific historical or documental time-space. This also indicated that the large-scale social networks extracted from online encyclopedia had great value to support digital humanities research and improve researchers’ perception of historical person age in reality and virtual characters in literary and artistic works. 8 figs. 6 tabs. 39 refs.

KEY WORDS

Social network extraction. Social network analysis. People social network. Online encyclopedia. Digital humanities.

0 引言

社会网络是社会成员(如人物或组织)之间通过一系列社会联系所形成的关系网络^[1]。对社会网络的研究始于20世纪30年代,早期的社会网络研究主要出现在心理学、社会学等领域,多使用问卷调查、行为实验和数学演绎等手段进行社会网络的构建^[2]。20世纪后半叶,复杂网络的兴起为社会网络研究提供了新的视角。

复杂网络具有网络规模巨大、节点联系错综复杂的特性,其社区结构及其动力学原理吸引了多领域学者的关注。在这一阶段,计算机开始被用于生成复杂社会网络并进行社会网络分析(Social Network Analysis, SNA)。近年来,随着网络大数据的兴起,面向在线数据的社会网络抽取成为社会网络研究的新兴热点。社会网络抽取(Social Network Extraction, SNE)是指基于数据集中明确或隐含的信息,从中自动抽取社会成员及其联系的技术^[3]。在网络环境下,

如何利用各类在线社区及其海量、异质、多样的开放数据构建出大规模社会网络,是值得探讨的一大课题。

在线开放数据的一大重要来源是在线百科。在线百科是由网络用户协同编辑形成的在线知识库。自2001年首个在线百科——维基百科发布以来,在线百科已逐渐发展为重要的知识载体:截至2018年7月,维基百科已拥有300余种语言编写的4 535万个词条^[4],最大的中文在线百科——百度百科的词条规模亦达到1 543万条^[5]。在线百科的词条通常对应于一定的实体或概念,其中人物类词条是在线百科词条的重要类型之一,此类词条提供了关于历史人物、当代人物及艺术作品中虚拟人物的各种介绍。通过对在线百科词条中人物及其联系的抽取,可以构建出规模庞大的社会网络,进而为人文研究或社会计算提供数据基础。

在线百科词条中蕴含的人物社会网络(以下简称“百科人物网络”)具有不同于传统社会网络研究中人物社会网络(以下简称“传统人物网络”)的特征:首先,传统人物网络通常来自现实世界或文学作品,多为受限于一定时代截面的横向网络,而百科人物网络则是一个涵盖数千年人物的纵向网络;其次,传统人物网络中的节点通常仅涉及现实人物和虚拟人物二者之一,而百科人物网络既涉及现实人物,也涉及虚拟人物;最后,传统人物网络多为规模有限的静态网络,而百科人物网络是一个包含上百万个节点和联系,且数据源可随词条编辑不断完善的巨大动态网络。鉴于百科人物网络蕴含知识的丰富程度,其在推荐系统、问答系统和知识工程等领域具有重要的应用价值。此外,社会网络构建与分析也是数字人文研究的一种典型形式,当前数字人文研究中的社会网络多提取自纸质档案资料或文学作品^[6],其人物覆盖的广度较为有限,而涵盖多时空、多领域人物关系的百科人物网络可形成数字人文研究的基础性数据。值得注意的是,尽管百科人物网络具有较高的利用价值,但其网络规模和复杂性也给网

络构建与分析带来了困难。

本研究提出一种在线百科人物网络抽取的新方法,该方法将排序学习和人名在搜索引擎中的共现频率应用于百科人物相关度计算中,并利用词条中的时空特征估计相关联人物的时空耦合关系(即人物生存时空的彼此交叠现象)。基于上述方法,本研究从中文百科中首次提取出完整的、具有时空特征的人物社会网络,并对其结构特征进行分析。全文结构如下:第一节对相关研究进行回顾和梳理;第二节以中文百度百科为例阐述社会网络抽取的流程,涉及初始网络生成、关系权重计算和人物时空分析三个主要阶段;第三节采用社会网络分析法对百科人物网络的结构特性进行分析;第四节探索百科人物网络的可视化分析方法及其在数字人文研究中的应用模式;最后提出结论与展望。

1 相关研究

对社会网络抽取的研究可追溯至20世纪90年代Kautz等人开发的Referral Web社会网络抽取系统^[7]。此后大量学者对社会网络抽取问题进行了探索,相关研究主要涉及三个领域:①关注网络抽取:基于社会化媒体中的用户关系数据,根据用户的订阅关系、关注列表或用户间交互关系构建用户间的关注网络^[8-10];②合作网络抽取:根据学术文献中作者(或机构)间的合作和引用关系构建作者(或机构)的合作网络^[11-12];③共现网络抽取:根据文献(如小说、评论、档案、网页)中人物名称在文章、句子或对话中的共现关系,构建人物共现网络^[13-15]。

近年来,越来越多的学者开始关注在线百科中的社会网络抽取问题。Laniado等人根据词条编辑历史计算出词条贡献者对某一词条的编辑数量和质量,以此确定对各词条的高贡献度作者,然后根据词条的共作者情况生成一个社会网络^[16];Massa根据维基百科词条讨论页的用户交互情况生成社会网络^[17];Iba关注到词条

编辑的时序特征,将对词条的连续编辑关系转换为词条贡献者的关联,以此构建出社会网络^[18]。然而,上述研究仅涉及在线百科编辑者的社会网络,其本质仍是合作网络抽取的一种特殊形式。

自2009年开始,有研究者发现百科词条内容所蕴含的社会关系网络(即词条本身所描述的人物及其相互间关系)具有的潜在价值,并逐步开始此方面的研究^[19]。相关研究多以维基百科作为数据源进行人物关系提取,如卢克从维基百科发布的XML数据集中抽取了千余个现实人物信息,并根据两个人物间的结构化信息框(Infobox)键值对的文本相似度大小在人物间建立联系^[19];孟新萍使用290篇维基百科台湾地区人物词条信息框中的“人物姓名—属性—属性值”三元组抽取人物关系,并以人物间的相关度为关系权重,构建出一个无向带权的百科人物网络^[20]。

社会网络抽取存在两大难点问题:一是如何区分多个同名人物,即人名消歧问题;二是如何确定人物间关系的权重与类型。在百科人物网络抽取中,由于在线百科存在着同名人物的自区别机制,即多个同名人物分别对应于不同词条(如李密[隋唐时期群雄之一]、李密[西晋初年官员]),故百科人物网络的抽取通常无需专门进行人名消歧,其网络构建的主要难点在于人物间关系权重与类型的判断。由于百科人物关系类型的复杂性,当前研究侧重于人物关系权重计算而非类型识别。一些研究者从词条文本相似度、人物文本距离等角度提出人物关系权重计算方法。如Yang提出了一种基于相似度的维基百科人物关系权重算法,即对于每对人物关系,分别计算两个人物的词条信息框中所包含的个人信仰、出生年、出生地和人物经历信息以及词条正文的相似度,然后通过系统相似性度量法(Systematic Similarity Measurement)将上述相似度值综合成统一的相似度指标作为人物关系权重^[21];Geiß在人物关系权重计算中考虑了文本距离因素,将两个人物在词条中间隔的

句子数视为人物距离,然后利用指数衰减函数将人物距离转换为关系权重^[22]。

综上所述,当前社会网络研究开始关注到百科人物网络的抽取,并在抽取过程中采用了文本相似度、文本距离等特征对人物间相互关联的程度予以量化,但这些研究仍有一定的局限性:①当前研究在人物关系权重计算方面所考虑的特征类型仍较为有限,主要局限于词条结构化数据和正文的文本特征,且在综合多种特征时多采用系统相似性度量法,该方法需要依靠经验为各特征赋予权重,具有较大的主观性;②百科人物网络是一个跨越多时空环境的社会网络,但人物的时空特征在当前研究中被忽视,尤其是百科人物有现实人物和虚拟人物之分,但未见有研究在网络构建时对二者进行有针对性的处理;③当前百科人物网络抽取的数据源较为单一(仅针对维基百科),且所抽取的人物网络多为局部网络。通过对主流文献数据库的检索,我们未发现目前有利用中文在线百科进行整体人物网络抽取和分析的研究。

针对上述不足,本研究以百度百科为数据来源,对中文百科词条中大规模人物网络的抽取进行了探索:一方面,关注百科人物中的时空特性,对现实人物和虚拟人物进行了区分,并进行人物生存时空估计;另一方面,基于百科词条的非结构化数据设计了多种人物相关度特征,提出了一种基于多特征的人物关系权重学习方法,该方法无需人工设置特征权重,较系统相似性度量法更具客观性;在此基础上,对抽取出的中文百科人物网络进行网络结构特征分析和可视化分析。

2 在线百科人物网络的抽取

本研究抽取的百科人物网络是以百科词条中的人物实体为节点、人物间的联系为边的无向有权网络。整个社会网络抽取框架如图1所示,包括三个主要阶段:初始人物网络生成、关系权重计算和人物时空分析。通过人物网络抽

取,形成“带权跨时空人物网络”和“时空耦合人物网络”两个网络,二者为后续的网络分析与可

视化提供数据支撑。本节主要阐述上述三个阶段的实现方法和技术。

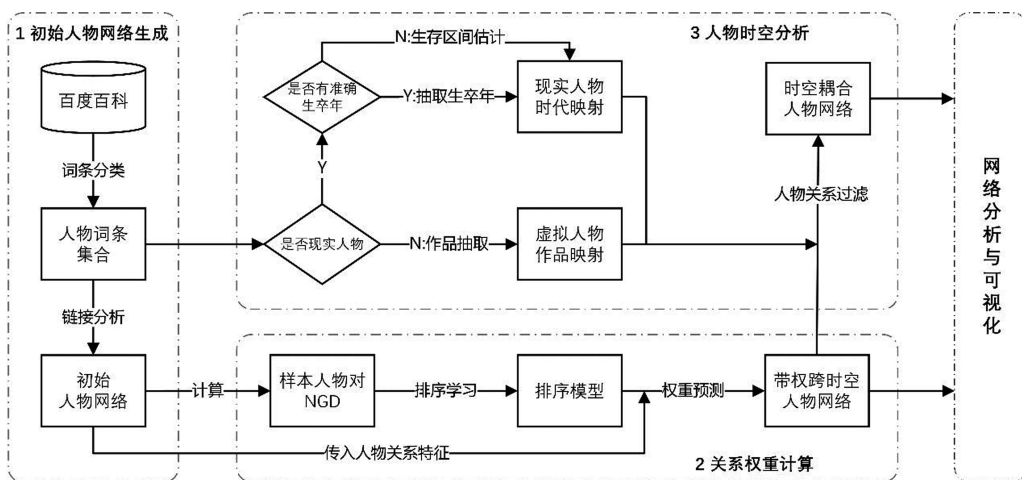


图1 在线百科人物网络抽取框架

2.1 初始人物网络生成

初始人物网络生成这一阶段的目的是从在线百科中筛选出人物类(含现实人物与虚拟人物)词条,并基于人物词条间的链接结构构建出一个无权重的初始人物网络。

本研究采用“基于规则+基于机器学习”的混合方法筛选人物类词条。虽然百度百科提供基于大众分类法(Folksonomy)的用户标签用于标注词条类别,但标注结果具有较强的主观性和随意性,仅以此作为分类依据将会导致准确率偏低。相比用户标签,词条的信息框数据能更加准确反映词条的类型信息,如人物词条信息框常包含“国籍”“出生地”等字段,故我们基于此类数据编写分类规则,若词条信息框中具有人物类别倾向的字段数超过一定阈值,就将该词条归入人物词条。分类规则的构建以保证较高的查准率为原则,暂不考虑查全率。由于并非所有词条均具备足够用于分类的信息框字段数据(少数词条甚至不包含信息框),对于此类词条,则采用基于机器学习的文本分类方法识别其类别,以提升分类的查全率。分类器的

构建以通过规则识别出的人物和非人物类词条数据为训练语料,利用有监督的机器学习算法训练出文本分类器。

我们通过爬虫抓取到百度百科中1 509万余个词条数据,其中,36万余个词条仅为其他词条的同义别名,故具有实质意义的词条有1 473万余个。通过基于规则的分类,共识别出190万余个人物词条和837万余个非人物词条,另外446万余个词条的信息框信息量过少,无法基于规则判定其类别。针对这类词条,以已分类词条正文和信息框文本为训练样本,采用开源文本分类工具fastText^①训练文本分类器,然后利用该分类器预测未分类词条的类别。通过上述混合方法,共识别出227万余个人物词条。百度百科中虚拟人物词条通常带有专门的信息框字段(如“登场作品”)和词条标签(如“文学形象”),我们基于此对现实人物和虚拟人物进行了区分,共得到219万余个现实人物词条和8万余个虚拟人物词条。此外,少数人物同时具有现实和虚拟身份,即一些文学作品将现实人物

① Facebook公司开发的基于词嵌入(word embedding)技术的词向量计算和文本分类工具。

作为虚拟人物形象加以演绎(如《三国演义》中的诸葛亮)。但在百度百科中通常仅将这些人物视为现实人物,并不提供“登场作品”字段。对于此类人物,我们设计了一种双重身份人物识别机制。首先判断每一部文学作品是否具有真实的历史背景,判断依据是:若一部作品中所含的全部虚拟人物的百科词条中出现的年份集中分布于某一特定历史区间(判断阈值根据经验设置为年份均值小于1900且标准差小于150),则说明该作品具有较强的真实历史背景。针对识别出的每一部具有真实历史背景的文学作品,将同时符合三个条件的现实人物标记为双重身份人物:①人物生存时空与该作品历史背景重叠(人物生存时空估计方法见2.3.1节);②词条正文中提及该作品;③与该作品中已识别的虚拟人物存在链接关系。利用这一识别机制,我们

共标注出4449个具有双重身份的人物。

为了检测人物词条识别的准确率,我们从全部1473万余个百度百科词条中由计算机随机抽取出500个词条作为测试样本,并对测试样本的类别进行人工标注,共识别出81个人物词条和419个非人物词条。通过将机器自动识别的词条类型和人工标注相对比,计算出词条分类的查全率、查准率和F1值(见表1)。可以看出,基于规则的词条自动分类具有较高的查准率(98.51%),但查全率较低(81.48%);在其上叠加机器学习方法能使查全率有较大提升(从81.48%上升到98.77%),但由于机器学习不可避免地会带来一定的预测误差,会导致查准率稍有下降(从98.51%下降到94.12%),但总的分类效果还是得到较大提升(F1值从89.19%上升到96.39%)。

表1 人物类词条识别准确率评测

分类方法	查全率(%)	查准率(%)	F1-measure(%)
基于规则	81.48	98.51	89.19
基于规则+基于fastText机器学习	98.77	94.12	96.39

在识别出人物词条后,我们基于人物词条间的内部链接构建初始人物网络。内部链接是反映在线百科词条相关关系的一种机制,若词条作者认为另一词条对理解本词条内容有帮助,可在本词条文本上构建指向目标词条的链接。针对人物类词条,这种词条链接反映了不同人物实体间存在的内在关联,可据此构建出一个初始人物关系网络。具体方法为:在图数据库(本研究采用Neo4j^①数据库)中,将每一个人物类词条所对应的人物设立为节点,并设置值为real或virtual的节点标签,用于区分现实人物和虚拟人物(双重身份人物同时具有两类标签);若任意两个人物词条间存在内部链接关系,则在相应的两个人物节点间建立名为is_related_to的节点联系(即边),从而生成一个初始

人物网络。由于两个人物词条间可能存在多对链接,初始人物网络仅在此类人物节点间建立一条无向边,且暂不考虑边的权重问题。由于百度百科是一个开放编辑的在线百科,故存在大量用户创建的描述其个人的词条(在人物类词条中占比76.3%),这些人物多数未与其他人物发生关联,在网络中成为孤立节点,且对百科人物网络的贡献非常有限,故在初始人物网络中去除此类人物,只保留53.9万余个非孤立人物作为节点(含46.9万余个现实人物、6.4万余个虚拟人物和0.4万余个双重身份人物)。在所抓取的所有百度百科词条间共存在5846.7万余对内部链接,其中链接源与链接目标均为人物类词条的链接数为346.2万余对,合并人物对间的多重链接后得到222.3万余对人物链接。由此,

① Neo4j为Inc.开发的开源图数据库(graph database)管理系统。

构建出的初始人物网络是一个包含了 53.9 万余个节点、222.3 万余条边的无向无权重网络。

2.2 关系权重计算

初始人物网络并未反映出人物间关系的密切程度。关系权重计算的目的是利用人物词条间的多种相关性特征,通过排序学习得到用于计算人物关系密切度的排序模型,从而预测出任意两个存在联系的人物节点间的相关度,即关系权重。

2.2.1 人物相关性特征

根据我们对人物类百科词条的观察,针对人物间的关系给出以下三个假设:①关系密切的人物常具有相同的时代、组织或地理区域背景,故相关度较高的人物词条中更有可能出现相似的词汇;②关系密切的人物词条间通常存在较高频的相互链接与人名共现关系;③关系密切的人物常具有较高的同配性^①。基于以上假设,我们设计了 5 种特征用于度量两个相连人物节点 x 与 y 之间关系的密切程度(即相关度)。

特征 1: x 、 y 人物词条正文的余弦相似度。使用词袋(Bag-of-words)法分别将分词后的 x 、 y 词条正文表示为基于向量空间模型的权重向量 $\vec{d} = [w_{1,d}, w_{2,d}, \dots, w_{N,d}]^T$, 其中 $w_{t,d}$ 为任一词汇 t 在文本 d 中的 TF-IDF 值,然后计算两个向量间的余弦相似度。

特征 2: x 、 y 人物词条间的相互链接次数。

特征 3: x 、 y 人物词条共同被第三方词条链接的总次数。

特征 4: x 、 y 人物名称在对方词条中出现的平均次数。为了识别同一人物的不同名称,我们为初始人物网络中的百科人物建立了一个包含 84 万余个键值对的名字字典,其涵盖了每个百科人物对应的各种不同名称(规范名称、别称、译名、缩略名等)。然后,使用以下公式计算

名称的平均互现次数:

$$NF(x, y) = (\sum_i TF(x, i, y) + \sum_i TF(y, i, x)) / 2$$

其中, $TF(x, i, y)$ 为 x 人物的第 i 个名称在 y 人物词条中出现的词频, $TF(y, i, x)$ 为 y 人物的第 i 个名称在 x 人物词条中出现的词频。

特征 5: x 、 y 人物的同配性特征。使用以下公式度量人物的同配性:

$$AS(x, y) = \frac{\min(IN_LINK(x), IN_LINK(y))}{\max(IN_LINK(x), IN_LINK(y))}$$

其中 $IN_LINK(x)$ 与 $IN_LINK(y)$ 分别为 x 、 y 词条被其他人物词条链接的次数(即入链数)。

2.2.2 人物相关度学习

本研究利用排序学习算法 LambdaMART^[23] 基于上述相关性特征生成统一的人物相关度指标。排序学习(Learning to Rank, L2R)是利用机器学习手段训练排序模型以解决排序问题的方法统称,常用于搜索引擎排序结果的优化。人物关系权重计算也可被视为一种排序问题:给定任一人物 x 以及与 x 存在联系的人物集合 $P = \{p_1, p_2, \dots, p_n\}$, 利用已训练好的排序模型基于相关度特征对集合 P 中人物 $p_1, p_2, \dots, p_n$ 与 x 的相关度进行排序,可得到一个带有偏序的人物列表,人物在列表中的次序反映了该人物与 x 间关系的密切程度。

标注训练数据是排序学习的关键任务,其目标是生成一系列已带有偏序的训练样本用于训练排序模型。我们借鉴了 Referral Web 的设计思路,通过人名在搜索引擎中的共现频率来计算人物间关系权重^[7],以此自动标注训练数据。人名共现频率使用基于搜索引擎的词汇语义相似性测量指标“归一化 Google 距离(Normalized Google Distance, NGD)”^[24]来度量,其计算公式为:

$$NGD(x, y) = \frac{\max(\log f(x), \log f(y)) - \log f(x, y)}{\log N - \min(\log f(x), \log f(y))}$$

① 在社会网络理论中,同配性(Assortativity)指度值相近的顶点倾向于互相连接,反映人物间交往的“物以类聚,人以群分”现象。

其中 $f(x)$ 与 $f(y)$ 分别为搜索引擎单独检索人物名称 x 及 y 的检中网页数; $f(x,y)$ 为搜索同时包含 x 及 y 的检中网页数; N 为搜索引擎索引的网页总数,通常可通过检索某一高频词(如“a”)估计得到。NGD的值域为 $[0, +\infty)$,两人名的NGD值越低表示两者具有越高的相关性。

在本研究中,我们拟从初始人物网络中抽取部分人物(称为种子人物)及其相关人物构成训练样本,并通过计算人物间的NGD值来对其进行标注。鉴于种子人物的相关人物数量过少将无法在后续步骤中进行有效的训练和测试,故我们设定种子人物在初始人物网络中的度中心性须大于30。首先,从初始人物网络中利用计算机随机抽取100名度中心性大于阈值的人物作为种子人物,然后对于每一个种子人物,再从初始人物网络中抽取30名与之有直接联系的相关人物,共形成3 000个人物关系对。对于每一个人物对,通过程序自动向Google发送检索请求并获取检中页面的数量,从而计算出每对人物规范名称的NGD值,并将人物对的关系权重定义为 $1/\text{NGD}$,即得到一组带有偏序的人物关联列表(见表2)。

表2 人物关系权重值计算实例

人物1	人物2	NGD	$1/\text{NGD}$	排序
贾宝玉	薛宝钗	0.255 3	3.916 1	2
贾宝玉	尤三姐	0.346 8	2.883 3	4
贾宝玉	秦可卿	0.302 2	3.308 6	3
贾宝玉	林黛玉	0.198 8	5.027 8	1

接下来,采用2.2.1节的方法计算出每一个人物对的5个相关性特征值,再将其关系权重与特征组合表示成如下特征向量,以此作为一个排序学习样本:

$$(w_{ij}, i, f_{ij1}, f_{ij2}, f_{ij3}, f_{ij4}, f_{ij5}), i = 1, 2, \dots, 100, \\ j = 1, 2, \dots, 30$$

其中 i 为种子人物的ID, j 为相关人物的ID, w_{ij} 为每一人物对的关系权重(即 $1/\text{NGD}$), $f_{ij1} \sim f_{ij5}$ 为每一人物对的5个相关性特征值。

使用上述方法共生成3 000个排序学习样本,这些样本按照种子人物的不同分为100组,将这100组样本按6:2:2的比例分为训练集、验证集和测试集,利用支持LambdaMART排序学习算法的开源工具包Ranklib^[25]在训练集上训练出排序模型,训练过程中由程序自动基于验证集确定超参数的最优取值。为了测试不同特征组合在排序学习中的有效性,将特征1(余弦相似度特征)作为基准(baseline),逐一叠加其他四种特征,共训练出5个排序模型,并利用测试集对人物排序准确率进行评测。评测指标采用归一化折损累积增益(Normalized Discounted Cumulative Gain, NDCG)^[26],NDCG的值域为 $[0, 1]$,其取值越接近1,表示一组排序具有越大的相对增益,即通过模型预测的样本权重次序更接近测试集标注的理想权重次序。我们分别统计了测试集相关度排序的平均NDCG@10、NDCG@20、NDCG@30(即相关度最高的前10、20、30个人物的NDCG)取值,测试结果见表3。

表3 不同特征叠加下的人物排序准确率

模型	排序模型特征	NDCG@10	NDCG@20	NDCG@30
1	特征1(baseline)	0.535 8	0.639 0	0.715 9
2	特征1+2	0.577 4	0.662 1	0.734 0
3	特征1+2+3	0.680 2	0.760 0	0.802 7
4	特征1+2+3+4	0.703 0	0.772 7	0.814 3
5	特征1+2+3+4+5	0.759 6	0.806 6	0.843 5

评测结果显示,各特征叠加后 NDCG 均比未叠加前出现较明显提升,说明各相关度特征均对相关度计算具有一定效用。表现最好的第 5 组特征组合,其 NDCG@k 准确率介于 0.759 6—0.843 5 之间,说明预测的排序结果接近人物对的真实 NGD 次序,具有较好的预测效果。

最后,针对初始人物网络中的每一个人物,利用训练出的最佳排序模型(即模型 5:基于所有相关度特征)对其所有相关人物进行排序,然后基于排序位次计算出两个人物节点间的边权重 w ,计算公式如下:

$$w = \min(\text{Percentile}(\text{Rank}_{xy}), \text{Percentile}(\text{Rank}_{yx}))$$

其中, x 和 y 为边两端的人物节点, Rank_{xy} 代表模型预测出的 x 人物在 y 相关人物中的位次, $\text{Percentile}(\text{Rank}_{xy})$ 为该位次在 y 所有相关人物位次中的百分位数;同理, $\text{Percentile}(\text{Rank}_{yx})$ 代表 y 的位次在 x 所有相关人物位次中的百分位数。在真实社会中,两个人物间关系的密切程度往往取决于较疏远的一方,因此本研究取两个百分位数中的较小值作为边权重,其值域为 $[0, 100]$ 。

由此,通过关系权重计算,无权的初始人物网络被转换为一个带权的百科人物网络,其包含 53.9 万余个人物节点,222.3 万余条边,边权重反映出两个相关联人物的相关性大小。在该网络中,一个人物可与任意时空的另一个人物产生直接的跨时空联系,譬如,不同时期的历史人物朱熹与孔子间可存在直连的边;小说《红楼梦》的作者曹雪芹与小说中的虚拟人物贾宝玉亦存在直接关联。因此,其实质是一个相邻人物间允许存在跨时空联系的网络。为了与下一节中的“时空耦合人物网络”相区别,本文将该人物网络称为“跨时空人物网络”。

2.3 人物时空分析

一个百科人物,无论是真实的还是虚拟的,均具有时空特性:对于真实人物,时空特指其所处的时代背景;对于虚拟人物,则特指其所处的艺术作品(含小说、漫画、戏剧、影视作品等)背

景。在前文中,我们基于百科人物词条间的链接关系生成了一个跨越时空的人物网络,但在传统社会网络研究中,主要关注处于同时空域人物间的关联,即与一个现实人物产生直接关联的现实人物应同属一个历史时期;与一个虚拟人物直接关联的虚拟人物应同属于同一部文学作品。在社会网络研究中,对跨时空人物关联与同时空人物关联进行研究具有各自不同的价值:通过对跨时空人物关联进行研究,可发现人物间可能的思想传承和渊源关系,而同时空人物关联则对于某一历史时期或文献作品的微观研究具有辅助作用。因此,在本节中,我们通过跨时空人物网络中人物所处的时空进行分析估计,从中衍生出一个时空耦合网络,可对同时空人物关联的研究提供支持。在系统科学中,“耦合(Coupling)”指两个或两个以上的系统彼此影响以致联合起来的现象^[27]。在本研究中使用“时空耦合”一词指代相关联人物生存时空的彼此交叠现象。例如,雍正(1678—1735)与乾隆(1711—1799)、乾隆与和珅(1750—1799)这两对相关人物间分别存在着时空的耦合;《围城》中的方鸿渐与唐晓芙之间具有直接的社会关系,且同属一部作品,故亦存在时空的耦合。通过分析人物的生存时空,可从前述跨时空人物网络中提取出大量存在时空耦合的人物对,并将它们相互联结,从而衍生出一个时空耦合人物网络。该网络与跨时空人物网络的区别在于,任意两个存在直接关联的人物不允许存在跨时空的联系。通过该时空耦合网络,可以快速得到与任一现实人物或虚拟人物处于同时空的相关人物;同时,对该网络进行切片,可以获得不同历史时空或作品时空的子图,这些子图反映出同时处于某一时空域的人物社会网络(即同时空人物网络)。

对人物生存时空进行估计是构建时空耦合人物网络的关键问题,本节首先讨论百科人物生存时空的估计方法,然后说明如何利用该方法构建时空耦合的人物网络。

2.3.1 人物生存时空估计

本研究根据现实人物和虚拟人物的特点,

对两类人物采用了不同的时空估计方法。

(1) 现实人物生存时空估计

在百度百科中,根据现实人物生卒信息的完整程度可将现实人物分为两类:①人物具有完整的生卒年,并以“出生日期”“逝世日期”等字段在信息框中体现,经统计此类人物在现实人物中占比仅 14.3%;②人物无生卒年信息或仅具有生卒年其一,多数人物属于此类。对于第一类人物,可利用正则表达式从信息框中抽取出生准确的生卒年;对于第二类人物,需要对其生卒年进行估计,为此我们设计了一种基于统计的生存区间估计方法。该方法利用正则表达式从每个百科人物词条正文中抽取所有公元年份(若存在年号纪年,则将其提取并映射为公元纪年),并将一个词条中提取出的所有公元年份作为该人物生存区间年份的一组样本观测值,据此估计最可能的人物真实生存区间。

多数人物词条在人物的真实生存区间具有

较高的年份频率分布,整体呈钟形曲线形态,但也有部分人物的年份分布呈多峰形态,出现该情况的原因多为历史人物词条中提及后代对该人物的研究、影视作品或纪念活动,导致人物逝世后仍出现较小的波峰。譬如,根据郑板桥(1693—1765)及林则徐(1785—1850)词条中提取的年份观测值所生成的年代分布均呈现这一特征(见图2)。根据上述特点,我们引入了活跃区间(Active Interval)这一概念,将人物年代分布中年份频数最高的一段时间窗口认定为该人物的活跃区间,其作用是从多峰分布中析出人物实际生存期所在的大致区间。根据人类的一般寿命,我们将活跃区间的窗口大小设定为 100 年,并用程序将窗口以滑窗形式从坐标轴上的最小年份观测值逐年滑动至最大观测值,以匹配出年份频数最高的活跃区间。落在人物词条活跃区间中的年份观测值被视为是与人物真实社会活动年份有关的一组样本。

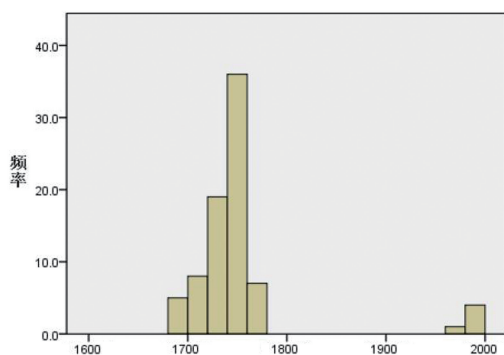


图 2a 郑板桥词条中年份分布直方图

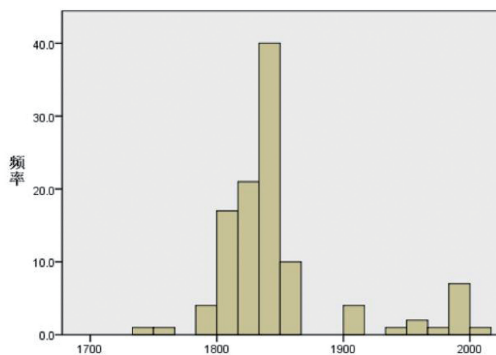


图 2b 林则徐词条中年份分布直方图

我们假设,活跃区间内年份分布的离散程度与人物寿命长度具有相关性,即寿命较短的人物其年份分布较为集中。为了验证上述假设,我们随机抽取了 10 000 个信息框中带有完整生卒年的人物,将这些人物按 4 : 1 分为训练集和测试集,并计算每个人物活跃区间内年份的标准差 s 以及真实寿命长度 y (基于信息框中逝世年份—出生年份)。在训练集中,利用 SPSS 的双变量相关性分析模块检验年份标准差 s 与人物寿命

y 的相关性,结果显示 Pearson 相关系数为 0.673 ($P < 0.01$),说明两者存在显著正相关。接下来,利用 SPSS 的线性回归分析模块,得到年份标准差 s 和人物寿命 y 这两个变量的散点图并进行线性拟合(见图3),所拟合的回归模型为 $y = 25.2 + 1.88s$ 。该模型的拟合度 $R^2 = 0.452$, R^2 越接近 1 说明模型的解释能力越强,一般认为在社会科学中 $R^2 > 0.25$ 即具有良好的拟合优度^[28],故该模型可被用于估计人物寿命长度。

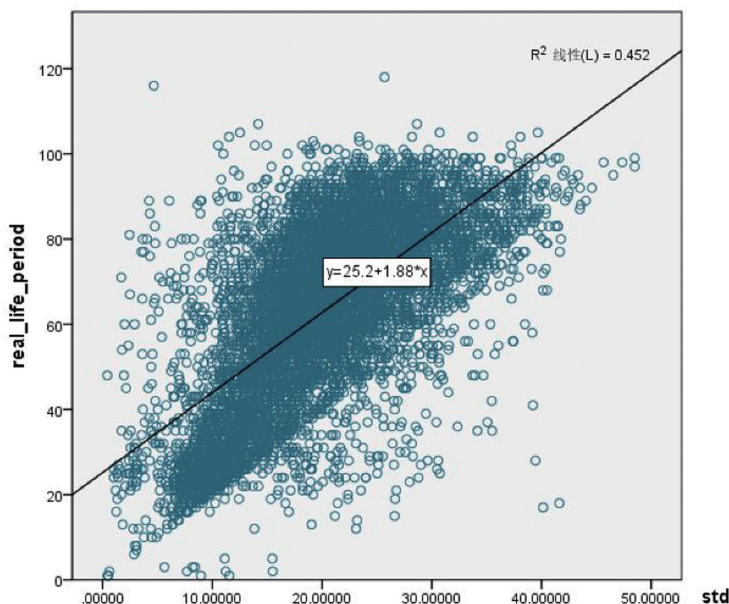


图3 活跃区间年份标准差与人物真实寿命关系拟合

得到人物寿命长度后,若能再取得人物生存区间的中值年份,即可估算人物的实际生存区间,即令人物生存区间= $m \pm 0.5y$ 。其中, m 为生存区间的中值年份,其值为(实际出生年份+实际逝世年份)/2, y 为寿命长度。对于信息框中无完整生卒年的人物,其 m 值需通过估计的方式得到。通常, m 与人物社会活动的活跃年份均值 μ 较为接近。通过将活跃区间内的年份均值 $\bar{x}$ 作为 μ 的点估计量,即可间接估算出 m 的取值。我们计算出训练集中所有人物的活跃区间年份均值 $\bar{x}$,并将其与各个人物实际生存区间中值年份 m 进行对比,结果显示两者离差($\bar{x} - m$)的中位数为8.34年,说明 $\bar{x}$ 与 m 较为接近且 $\bar{x}$ 略大于 m 。考虑到人物幼年社会活动在历史中通常少有记载,故有记载的人物社会活动年份均值理应略大于 m ,因此,将 $\bar{x}$ 作为 m 的估计量具有合理性。综合上述指标,我们使用公式 $(\bar{x}-c) \pm 0.5y$ 估计人物的生存区间,其中 c 为常数项,代表人物幼年期所导致的 $\bar{x}$ 与 m 间的偏移量,本研究中设 $c=8.34$; y 为采用回归模型估算的人物寿命长度。

为了测试估计的人物生存区间的准确程度,我们采用测试集中的人物进行测试。为保证测试的公平可靠,首先对测试集进行预处理,去除了词条正文中可提取年份数过少(<3)的354个词条(视为无法估计生存区间的人物),并将剩余词条中人物的真实生卒年去除。接下来,将估计的生存区间与真实生存区间进行对比,评估指标使用Dice系数,其公式为:

$$Dice(\hat{I}, I) = \frac{2 \times |\hat{I} \cap I|}{(|\hat{I}| + |I|)}$$

其中, $\hat{I}$ 和 I 分别代表经离散化的估计生存区间与真实生存区间中的年份集合。 $Dice(\hat{I}, I)$ 越接近1,代表两区间重叠程度越高。经统计,测试集样本Dice系数的中位数为0.83,说明估计的区间与人物真实生存区间较为接近,故本研究所采用的现实人物生存区间估计方法具有有效性。针对初始人物网络中无准确生卒年的现实人物和双重身份人物,我们使用该方法估计出其中37.1万余个人物的生存区间,另有3.5万余个人物因样本年份数过少无法估计。

(2) 虚拟人物生存时空估计

虚拟人物的生存时空即为其相关的作品。不同于现实人物,一个虚拟人物可以出现于多个时空中,譬如“二郎神”这一角色在《西游记》《封神演义》中均有出现。通过在人物词条正文和信息框中匹配所有书名号内字符,即可得到与虚拟人物相关的作品集合。对于每个虚拟人物,抽取其全部相关作品,并根据作品名称在人物词条中出现的频率确定该人物相关作品的重要性排序。我们从初始人物网络的所有虚拟人物和双重身份人物词条中识别到共计 5.5 万余种不同的作品时空,另有 0.1 万余个虚拟人物因词条中无任何作品信息无法对其时空进行估计。

2.3.2 时空耦合网络的构建

通过时空估计得到人物的生存时空后,即可根据人物的时空重叠关系构建出时空耦合的人物网络。对于每一对生存时空重叠(即时空耦合)的人物节点,在两者之间创建一条名为 has_overlap_with 的关系(即边),其关系权重与

跨时空人物网络中两者间的关系权重相同。

图 4 为时空耦合人物网络中现实和虚拟人物时空耦合关系的微观实例。Neo4j 等图数据库允许为关系创建属性,我们为现实人物间的 has_overlap_with 关系添加了名为 upper_limit 和 lower_limit 的两个属性,用于识别人物对重叠生存时空的上下限年份;为虚拟人物间的 has_overlap_with 关系添加了名为 work 的属性,以区分每对耦合关系所涉及的作品时空。需要说明的是,由于一个虚拟人物可存在于多个作品时空,故任意两个虚拟人物间可能存在多条 has_overlap_with 关系,表示两人在多部作品中同时出现。例如,“郭靖”与“郭啸天”这一人物对仅同时出现于作品《射雕英雄传》中,故两者间存在一条 has_overlap_with 关系,而“郭靖”与“黄蓉”则同时出现于《射雕英雄传》和《神雕侠侣》两部作品中,故两者间存在两条 has_overlap_with 关系,每条 has_overlap_with 关系通过 work 属性对作品时空进行了区分。

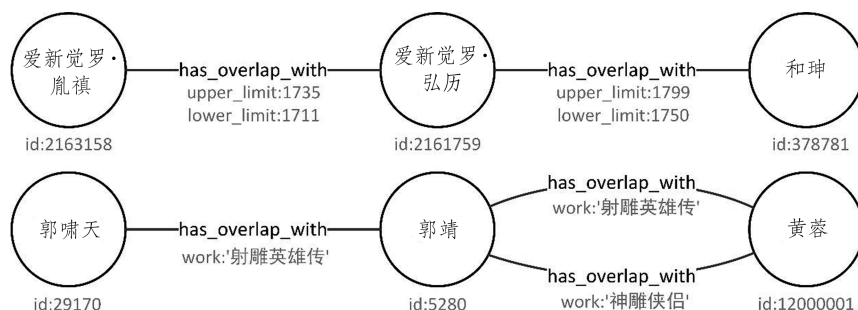


图 4 现实和虚拟人物的时空耦合关系微观实例

基于上述策略,我们从跨时空人物网络中衍生出一个时空耦合的人物网络,该网络由 44.5 万余个节点,169.5 万余条边构成。在时空耦合网络中,任意两个直接关联的人物均存在生存时空的重叠。通过判断网络中边的 upper_limit 和 lower_limit 属性值是否涵盖特定年份,以及 work 属性值是否与特定作品名一致,即可从该网络中快速识别出在某一特定历史或虚拟时空存在关联的人物对,基于这些人物对即可生

成该时空域的人物社会网络子图。

3 在线百科人物网络的结构特征分析

社会网络分析是对社会网络的结构及其特征进行分析的过程。从整体层面对百科人物网络的拓扑结构特征进行分析,有助于考察其与现实社会人物网络的异同,同时是对其进行深入和有效应用的前提。社会网络分析理论认

为,现实社会网络通常具有小世界和无标度特性,并存在清晰的社区结构^[29]。本节拟对前文中从百度百科所抽取的跨时空人物网络和时空耦合网络的这些结构特性进行分析。

小世界网络 (Small-World Networks) 是一类特殊的复杂网络结构,其多数节点并不直接相连,但大部分节点间可经由少数几步跳转到达^[30]。和同等规模的随机图相比,小世界网络的聚类系数 (Clustering Coefficient) 远大于随机图,但平均路径长度 (Average Path Length) 与之相当。本研究使用 Humphries 等人提出的“小世界系数”^[31]对网络的小世界性进行计量:

$$\sigma = \frac{c/c_r}{L/L_r}$$

其中, c 与 c_r 分别为待测量网络和同等规模随机图的聚类系数, L 与 L_r 分别为两者的平均路径长度,若 $\sigma > 1$ 则可将该网络定义为小世界网络。为了与跨时空和时空耦合人物网络相对比,我们通过 igraph^①生成同等规模(节点数和边数均相同)的两个 E—R (Erdős—Rényi) 随机图,并计算出跨时空和时空耦合网络的聚类系数 c 和平均路径长度 L 、同等规模随机图的聚类系数 c_r 和平均路径长度 L_r ,以及基于上述指标计算的 σ ,结果如表 4 所示。

表 4 百科人物网络的小世界性

网络类型	c	c_r	L	L_r	σ
跨时空网络	0.200 8	1.58E-5	5.607 1	6.496 3	14 771.59
时空耦合网络	0.237 9	1.57E-5	7.996 2	6.640 4	12 559.31

分析结果可知,两网络相比同等规模的 E—R 随机图均具有更大的聚类系数,说明两网络的集聚程度较高;两网络的平均路径长度与随机图处于同一量级,并显示出社会网络中典型的“六度分隔”特性,其中时空耦合网络的平均路径长度略大于跨时空网络,这反映出在时空耦合网络中去除人物对间的跨时空联系后导致任意两人物间隔增加;两网络的 σ 均远大于 1,说明两网络均为典型的小世界网络。

无标度网络 (Scale-free Networks) 是指节点的度分布服从幂律分布的一类复杂网络,可使用公式 $P(k) \propto k^{-\gamma}$ (即 $P(k)$ 正比于 k 的某个负幂次 $-\gamma$) 判断网络是否具有无标度特性^[32]。其中, $P(k)$ 为网络中节点的度为自然数 k 的概率, γ 为度指数,其典型值为 $2 < \gamma < 3$ ^[33]。在双对数坐标系^②中,无标度网络的节点度分布表现为一条斜率介于 -2 — -3 的直线。对于跨时空人物网络和时空耦合网络,首先利用程序计算并输

出网络中所有人物节点的度,并以散点图形式绘制在双对数坐标系中,然后通过线性回归分析得到二者的拟合直线方程,如图 5 所示。由图 5 可以看出,两个网络的度分布十分相似,且两个回归方程的决定系数 R^2 均接近 1,反映出两个网络的度分布在双对数尺度下均呈现出明显线性趋势。度指数 γ 分别为 2.034 和 2.004,均属于 γ 的典型值范围内,说明两个网络具有一定的无标度特性。

在无标度网络中,节点的度分布并不均匀:多数节点的中心度较低,而少数节点(称为集散节点)具有高中心度,这些集散节点对整个网络的凝聚具有重要作用。表 5 列出了两个网络中的高中心度节点。由表 5 可以看出,近现代中国政治人物在百科人物网络中具有较高的度中心性,其次为当代娱乐人物和历史文化人物。虚拟人物的度中心性总体低于现实人物,故未在 TOP10 人物中出现。

① 基于 C 语言的高性能网络分析库,支持对数百万级节点的复杂网络进行网络分析。

② 双对数坐标系是将两坐标轴单位长度进行对数变换后得到的平面坐标系。

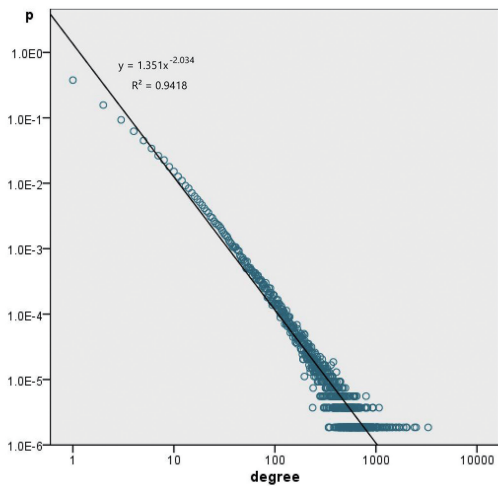


图 5a 跨时空网络的度分布

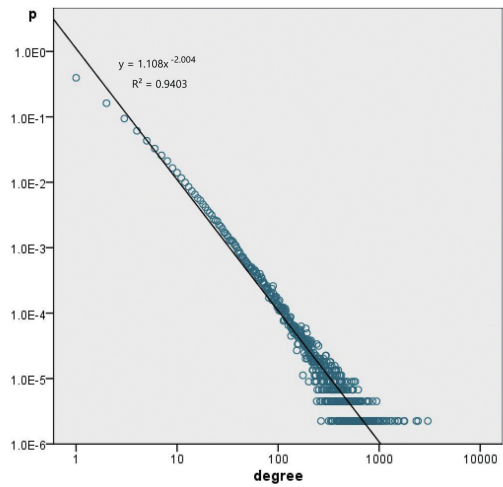


图 5b 时空耦合网络的度分布

表 5 百科人物网络中度中心性 TOP10 人物

跨时空网络 TOP10	度中心性	时空耦合网络 TOP10	度中心性
毛泽东	3 288	毛泽东	3 028
周恩来	2 499	周恩来	2 425
蒋介石	2 402	蒋介石	2 334
孙中山	1 991	刘德华	1 781
刘德华	1 874	孙中山	1 755
齐白石	1 708	袁世凯	1 561
袁世凯	1 689	成龙	1 526
成龙	1 602	邓小平	1 434
王羲之	1 558	周杰伦	1 327
鲁迅	1 450	齐白石	1 285

社区发现 (Community Detection) 是从社会网络中发现社区结构的聚类算法。社区发现算法可分为图分区、结构聚类、贪婪算法、标签传播等十余种类型^[34]。其中, 标签传播 (Label Propagation) 算法是 Raghavan 等人提出的一种高效的社区发现算法, 其优点在于无需预先给定社区的数量, 且迭代的时间复杂度接近于线性^[35]。我们通过标签传播算法对跨时空和时空耦合人物网络进行了社区探测, 分别从这两个

网络探测到 43 131 和 47 724 个社区。表 6 列出了两个网络中探测到的人数最多的 5 个社区, 可以看出百科人物网络中的社区划分与常识较为一致, 这可从一个侧面反映出两个网络社区结构的有效性。

此外, 我们还使用 Newman 和 Girvan 提出的模块度^[36] (Modularity) 指标测试了社区划分的质量。模块度的值为 $[-1, 1]$, 其取值越接近 1, 代表网络社区结构的强度越强。一般认为,

表6 百科人物网络人数 TOP 5 社区

	排序	人数	核心人物 (按度中心性降序排列)
跨时空网络	1	256 198	毛泽东 周恩来 蒋介石 孙中山 齐白石 王羲之 鲁迅...
	2	18 198	哈利·戴恩·斯坦通 海伦·米伦 约翰尼·德普 汤姆·克鲁斯...
	3	9 050	小栗旬 佐藤浩市 斋藤工 木村拓哉 成宫宽贵 阿部宽...
	4	7 173	李宏图 董圆圆 靳学斌 马连良 荀慧生 景荣庆 吕扬...
	5	6 501	宋慧乔 朴信惠 河智苑 苏志燮 宋承宪 金喜善 韩彩英...
时空耦合网络	1	165 294	毛泽东 周恩来 蒋介石 孙中山 袁世凯 邓小平 齐白石...
	2	8 419	小栗旬 斋藤工 佐藤浩市 木村拓哉 阿部宽 成宫宽贵...
	3	5 469	宋慧乔 朴信惠 河智苑 宋承宪 金喜善 苏志燮 权相佑...
	4	4 636	李自成 崇祯 吴三桂 张献忠 董其昌 郑成功 洪承畴...
	5	4 285	曹操 刘备 诸葛亮 曹丕 袁绍 关羽 刘表 孙策...

模块度大于 0.3 则代表网络具有明显的社区结构,真实世界社区的模块度通常介于 0.3—0.7 之间^[37]。本研究所构建的跨时空网络的模块度为 0.387 8,时空耦合网络的模块度为 0.596 1,两者均大于 0.3,说明两个网络均具有良好的社区结构,且时空耦合网络的社区划分具有更高的结构强度。

4 百科人物网络的可视化分析及应用

可视化分析是对抽象数据进行交互式的可视化表示,从而增强人类对信息认知能力的分析方法^[38]。通过对百科人物网络进行可视化分析,有助于发现人物关联中隐含的信息和规律,进而可辅助用户利用大规模人物网络开展学术研究。同时,百科人物网络的可视化也可被用于文献推荐系统和知识发现系统等应用,以提升用户对知识单元间关联的感知。本节以从百度百科词条中抽取出的跨时空和时空耦合这两个人物社会网络为例,通过我们设计的一套名为 EncycloGraph^①的可视化原型系统,阐述百科人物网络在数字人文研究中的一种应用模式。

EncycloGraph 中设计了点状、链状和网状三种人物可视化分析模式。

(1) 点状分析

点状分析是以百科人物网络中的任一人物(无论现实或虚拟)为中心,展示该人物的同时空或跨时空相关人物节点,节点间的距离反映出它们相关性的大小(基于 2.2 节的人物相关性计算方法)。图 6a 所示为以民国著名人物“陆小曼”为核心的时空耦合网络,涉及与其存在时空重叠的 45 个现实人物;若选择“允许跨时空联系”,则该人物网络可扩展至包含陆小曼的当代研究者、陆小曼的形象饰演者等 55 个人物。图 6b 所示为虚拟人物“贾宝玉”的时空耦合网络,该网络涉及《红楼梦》中与贾宝玉相关的 111 个人物形象,其跨时空网络除包含上述虚拟人物外,还包含“欧阳奋强”等与贾宝玉人物形象有关的现实人物。在点状分析视图中,研究者可通过双击图中的任意人物跳转至以该人物为中心的视图,实现“滚雪球”式的人物关联分析。

(2) 链状分析

链状分析是根据任意指定的两个人物生成并展示人物间的最短路径。本研究使用 Dijkstra

① <http://e.dhgraph.com>

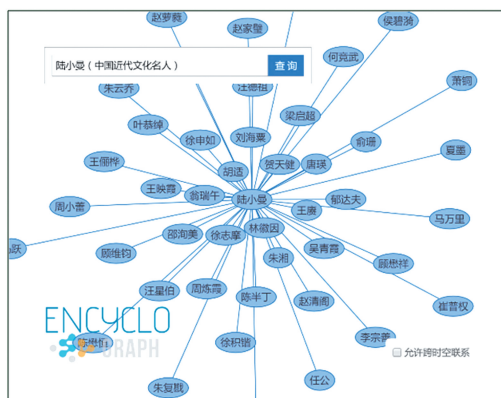


图 6a 现实人物点状分析实例

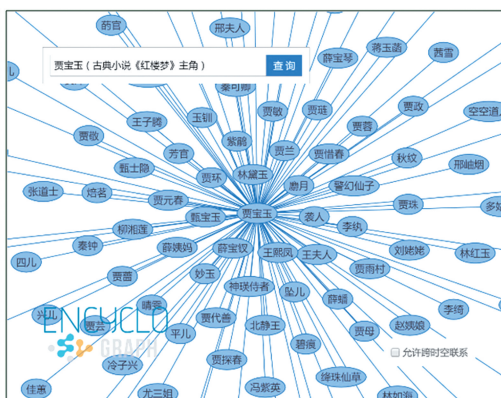


图 6b 虚拟人物点状分析实例

算法计算最短路径,该算法是一种基于广度优先搜索的、面向赋权图的经典最短路径算法^[39]。图 7a 所示为民国时期著名女性作家“张爱玲”和“冰心”之间的最短路径,可以看出两人可通过“胡适”和“梁实秋”两个中介节点相互连接,查询相关资料可知,张爱玲与冰心两人并不直接结识,但张爱玲与胡适、胡适与梁实秋、梁实秋与冰心互为好友,故张爱玲与冰心两人存在

间接的社会联系;图 7b 所示为《红楼梦》人物“林如海”和“林之孝”的分析视图,若读者试图快速判断两同姓人物是否存在直接的亲戚关系,可从图中发现两人物间存在两个中介节点,表明二者关系较为疏远,进一步检索文献可知林如海与林黛玉为父女关系,林黛玉和王熙凤存在亲戚关系,而林之孝为王熙凤手下的管家,故可确认两人并不存在直接联系。

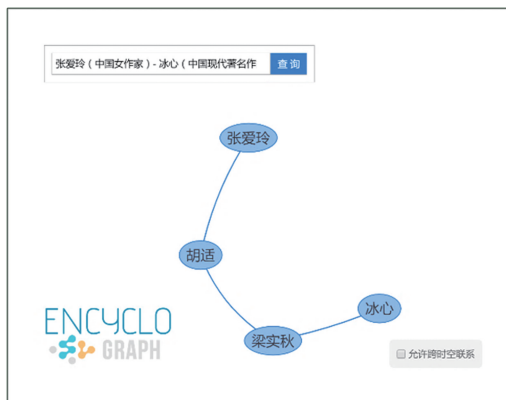


图 7a 现实人物链状分析实例

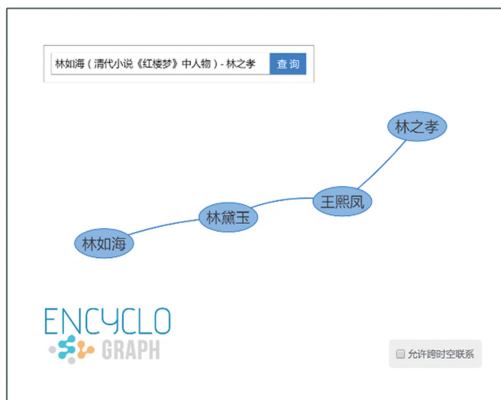


图 7b 虚拟人物链状分析实例

(3) 网状分析

网状分析是根据指定的现实时空(指定起止公元年份)或虚拟时空(指定特定的作品名称),生成该时空截面中的重要人物节点及其联系,其中节点的距离和大小分别反映人物的

相关度和度中心性的大小。图 8a 所示为公元前 202 年至公元 8 年(即西汉时期)这一现实时空中的重要人物及其相互关系;图 8b 所示为《西游记》这一虚拟时空中的重要人物及其相互联系。在网状分析中,由于关系密切的人物具有

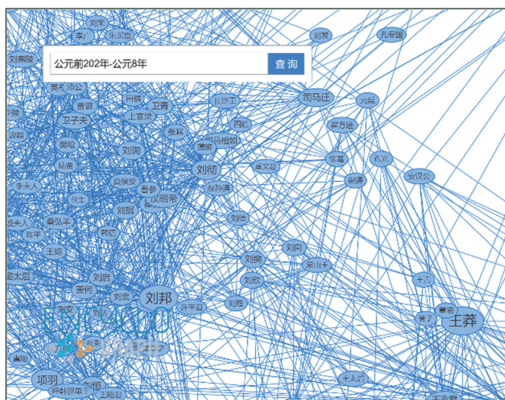


图 8a 现实时空网状分析实例

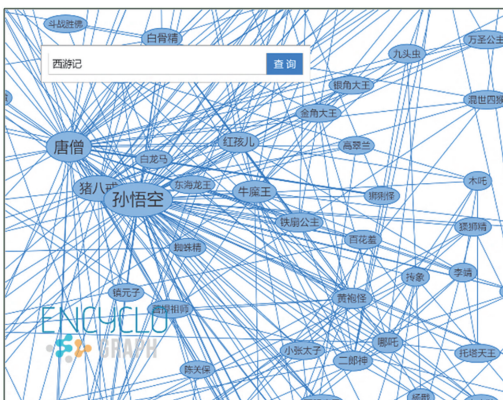


图 8b 虚拟时空网状分析实例

较小的节点距离,故一组高相关度人物可在视觉上集聚成簇。

通过上述可视化分析实例可知,点状分析适用于人文研究中针对某个待研究的现实人物或艺术作品人物,研究者试图通过发掘相关人物以扩展其研究范围的应用场景;链状分析可帮助研究者判断人物间是否存在直接或间接的关联,其中人物间的间接联系往往对人文研究的创新具有一定启发性;网状分析则能够揭示出特定历史时空与文献时空中人物间的集聚现象,并发现时空中的高中心度人物。借助上述可视化分析技术,可动态地从百科人物网络中提取并呈现面向不同人文研究领域的子图,从而为人文学者开展特定领域的研究工作提供参考。

5 总结与展望

本研究以百度百科为例,对基于在线百科的大规模人物社会网络抽取机制进行了探索,提出了一种新的在线百科人物网络抽取方法,其由初始人物网络生成、关系权重计算和人物时空分析三个阶段构成。该方法的主要创新点在于:①将排序学习应用于人物关系权重计算,综合5种人物相关度特征进行人物相关度排序,并基于排序位次计算人物关系权重;②在百科人物网络中区分现实人物和虚拟人物,并分

别针对现实人物和虚拟人物提出了人物时空估计方法,以此确定人物的时空耦合关系。通过此方法,我们首次从中文百科中抽取出完整的带权重跨时空人物网络,并通过进一步对人物所处时空进行分析估计,从中衍生出一个时空耦合的百科人物网络。通过对这两种百科人物网络的拓扑结构特征进行网络分析,发现百科人物社会网络具有小世界特性、无标度特性和良好的社区结构。此外,我们还提出了三种基于百科人物网络的可视化分析方法,以此展示百科人物网络在数字人文研究中的可能应用模式。

本研究目前还存在一些局限性:①百科人物间的关系具有多种类型(如亲人、朋友、敌人),尽管我们区分了跨时空和时空耦合关系,并通过权重来反映关系的密切程度,但暂未对人物关系的类型进行进一步区分;②尽管百科人物网络可揭示数十万现实人物和虚拟人物间的联系,但仍无法涵盖人文研究所涉及的海量人物及其关系。针对上述问题,在下一步工作中,我们拟通过关系抽取机制识别百科人物间的关系类型,以丰富百科人物网络的语义;其次,探索基于非结构化文献的人物网络抽取,实现对百科人物网络的补充和扩展,并利用文献信息为人物关系识别提供事实依据;最后,对基于在线百科的非人物实体(如组织机构、地点)社会网络抽取机制进行探索。

参考文献

- [1] Thai M T, Wu W, Xiong H. Big data in complex and social networks[M]. London: CRC Press, 2016: 67.
- [2] Borgatti S P, Mehra A, Brass D J, et al. Network analysis in the social sciences[J]. Science, 2009, 323 (5916): 892-895.
- [3] Arif T. Social network extraction using web mining techniques[D]. BGSB University, 2015.
- [4] Size of Wikipedia[EB/OL]. [2018-07-05]. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.
- [5] 百科词条数统计[EB/OL]. [2018-07-05]. <https://baike.baidu.com>. (Size of Baidu Encyclopedia[EB/OL]. [2018-07-05]. <https://baike.baidu.com>.)
- [6] 林泽斐, 欧石燕. 美国高校数字人文研究项目研究内容解析[J]. 图书情报工作, 2017, 61(22): 52-58. (Lin Zefei, Ou Shiyan. An analysis of the research contents of digital humanities research projects in American universities[J]. Library and Information Service, 2017, 61(22): 52-58.)
- [7] Kautz H, Selman B, Shah M. Referral web: combining social networks and collaborative filtering[J]. Communications of the ACM, 1997, 40(3): 63-65.
- [8] Catanese S A, De Meo P, Ferrara E, et al. Crawling Facebook for social network analysis purposes[C]// Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011: 52-60.
- [9] Purohit N S, Angadi A B, Bhat M, et al. Crawling through web to extract the data from social networking site-Twitter[C]// Proceedings of 2015 National Conference on Parallel Computing Technologies. IEEE, 2015: 1-6.
- [10] Rieder B. Studying Facebook via data extraction: the Netvizz application[C]// Proceedings of the 5th Annual ACM Web Science Conference. ACM, 2013: 346-355.
- [11] Elmacioglu E, Lee D. On six degrees of separation in DBLP-DB and more[J]. ACM SIGMOD Record, 2005, 34 (2): 33-40.
- [12] Tang J, Zhang J, Yao L, et al. Arnetminer: extraction and mining of academic social networks[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 990-998.
- [13] Elson D K, Dames N, McKeown K R. Extracting social networks from literary fiction[C]// Proceedings of the 48th Annual Meeting of ACL. ACL, 2010: 138-147.
- [14] Agarwal A, Kotalwar A, Rambow O. Automatic extraction of social networks from literary text: a case study on alice in wonderland[C]// Proceedings of the Sixth International Joint Conference on Natural Language Processing. IJCNLP, 2013: 1202-1208.
- [15] 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(2): 99-106. (Zhao Jingsheng, Zhang Li, Zhu Qiaoming, et al. Extracting and analyzing social networks from Chinese literary[J]. Journal of Chinese Information Processing, 2017(2): 99-106.)
- [16] Laniado D, Tasso R. Co-authorship 2.0: patterns of collaboration in Wikipedia[C]// Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia. ACM, 2011: 201-210.
- [17] Massa P. Social networks of Wikipedia[C]// Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia. ACM, 2011: 221-230.
- [18] Iba T, Nemoto K, Peters B, et al. Analyzing the creative editing behavior of Wikipedia editors: through dynamic social network analysis[J]. Procedia-Social and Behavioral Sciences, 2010, 2(4): 6441-6456.
- [19] 卢克. 基于 Wikipedia 的社会网络挖掘[D]. 哈尔滨: 哈尔滨工业大学, 2009. (Lu Ke. Social network mining based on Wikipedia[D]. Harbin: Harbin Institute of Technology, 2009.)
- [20] 孟新萍. 维基百科人物属性自动获取及社会网络构建方法研究[D]. 沈阳: 东北大学, 2011. (Meng Xinping. Research on automatic person attributes extraction and social network construction from Wikipedia[D]. Shenyang: Northeastern University, 2011.)

- [21] Yang F, Xu Z, Li S, et al. Social network mining based on Wikipedia[C]// 2010 International Conference on Asian Language Processing. IEEE, 2010: 223-226.
- [22] Geiß J, Spitz A, Gertz M. Beyond friendships and followers: the Wikipedia social network[C]// 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2015: 472-479.
- [23] Burges C J C. From Ranknet to LambdaRank to LambdaMART: an overview[R]. Redmond: Microsoft, 2010: 1-19.
- [24] Cilibrasi R L, Vitanyi P M. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383.
- [25] Van D. RankLib[EB/OL]. [2018-09-05]. <https://sourceforge.net/p/lemur/wiki/RankLib>.
- [26] Wang Y, Wang L, Li Y, et al. A theoretical analysis of NDCG ranking measures[C]// Proceedings of the 26th Annual Conference on Learning Theory. Princeton, 2013: 1-30.
- [27] 刘景燕. 科学方法论[M]. 北京: 中国财富出版社, 2014: 91.(Liu Jingyan. Scientific methodology[M]. Beijing: China Fortune Press, 2014:91.)
- [28] 戴维 R 安德森, 丹尼斯 J 斯威尼, 托马斯 A 威廉斯. 商务与经济统计[M]. 北京: 机械工业出版社, 2015: 323-324. (Anderson D R, Sweeney D J, Thomas A W. Statistics for business and economics[M]. Beijing: China Machine Press, 2015: 323-324.)
- [29] 杜海峰, 李树茁, Marcus W F, 等. 小世界网络与无标度网络的社区结构研究[J]. 物理学报, 2007(12): 6886-6893. (Du Haifeng, Li Shuzhuo, Marcus W F, et al. Community structure in small-world and scale-free networks[J]. Acta Physica Sinica, 2007(12): 6886-6893.)
- [30] Salama M, Panda M, Elbarawy Y, et al. Computational social networks: security and privacy[M]// Abraham A. Computational social networks. Berlin: Springer, 2012: 136.
- [31] Humphries M D, Gurney K. Network 'small-world-ness': a quantitative method for determining canonical network equivalence[J]. PloS One, 2008, 3(4): e0002051.
- [32] Pastor-Satorras R, Rubi M, Diaz-Guilera A. Statistical mechanics of complex networks[M]. Berlin: Springer, 2014: 47.
- [33] Artikov A, Dorodnykh A, Kashinskaya Y, et al. Factorization threshold models for scale-free networks generation[J]. Computational Social Networks, 2016, 3(1): 4-24.
- [34] Khan B S, Niazi M A. Network community detection: a review and visual survey[EB/OL]. [2018-10-11]. <https://arxiv.org/abs/1708.00977>.
- [35] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.
- [36] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.
- [37] Green G D, Liu J, Abbass H. Dual phase evolution[M]. Berlin: Springer, 2014: 50-51.
- [38] Earnshaw R. Research and development in digital media[M]. Berlin: Springer, 2018: 32.
- [39] Heineman G T, Pollice G, Selkow S. Algorithms in a nutshell: a practical guide[M]. O'Reilly Media, Inc., 2016: 152-154.

林泽斐 南京大学信息管理学院博士研究生, 福建师范大学社会发展学院讲师。江苏 南京 210023。

欧石燕 南京大学信息管理学院教授, 博士生导师。江苏 南京 210023。

(收稿日期: 2019-02-09; 修回日期: 2019-04-02)