社会结构的文本大数据测量*

——以中国社会职业地位变迁为例(1940—2015)

陈　茁

提要:基于问卷调查的社会结构定量测量存在时间跨度有限、测量维度单一、隐性指标不足等问题。为此本文阐述了一种基于文本大数据,运用自然语言处理算法来测量不同时期的话语结构,进而反映社会结构及其变迁规律的方法。以中国社会职业地位的历史变迁(1940—2015)为例,本文基于书籍大数据,从财富、权力、文化、声望四个维度刻画了职业地位和职业地位结构的历史变迁图景。该方法对传统问卷测量方法形成了重要的补充,为缺乏问卷资料的场景,特别是大历史跨度下的主观观念结构和客观社会结构变迁的测量提供了新的计算社会学工具。

关键词:社会结构　职业地位　计算社会学　大数据　机器学习

一、导　言

社会结构是社会学研究的关键概念和重点研究对象。广义上说,社会结构的概念十分宽泛,指社会行动者在互动基础上形成的相对稳定的社会关系协调体系(郑杭生、赵文龙,2003)。既有研究除了对社会结构进行基于经验的观察、归纳和抽象分析外,也在探索以数量分析生成测量和描述社会结构变化的关键指标。通过数据分析社会结构能够清晰地看到社会结构不同层面的变动时序和轨迹,是对"传统—现代"简单二分法的有益补缺(李培林,1992)。

如何测量社会结构?一方面,研究者可以直接依据宏观社会经济统计指标获得社会结构某一方面的测量,如收入分配、城镇化率、人口年龄分布等;另一方

*　本文系首批教育部哲学社会科学创新团队(南京大学"中国式现代化的社会治理数智研究创新团队")成果。特别感谢陈云松老师的指导,感谢谢宇、董浩等师友的批评建议,感谢匿名评审专家的宝贵意见。文责自负。

面,研究者可以在科学抽样的前提下通过调查问卷刻画社会总体的职业地位、生活方式、文化消费结构等。尽管这些尝试丰富并繁荣了一代社会科学研究,但现有的社会结构测量方式总体上依然存在以下三大问题。

第一,时间跨度有限。特别是对于发展中国家来说,学界对社会结构的数据收集和分析测量往往起步较晚,如中国社会结构调查在20世纪90年代后才逐渐普及与成熟。这意味着,关键指标的生成实际仅限于回溯近30年的历史变迁。

第二,测量维度单一。由于统计口径和社会调查的时间、成本及操作限制,相关社会结构的测量往往聚焦于少数核心指标,或者含义高度聚合的综合性指标,较少从多维度更细致地进行详细的数据收集。

第三,隐性指标不足。社会结构的传统问卷测量更多倚重显性的客观事实性指标,对隐性的文化观念指标往往缺乏丰富的测量手段,而观念本身也是社会结构的重要组成部分。

那么,有没有一种新的测量方式,能让社会结构的测量突破时间、维度和隐性测量的限制,相对准确地反映社会结构图景和变化过程?本文认为,基于文本大数据和机器学习方法可以构成测量社会结构的计算社会学路径,其基本思路是:运用图书、新闻、史料、政策等文本语料,通过词向量等自然语言处理方法,对不同时期的文本话语结构进行量化,进而生成测量社会结构的具体指标。

不难发现,该方法的逻辑前提是社会话语和社会现实之间、话语结构与社会结构之间的紧密关联。一方面,话语反映社会结构,记载着社会现实、文化认知、集体心态和价值观念;另一方面,话语建构社会秩序,它直接生产、调节和控制各种形式的实践、社会关系和社会结构(辛斌,1996;黄国文、徐珺,2006)。同时,话语是历史性的,不同时期的话语存在差异。这意味着,如果能对不同时期相对稳定的话语结构进行测量,就可以在认知层面直接测量主观社会结构,同时在相当程度上间接折射真实的社会结构及其变迁过程。

由于社会结构含义宽泛,不少研究将职业地位结构看作社会结构的核心组成部分(刘欣、田丰,2018)。因此,本文将以职业地位结构为例进行具体的方法演示。本文第二部分将探讨使用文本大数据和机器学习测量社会结构的方法论基础;第三部分将提出基于语义分析方法测量职业地位的具体策略,并验证方法有效性;第四部分将通过社会地位的描述总结中国近几十年职业地位和职业地位结构的变迁趋势;第五部分将对这一分析方法的优势和局限进行反思。

二、方法论基础:基本前提和分析框架

(一)话语结构与社会结构

使用文本大数据来测量社会结构的关键前提是:对特定语言的符号意义测量能够产生对宏观社会结构的正式描述。根据索绪尔(Ferdinand de Saussure)的观点,语言并不是主观意图的直接产物,而是一种宏观结构(de Saussure,2004)。符号、图像、动作或事物的意义或多或少是固定在社会、空间、时间或历史符号秩序中的(Keller,2011)。个人所表达的符号取自社会历史生成的集体知识库,语言的使用是与广泛的社会文化过程联系在一起的,社会创造了语言,因此参照这种惯例来描述语言单位,也是在考察真实的社会结构(Lemert,1979)。

而话语就是规范、结构化的符号使用实践。它是特定历史时期谈论特定话题的表述方式,它组织并构造如何谈论某一话题、对象和过程,它为社会和人的行为做出描述、规定、许可和限制(张铮等,2021)。福柯将话语视为社会历史的实践,将话语分析从特定的语言问题中解放出来(Keller,2011)。批评性话语分析学派进一步将话语是社会的构成要素、语言反映现实并影响社会过程、话语具有历史性等作为话语分析的核心观点(Fairclough,2013)。与内容分析将文本看作单纯的信息流通过程不同,话语分析倾向于把文本置于历史和社会语境中,揭示其背后的意义生产模式、意识形态和文化内涵(胡春阳,2007)。

因此,分析文本中的话语结构也就跟分析社会结构联系在了一起。而从话语结构到社会结构的逻辑实际上可以细化为两个步骤:其一,直接测量主观社会结构。作为人类文化载体的文本大数据实际上蕴含了人类普遍的观念结构,而对话语结构的测量也就能直接反映人们对社会结构的主观认知。其二,间接反映客观社会结构。主观社会结构既是对客观社会结构的反映,又对客观社会结构存在导向和形塑作用,因此能在很大程度上折射出客观的社会结构状态。换言之,文化架起了通往整体社会结构的桥梁(陈云松,2022a),通过对文本大数据进行话语形态的变迁分析,能够为主客观社会结构及其变迁的多维度测量提供另一种可能性。

(二)词语关系与话语结构

如何测量话语?目前主流的话语研究方法主要是对实际文本进行定性分析:基于少量话语,依赖研究者的主观介入,采用自上而下的方法,在全文及语境下进行细致、深入的分析(Biber et al., 1998)。但传统话语分析模式存在一定的

局限,包括分析范围小、主观解释存在偏见、微观分析难以支持宏观社会视角、研究结果不可重复等问题(Biber et al., 1998;辛志英,2020)。换言之,基于质性研究的话语分析方法固然有其优势,但并不擅长利用文本大数据客观地从宏观层面把握整体社会结构及其变迁趋势。

除了定性话语分析之外,近年来语料库话语分析迅速发展(Kennedy, 2014)。语料库话语分析将词汇作为研究的中心,通过基于词汇的量化统计找到术语之间的统计相关性,常用的方法包括语义编码、词频分析、词语共现和搭配模式分析等。通过大批量、标准化的数据分析模式,语料库话语分析能够帮助揭示直觉难以察觉的模式和规律,减少研究者的主观偏见(McEnery & Hardie, 2012)。但限于时代和技术发展,这样来测量话语结构依然存在局限:词频、词语共现和搭配的统计只能揭示语言使用的表面模式,难以捕捉深层次的语义关系;通过人工语义标注的方法(如标注词语意义、近反义词关系、上下位词关系)只能覆盖有限的词汇和短语,难以面向随时间变化的词语意义和话语结构。

计算社会科学的发展为量化分析话语提供了新的可能性。根据语言学的理论,一个词的含义由它所处的上下文语境来决定;两个词的使用语境越相似,也就越倾向于表达相同的含义(Harris,1954)。因此,词汇的意义和功能可以通过它们在语言中的分布模式来推算,如果两个词在类似的上下文中频繁出现,它们很可能有相似的意义和功能。这一分布式语义假设成为目前自然语言处理的基础理论之一,即机器学习可以通过分析大量文本数据学习每个词在各种上下文中出现的模式,然后使用这些模式来表示词的语义。分布式语义分析能够基于海量数据并根据词语之间的分布关系来“自下而上”地自动学习词汇的含义,处理随着时间变化而新出现的词汇和话语结构,反映语言的变化。

(三)测量社会结构的方法框架

在以上两个前提的基础上,测量社会结构的基本思路是:基于文本大数据,通过词语之间的关系和词表设定发现关于特定对象或话题的普遍表述模式,进而通过对话语结构的测量反映社会结构。这具体可以细分为以下五大步骤。

一是收集文本大数据。文本大数据是反映社会结构的一手资料,大数据的选择关系到测量何种形式的话语以及社会结构。需要注意不同类型的文本反映了不同对象或者场域的话语结构,比如政府文件反映官方话语,新闻报道反映媒体话语,书籍小说反映大众话语,消费广告反映商业话语等。研究者可以收集不同类型的大数据进行互相验证,也可以比较不同叙事主体的话语差异。另外,收

集不同时期的大数据能够分析社会结构的历时性变迁;收集不同地域的资料如地方志、跨国书籍等,能够对不同区域的社会结构进行横向比较。

二是训练自然语言模型。借助机器学习算法,研究者可以依据文本中词和词之间的关系,自下而上地将词语转化为能够表征词汇意义的向量和指标。这些向量和指标反映了语料中最普遍的用语模式,即话语结构。研究者可以根据分析目标选择不同的模型:如果侧重于进行多语料类型、跨时期、多空间的比较分析,则推荐直接基于不同属性的子语料分别训练模型;如果分析侧重于描述人类最普遍的话语结构,则推荐使用生成式大语言模型等已经基于超大规模语料预训练好的模型,以提供"整体事实"的视角来反映人类的普遍观念(梁玉成,2024)。

三是制定社会结构的测量策略。社会结构表现为方方面面,研究者需要针对具体的研究议题制定具体的指标测量策略。首先,研究者应该明确分析的对象,如职业、性别、国家、组织、亚群体、概念等,构建分析对象词表。其次研究者应该明确分析的维度,构建分析维度词表,也即需要对研究对象的哪些方面进行测量。最后,研究者需要确定社会结构指标的计算方式。指标的操作化过程比较灵活,一般来说,可以先计算分析对象词汇与不同分析维度词汇之间的空间距离,将分析对象投射在对应的分析维度上,再根据需要进行更灵活和细致的计算分析。①

研究者如何构建分析对象和分析维度的词表?一方面可以查阅词典,或使用既有研究编撰的特定领域词表。另一方面也可以采用数据驱动的方法,根据训练好的词向量寻找同义词,或者通过人工排查所有词表的方式来完成。需要注意的是,如果要对基于不同语料训练的多个模型开展比较性分析,应尽量先在同一个模型内部生成指标的标准化值,再进行跨模型比较。

四是验证分析有效性。为保证对社会结构指标测量的效度,研究者需要验证分析的有效性。在理想情况下,如果有社会调查数据,可以直接对调查数据和基于文本大数据测量的社会结构指标进行比照;但使用文本方法测量社会结构的情况大多是因为社会调查数据缺失,这时可以采用局部验证整体的思路,对能够与问卷数据相匹配的部分年份、部分维度指标进行比照,进而推定整体模型。当完全缺乏社会调查数据时,可以选择多来源的文本大数据,进行多重交叉验证。

五是描述社会结构。研究者使用验证过的分析策略生成社会结构的具体操作化指标,对社会结构进行描述,呈现研究对象在不同维度下的现状、在不同时期的变迁趋势以及在不同区域下的结构性差异。同时,研究者可以将该指标与

① 值得注意的是,由于对单个词的分析可能并不总是稳健的,对同一个对象和含义的表述往往存在多个词,因此建议选择多个词语进行距离计算,最后取均值。

其他社会宏观指标进行链接,以便进一步探讨社会结构的影响机制。使用文本大数据测量社会结构的方法框架见图1。

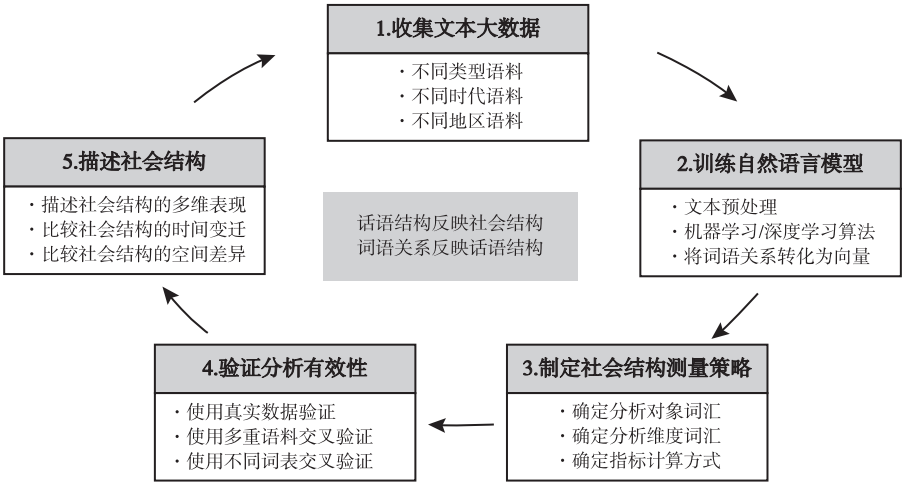

图1 使用文本大数据测量社会结构的方法框架

事实上,已有一批研究运用类似方法对多个议题进行了测量。例如,科兹洛夫斯基(Austin C. Kozlowski)等人提出文化几何学,量化了阶级与财富、就业、教育、修养等维度的关系和历史变化趋势,以此来理解阶级在不同历史阶段的含义(Kozlowski et al., 2019);加格(Nikhil Garg)等运用类似的方法对100年来英语文化中的性别和种族观念进行了量化(Garg et al., 2018)。在此基础上,本文聚焦于这些研究通用的方法论价值,探讨使用文本大数据和语义分析测量社会结构的可能性、优势和局限。

三、职业地位的文本大数据测量:中国案例

社会结构含义宽泛,可以细化为地位结构、人口结构、家庭结构、组织结构、城乡结构、消费结构等(陆学艺,2010;李培林,2011;刘欣、田丰,2018)。由于职业地位常常被社会学家用来测量社会结构分化的程度和形态分布,同时20世纪80年代涌现了一大批关于中国职业地位的调查研究,这为验证方法有效性提供了参照标准,因此本文将以"职业地位"为例详细演示如何使用上述方法框架对大历史跨度下的职业地位及其结构变迁进行测量。

(一)职业地位的传统测量方法

职业声望是人们对各种职业做出的主观评价,是最早和最广泛使用的测量职业地位的指标之一。测量职业声望一般采用主观评价法,即借助社会调查获取访问者对一些职业的评分,对职业进行打分、分级或者排序(高顺文,2005)。在中国,关于职业地位测量的研究绝大多数采用主观评价法(Lin & Xie,1988;李强,2000;许欣欣,2000,2005;李强、刘海洋,2009)。主观评价法的测量方式简单直接,且能够捕获更广泛的社会文化因素和价值观念。但由于数据收集方式的限制,这种方法只能测量少部分职业,因而也难以描述和分析整个职业地位的分层体系(李春玲,2005a)。不止于此,职业声望也并不总是与收入、教育等职业属性相关(Hauser & Warren,1997)。

另一种常用的测量方法是以社会经济地位指数为代表的客观测量方法,其基本思路是依据教育、收入等客观指标,通过拟合模型和权重分配等方法,建立职业地位分数的方程模型。典型的研究是邓肯(Otis Dudley Duncan)基于职业收入和教育程度来计算社会经济地位指数(Duncan,1961);在中国,边燕杰、李春玲等学者也基于收入、教育等因素构建了职业地位评价得分(Bian,1996;李春玲,2005a)。相比于主观测量法,客观测量法能依赖少量职业对大量甚至全部职业的地位指数进行计算和推测,且同时考虑职业相关的多个维度;但其缺点在于十分依赖权重和测量维度的科学设定,同时容易忽略文化价值观念等难以量化的非经济因素。

不论是主观还是客观测量,测量职业地位事实上都绕不开一个关键问题,即职业地位包含哪些维度。社会学先驱韦伯最先提出社会地位三分法:由财产占有不同产生的经济地位差别,由权力占有而产生的政治地位差别,以及由社会评价和荣誉占有不同而产生的社会地位差别(韦伯,2010)。布迪厄的文化资本理论认为经济资本、文化资本和社会资本共同决定了一个人的社会地位,并将文化资本进一步纳入社会地位的分析维度(Bourdieu,2018)。同时,赖特(2006)特别强调了专业技术与劳动过程的关系,认为专业技能不仅影响个体的职业地位,也决定了劳动者在生产体系中的阶级属性。这些研究表明,职业地位的测量并非仅仅包含一个单一指标,而是包含着内涵丰富的子维度。

表 1 列举了 11 项测量职业地位的代表性研究。如该表所示,中国最早的关于职业地位的测量可以追溯到 1983 年,20 世纪 90 年代和 21 世纪初的相关调查与研究非常丰富。总体看来,尽管职业地位的相关研究已经较为成熟,但仍然存在需要补足的空间。第一,从时空范围上看,多数研究都是基于特定年份和一定地域的静

态研究,缺乏超大时间跨度和空间尺度的整体宏观测量和变迁研究。第二,从测量的维度看,这些研究在测量分析时多采用一个综合指标,对各个职业地位维度分项进行独立分析的研究几乎没有。然而,职业地位的各个维度是相互关联但不完全重合的,更好的选择是独立地分析每一个维度(Hauser & Warren,1997)。第三,从测量的隐性指标看,客观法直接剥离了职业地位的主观评价维度,主观法可能由于测量指标相对抽象,并不总是与收入、教育程度等客观属性相关。

表 1 **职业地位代表性研究方法统计**

研究	年份	地域	职业数量	测量方法	测量维度
Lin & Xie,1988	1983	北京	50	主观法	单一指标
Bian,1996	1988	天津	46	客观法	教育、收入
许欣欣,2000	1987	北京、沈阳	85	主观法	单一指标
陈婴婴,1995	1993	百县市	100	主观法	单一指标
蔡禾、赵钊卿,1995	1994	广州	102	主观法	单一指标
李强,2000	1998	北京	100	主观法	单一指标
仇立平,2001	1999	上海	50	主观法	收入、权力、声望
李春玲,2005a	2001	全国	161	主客观	收入、教育、权力、部门、社会歧视
迟书君,2003	2001	深圳	81	主观法	单一指标
许欣欣,2005	2002	全国	51	主观法	单一指标
李强、刘海洋,2009	2009	北京	99	主观法	单一指标

(二)职业地位的文化测量:数据、方法和效度

1. *数据来源*

书籍作为人类文化观念表达和传承的主要方式,反映着社会的文化认知、集体心态和价值观念,为职业地位的测量提供了另一种可能性。已有相关学者基于书籍这一重要文化载体量化地测量和分析社会普遍的文化认知、集体心态和价值观念(陈云松,2015)。

本文使用目前全球最大的数字化工程项目即谷歌图书的中文图书语料库,来测量职业地位。截至 2012 年的官方统计数据,谷歌图书数据库已经囊括人类全部印刷总量的 6%,其中涉及中文图书 30 万部,中文词汇 268 亿余个(Lin et al., 2012),并且依然在不断补充和更新。为了规避可能的伦理问题,该数据库开发了 N-gram 工具,对文本数据进行直接切分、断句,实现以词语或词组为单位的统计分析。我们借鉴科兹洛夫斯基等人的研究思路(Kozlowski et al.,

2019),选取所有 5-gram 的中文词组作为分析原语料。[①] 5-gram 是指由 5 个词语组成的序列,可以被理解为一个词组。根据 N-gram 的筛选规则,5 个词语在一起使用超过 40 次才会被作为 5-gram 统计。

而以谷歌图书的中文 5-gram 语料为基础来测量职业地位主要是基于以下三个原因。第一,作为最大的数字化图书数据库,它不仅体量大,而且还包含了多种类型的文档,如小说、政府文件、科学文本、调查报告等,能尽可能地反映一定时期内人们对职业的普遍观念和文化共识。第二,超长时间的书籍可追溯性保证了大历史跨度测量的可能性。第三,书籍生动具体的文本语境信息提供了灵活多样的分析维度。一方面,我们可以对所有在书籍中出现过的职业进行测量,而不必拘泥于调查问卷中涉及的少量主要职业;另一方面,我们可以灵活地选择分析维度,构建涉及财富、权力、文化、声望等多方面的立体化测量指标。

2. 模型训练

近年来,词嵌入方法已被广泛用于测量语义信息(龚为纲等,2019;刘河庆、梁玉成,2021),本文采用该方法来衡量职业的文化意义。嵌入空间中词语之间的距离通常用词语向量间夹角的余弦值来评估,代表两个词语语义联系的紧密程度。如果一些词语有类似的上下文,或者经常出现在一起,它们的向量表示在高维空间中将会靠得很近;反之,如果一些词汇并不常共同出现,且并不享有同样的上下文,它们的向量距离则会较远。例如,如果“科学家”和“发明”两个词的余弦相似度很高,则代表着两个词常常出现在同一语境中。

值得一提的是,在模型的训练过程中,目标词及其上下文的配对依赖动态滑动窗口的大小。从理论上来说,更大的滑动窗口会涵盖更多的上下文信息,从而有更高的准确度。但由于数据的限制,本文使用的数据为书籍中 5 个词语组成的词组(5-gram),这意味着模型训练窗口最大只能设定为 4。而既有研究和文献已从多角度证明了使用 5-gram 捕捉语义关系的有效性(Garg et al., 2018; Kozlowski et al., 2019)。事实上,Skip-gram 模型训练过程中默认的滑动窗口也仅为 5,且距离更远的词汇在训练时也会被赋予更低的权重。[②] 不止如此,一些研究者也比较了滑动窗口设置为 2、5 和 10 三种情况下的模型效果,在数据量够

① 值得一提的是,谷歌图书的中文语料构建是基于简体中文的。尽管 1956 年才开始正式推行大规模的使用简体字,但新中国成立前国民政府曾多次推行汉字简化,共产党领导范围内的油印刊物也采用并创造了很多简体字。同时,早期中文语料中可能混入少量越南、韩国等地区的出版资料,但这种情况随着 20 世纪 30、40 年代越南、韩国相继颁布废除汉字的法令而消失。

② 例如,如果滑动窗口设置为 5,模型则会利用目标词前后各 5 个词语作为输出,且五个词的权重将分别为 5/5,4/5,3/5,2/5,1/5。

大的情况下,三种效果并不存在显著的效度差异(Levy et al., 2015)。

为了追踪随着时间推移的语义变化,本文以 10 年为时间范围,5 年为滑动间隔,将 1940 年到 2015 年共计 75 年的书籍大数据划分为 14 个子语料库,[①]并在此基础上分别训练了 14 个 Word2Vec 模型。[②] 这些模型分别代表了 1940—1950、1945—1955、1950—1960 等以此类推的阶段性话语结构。遵照相关标准(Mikolov et al., 2013),本文将词向量维度设置为 300,最大窗口设置为 4,训练时删去了出现次数小于 50 的词。表 2 显示了 14 个词向量模型的基本情况。

表 2　　14 个词向量模型的基本情况

模型	语料范围	文档数	词语数	模型	语料范围	文档数	词语数
1945	1940—1950	685405	13165	1980	1975—1985	291600225	39637
1950	1945—1955	1530054	13435	1985	1980—1990	63593881	44657
1955	1950—1960	17400030	23348	1990	1985—1995	768691103	44259
1960	1955—1965	27649733	24060	1995	1990—2000	713278626	45192
1965	1960—1970	16845319	21861	2000	1995—2005	660748454	44801
1970	1965—1975	17265600	22454	2005	2000—2010	482594943	44898
1975	1970—1980	61062870	29954	2010	2005—2015	160253622	41649

3. 职业地位的测量策略

依据词嵌入和职业地位不同维度的单词列表,我们可以测量一组职业的地位得分。具体来说,参照韦伯的社会地位三分法和布迪厄的资本理论(韦伯,2010;Bourdieu,2018),本文将职业地位的测量细分为四个维度:财富、权力、文化和声望,代表四个场域资源在不同职业中的分配情况。我们参照加格等人测量性别刻板印象偏差的方法(Garg et al.,2018)构建职业地位的测量策略。

首先,对于每一个维度分别构建两组词,一组用来形容职业高资源集聚的状态,如“富有”“博学”等;另一组用来形容职业低资源集聚的状态,如“贫穷”“文盲”等,以此作为该维度的两极。其次,以职业为中心词,分别计算职业与高资源组词和低资源组词的距离,用以衡量该职业在多大程度上倾向于出现在高资源语境或者低资源语境中。最后,使用该职业与高职业组词的平均距离减去该职业与低职业组词的平均距离,以此作为该维度的具体得分。如果该得分为正

① 谷歌图书收录的中文图书可追溯的时间范围要早于 1940 年,但 1940 年之前的数据相对较少,原因可能是词语低频或遗漏而带来的测量偏差。为保证严谨性,本文从 1940 年开始进行分析。

② 以 10 年为语料范围是因为语义信息的准确计算需要较大的文本数据作为基础,10 年间的书籍数据能够较好地保证充足的训练语料。滑动间隔(sliding interval)是一种处理时间序列数据的方法,专门保留一些数据重叠,用于平滑数据和捕捉更细微的变化。

值,意味着该职业更常出现在高资源语境中,人们对这一职业的认知和评价更高;反之则评价较低。四个维度得分的均值计为总体职业地位得分。

接下来,本文以财富维度为例更具体地说明计算过程。考虑现有的“富有”和“贫困”这两个词,计算“科学家”和“富有”两个词向量的余弦距离记为 D1,计算“科学家”和“贫困”两个向量的余弦距离记为 D2,而 D1 和 D2 则分别代表“科学家”与“富有”或“贫困”共同出现的概率高低。由于单个词语的计算难免会带来偏差,我们进一步考虑两组词:一组是形容财产较多的词,如“富裕”“富有”“宽裕”等;另一组是形容财产较少的词,如“贫穷”“拮据”等。而后,我们分别计算每组的每一个词与“科学家”的距离,并计算两组词与“科学家”的距离均值 M1 和 M2。最后,M1 - M2 的值即为科学家在财富维度上的地位得分。

如果用公式表示,每一维度的职业地位得分即为:

$$S_h = \frac{1}{n}\sum_{i=1}^{n}\cos(W_O, W_{hi})$$

$$S_h = \frac{1}{n}\sum_{i=1}^{n}\cos(W_O, W_{li})$$

$$O_p = S_h - S_l$$

其中,S_h代表职业与高地位词语的平均距离,S_l代表职业与低地位词语的平均距离,W_O代表职业词汇,W_{hi}代表高地位词汇,W_{li}代表低地位词汇,O_p代表该维度下某个职业的地位得分。

接下来,本文采用人工设定和数据驱动相结合的方法确定词表。

(1)职业词语的选择

对于职业词汇,最直接的办法是参考《中华人民共和国职业分类大典》等官方文件,然而,将这些职业词汇运用在本研究中存在两个问题:其一,官方职业分类所用名称十分规范,如“人民法院负责人”等,但书籍使用的词汇偏向于更通俗的称呼;其二,职业会随时间改变,我们难以使用当今的职业分类大典去衡量跨越长时间段的职业情况。因此,本文通过人工判断的方法,以所有模型中包含的词语为依据,从四万余个不重复的词语中挑选出与职业有关的词汇作为职业选词,共计 382 个。①

(2)维度词语的选择

对于财富、权力、文化和声望四维度词语的选择按照如下步骤:首先,查阅现代汉语词典,尽可能多地囊括能体现各维度下评价高低的词汇,建立初步词库。

① 鉴于部分职业的文本适用性限制,职业选词中使用部分职务称呼来代替职业称呼,如使用“局长”来代指广义的“国家机关负责人”相关职业。

但由于一些词语可能存在一词多义、特殊用法、词义发生较大变动、出现频率低等情况,将其纳入会进一步带来测量噪音,因此还需要对这些词语做进一步筛选。具体的做法是,以社会调查计算的真实职业地位作为参照标准,如果依据某个词汇测量的结果与真实数据完全不相干甚至反相关,则将其作为噪声剔除。经过人工筛选和数据驱动后,每个维度的测量词表如表 3 所示。同时我们还构建了一个每维度只包含 10 个词语的子词表进行稳健性检验,其设计原理在于,同一维度下往往存在很多词语且难以穷尽,如果只使用部分词能够得出类似的效果,则意味着尽管词语未完全罗列,依然能够得到准确的操作化结果。

表 3　　职业地位四维度测量词表

财富	殷实、富饶、富裕、昂贵、高档、阔气、豪气、富有、余裕、豪华、富足、宽裕、富庶、奢侈、富人
	穷困、清贫、穷苦、低收入、清苦、赤贫、贫寒、乞讨、贫民、一贫如洗、贫困、穷人、贫穷、贫苦、拮据、廉价
权力	领导、决策、监察、管理、控制、强制、统领、审查、协调、任免、拍板、管辖、指挥、统治、特权、掌控、权势、要求
	勉强、无奈、忍让、受气、冤屈、委曲求全、受屈、逆来顺受、服从、隐忍、妥协、忍耐、被迫、听命、依附
文化	才华、学识、广博、栋梁、知识、贤明、智慧、有才、渊博、英明、饱学、研习、博学、才干、学习、读书、才子、鸿儒、专业、精通、娴熟、技术、技能
	愚钝、孤陋寡闻、笨拙、笨蛋、白痴、愚笨、愚昧、文盲、半文盲、不识字、辍学、无知、弱智、肤浅、退学、没文化、失学、愚蠢、落后、无能、平庸、拙劣
声望	驰名、尊敬、楷模、敬重、崇敬、佩服、荣誉、知名、榜样、拥戴、杰出、著名、爱戴、景仰、敬仰、仰慕、名望、敬佩、敬意、威望、闻名、敬爱、崇拜
	卑微、看轻、俗人、庸人、鄙夷、下贱、底层、下层、卑贱、难堪、丢人、不体面、鄙视、蔑视、侮辱、庸碌、轻视、轻蔑、歧视、窝囊、看不起

4. 职业地位的测量效度

本研究所用的测量策略能在多大程度上反映真实社会的职业地位?本文将从书籍中测量的职业地位和基于社会调查的真实职业地位进行了比较。我们以从 1982 年到 2009 年、跨度为 27 年的 6 篇职业地位的实证调查为基准(Lin & Xie,1988;陈婴婴,1995;许欣欣,2000,2005;李春玲,2005a;李强、刘海洋,2009),根据每项调查实施的年份,测试了 1980 年至 2015 年期间的六个模型。

图 2 以 2001 年李春玲测量的职业地位分数为基准,详细反映了调查数据和书籍数据测量①的相关性情况。我们使用书籍测量结果对真实结果拟合了一条

① 由于李春玲的调查于 2001 年开展,因此我们选取时间跨度为 1995—2005 年的模型计算职业地位,比较两者的相似性。李春玲的研究列出了 81 个职业的声望,经过删选只有部分职业可与书中的词汇相匹配。原因是部分职业分类较细,被合并为职业大类(如中学教师、小学教师等被合并为教师),同时部分职业在书籍中不存在(比如市民主党派负责人)。

直线，大部分职业集聚在拟合线的两侧，书籍测量的职业地位与真实数据测量的职业地位显著正相关（$P=0.000$；$R^2=0.509$）。同时，两者的皮尔逊相关系数为0.712。这意味着，基于图书的测量能够在很大程度上反映真实社会情况。

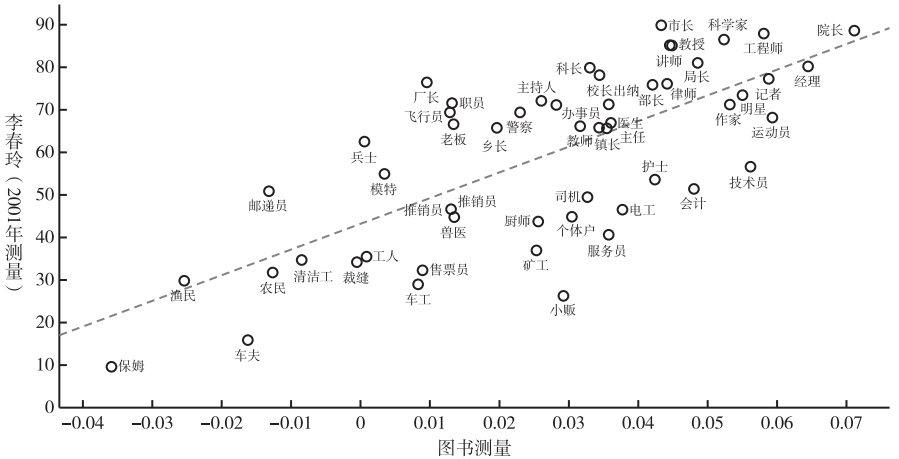

图2　书籍与调查数据计算的职业地位分数相关性

图3反映了客观调查与大数据测量的相对应时间段职业地位间的皮尔逊相关系数，包括全部选词和部分选词的稳健性分析结果。事实上，基于大数据的测量策略所得的结果与所有调查分析结果的关系都十分显著，相关系数均在0.5以上，意味着该方法有较大的效度。但基于书籍的测量又确实与客观调查数据存在一定偏差：前者测量的实际上是人们对于职业地位的主观认知结构，而通过观念来折射客观社会结构仍可能与真实社会存在差异。

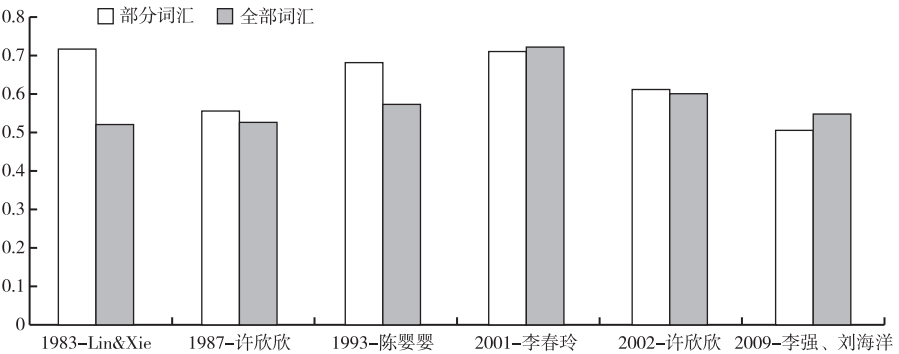

图3　书籍职业地位分数与客观调查数据的皮尔逊相关系数①

① 横轴的年份为职业地位的调查年份，而非论文发表年份。

四、历史变迁中的职业地位和职业地位结构

(一)职业地位的历史变迁

1. 宏观层面:量化职业地位的时代差异

从宏观层面看,近几十年来职业地位是否存在明显的时代变化?图4展示了基于各时期子语料计算的职业地位得分两两间的皮尔逊相关系数,系数越高颜色越深,代表这两个时期的整体职业地位得分越相似,职业地位变化越小。[①]举例来说,对角线上为两个相同时期的职业地位,故相似度最高为1;1945年的职业地位与1950年的职业地位相似度最高,相关系数达到0.77。

总的来看,中国的职业地位变化并不总是稳定的,在1980年前后存在非常明显的分界线。1980年之前,职业地位总体不稳定,仅仅相邻年份的相似度比较高,职业地位处在不断变化之中。特别是1945—1950年前后的新中国成立时期、1970年前后的"文化大革命"时期,社会地位变化极大。但同时,在1955—1960的社会主义革命时期临近年相关系数达到0.88,预示着一个短暂的相对稳定期。到1980年后,整体职业地位变化趋于稳定,该时期职业地位相似度明显呈现一个颜色相近的"矩阵块",职业地位相似度基本高于0.8。

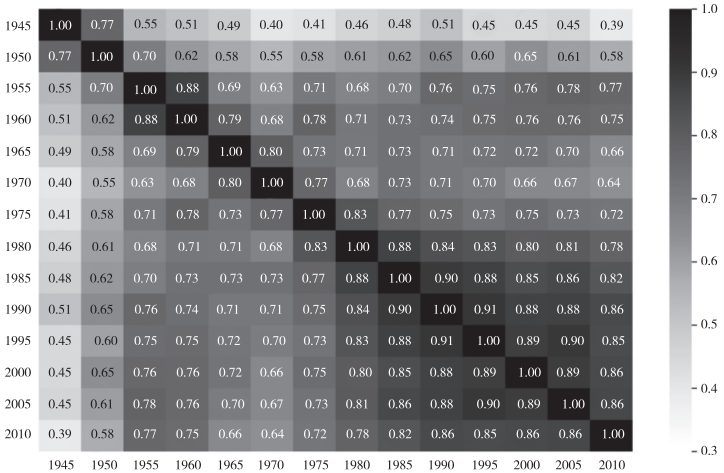

图4 职业地位(四维度均值)时代变化的相关系数

① 考虑到不同时期的语料中涉及的职业数量不一致可能带来潜在偏差,本文专门针对在所有年份中都出现的85个职业做了稳健性检验,结果显示跟全部职业的分析结果具有非常高的一致性。

图 5 展示了四维度职业地位在不同时期两两间的相关系数。可以发现,在四维度地位中,权力地位维度的时代变化最为稳定,财富地位最不稳定。其中,财产地位的剧烈变动集中于 20 世纪 60—70 年代;文化和声望地位的剧烈变动集中于 20 世纪 70 年代。20 世纪 80 年代之后,四维度地位也均趋于稳定,形成明显的颜色更为接近的深色矩阵块;但相比财富和文化维度而言,职业地位在声望和权力维度的变化更为稳定。

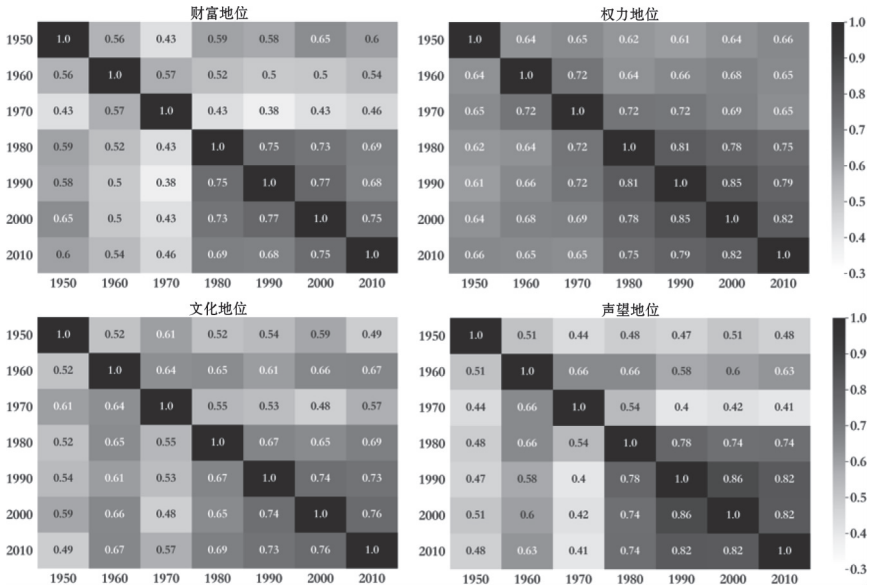

图 5　四维度职业地位时代变化的相关系数

上述职业地位的时代变迁说明了什么？回看以往职业地位的研究,一种流行的观点是:生活在不同时代、不同社会、同一社会不同的群体成员对职业地位的评价可能非常相似。特雷曼(Donald J. Treiman)对 60 个国家 85 项职业声望的研究结果表明,声望在时空上基本是不变的(Treiman,1977),这种稳定性被称作特雷曼常数。本文的实证结果发现,从宏观大历史尺度看,特雷曼常数并不完全适用于经历了反帝反封建、社会主义改造和改革开放等重大变革的近现代中国社会,这挑战了特雷曼关于职业声望在时空上具有普遍稳定性的观点。同时,本文拓展了测量变迁的时间和维度边界,量化了职业地位在不同时代的变迁程度以及重要时间转折点,同时明确了财富、权力、文化和声望变迁的不同稳定程度。

2. 微观层面:多职业、多维度的变迁轨迹

中国的职业地位在特定时期具体是如何变动的?本部分将基于部分典型职业的微观分析进一步展示变动的具体方式。为便于不同时期模型的比较,我们将同一时期的所有职业地位得分缩放标准化为0~100的得分。① 图6详细展示了部分典型职业地位在四大维度上的时间变化趋势,②其中,纵坐标代表职业地位标准化得分,横坐标代表年份。微观职业地位变迁特点主要有如下表现。

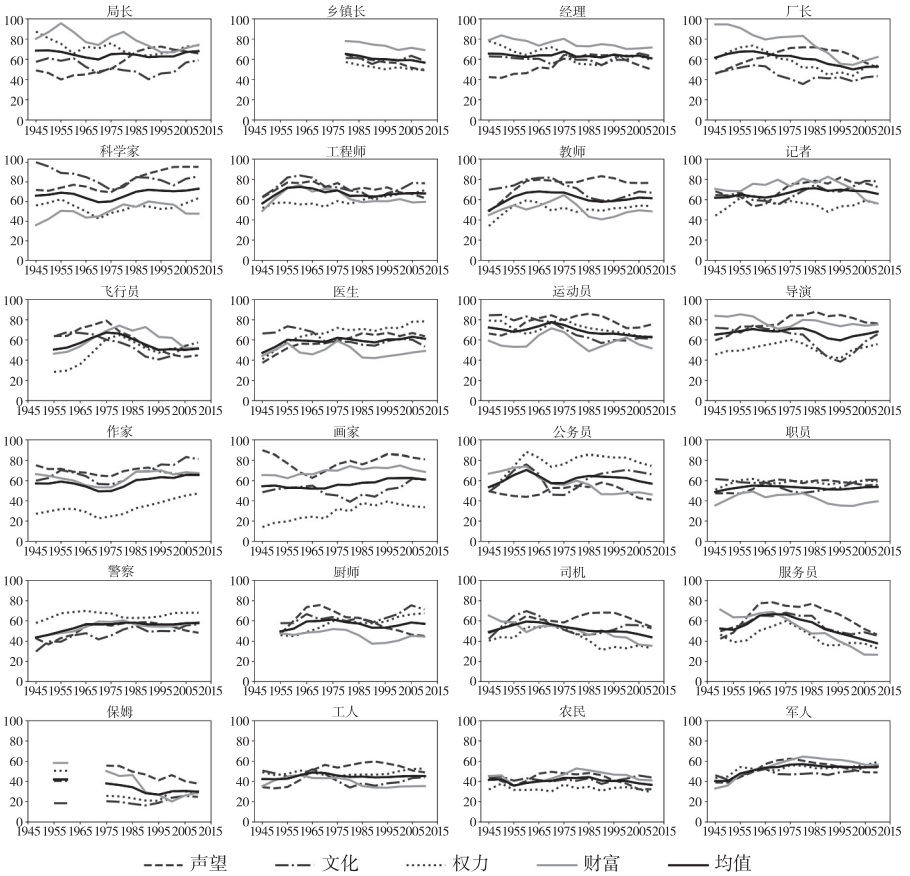

图6 典型职业各维度地位的变迁情况

① 标准化之后的值实际上是每个时期内所有职业的相对地位,相对地位在不同时期的变化并不代表绝对地位的增加或减少,而是说相对于该时期的其他所有职业的变化特征。之所以不用职业地位的绝对数值,是因为绝对数值的变化在很大程度上反映的是职业地位的整体结构性上升或者下降。

② 限于篇幅,本文基于《中华人民共和国职业分类大典(2022年版)》划分的七大类职业(不包括第八类“不便分类的其他从业人员”),在每一类中选择部分典型代表性职业进行展示。由于职业称呼中部分职业之间存在嵌套或重叠,本文对这些职业称呼进行了合并和均值计算,例如“警察”实际上是对“警察、武警、警卫、民警、交警、巡警、刑警”等多个词语计算后得出的均值。

20 世纪 50—60 年代,职业地位明显提高的职业包括工程师、公务员、教师、军人、工人等。伴随着新中国工业化建设以及城乡二元结构的出现,体制内外的社会分割强化,城市居民、体制内人员的职业地位提升,农民地位在 20 世纪 50 年代略有下降。

20 世纪 70 年代是职业地位变动最剧烈的时期之一,剧烈变动典型表现为科学家、工程师、公务员、画家、作家、教师等职业地位下降,飞行员、运动员、警察、服务员、厨师等职业地位上升。其中,知识、文化类职业地位由于“文化大革命”而受到打击;20 世纪 70 年代飞行员和运动员职业地位的变化与我国航天事业、国家外交和综合国力联系紧密。

20 世纪 80—90 年代,改革开放后的职业地位变化表现为:首先,专业技术类职业的地位重新上升,科学家、工程师、教师、记者、医生、画家、作家等职业地位重新提高。其次,飞行员、运动员职业地位相对下降,可能与这两个职业的政治、外交属性相对剥离相关。服务员、司机、保姆等传统服务业职业和农民的职业地位也有不同程度地下降。

21 世纪初,职业地位变化总体保持稳定,其中公务员职业地位随着市场化改革的深化而小幅度下降,官本位呈现淡化趋势;医生、导演、作家等专业技术和文化艺术类职业地位小幅度上升;司机、服务员等传统服务类职业的地位继续下降。

上述职业地位在不同时期呈现变化的内在动力是什么?本文尝试对上述描述性结果进行推测性解释,将成因归纳为三大轴线。一是政治力量主导的轴线。从新中国成立、三大改造、“文化大革命”到改革开放,政治力量几乎是 20 世纪 80 年代前职业地位变化的最主导因素。高知型和文体类职业是这一时期受到影响最大的职业类型,同时工程建设、航空航天、体育事业等相关职业也在很大程度上因与国家综合实力和外交声誉联系紧密而被赋予高社会地位。二是产业结构发展的轴线。产业发展刚起步时相关职业往往被赋予较高的职业地位,如 20 世纪 60 年代工业发展初期的工程师、厂长和工人,20 世纪 70 年代服务业发展初期的司机、厨师和服务员。改革开放之后,市场要素配置带来的最直接影响是专业技术职业地位的提升,第一、第二产业和传统服务业的职业地位下降或趋于稳定,文娱类(如导演、作家)职业地位提升。三是传统文化观念的轴线。在尊师重教和官本位文化传统的影响下,知识型和公职类职业地位一直较高,且知识型职业具有极高的职业声望。

通过分析职业内部四维度职业地位的分化与重叠,同样可以发现有趣的现象。从实证分析结果看,职业的财富、权力、文化、声望地位并不总是一致的。例

如,科学家、教师等知识型职业的典型特点是声望和文化地位高,而权力和财富地位相对较低。尊师重教的中华文化传统让人们对这些职业充满敬重和赞誉,同时这些职业也被赋予无私奉献的理想化形象。又如,文化类职业(导演、画家、作家、记者等)表现为高声望、高文化、高财富、低权力的分化。服务类职业(司机、服务员、保姆等)虽然财富、权力、文化地位较低,但声望地位相对较高。

四维度地位的一致性分析可以为测量社会结构提供一个具体、连续的量化视角。关于中国的社会结构特征,学界一直以来存在"碎片化"和"结构化"的争论(李春玲,2005b)。持"碎片化"观点的学者认为当前的社会是一个多元分化的社会,利益群体在不同分化坐标上是相互交叉的,不存在绝对的分界线(李强,2008)。持"结构化"观点的学者认为多维度的地位分化趋于一致,特别是经济地位的差异扩散到其他领域,各维度资源的叠加形成整体性社会聚合体(李路路,2003)。实际上,碎片化和结构化都是描述社会结构分化的理想类型,本文的分析结果显示,社会分化表现出碎片化和结构化并存的局面:职员、工人、农民、军人等职业表现出更高的地位一致性;而其他职业特别是专业技术类职业多存在一定程度的维度分化。

(二)职业地位结构的历史变迁

在刻画完职业地位变迁的基础上,本部分将进一步结合职业地位得分和职业频率,刻画社会整体职业地位结构的变迁情况。① 图7(A)反映了1940—2015年14个子语料库下的职业地位结构,每一时期的地位结构图包含左右两个部分,左边为地位分布曲线,高度为职业地位的得分,宽度为某一职业在书籍中出现的频率;右边为地位分布箱线图,反映了所有职业地位的四分位分布。图7(B)进一步呈现了反映地位结构的具体指标,一是反映职业地位的最大分化程度的极差,即该时期最高和最低职业地位的差值;二是反映职业多样性和均衡程度的职业熵,由信息熵公式计算得出,如果该时期文本大数据中出现的职业种类越多样,不同职业间分布的数量越均衡,则职业熵越大。

各时期职业地位箱线图的中位数值(箱子中的黑色线)反映了职业人群的中间值所处的地位水平。总的来看,近几十年我国的职业地位结构经历了整体的结构性上移。进一步,不同时代职业地位结构也呈现不同的分布特点。

20世纪50—60年代初的职业地位结构表现为多层分化的塔型结构。战后

① 考虑到不同时期的语料中涉及的职业类别不一致可能带来潜在的偏差,本文专门针对在所有年份中都出现的85个职业做了稳健性检验,结果显示跟全部职业的分析结果具有非常高的一致性。

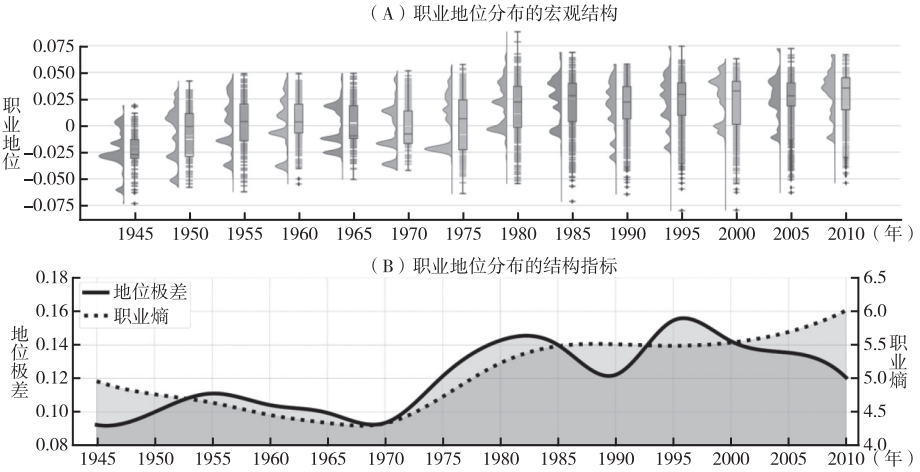

图 7　职业地位结构的历史变迁

经济恢复时期的职业地位极差存在短暂的小幅度上涨，但职业的多样性和均衡程度则不断下降。在过渡时期，由于一系列社会改组措施，职业地位结构呈现不连续多峰分布，意味着职业阶层分化明显；同时地位结构的中下部分更宽，呈现整体的塔型分布。

20 世纪 60—70 年代的职业地位结构表现为断裂的平均主义结构。该时期职业地位结构仅存在两到三个明显的波峰，且波峰间存在巨大的鸿沟，职业地位结构断裂明显；而职业地位极差和职业熵在这一段均达到最低峰值。伴随着人民公社化运动、“文化大革命”等社会运动的开展，对阶级和身份制度差异的强调与身份群体内部的地位均等化并存。

20 世纪 80 年代后的职业地位结构表现为相对连续的纺锤型结构。改革开放初期，社会结构迅速变化，不同职业间的区隔程度渐趋减小，地位极差和职业熵不断增大，意味着社会分化迅猛加剧，不平等程度拉大，职业也变得多样。21 世纪以来，社会结构保持纺锤型并向更健康的形态发展。一方面，四分位距更为集中，社会结构的中间部分愈发聚集；另一方面，职业地位极差逐渐缩小，特别是极低值不断提高，尾部分布不断稀疏，职业熵也显示职业分布变得更多样和均衡。

社会结构分层形态一直是社会分层研究的经典议题，学者对不同时期的社会分层形态做过各种概括和解读。例如，孙立平提出“断裂化”观点，认为社会分化成相互隔绝、差异鲜明的两部分（孙立平，2003）；陆学艺等人提出“中产化”观点，认为社会中间层日益壮大，底层和顶层逐渐缩小（陆学艺，2002）。本文的

结果显示,在改革开放初期,社会分化加大且存在明显的分层,这在21世纪初期变得尤为明显。但伴随着改革开放的深化、社会保障的完善和更合理的收入分配,社会分化得到有效控制,中间群体越来越庞大。

五、结 语

社会学对社会结构的量化测量长期以来多倚靠宏观社会经济指标或微观社会调查。但由于时间、成本和测量方式的限制,指标的测量依然存在时间跨度有限、测量维度单一、隐性指标不足等问题。而大数据和机器学习为测量社会结构提供了一条新的计算社会学路径:基于图书、史料、政策文件等文本大数据,使用词嵌入等自然语言算法对不同历史时期相对稳定的话语形态进行量化测量,进而反映深层次的社会结构、集体认知和社会观念。基于"话语结构反映社会结构"和"词语关系反映话语结构"的前提,本文对使用文本大数据测量社会结构的方法合理性进行了论述,并提出了通用的测量方法框架。

本文以"职业地位"为例,展示了1940—2015年中国社会结构变迁的具体图景。分析结果显示,中国职业地位的变化呈现明显的时代特征,改革开放之后才出现相对稳定的变化趋势。从不同维度资源的整合情况看,社会地位的四大维度不总是一致的,中国的职业结构呈现结构化和碎片化并存的特点。从宏观的社会地位结构变化看,中国的整体地位结构逐渐从新中国成立前后的"多层分化的塔型结构"转变为特殊时期的"断裂的平均主义结构",再发展成改革开放后的"相对连续的纺锤型结构",总体结构向更合理的方向迈进。而这些发现都为以调查问卷为主的传统职业地位测量方法提供了重要的补充。

基于文本大数据和机器学习的计算社会学路径实际上为社会结构的测量提供了双重价值。第一,直接测量主观社会结构的价值。人类的文化认知、价值观念、意识形态本就是社会结构的一部分,但常常隐藏在非自觉的深层集体观念下难以被有效测量。作为人类文化观念的载体的文本大数据能直接反映人类的观念结构。第二,间接测量客观社会结构的价值。由于观念结构在很大程度上是对客观社会现实的反映,并且会进一步形塑客观社会结构,因此可以基于主观社会结构的折射来间接测量客观社会结构。这实际上能够在缺乏真实数据的情况下为客观社会结构的测量提供很好的补充。而不管是对主观还是客观社会结构的测量,都可以基于文本大数据的属性,进一步在时间、空间和情境上拓展:借助

文本时间标签,对大历史跨度下的宏观结构变迁进行趋势分析;借助文本空间属性,对跨区域的社会结构差异开展比较分析;基于文本的内容属性,对多维度、多情境的社会结构进行立体深描。

但任何方法都有局限性,我们同样需要辩证地反思计算方法测量的挑战。首先,文本中反映的主观社会结构和客观社会结构之间存在一定张力。主观社会结构可以间接折射客观社会结构,但并不能够完全代表真实的社会结构。文化观念是现实的反映,但同样受到文化传统、政治力量、经济环境、媒体建构等多方面的影响,从而导致主观认知与客观现实之间的偏差。从另一方面来说,研究和讨论主观结构和客观结构之间的差异,本身也是社会学的关键议题。其次,文本大数据的类型、叙事方式和时代发表特征可能带来潜在问题。文本数据是分析的基础,不同文本类型代表不同的话语场域和叙事主体,如政策文本、学术文本、小说文本、新闻文本代表着不同群体的视角和叙事方式。研究者也可以比较不同叙事主体和话语场域下的结构差异。同时,文本数据可能存在早期语料稀缺和出版滞后问题,从而导致在时间回溯和新现象挖掘上的不足。最后,以词语为分析单位的测量方法可能存在偏差。一方面,分析词表的选择需要准确细致,但存在一定的主观性;另一方面,不同时代、区域存在特殊用词,且部分词语一词多义,含义也可能随时间变化。因此,分析应尽量选择意义单一、词义稳定的词表,同时使用多个大样本词语来保证分析的稳健性。

本文并不主张使用文本大数据和社会计算的方法来代替传统社会调查的价值。作为一种间接的测量手段,该方法可能不如调查那样直接和准确,但却能够在缺乏问卷数据的情况下为社会结构的测量提供很好的补充,特别是能够在宏观大时空尺度下对多维度社会结构进行全局测量和比较分析(陈云松,2022b)。放眼看来,中国近百年间经历了一系列重大社会变革,社会结构的变迁涉及各个领域,这在人类发展史上也是绝无仅有的宝贵案例。开展以中国为对象、以中国为方法、以中国之治为旨趣的研究,是建构中国自主知识体系的重要议题(陈云松,2025)。但中国的量化社会调查起步较晚,许多历史社会指标时常无迹可寻。这份独特的宝贵财富不能因此被埋没,扎根于丰富文本大数据的社会结构测量方法无疑为处于知识生产后发进程中的中国社会学建设提供了新的可能性。

参考文献:

蔡禾、赵钊卿,1995,《社会分层研究:职业声望评价与职业价值》,《管理世界》第 4 期。

陈婴婴,1995,《职业结构与流动》,北京:东方出版社。

陈云松,2015,《大数据中的百年社会学——基于百万书籍的文化影响力研究》,《社会学研究》第 1 期。

陈云松,2022a,《社会计算在文化社会学中的运用》,《学术月刊》第1期。
——,2022b,《当代社会学定量研究的宏观转向》,《中国社会科学》第3期。
——,2025,《“社会学想象力”的重思与拓展——建构中国自主知识体系的方法自觉》,《中国社会科学》第1期。
迟书君,2003,《深圳人职业声望评价的特点》,《社会学研究》第4期。
高顺文,2005,《我国职业声望研究二十年述评》,《华中科技大学学报(社会科学版)》第4期。
龚为纲、朱萌、张赛、罗教讲,2019,《媒介霸权、文化圈群与东方主义话语的全球传播——以舆情大数据GDELT中的涉华舆情为例》,《社会学研究》第5期。
胡春阳,2007,《话语分析:传播研究的新路径》,上海:上海人民出版社。
黄国文、徐珺,2006,《语篇分析与话语分析》,《外语与外语教学》第10期。
赖特,埃里克·欧林,2006,《阶级》,刘磊、吕梁山译,北京:高等教育出版社。
李春玲,2005a,《当代中国社会的声望分层——职业声望与社会经济地位指数测量》,《社会学研究》第2期。
——,2005b,《断裂与碎片——当代中国社会阶层分化实证分析》,北京:社会科学文献出版社。
李路路,2003,《再生产的延续:制度转型与城市社会分层结构》,北京:中国人民大学出版社。
李培林,1992,《另一只看不见的手:社会结构转型》,《中国社会科学》第5期。
——,2011,《中国改革以来阶级阶层结构的变化》,《黑龙江社会科学》第1期。
李强,2000,《转型时期冲突性的职业声望评价》,《中国社会科学》第4期。
——,2008,《从“整体型社会聚合体”到“碎片化”的利益群体——改革开放30年与我国社会群体特征的变化》,《新视野》第5期。
李强、刘海洋,2009,《变迁中的职业声望——2009年北京职业声望调查浅析》,《学术研究》第12期。
梁玉成,2024,《基于生成式大语言模型的“测试社会学”》,《探索与争鸣》第11期。
刘河庆、梁玉成,2021,《政策内容再生产的影响机制——基于涉农政策文本的研究》,《社会学研究》第1期。
刘欣、田丰,2018,《社会结构研究40年:中国社会学研究者的探索》,《江苏社会科学》第4期。
陆学艺,2002,《当代中国社会阶层研究报告》,北京:社会科学文献出版社。
——,2010,《当代中国社会结构》,北京:社会科学文献出版社。
仇立平,2001,《职业地位:社会分层的指示器——上海社会结构与社会分层研究》,《社会学研究》第3期。
孙立平,2003,《断裂——20世纪90年代以来的中国社会》,北京:社会科学文献出版社。
韦伯,马克斯,2010,《经济与社会》,阎克文译,上海:上海人民出版社。
谢宇,2018,《走出中国社会学本土化讨论的误区》,《社会学研究》第2期。
辛斌,1996,《语言、权力与意识形态:批评语言学》,《现代外语》第1期。
辛志英,2020,《话语分析:理论、方法与流派》,厦门:厦门大学出版社。
许欣欣,2000,《从职业评价与择业取向看中国社会结构变迁》,《社会学研究》第3期。
——,2005,《社会、市场、价值观:整体变迁的征兆——从职业评价与择业取向看中国社会结构变迁再研究》,《社会学研究》第4期。
张铮、吴福仲、林天强,2021,《“未来定义权”视域下的中国科幻:理论建构与实现路径》,《南京社会科学》第1期。

郑杭生、赵文龙,2003,《社会学研究中"社会结构"的涵义辨析》,《西安交通大学学报(社会科学版)》第2期。

Bian, Yanjie 1996, "Chinese Occupational Prestige: A Comparative Analysis." *International Sociology* 11(2).

Biber, Douglas, Susan Conrad & Randi Reppen 1998, *Corpus Linguistics: Investigating Language Structure and Use*. New York: Cambridge University Press.

Bourdieu, Pierre 2018, "The Forms of Capital." In Mark Granovetter & Richard Swedberg(eds.), *The Sociology of Economic Life*. New York: Routledge.

de Saussure, Ferdinand 2004, "Course in General Linguistics." In Julie Rivkin & Michael Ryan(eds.), *Literary Theory: An Anth*ology. Oxford: Blackwell Publishing.

Duncan, Otis Dudley 1961, "A Socioeconomic Index for All Occupations." In Albert J. Reiss(ed.), *Occupations and Social Status*. New York: The Free Press of Glencoe.

Fairclough, Norman 2013, *Critical Discourse Analysis: The Critical Study of Language*. New York: Routledge.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky & James Zou 2018, "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115(16).

Harris, Zellig S. 1954, "Distributional Structure." *Word* 10(2/3).

Hauser, Robert M. & John Robert Warren 1997, "Socioeconomic Indexes for Occupations: A Review, Update, and Critique." *Sociological Methodology* 27(1).

Keller, Reiner 2011, "The Sociology of Knowledge Approach to Discourse (SKAD)." *Human Studies* 34.

Kennedy, Graeme 2014, *An Introduction to Corpus Linguistics*. New York: Routledge.

Kozlowski, Austin C., Matt Taddy & James A. Evans 2019, "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 84(5).

Lemert, Charles C. 1979, "Language, Structure, and Measurement: Structuralist Semiotics and Sociology." *American Journal of Sociology* 84(4).

Levy, Omer, Yoav Goldberg & Ido Dagan 2015, "Improving Distributional Similarity with Lessons Learned from Word Embeddings." *Transactions of the Association for Computational Linguistics* 3.

Lin, Nan & Wen Xie 1988, "Occupational Prestige in Urban China." *American Journal of Sociology* 93(4).

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman & Slav Petrov 2012, "Syntactic Annotations for the Google Books Ngram Corpus." Paper presented at the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, July 8 – 14.

McEnery, Tony & Andrew Hardie 2012, *Corpus Linguistics: Method, Theory and Practice*. New York: Cambridge University Press.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean 2013, "Efficient Estimation of Word Representations in Vector Space." *ArXiv*(https://arxiv.org/abs/1301.3781).

Treiman, Donald J. 1977, *Occupational Prestige in Comparative Perspective*. New York: Academic Press.

作者单位:南京大学社会学院

责任编辑:刘保中