中国网络大众生育态度倾向变迁*

——兼论舆情大数据在人口学中的应用

李 婷 袁 洁 夏 璐 熊英宏 张露尹

【内容摘要】文章探索了如何利用网络舆情大数据来推断大众生育态度倾向的变迁。基于对网络评论的简单情绪分类、LDA 非监督分类和 fasttext 监督分类结果的比较，文章发现依托于计划行为理论框架的监督分类法能提供更准确也更有意义的分析结果。从趋势来看，自 2012 年以来，涉及生育的评论态度倾向从积极占多数快速转变为消极占主导，这个过程也伴随着话题维度的转变，从更多行为态度上的表达转变为主要陈述生育面临的客观限制。房价、子女教育、工作成为提及最多的限制条件。这些结果说明低生育意愿更多是受客观限制的影响，而非观念转变。分省市生育情绪与地方生育水平较强的相关性体现了生育态度倾向分析的有效性。文章借此进一步讨论了舆情大数据在人口学研究中的应用前景。

【关键词】生育态度倾向；舆情大数据；计划行为理论

【作者简介】李婷，中国人民大学人口与发展研究中心副教授；袁洁，中国人民大学信息学院硕士研究生；夏璐，中国人民大学统计学院本科生；熊英宏、张露尹，中国人民大学社会与人口学院本科生。北京：100872

The Transition of Fertility Intention Propensity of Chinese Web-Users: The Application of Public Opinion Big Data Analysis

Li Ting Yuan Jie Xia Lu Xiong Yinghong Zhang Luyin

Abstract: This study explores how to use the online public opinion big data to infer the transition of fertility intention propensity of Chinese web-users. By comparing the results from the simple sentiment analysis, the LDA un-supervised classification, and the fasttext supervised classification, it is found that the supervised classification based on the theory of planned behavior yields more accurate and meaningful results. A change from the positive to negative dominant mood of fertility intention has occurred since 2012, during which there is a shift of topics. The comments regarding perceived controls have replaced those related to behavioral beliefs as the most popular ones for fertility. Housing price, children's education, and work are the most mentioned key words in the category of perceived controls. The strong correlation between fertility intention propensity and fertility level among Chinese provinces validates our fertility intention analyses. We also discuss the prospect of applying public opinion big data in demographic research.

Keywords: Fertility Intention Propensity, Public Opinion Analysis, Theory of Planned Behavior

Authors: Li Ting is Associate Professor, Center for Population and Development Studies, Renmin University of China; Yuan Jie is Graduate Student, School of Information, Renmin University of China; Xia Lu is Undergraduate Student, School of Statistics, Renmin University of China; Xiong Yinghong and Zhang Luyin are Undergraduate Students, School of Sociology and Population Studies, Renmin University of China. Email: li. ting@ ruc. edu. cn

* 本研究受到霍英东教育基金会青年教师基金项目“大数据视角下大众二胎态度探析”(161085)和国家自然科学基金重大项目“特征、规律与前景——老龄社会的人口学基础研究”(71490731)的支持。

1　导言

自 20 世纪 90 年代总和生育率下降到更替水平以下，我国生育率一直在低水平徘徊。为了应对中国社会日趋严重的人口老龄化以及人口红利消减等问题，我国在近几年内对生育政策做出了系列调整。然而，随着 2016 年全面两孩政策的实施，我国生育率并没有出现预期的强势反弹，这引发了社会和舆论的普遍忧虑，认为中国已经陷入了严重的生育危机。学术界虽然普遍认同生育政策的放开并没有改变我国目前低生育水平的现状，但是却对中国人的生育观念是否发生彻底转变存在一定争论（杨菊华，2015）。

对于这个问题的研究，实质上是在探索生育政策放开以后，中国生育水平仍然无法有效提升的重要原因。生育率低下是因为年轻人普遍推迟婚姻和生育的进度效应引起的？还是由真实的生育意愿降低造成的？如果是缘于真实的生育意愿降低，那么这种降低是由于客观条件（经济、住房等）的限制，还是因为中国像很多西方国家一样正在经历第二次人口转变，生育不再成为个体的义务以及实现生命价值的手段，从而彻底动摇了多子多福的传统生育观念（Lesthaeghe，2014；刘爽等，2012）。理解这个问题有非常强的现实意义，只有了解生育率低下背后的真实逻辑，才可能制定有效政策促进生育率回升或者帮助家庭实现发展需求。

传统的人口学对于生育观念和态度的研究主要借助于问卷调查和田野访谈（侯佳伟等，2014）。虽然调查访问是社会科学领域最基本和有效的数据收集手段，但是在生育态度的研究上，存在一些重要的缺陷。一方面，观念和态度这类主观性强、抽象度较高的问题很难在问卷中得到准确量化。通常来说，这些问题会以五分或七分李克特量表的形式呈现，以区分态度从负到正的不同等级。但是由于个体对等级尺度的认知基础不同，难以对得分进行完全对等的比较。同时，简单的打分也很难测量个体对该问题内涵和外延的全面理解。生育态度虽然不是一个抽象性很高的概念，但是在实地问卷调查中，迫于各种各样的压力，获得的信息可能不是个体最真实的想法。另一方面，比起其他的态度观念，生育是一个可以快速转变的态度，它会受到周围的社会环境以及群体行为的影响。问卷调查特别是有全国代表性的调查通常时效性比较低，在生育政策变化频繁，生育态度快速转变的时代，很难及时追踪这种变化或者体现生育态度的不确定性。

近些年来，大数据的飞速发展及其在各个科学研究领域的全面渗透，为研究大众生育态度及其变化提供了一个新途径。大数据本身所具备的一些优点恰好可以弥补传统问卷调查和田野访谈中的不足。首先，大数据具有较高的时效性和经济性。因为数据即时抓取的特征，它们能随时反映最新发生的情况，解决大部分问卷调查的滞后性，并且大大降低了研究的成本（孙秀林，2015）。其次，大数据能弥补传统调查中对观念性、文化性、理解性等抽象特征测量的不足（Inati et al.，2014；孙秀林、施润华，2016）。在大数据系统中，我们可以利用个体的语言以及即时情绪上的反应来间接体现这些抽象信息。具体来说，我们可以通过爬虫获取网络上大众对生育相关问题的评论，经过文本分析辨析出个体对生育的态度倾向。这也是新闻传播学中网络舆情分析的一种（李弼程，2015）。

然而在社会科学中要利用大数据进行严谨的科学研究，也面临着很大的挑战，这与大数据本身的特性息息相关。从根本上讲，大数据并不是为研究专门设计，通常来自对大众日常活动中所产生的数字痕迹的追踪。这就导致了大数据具有非代表性、有效信息含量低、内容非结构化、缺乏逻辑范式等研究应用中的弱点（孙秀林、施润华，2016）。那么在严谨的社会科学研究中究竟能否应用大数据？本研究将以中国网络大众生育态度的舆情分析为例，一方面揭示我国网络民众生育态度的变迁；另一方面也就此探讨在人口学研究中应该如何利用大数据和舆情分析。

2　研究设计

有别于传统的社会学和人口学研究，本研究需要充分考虑到网络舆情大数据的优点和劣势，以便

制定适当的研究设计。依据一般网络舆情分析的步骤，第一步需要通过爬虫获取相关网络评论。第二步进行文本分析，将获得的评论进行分词处理，再借助语义分析中的情绪分析技术来获得对单条评论的情绪分类（快乐、生气、悲伤）或者正负态度倾向。最后一步在集合层面获得整体的情绪或者态度趋势。然而在针对生育态度的研究上，我们需要对以上的程序做更为精细的设计，以克服大数据的一些天然缺陷。

2.1 概念界定

研究设计首要的一项任务是对研究概念进行界定。在问卷调查中，生育意愿是一个比较容易被界定的概念，它是指人们对自己生育行为和结果的主观愿望和要求，涵盖理想的子女数量、性别和生育子女的时间等（顾宝昌，2011）。期望生育子女数、生育意向和生育计划是常用的测量指标（郑真真，2014）。然而在进行舆情分析时，绝大部分评论并不是针对生育意愿问题的回答，因而我们是无法直接获得对生育意愿的测量的。我们只能从每一条评论中判断个体对生育的态度，获得的更多是一种倾向判断。因此在本研究中，我们采用生育态度倾向这一概念，它反映了个体在评论中所体现出的对生育或者生育更多孩子的意愿/行为的赞成或者反对的倾向，本文也把它称为生育情绪。

对于这个概念还有两点需要澄清。首先，我们没有也无法限制生育态度是针对一孩、二孩还是多孩的。我们只能认定凡是认为“无论如何应该生育一个孩子”或者“孩子越多越好”都是正面情绪，而认为“不生孩子也挺好”或者“孩子质量比数量更重要”都是负面情绪。其次，我们也无法规定态度倾向是针对自身的生育意愿，还是从社会层面出发的生育态度。因此，“希望至少生两个孩子”“支持全面放开生育”“人多力量大”等都被纳入正面情绪中，“一个也不想生”“中国问题就是人太多了”等都被纳入负面情绪中。这种概念边界的模糊是由舆情大数据的完全开放性和数据收集的被动性特征所决定的。

2.2 基础数据爬取

基础数据获得的第一步是需要对网络爬取对象的范围进行界定，要考虑潜在的数据有效信息含量低的问题。大多数简单舆情分析都是针对特定事件，因而事件本身比较好界定，相应评论范围也好确定。但生育是一个较为广泛探讨的问题，特别是要考察一个相对比较长的时期的变迁，话题比较分散，搜索边界并不明确。我们尝试获取微博随机选取的 3 个月内所有涉及生育、生孩子、二胎等关键词的主题微博或者转发评论，共计 28 万多条，但是经过人工随机抽取检查，能体现个人生育态度的条数不到 1%。大量的无关信息对后续准确识别生育态度产生了巨大的干扰。这也就是大数据的信息稀薄问题。

为此，我们将搜索对象限定在新闻媒体发布的、与生育政策或者生育水平讨论有关的新闻网络评论上。我们将主要借助新浪微博和网易新闻这两个媒体中介。选取微博的好处在于，各大主流新闻媒体均在微博上有发布账号，方便新闻集合搜索，并且利用微博嵌入的评论功能，可以便捷地获得对相应新闻的网络评论。网易新闻也是较早开放新闻即时评论的媒体平台。除了自写新闻以外，网易也大量转载其他媒体新闻，可以作为微博评论的有力补充，避免使用单一平台的偏误。选用这两个平台的另一个优点在于，它们均很少对评论进行筛选处理，特别是对敏感性并不高的生育话题，不会存在很高的删除偏误。

最后，我们将信息爬取对象界定为在新浪微博和网易新闻平台上自 2012 年 1 月 1 日到 2018 年 11 月 10 日间所有与中国生育政策、生育水平或者生育意愿相关的新闻的网络评论。为了进一步排除信息稀薄的干扰，我们的实际操作分为三步：第一步爬取界定日期范围内所有以新浪微博媒体认证的账号发布的涵盖生育政策、生育水平或者生育意愿等关键词的主题微博，以及包含类似关键词的网易新闻标题，共计获得 794 条微博新闻和 995 条网易新闻；第二步对所获得的新闻内容进行筛选，确保

纳入的新闻主题是与中国生育政策、生育水平或者生育意愿等问题高度相关的，由此获得两个平台一共 167 条新闻；第三步是获取这 167 条新闻的全部评论，最终爬取 61090 条评论，作为分析的基础样本。除了评论本身，每个样本还获取了评论的时间、评论者网络标注的性别以及根据 IP 或者自填信息获得的评论者所在的省级行政单位这几个变量。

2.3　样本的代表性问题

在进入下一步样本分析之前，我们需要先来讨论非常重要的样本的代表性问题。大数据虽然数据容量大，但是却常常在数据代表性问题上存在巨大缺陷（孙秀林、施润华，2016）。这与很多大数据是被动数据的特性有关。也就是说，我们通常只能获取留下数字痕迹的样本，而这些样本在年龄性别结构、受教育程度等多方面存在巨大选择性。因此，在大多数情况下，使用网络大数据并不能对总体进行推断。

在本研究中，我们需要采取以下一些策略来应对样本代表性问题。首先，将研究总体界定为网络大众。当然，这也不能完全解决评论样本的自选择性，这就决定了在本研究中，我们的研究目的不是获得总体的生育态度分布，而是探索态度的变化趋势。其次，要保证趋势研究的可靠性，我们还需要对样本特征进行考察，因为不同年份趋势的可比性是建立在不同年份样本的可比性上的。根据新浪微博用户发展报告，虽然每一年份的用户构成都会发生一定程度的改变，但是微博的使用主体一直是 19 ~ 35 岁的人群，其占比在 70% 以上；男女性别比例相当；受过高等教育的用户比例稳定在 70% ~ 80% 之间；一二三线城市用户达到 70% 左右（新浪微博数据中心，2017）。而网易新闻用户主体是 25 ~ 40 岁的人群；男性用户占比为 60%；高等教育用户占比在 50% 以上；地级市及以上用户占到 70%（易观，2016；人人都是产品经理，2018）。可以看出，这两个平台使用者的特征在观察期间并没有发生显著改变，主要涉及 19 ~ 40 岁处在生育高峰期且有较高受教育程度的群体。因此，本研究生育态度变迁的主要推及人群为城市地区的中青年，而这部分人群也被普遍认为当前生育意愿较为低下（宋健、陈芳，2010），对这一人群的研究具有较强的现实意义。

2.4　态度倾向分类

态度倾向的分类识别是整个研究的核心。在获得评论样本之后，计算机需要对文字进行处理，变成机器可识别语言，进而再基于特定规则对语言文字进行分类判断。这个过程也就是计算科学与人工智能领域中的自然语言处理（Natural Language Processing，NLP）。NLP 本质上希望实现人与计算机之间用自然语言进行有效通信交流，包括分词、主题提取、语义和情绪分类等具体应用技术（王灿辉等，2007）。在本研究中，我们的态度识别实际包含了文本预处理和态度倾向分类这两步。

文本预处理也就是数据清理，在本研究中，我们要先进行分词处理，即将评论转化成由单个词组成的词向量。在获得分词结果之后，我们还需要再移除标点符号、过滤无意义词，最终得到可供分析的语料矩阵。

2.4.1　简单情绪分析

在进行最重要的态度倾向分类这一步时，我们有几种选择。第一种是将从文本样本中提取的主要词汇与已有的情感语料库进行比对打分。这些情感语料库通常是已经训练好、代表一般意义下的词的情绪。比如，出现“开心”这个词就会对这个评论做积极情绪的判断，出现“失望”则会做消极判断。我们采用 Python 的中文文本分析程序包“snownlp”，并利用其默认的情感语料库来进行分析，得出每一个评论呈现积极情绪的概率。

这种方式虽然操作比较简单，但是只能判断话语本身的情绪，而语言本身的情绪并不完全等同于态度，情绪的积极并不代表态度的赞同。同时，态度本身除了积极与消极这一个维度外，还有着更为

丰富的内涵。这就要求我们对态度倾向的分类超越积极和消极这一单一维度，接下来的两种选择可以在一定程度上实现这一目标。

2.4.2 非监督分类

首先是非监督分类法。这一类方法不要求提前设定分类类别，而是通过机器学习挖掘文本信息特征，自动划分成典型类别，再将文本进行归类。文档主题生成模型（Latent Dirichlet Allocation，LDA）是一种典型的非监督分类法，它通过机器学习来识别文本或文档集中潜藏的主题信息。简单来说，该方法利用文本预处理所得到的词频向量，将文本信息转化为易于建模的数字信息。通过统计词频的方式，形成主题单词构成和评论主题构成的多层概率分布，进而实现分类（邹晓辉、孙静，2014）。也就是说，我们提取所有评论中关于生育态度倾向的主题单词，形成评论主题类别及其概率分布，再依据主题类别对评论进行分类。我们利用 Python 里的“gensim”模块来实现 LDA 分析。

2.4.3 监督分类

另一种方法就是监督分类了。与非监督不同，监督分类需要提前设定好预期的分类类别以及训练样本，经过对训练样本的机器学习建立分类器，再实现对所有样本的分类。这种分类方法面临的最大挑战在于如何设定类别。类别的划定从本质上来讲是建立理解生育态度倾向的视角，一种更好的选择是依托于已有的理论框架。一个经典的探索生育意愿行为的框架为 Ajzen（2000）提出的计划行为理论。这个理论不仅展示了从生育意愿、生育计划再到生育行为的逻辑链条，同时提出了行为意愿背后的 3 个认知维度：行为态度（Behavioral Beliefs）、主观规范（Subjective Norms）和知觉行为控制（Perceived Controls）。其中，行为态度是指个体对特定行为（这里指生育）的总体评价，是基于个人的价值观念所反映的对该行为的喜好程度。例如，个体认为“孩子越多老年越有保障”代表对生育的正面态度，而认为“生孩子不再是个体实现生命价值的方式”则代表了在该维度上的负面态度。主观规范则反映个体在决定是否执行该行为时所感知到的社会压力，它可以包括感受到的家人的催生压力、朋友们的生育态度或是其他人的生育行为。知觉行为控制是指个体对执行该行为的难易程度的感知。在针对生育的问题上，像“没人帮带孩子”“现在养孩子太贵”“房价太高”之类的表述都属于知觉行为控制上的负面态度。

值得注意的是，在对评论样本进行初步检索之后，我们发现除了上述 3 个认知大类外，还有一组比较明显的类别。这些评论虽然涉及对生育的态度，但是其出发点并不是个体的生活抉择，而是更为宏观的视角，例如，认为“中国人已经太多了”是典型的负面态度，认为“中国老龄化问题严重应该鼓励生育”则是正面态度。我们将这类评论单独剥离出来，将其命名为宏观态度以区别于从个体视角出发的行为态度类别。相应的，我们把后者命名为个体态度以示区别。

在以往的研究中，这些认知维度都是通过问卷中专门设计的量表来测量的（Ajzen and Klobas，2013）。在本研究中，我们将这 4 类认知作为划分生育态度倾向的标准。也就是说，我们希望通过对评论的分析，辨析出该条评论所涉及的态度是否属于这 4 类认知维度中的一类。在每一类别中，我们又把态度划分为正向、中性、负向与无关这 4 种情绪类别，分别对应于在该类认知维度上倾向于生育或者认为应该多生育、对生育多少持无所谓或者自由态度、不倾向于生育或者反对多生育以及与该类别态度无关的评论。在实际操作中，这 4 类认知维度并不是完全互斥的，也即同一条评论可能同时涉及多个类别的认知。因此，我们会建立多个分类器，依次辨别同一个评论在每个认知维度上的情绪取值。

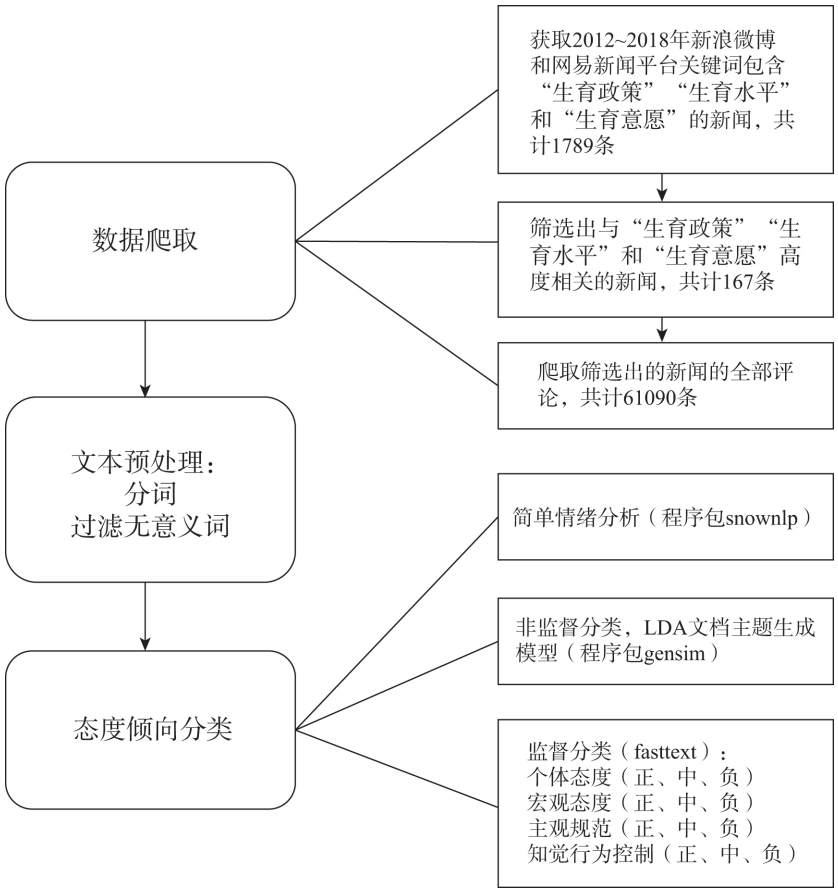
利用计划行为理论的认知基础来实行监督分类除了提供一个分类划分标准之外，更重要的是这个框架能帮助我们更理论、更系统地来审视生育态度倾向的可能变迁。行为态度更多反映了个体的生育价值观，主观规范代表生育文化环境，而知觉行为控制则揭示生育行为可能面临的困境。根据西

方经典的第二次人口转变理论，西方发达国家陷于低生育率的主要原因是由个人价值观念转变所引起的生活方式抉择的变化(Lesthaeghe，2014) 。一直以来，关于中国特别是城市地区是否也在经历第二次人口转变存在争议(刘爽等，2012) 。一个值得探究的问题是近些年来即便有了生育政策的调整，中国的生育率也仍旧维持在较低水平的原因。是个体价值观念的转变，还是生育文化环境的变化，抑或是现实条件的抑制在起最主要的作用？依据这个分析框架，可以为我们理解这个问题提供线索。

在本研究中，我们利用 facebook 开源发布的 fasttext 工具来实现监督分类。fasttext 能提供简单而高效的文本分类和表征学习的方法(Joulin et al. ，2016) 。在分析的过程中，我们首先需要选择一批训练样本，按照上面提到的分类框架对它们进行人工分类标记，再利用标记好的数据训练模型分类器，最后使用模型分类器对其他评论进行分类预测。我们在样本库中选择了新浪微博 2012 ~ 2018 年每年 1 条微博(共计 7 条) 的全部评论作为训练样本，一共人工标记 8962 条评论。研究的全部执行流程可以参见图 1。

图 1　研究的设计和执行流程

Figure 1　Research Design and Implementation Flow



3　研究结果

3.1　简单情绪分析结果

如上文所述，我们利用 snownlp 程序包对获取的评论进行简单情绪判断，输出的结果是每一条评论是积极情绪的概率。对所有的样本来说，得到积极情绪的平均概率是 60.5% 。为了检验情绪

分析的效果，我们将在监督分类中人工标识出的 8962 条样本内部进行结果比对。我们将 snownlp 概率大于 80% 的训练样本设为正向情绪，共得到 2576 条“正向”评论。然而，这些“正向”样本中仅有 699 条被真实标记为正向，正确率不足 30%。我们继续将正向概率阈值提高到 90%，正确率也没有显著变化。这个结果说明，利用常规的情感语料得到的情绪分类，并不能体现个体对像生育这样复杂情绪的态度。简单的情绪分类分析并不适用于大多数涉及复杂概念的社会科学研究。

3.2 非监督分类结果

LDA 分析首先给出了评论的词频统计，除了比较中性的“人口”“生育”“计划生育”“国家”之外，“不生”“养不起”“不想”等带明显负面生育情绪的词汇出现频率很高。这是因为消极生育态度倾向容易通过否定词捕捉，但是积极态度就很难通过词汇本身判断了。

我们利用 LDA 的主题提取功能获得了 5 个主要的主题聚类，然而从这些关键词中，我们无法明确分辨出每个主题的意义及其相互间的差异，例如，主题 1 和主题 2 分别出现最多的词是“养不起”和“不生”，都代表负面生育态度倾向。由此可见，利用 LDA 进行非监督分类对生育态度情绪的辨识也不理想，无法分辨提取主题的特征意义，所得的分类也就变得无意义。LDA 非监督分类效果不佳的原因，一是因为生育相关的评论话题比较开放，二是因为主题提取一般被证明对文档分析比较有效，评论通常以短句为主，使得样本中关键词比较少，无法发挥 LDA 的真正效果。

3.3 监督分类结果

按上文所述，监督分类对每一条评论在 4 个认知维度上分别执行分类，获得在每一维度上正向、中性、负向和无关的赋值。我们把训练样本划分为 4∶1 的训练集和测试集，以验证 fasttext 的分类效果。结果发现测试集中个体态度、宏观态度、主观规范和知觉行为控制的正确率分别为 70%、94%、92% 以及 87%。除了行为态度的正确率稍低以外，其他类别的正确率都维持在比较高的水平。对于考察生育态度这么复杂的社会观念来说，这个程度的正确率是可以接受的。

基于训练集生成的分类器，我们对所有样本进行分类预测。除了 4 类认知维度外，我们还需要计算一个总体态度倾向。我们将每条评论在单个维度上的正向态度赋值为 1，负向态度赋值为 -1，中性态度赋值为 0。当各个维度加总分为正时，总体态度确定为积极；加总分为 0 时，总体态度为中性；加总分为负时，总体态度为消极。如果在所有维度上的判断都是无关，则该条评论被划为无关评论。在全部的 61090 条样本中，我们最终获得 25590 条可辨别态度的样本，只占整体样本的 42%。可见，虽然我们对样本获取做了诸多聚焦处理，仍然要面对大数据的信息稀薄问题。

表 1 展示了样本在 4 个认知维度和总体态度上的倾向分布。涉及这 4 个维度的评论分别占总样本的 46.2%、3.6%、4.2% 和 54.3%，其中，表达个体态度和知觉行为控制的评论最多。需要注意的是，这些百分比的总和不等于 1，因为一个评论可能会涉及两个或多个维度的认知态度。具体来看，总体生育态度中有 78.5% 是负向的；个体态度中有超过 50% 是负向；提到知觉行为控制的评论基本都是负向态度；在涉及宏观态度的评论中，有接近 80% 认为中国人太多，不应该放开或鼓励生育；主观规范中中性占比较多，代表感受周边的态度是两极分化的。由于样本并不具备严格的代表性，我们不能从样本的态度分布推断总体的态度分布，这里更有意义的是观察认知维度和态度倾向的变化趋势。

图 2 展示了各类认知维度占比随年份的变化趋势。可以看出，在评论中涉及个体态度的占比从 2012 年的 60% 多下降到 2018 年的 40%，而谈论知觉行为控制的占比从 2012 年不到 15% 上升到 2018 年的 60%。宏观态度虽然整体占比不高，但是在整个区间内也呈下降趋势。主观规范在整个区间内呈现波动变化，在“单独二孩”和“全面两孩”政策提出的 2012 和 2015 年占比相对较高。

表 1　生育态度倾向在 4 个认知维度以及总体上的分布

Table 1　The Distribution of Fertility Intention Propensity by Cognitive Dimension

	个体态度	宏观态度	主观规范	知觉行为控制	总体态度倾向
正向(%)	39.7	21.0	25.3	0.0	17.2
中性(%)	3.2	0.3	32.4	0.0	4.3
负向(%)	57.1	78.6	42.3	100.0	78.5
合计条数 (占总样本百分比)	11813 (46.2)	932 (3.6)	1063 (4.2)	13885 (54.3)	25590

资料来源:表中数据为本研究依据从新浪微博和网易新闻中爬取的数据计算得到,后文图表资料来源同表 1。

图 2　各类认知维度占比随年份的变化趋势

Figure 2　The Change of Cognitive Dimension Distribution by Year

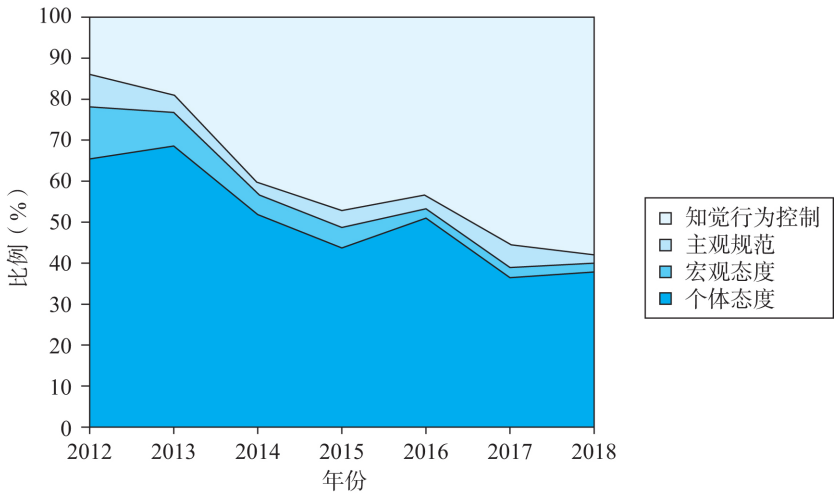


图 3 展示了总体态度倾向随年份的变化趋势。从中可见,积极生育态度倾向的占比从 2012 年的 50% 以上快速下降到 2018 年的 10%,而消极态度倾向在 2018 年则占据了几乎 90% 的有关生育的评论。这种总体态度的变化与所涉及的评论维度的变化有一定关系。例如,提及知觉行为控制的评论基本都是负面态度表达,随着知觉行为控制的占比提高,负面态度倾向也会相应显著提升。当然,态度本身的变化也是导致总体态度倾向发生转变的重要原因。由于宏观态度和主观规范在整体评论中占比很低,并且这两个维度内部态度倾向变化不大,因此我们需要重点关注在个体态度上的情绪变化。

图 4 展示了在个体维度上生育态度倾向的时期变化。如图 4 所示,个体维度上正向态度的占比从 2012 年超过 60% 滑落到 2018 年低于 30%。其中,2013 ~ 2015 年的降低幅度最大。这种态度的转变一方面可能与参与的话题有关,2015 年以前更多是针对可能放开生育的新闻,参与的评论更有可能是表达支持放开等积极的生育态度,而 2015 年以后更多是针对生育政策效果和生育水平的报道,参与的评论大部分讨论为什么生育率没有实现大幅度反弹,相应态度也会更消极。但是另一方面,我们也不可否认个体对生育本身的抉择态度也在变得更加消极,其在 2016 ~ 2018 年又有一个明显的下降。

综合来看,总体的生育态度倾向持续变得消极是由提及知觉行为控制的评论增多以及个体行为态度转变为负面共同导致的。这两种变化均表明在谈及生育这个话题时,近些年来的网络情绪转向了消极。

图 3 总体态度倾向随年份的变化趋势

Figure 3 The Change of Overall Fertility Intention Propensity by Year

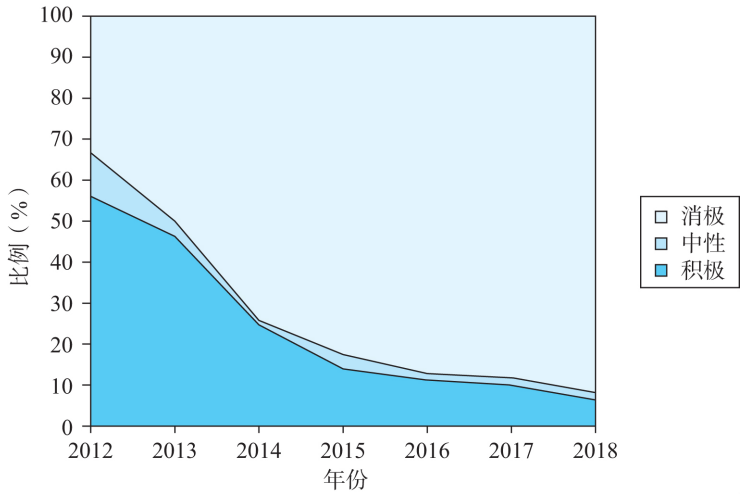
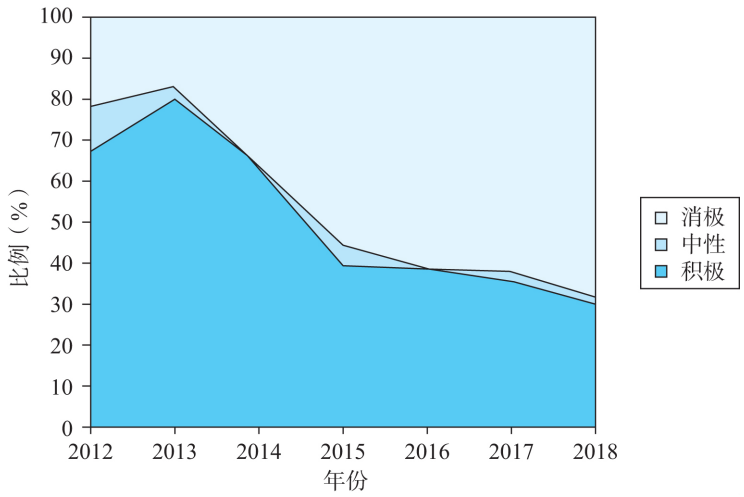


图 4 个体态度倾向随年份的变化趋势

Figure 4 The Change of Fertility Intention Propensity of Individual Behavioral Beliefs by Year



既然知觉行为控制成为近几年来主导生育态度倾向的认知维度，我们有必要进一步对该维度进行词频分析并制作词云图，以探索是哪些因素成为个体最常感知到的对生育的现实阻碍。如图 5 所示，除了直接表达态度的“养不起”“不生”之外，被提及最频繁的生育抑制条件是“房价”“房子”这两个跟住房相关的词，其次是“教育”和“幼儿园”这些跟孩子培养相关的词，再其次就是“生活”“工作”和“发展”这些代表家庭和个人发展关系的关键词。综合来看，网民在对生育限制条件的感知中，房子、孩子教育和工作发展成为其中最重要的现实考量。住房成为生育的首要感知限制，这可能跟住房购买对经济资源的强大挤占效应有关（唐重振、何雅菲，2018）。高企的房价大量消耗了年轻人的生活成本，从而降低了他们对生育投入的资源和信心。

我们下一步再来看分性别的认知维度和态度倾向的差异（见图 6）。从认知维度来看，女性有显著更多的评论涉及知觉行为控制和主观规范，说明女性更多地感受到生育的客观条件限制并同时受主观环境规范的影响更大。相对而言，男性评论会涉及更多的宏观视角。就生育态度倾向而言，女性

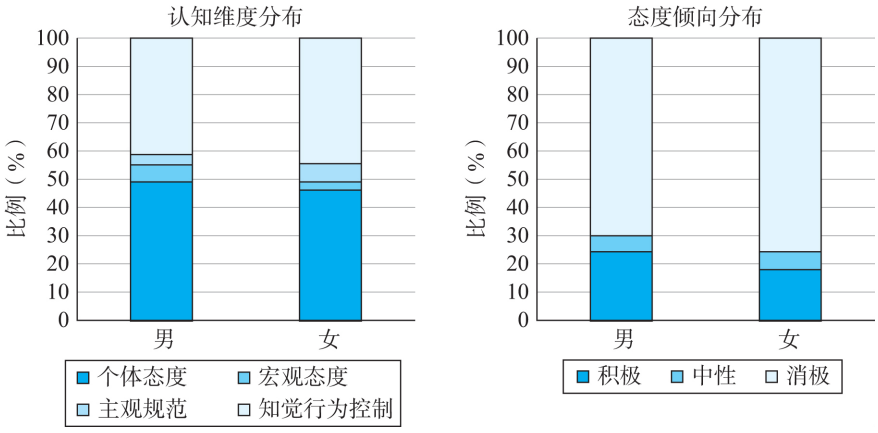
显著更为消极。性别差异的结果比较符合认知预期，这也从侧面证明了大数据舆情分析的有效性。

图 5　知觉行为控制词云图

Figure 5　Word Cloud of Perceived Controls

图 6　分性别认知维度和态度倾向分布

Figure 6　The Distribution of Congitive Dimension and Intention Propensity by Sex



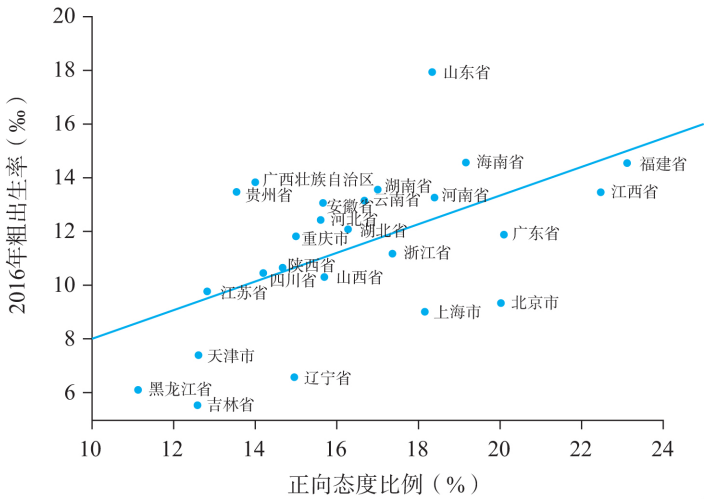
大数据分析的实际应用意义是要建立在有预测价值的基础上的。由于数据涉及的年份有限，并且生育政策在过去几年变动较大，我们无法检验态度倾向与全国生育水平之间的关系。但是我们可以借助评论的地理变量，在一个截面时期上观察各省级行政单位生育态度倾向与实际生育水平之间的关系。这 25590 条有效评论，除去地域是海外和其他无法识别地区的 3500 多条外，覆盖了我国 31

个省(市、自治区)以及港澳台地区。在排除掉港澳台地区之后,我们继续删除了西藏、新疆、宁夏、内蒙古和甘肃这几个省(市、自治区),一来因为这些地区少数民族人口较多,生育政策和生育观念与其他地区有较大差别,二来因为这些地域的评论条数较少(小于 120 条),显著低于其他省(市、自治区)的样本量(平均在 800 条),可能产生较大样本偏差。对于剩余的样本,我们考察各省(市、自治区)生育态度倾向中正向情绪占比与 2016 年各省(市、自治区)人口粗出生率之间的关系,并做简单线性回归。

如图 7 所示,各省(市、自治区)的正向态度倾向占比与 2016 年的人口粗出生率呈现较强的线性关系,回归的 R^2 达到 0.3。对于单个变量来说,预测的效果还是可以接受的。可以看出,出生率持续低迷的东北三省,网络评论的负面情绪也比较高,而生育文化更传统的福建、江西、广东、河南等地,正向情绪占比相对领先于其他省(市、自治区)。这些结果都符合预期判断,说明通过生育情绪分析来推测生育水平变动具有一定的可行性。我们再来考察那些对回归直线偏离较大的样本,其中,山东较高的出生率对应于二胎政策后生育的强势反弹。另外比较突出的是上海和北京这两个直辖市,它们的出生率一直处于较低水平,然而其正向态度比例却相对较高。可能的原因是使用北京和上海 IP 地址的网民涵盖大量流动人口,其生育行为可能并不发生在这两个城市中。此外,如果以自填地区来看,也会有不少非北京和上海的常住人口倾向于把个人定位设置为这两个城市,由此会产生样本偏误。当我们除去北京和上海这两个城市后,回归的 R^2 上升到 0.42。

图 7 各省(市、自治区)2016 年粗出生率对正向生育态度倾向占比的线性回归

Figure 7 Regression of Crude Birth Rate on Positive Fertility Intention by Province



为了验证结果的稳健性,我们又将各省(市、自治区)的正向态度倾向占比与 2015 年的粗出生率做回归。利用全部样本和除去北京、上海样本的回归所得 R^2 分别为 0.26 和 0.43,显示了结果的稳健性。

4 结论与讨论

本研究探索如何利用网络舆情大数据来推断大众生育态度倾向的变迁。基于对简单情绪分析、LDA 非监督分类和 fasttext 监督分类结果的比较,我们发现依托于计划行为理论框架的监督分类法能提供更准确也更有意义的分析结果。从整体的趋势来看,自 2012 年以来,涉及生育的评论态度倾向从积极居多快速转变为消极占绝对主导,这个过程也同时伴随着话题维度的转变,从更多表达行为态度到主要陈述生育面临的客观限制。这也可以帮助我们回答一开始提出的问题,中国生育水平在生

育政策调整之后并没有实现强势反弹，是生育观念的转变还是现实条件的抑制作用？从本研究的结果来看，虽然生育观念有一定程度的转变，但是在提及生育问题时，更多人的首要反应是谈及实现生育所面临的客观限制。也就是说，现实条件的抑制作用是阻碍生育率反弹的最大原因。这也说明如果下一步的政策措施得当，能够在一定程度上削弱现实的抑制条件，生育水平是有可能得到回升的。

在对所有涉及知觉行为控制的评论做词频分析之后，我们发现提及最多的限制关键词是“房子”“教育”和“工作”，说明高企的房价、子女的教育负担、生育所面临的工作和时间成本是阻碍生育率提升的重要因素。这些发现为进一步完善生育服务政策提供了重要依据。例如，抑制房价上涨，为年轻人提供可负担的住房可能是提升生育率的有效手段。其次，减轻儿童和家庭教育负担，引导树立正确的子女教育观念，有利于缓解对生育的忧虑情绪。最后，积极的家庭生育支持政策，比如托育服务的提供可能会对提高生育率产生效果。

本研究另一个重要目的是通过对生育态度倾向的分析，来探讨舆情大数据分析在人口学研究中的应用前景。具体来说，我们试图回答两个问题 “为什么需要”以及“怎么用”。关于第一个“为什么”的问题，实际上也是我们在导言中讨论过的大数据的优点。首先是信息收集的时效性和经济性很强。大部分大数据都可以做到实时收集和监控，并且不用耗费大量的人力物力。其次，大数据更加开放的数据收集框架和文本分析技术，使得我们能更深入挖掘和获取个体对抽象议题的综合认知，从而更加全面地理解特定现象的发展变化规律和背后的逻辑。基于时效性和全面性的特征，舆情大数据能够帮助建立更好的人口预测和预警辅助系统。

然而从“需要用”到“能够用”并不是一件自然发生的事情，如上文讨论的一样，由于大数据本身的缺陷，我们必须有针对地处理这些问题，才能有效利用大数据。这也就涉及“怎么用”的问题。

第一个是数据的代表性问题。对于网络舆情数据来说，无论数据量有多大，所获得的样本肯定是有偏的。这是由大数据“被动”采集的特征所决定的，即我们只能捕捉到留下数字“印记”的个体，而这些个体是高度自选择的。为了应对代表性问题，我们需要小心界定总体，同时明确很难基于这些数据去推断总体特征。对于网络大数据更合适的是去做趋势分析，追踪社会人口现象的发展变化规律，或者进行群体间的相对比较。当然，趋势分析或者群体比较成立的前提还是样本特征的稳定性。在本研究中，我们注重生育态度倾向变化趋势的探索，同时我们也需要意识到这些结果可能只对居住在城镇的年轻人群有一定的代表性。此外，我们也必须认识到，样本提供的信息不能代表个体对该议题全面的态度。比如，大部分评论虽然只涉及 1 ~ 2 个认知维度，但是并不代表个体在别的维度上没有观点和看法。他们只是“选择性”呈现出个人最首要的认知情绪。所以从某种意义上来说，本文也只是对个体最突出的生育情绪表达的分析，这意味着基于舆情数据发现的规律是需要谨慎看待的，我们要认识到其局限性。

第二个是数据的稀薄性问题。同样由于数据收集的被动性，导致在很多时候缺乏问题的框架限制，造成实际数据中有效信息含量很低。有效信息含量低不仅会影响研究的效率，更会因为掺杂太多无关信息而干扰后续分析。在本研究中，我们尝试过在更宽泛的话题框架下收集数据，而实践证明我们无法对这种有效信息含量很低的数据进行有意义的分析。因此，我们需要在研究中尽量设计多重筛选程序来限制信息的发散性。大数据的稀薄性还体现为能获得的个体变量非常有限，在本研究中，我们只能探讨性别和地域这两个特征与生育态度倾向的关系，这显然限制了分析的深入性。

第三个值得探讨的问题是在大数据分析中是否应该引入理论框架。大数据的广泛使用曾被认为会对现有社会科学的研究范式造成冲击。大数据的教父级人物维克托·迈尔·舍恩伯(2013)甚至认为大数据可以依据事物之间的相关关系建议模型，实现预测功能，从而跳过对因果推断过程的理论需求。然而在本研究中，我们展示了要构建有意义的预测指标是需要有理论框架的协助的。就本文来

说，计划行为理论为数据提供了更为合理和高效的分析框架，使我们能够获得有意义的态度倾向测量。从某种意义上讲，理论框架能够帮助研究者对大数据进行结构性梳理。相比于常规的调查数据，大数据更加开放发散但也更有可能缺乏内在的结构性。因此，在实际操作中，大数据的分析和建模通常面临比较高的风险，而好的理论框架的存在就大大降低了失败的可能。同时，理论框架的存在也能引导我们在数据分析中发掘和回应理论关心的问题。譬如在本文中，我们就试图去理解中国低生育水平背后大众生育态度变迁的逻辑，是生育观念的转变还是现实条件的限制作用更大。总结来说，一个有意义的大数据研究并不是要打破原有的社会科学研究范式，反而是在融入已有研究范式之后更能发挥其效用。由此引申来看，网络舆情大数据并不能完全替代传统的问卷调查数据，它与传统数据可以互为补充，弥补各自存在的缺陷。从某种意义上来说，除了在收集渠道和收集方式上存在差异之外，网络舆情大数据与传统数据并没有本质上的不同，因而网络舆情大数据也像传统数据一样需要融入特定的研究范式和理论框架中，才能更好地发挥价值。

最后，舆情大数据分析的应用前景是建立在大数据采集和分析技术的发展程度上的。目前，学术研究中对大数据的应用还比较粗浅，主要停留在描述、简单预测和辅助分析上。这些应用并没有发挥大数据最重要的一个优点，即可以揭示个体之间存在的网络结构（孙秀林、施润华，2016），并利用这种网络结构去理解社会人口现象背后的机制。可以预见，随着技术的进步、研究思维的拓展和更多数据的开发，网络舆情大数据以及其他形式的大数据将在社会科学研究中发挥更加重要的作用，形成对现有研究体系的有益补充。

参考文献/References:

1 杨菊华. 中国真的陷入生育危机了吗？人口研究，2015；6：44－61
Yang Juhua. 2015. Has China Really Fallen into Fertility Crisis? Population Research 6：44－61.

2 Lesthaeghe，R. 2014. The Second Demographic Transition：A Concise Overview of Its Development. Proceedings of the National Academy of Sciences 51：18112－18115.

3 刘爽，卫银霞，任慧. 从一次人口转变到二次人口转变——现代人口转变及其启示. 人口研究，2012；1：15－24
Liu Shuang，Wei Yinxia and Ren Hui. 2012. From First Demographic Transition to Second Deomographic Transition：Reflections on the Modern Demographic Transition. Population Research 1：15－24.

4 侯佳伟，黄四林，辛自强，孙铃，张红川，窦东徽. 中国人口生育意愿变迁：1980—2011. 中国社会科学，2014；4：78－97
Hou Jiawei，Huang Silin，Xin Ziqiang，Sun Ling，Zhang Hongchuan and Dou Donghui. 2014. A Change in the Desired Fertility of the Chinese Population：1980－2011. Social Sciences in China 4：78－97.

5 孙秀林，施润华. 社会学应该拥抱大数据. 新视野，2016；3：36－41
Sun Xiulin and Shi Runhua. 2016. Sociology Should Embrace Big Data. Expanding Horizons 3：36－41.

6 Inati，R.，Halford，S.，Carr，L. and Pope，C. 2014. Big Data：Methodological Challenges and Approaches for Sociological Analysis. Sociology 48：663－681.

7 孙秀林. 社会科学中的空间分析：概念、技术和应用实例. 山东社会科学，2015；8：63－70
Sun Xiulin. 2015. The Spatial Analysis in Social Science：Concepts，Techniques，and Applications. Shandong Social Sciences 8：63－70.

8 李弼程. 网络舆情分析. 国防工业出版社，2015.
Li Bicheng. 2015. Online Public Opinion Analysis. National Defense Industry Press.

9 顾宝昌. 生育意愿、生育行为和生育水平. 人口研究，2011；2：43－59
Gu Baochang. 2011. Fertility Intention，Fertility Behavior，Fertility Level. Population Research 2：43－59.

10　郑真真. 生育意愿的测量与应用. 中国人口科学，2014；6：15－25，126
Zheng Zhenzhen. The Measurement and Application of Fertility Intention. Chinese Journal of Population Science 6：15－25，126.

11　新浪微博数据中心. 2017 新浪微博用户发展报告. 新浪微报告. http：//data. weibo. com/report/reportdetail？id＝404&sudaref＝www. baidu. com，2017－12－25
Sina Weibo Data Center. 2017. 2017 Sina Weibo User Development Report. Sina Microdata Repor. http：//data. weibo. com/report/reportdetail？id＝404&sudaref＝www. baidu. com，Dec. 25th.

12　易观. 网易新闻客户端用户画像专题研究报告 2016. https：//www. analysys. cn/article/detail/1000108，2016－06－24
Analysys. 2016. Netease News Client User Profile Research Report 2016. https：//www. analysys. cn/article/detail/1000108，June. 24th.

13　人人都是产品经理. 网易新闻产品分析报告. http：//www. woshipm. com/evaluating/1358095. html，2018－09－04
WWW. WORSHIPM. COM. 2018. Netease News Product Analysis Report. http：//www. woshipm. com/evaluating/1358095. html ，Sep. 4th.

14　宋健，陈芳. 城市青年生育意愿与行为的背离及其影响因素——来自 4 个城市的调查. 中国人口科学，2010；5：103－110，112
Song Jian and Chen Fang. 2010. Reproductive Behaviors and Preferences of China's Urban Youths：Deviation and Determinants. Chinese Journal of Population Science 5：103－110，112.

15　王灿辉，张敏，马少平. 自然语言处理在信息检索中的应用综述. 中文信息学报，2007；2：35－44
Wang Canhui，Zhang Min and Ma Shaoping. 2007. A Survey of Natural Language Processing in Information Retrieval. Journal of Chinese Information Processing 2：35－44.

16　邹晓辉，孙静. LDA 主题模型. 智能计算机与应用，2014；5：105－106
Zou Xiaohui and Sun Jing. 2014. Latent Dirichlet Allocation Topic Model. Intelligent Computer and Applications 5：105－106.

17　Ajzen，I. 2002. Perceived Behavioral Control，Self-Efficacy，Locus of Control，and the Theory of Planned Behavior. Journal of Applied Social Psychology 4：665－683.

18　Ajzen，I. and Klobas，J. 2013. Fertility Intentions：An Approach Based on the Theory of Planned Behavior. Demographic Research 29：203－232.

19　Joulin，A.，Grave，E.，Bojanowski，P.，Douze，M.，Jégou，H. and Mikolov，T. 2016. Fasttext. zip：Compressing Text Classification Models. arXiv preprint arXiv：1612. 03651.

20　唐重振，何雅菲. 住房负担与生育意愿：正向激励还是资源挤出. 广西师范大学学报（哲学社会科学版），2018；4：66－72
Tang Chongzhen and He Yafei. 2018. Housing Burden and Fertility Desire：Positive Incentive or Resource Extrusion. Journal of Guangxi National University. Philosophy and Social Sciences Edition 4：66－72.

21　维克托·迈尔·舍恩伯，肯尼思·库克耶著. 盛杨燕，周涛译. 大数据时代. 浙江人民出版社，2013
Mayer-Schönberger，V. and Cukier，K. 2013. Big Data. Translated by Sheng Yangyan and Zhou Tao. Zhe Jiang People's Publishing House.

（责任编辑：陈佳鞠　收稿时间：2019－01）