DOI：10.13998/j.cnki.issn1002-1248.2019.10.17-0917

基于 Python 的农村土地流转新闻文本采集与分析

刘鑫东，邬锦雯，刘丁扬，周巧怡

（华南师范大学经济与管理学院，广州　511400）

摘　要：［目的 / 意义］为快速获取网页中土地流转资讯的海量数据，借助当前数据采集和分析的热门手段——网络爬虫技术，从新闻资讯的视角挖掘和整理国内对农村土地流转关注的新热点。［方法 / 过程］以人民网为目标网页，通过对网页的关键词检索，利用 Python 语言编写、设计程序对网页中的农地流转的新闻资讯进行爬取，将获取的 2 500 多条新闻作为数据样本，利用文本分析方法以及绘制 Python 词云图分析农地流转现状、内容及问题。［结果 / 结论］对网页新闻资讯进行爬取，通过对采集的新闻内容进行统计分析，热点挖掘后发现，当前国内对农地流转的宣传和重视程度逐渐加深，农地流转建设与精准扶贫、乡村旅游和合作社有密切的发展联系。从而对这 3 个方面现存的问题进行分析并提出建议。

关键词：网络爬虫；文本采集；文本分析；土地流转

中图分类号：G354.2　　**文献标识码：**A　　**文章编号：**1002-1248（2020）05-0076-06

引用本文：刘鑫东，邬锦雯，刘丁扬，周巧怡. 基于 Python 的农村土地流转新闻文本采集与分析[J]. 农业图书情报学报，2020，32(5)：76-81.

Collection and Analysis of Rural Land Circulation News Text Based on Python

LIU Xindong, WU Jinwen, LIU Dingyan, ZHOU QiaoYi

(School of Economics & Management, South China Normal University, Guangzhou 511400)

Abstract: [Purpose / Significance] In order to quickly obtain a massive amount of data of land transfer on the Internet, with the help of the current popular method of data collection and analysis-web crawler technology, the new hot spots of rural land transfer in China are excavated and sorted out from the perspective of news and information. [Method / Process] Taking *People's Daily* as the target web page, we search keywords on the web page and use Python to write and design programs to crawl the news documents of farmland circulation on the web page. and We then use the obtained more than 2 500 pieces of news as data to analyze the status, content and problems of agricultural land transfer by using text analysis methods and drawing Python word cloud diagrams. [Results / Conclusions]

收稿日期：2019-10-17

基金项目：广东省学位与研究生教育改革项目（项目编号：2018JGXM31）

作者简介：刘鑫东（1994-），女，硕士研究生，研究方向：信息政策，文本分析。邬锦雯（1969-），女，博士，教授，研究方向：信息政策。刘丁扬（1996-），男，硕士研究生，研究方向：信息政策。周巧怡（1993-），女，硕士研究生，研究方向：信息政策。

Crawling the web news information, through statistical analysis of the collected news content, after hot spot mining, it is found that China is attaching more importance to the transfer of agricultural land and increased publicity in this regard, The construction of agricultural land transfer and precision poverty alleviation, and rural tourism and cooperatives are closely connected with one another. At last we analyze the existing problems in above-mentioned three aspects and make recommendations.

Keywords: web crawler; text collection; text analysis; land transfer

1 引言

当今的信息技术时代，爆发式的数据经过整合和处理逐渐展现其量变到质变的价值，成为各行业把握行业发展趋势，探索时代未来必不可少的部分，农村发展相关信息亦然。中国乡村振兴战略开得如火如荼，农村土地改革也走向了新时代，农村土地流转信息是政府信息数据公开的一部分，对社会和人民都有非同寻常的意义，特别是对政府信息数据的公开和共享能够极大地促进政府部门和社会行业的发展与进步[1]。新闻资讯往往是信息传播的一种快速、高效、直观的呈现方式，也是可以集中获取行业热点的数据源。因此，笔者从新闻资讯的角度，以人民网平台为数据源，利用 Python 语言编写的面向主题爬取的程序设计，将 2007 年—2018 年农村土地流转的新闻报道作为研究数据，进行文本分析和热点挖掘，以挖掘中国关于乡村土地流转关注的热点和未来的研究方向，为农村相关行业提供发展新思路加快农地流转进程，推动农村经济社会的良好发展。

2 相关研究现状及方法

（1）网络爬虫技术研究。通常来讲，网络爬虫是一种可以自动爬取网页上大量数据信息，帮助用户进行计量和文本分析的程序，是构建搜索引擎的支撑技术之一[2]。早期的爬虫结构主要由网页下载模块、内容分析模块和 URL 地址去重以及 URL 地址分配组成[3]。随着互联网的高速发展，学者们对网络爬虫研究逐渐深入，并将新的爬虫方式运用到网页内容的爬取过程中，如周中华[4]等人提出了优于早期串行爬虫的并行架构网络爬虫，并通过抓取新浪微博的数据得到验证；随后兴起的广域网分布式 web 爬虫又融合了分布式系统等多种主题，由分布在在广域网中不同的地址位置和网络位置的代理节点共同完成并行计算的任务[5]，优势更加明显，受到众多研究者的喜爱。在网络爬虫技术中，还有一些针对特定范围爬取数据的爬虫分类——主题爬虫。通过对爬取范围以主题的方式聚焦，通过过滤不相关页面，提高了爬取的准确性，笔者采用主题爬虫的形式，根据用户输入的关键词，基于 Python 语言，获取数据。

（2）基于 Python 的网络爬虫技术研究。Python 语言诞生于 20 世纪 90 年代，语言简洁明了，扩展性强且具有强大又丰富的标准库，满足了海量数据的挖掘和分析需求。基于 Python 的网络爬虫首先就是通过编写程序，进行关键词的检索和抓取；其次，对网络爬虫抓取的内容进行数据分析，通过正则匹配，与目标 URL 地址建立连接。另外，Python 具有的强大标准库吸引 jieba 和 WordCloud 的加入，方便了对采集的文本进行分词和制作词云图。

（3）农村土地流转的相关研究。农地流转问题是“三农”经济稳定与高效持续发展的关键问题，涉及农民的核心利益。农地流转面积的增多，流转形式的多样，流转体系的健全使得农地流转正成为乡村振兴下关注的焦点，中国自改革开放以来就致力于农业信息化的建设和发展，2011 年颁布《全国农业农村信息化发展“十二五”规划》，2008 年明确提出要积极推进“农村信息化”建设，随后又颁布《全国农业农村信息化发展“十二五”规划》等一系列推进政策，不断优化农村信息化发展环境，对农地流转进行信息化建设

DOI：10.13998/j.cnki.issn1002-1248.2019.10.17-0917

也是农业信息化发展的重要组成部分[6]。目前，国内关于农村土地流转的研究主要集中在对政策演变的研究、流转动因、流转过程存在的问题、影响因素、构建评价体系对区域性农村土地流转进行评价等方面。特别在对阻碍因素研究中，宏观层次上体现在政策扶持和市场机制的不完善，农地流转过程中法律法规的漏洞以及技术的制约使得农民的利益容易受损[7]，微观层次上，有学者认为农民流转意愿不高更多的可能是获得感不足，加强农村的“三产融合”，“三位一体”（农业，文化和旅游）的空间重构[8]。学者们结合实证研究，根据地区发展的状况，探讨和总结可操作的建议措施，较少有学者从全国整体的范围来看，关注和利用社会中已发布和传播的新闻信息作为研究农村土地流转的一个方面，从大量信息中挖掘热点，探寻国家和人民对农村土地流转更为关心的问题和角度。

3 研究过程

3.1 数据来源及方法

笔者选取人民网 2007 年—2018 年的关于土地流转的新闻资讯作为数据来源，采集项目包括：新闻标题、新闻 url、新闻时间、新闻作者、新闻正文。通过设计 Python 语言，参考信息检索的关联词推荐算法设计编写程序[9]，实现数据采集。页面爬取后，数据最终将以表格的形式保存。

3.2 数据采集

在目标网页中进行页面分析后，进行相关数据采集。部分代码设计如下：

```
# 导入程序中所用到的库
import requests
from bs4 import BeautifulSoup
import pandas as pd
# 获取网页的 url 以及 headers
web = 'http://search.people.com.cn/cnpeople/search.do?'
headers ={'User-Agent': 'Mozilla/5.0 (Windows NT 10.0) AppleWebKit/537.36(KHTML, like Gecko) Chrome/69.0.3497.100 Safari/537.36´, 'Cookie' :'JSESSIONID=1BC93E460E80EE2DA256BB31A5D88483; ALLYESID4=1150358B4B9AE1ED;sso_c=0;sfr=1; _people_ip_new_code=510000;wdcid=030d23960abb8ca6;_ma_tk=1dkrjdc61o8qdeim9id461jc32406e8l; _ma_is_new_u=1; _ma_starttm=1552744723828', }
# 查找新闻的标题、时间、正文
title = j.select('h1')[0].text
time = j.select('div.clearfix.w1000_320.text_title > div > div.fl')[0].text
content = j.select(´#rwb_zw > p´)
```

4 数据分析与结果

4.1 年代分析

从人民网上抓取关于土地流转新闻的数据共计 4 082 条，通过剔除不相关和无效新闻，最终得到 2 576 条新闻。图 1 显示的是从 2007 年—2018 年之间不同年份与土地流转的新闻数量随时间的变化曲线，自 2007 年起，新闻数量整体趋势是持续上升，可见国内对土地流转的关注度也是逐渐加深的。在这一时期的变化当中，2011 年和 2017 年均有一个小高峰，这是由于在 2010 年有关文件指出要“完善农地承包法律法规和政策”，“加强土地承包经营权流转管理和服务，健全流转市场”[10]引发新闻热度上升；从 2013 年开始，中国相继颁布一些为承包土地确权做好相应准备的文件，包括 2013 年 11 月出台的全面深化改革若干

图 1 不同年份与土地流转相关的人民网新闻总数量随时间的变化

重大问题决定，拓展了农村土地流转规范化的政策空间，赋予农民对承包地更多的权力和职能等[10]；2014 年 1 月，中共中央针对农业现代化将土地承包经营权进一步明确划分为承包权和经营权[11]，同年 11 月出台的《关于引导农村土地经营权有序流转发展农业适度规模经营的意见》对土地流转新政策做出了更详细的阐述[12]。到了 2016 年，为深化农村集体产权制度改革，做出了“明确农村土地承包关系永久不变的具体规定”，2017 年，随着农地流转逐渐走入正轨后，开始提倡要通过流转土地经营权等多元方式推动农业经营模式多形式展开。由此可见，中国土地流转的年代发展变化在很大程度上是受国家政策所影响，未来，随着农村土地流转会更加正规化和制度化，相关新闻数量也会逐渐减少。

4.2 新闻热点分析

词云图是是文本数据可视化一种常用的表现方式，通常以不同的颜色和字体的大小，将文本中出现频率或权重较高的词汇以可视化的形式展现出来。目前，制作词云图的主要方式包括 Python 编写程序和在线词云生成工具。基于新闻样本的含量，笔者通过两种方式实现热点的解读：利用 Python 软件，基于网上开源的中文词典并添加了部分农业词汇作为词典。主要步骤包括：提取关键词、设置单词权重、生成词云；利用词云生成软件对新闻主题进行分析。

经过对无意义词的筛选和剔除，绘制了图 2、图 3 两种图表展示新闻热点，新闻内容和新闻主题的人点展示大致相同，因此可以发现：近几年，新闻热点多集中于农村土地流转以及农村发展建设方面，其中也不乏其它高频词汇，笔者展示了前 10 的高频词统计(图 4)，下面主要选取农村扶贫、旅游、合作社 3 个热点角度解读与农地流转之间的发展关系。

（1）土地流转与精准扶贫。

消除贫困是中国社会发展的头等大事，为按期实现脱贫目标，通过农村土地经营权的流转，引导农村土地适度集中，则有利于提高地区土地的利用率和农民收入。当前，全国多地探索了土地流转推动精准扶

图 2　人民网 2007 年—2018 年土地流转新闻内容词云图

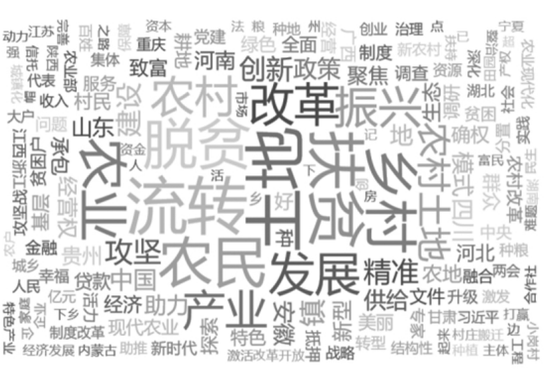

图 3　人民网 2007 年—2018 年土地流转新闻标题词云图

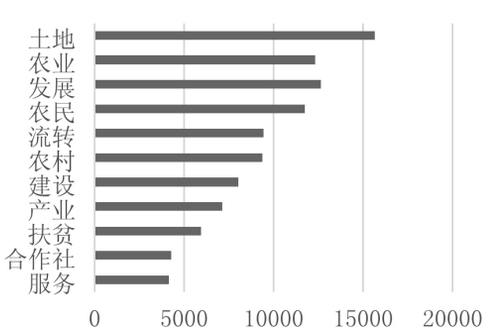

图 4　新闻内容高词频统计条形图

贫的模式，包括：种植大户转包模式、土地股份合作模式、农村土地银行模式、土地信托等[13]。通过各级政府的实践，土地流转带动“就近就业”，实施“产业扶贫”，已成为精准扶贫的一种最佳方式。

（2）土地流转和农民专业合作社。

2006 年，《农民专业合作社法》正式颁布以后，

DOI：10.13998/j.cnki.issn1002-1248.2019.10.17-0917

农民合作社开始以独立的法人资格身份在经济市场进行交易,走土地流转合作社的道路是中国实现农业现代化的必由之路[14]。因此，合作社的良好发展为现代化农业技术的发展与推广创造了条件，在农村土地流转和农业产业化中具有节约交易费用、规范程序、降低风险等优势[15]。随着农业现代化的快速发展，农民合作社也逐渐出现譬如入社流程不规范、建设功能单一、发展范围狭窄、融资难题制约合作社壮大等问题[16]，为此，国家于 2017 年再次对《农民专业合作社法》进行修订和调整，鼓励发展规模更大、组织程度更高的联合社。未来，农民专业合作社还将不断巩固和探索，其健康发展既是推动农村经济社会发展的动力，也是完成乡村振兴战略的目标途径。

（3）土地流转和乡村旅游。

近几年，乡村旅游作为乡村振兴的发展方式之一，受到了社会的关注。一方面乡村旅游是中国三农政策和供给侧改革下，农业、乡村社区和旅游业融合的产物，是一个特色鲜明的新业态旅游方式[17]；另一方面，它又对调整农村产业结构、吸引投资，增加当地收入等方面起到良好的推动作用[18]。面对乡村旅游的兴起，商业开发势必需要大量用地，土地不可避免成为乡村旅游发展的重要资源。十九大报告提出乡村振兴战略，土地流转政策的改革也为乡村旅游的发展解决了政策和法律上的限制。

5 农地流转存在的其它问题及建议

以上在讨论了农地流转与 3 个相关热点的正向解读，除此之外，农地流转中的乡村旅游和农村合作社等其它方面也存在急需解决的问题：

（1）农村合作社的发展易滞后。国家信息化高速发展的同时也带动农村信息化的发展，信息技术革命为乡村的振兴带来了重大的发展机遇。此前，农村合作社存在建设功能单一，流转管理不规范等问题逐步被完善、被解决，但依旧存在流转信息不及时，流转程序耗时的问题。当下，搭上信息化发展的浪潮，农村合作社的工作可以转移到平台端、移动端，实现农村土地流转政府服务的全流程信息化，包括解决因土地流转带来的土地确权登记变动，土地流转信息的互动，也为土地流转政务服务信息化提供了新的思路。

（2）开发乡村旅游的土地保护及利益分配无法保障。农地流转在乡村旅游发展虽然更多表现为正效应，推动农村旅游发展形式的创新，带动地区经济、生态效益，扩大交流和文化传承[19]；但是，农地流转新政也会对农村旅游发展所带来的大量资本会忽视农民利益，缺乏对乡村景观保护的认知则更易破坏农村环境[20]。农地流转补偿金较低，忽略土地未来预期增值，导致农民利益流失[21]。这需要政府在补偿金方面制定政策，实施扮演好监管职能。

（3）区域农村土地盲目流转。近年来，国家及地区鼓励农村将闲置土地根据农民意愿进行流转。事实上，中国土地资源差异很大，土地流转在城市周边的农村、土地连块地区等地具有较大价值，在偏远地区或山区的土地面积小且分散，价值有限。盲目流转，一方面致使部分承包土地的企业更改了经营模式；另一方面农民大批量涌入城市，引发城市规模、资源所能承载理想人口和人数增长失衡的矛盾，外出务工人员失业返乡时也没有了土地作为保障。流转既没有达到合理利用土地，提高效益的目的，反而损害了自身利益，阻碍了农村经济发展。因此，在各区域推进农地流转的过程中，推进农地适度规模流转需符合客观现实，可以通过根据构建农地流转适度规模的评级指标体系，分析流转规模影响大的指标，因地制宜确定该区域的适度流转规模[22]，才能更合理有效地实现农业与土地的供给侧改革，增加农民收入。

6 结语

综上所述，大数据时代，研究者们对各类数据需求很大，挖掘数据价值并有效地分析可为相关决策提供合理科学的依据。爬虫技术作为一种自动采集数据的手段，快速顺利地抓取所需数据，挖掘新的热点是利用数据的一种新兴方式，笔者利用爬取技术，采集数据并绘制热点词云图进行文本内容分析，证明了该