Hans Journal of Data Mining 数据挖掘, 2023, 13(4), 279-287
Published Online October 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.134028

Hans汉斯

基于文本挖掘的微信公众号文本内容传播特征研究

——以“北京农业”微信公众号为例

吕金宝

北京市农业农村宣传中心，北京

收稿日期：2023年8月18日；录用日期：2023年9月18日；发布日期：2023年9月26日

摘　要

目的/意义：从北京农业微信公众号推送文章出发，爬取文本内容，对微信推送文章的外部特征及传播特征进行了分析，以期能从微信公众号推文内容角度为农业知识传播提供参考与借鉴。方法/过程：利用Python语言编写爬虫代码及各种计算脚本，将获取的5103条推文作为数据样本，结合主题词抽取、文本聚类、以及高频词共现网络，分析了北京农业微信公众号推文现状及内容传播特征。结果/结论：北京农业微信公众号在农业知识传播的内容组织方面已较为全面，在深刻剖析传播特征的基础上，应着力从需求挖掘、创造热点等方面让推文更有热度。

关键词

微信公众号，文本挖掘，文本内容，传播特征

Research on the Dissemination Characteristics of the Text Content of WeChat Official Accounts Based on Text Mining

—Taking the WeChat Public Account of “Beijing Agriculture” as an Example

Jinbao Lyu

Publicity Center of Beijing Agriculture and Rural Affairs, Beijing

Received: Aug. 18th, 2023; accepted: Sep. 18th, 2023; published: Sep. 26th, 2023

文章引用: 吕金宝. 基于文本挖掘的微信公众号文本内容传播特征研究[J]. 数据挖掘, 2023, 13(4): 279-287.
DOI: 10.12677/hjdm.2023.134028

Abstract

Purpose/Meaning: Starting from the crawled Beijing Agricultural WeChat official account push articles, the external characteristics and dissemination characteristics of WeChat articles are analyzed, in order to provide reference for the dissemination of agricultural knowledge from the perspective of the content of the WeChat official account tweets. Method/Process: Using Python language to write crawler system and various calculation scripts, 5103 tweets obtained as data samples, combined with text clustering, LDA topic model and high-frequency word co-occurrence network, analyzed the public of Beijing Agricultural WeChat Status of Tweet No. and the characteristics of content dissemination. Results/Conclusions: The Beijing Agricultural WeChat official account should be more comprehensive in the content organization of agricultural knowledge dissemination. Based on the in-depth analysis of the characteristics of dissemination, efforts should be made to make tweets more popular in terms of demand mining and creation of hot spots.

Keywords

WeChat Subscription Accounts, Text Mining, Text Content, Transmission Characteristics

1. 引言

微信作为一个新兴的自媒体平台，它的发展异常迅猛，如何使其健康稳定地成长需要不断探索和实践。微信有别于其他自媒体平台的特点在于封闭性，这是它与微博等开放性的自媒体社交平台的最大差别[1]。微信的封闭性体现在获取微信公众号发布的内容较难。一是由于微信专注于移动客户端，从移动端获取数据有着严格的限制。二是由于微信公众号与用户之间存在被订阅的关系，只有订阅某公众号的用户才能接收其消息。本文利用 Python 爬取微信公众号平台的文本语料，按照时间维度对“农业知识”主题下的重要事件进行案例筛选，以突出案例的时间价值、事件追踪轨迹为标准。基于文本挖掘分析微信公众号文本内容的传播特征，并为优化微信公众号传播效果提供参考。

2. 研究对象与方法

2.1. 研究对象

研究的对象是基于真实的微信平台数据，针对本研究目前并没有开源的可供研究的数据集。因为微信公众号本身内容包括不同的部分，涵盖内容较多，在此基础上针对研究内容选取了“北京农业”这一公众号中的数据。通过对微信公众号的内容爬取，选取从 2018 年 3 月到 2022 年 12 月共 5103 条数据作为研究样本总数进行分析。

2.2. 数据获取方法

微信公众号中的文本获取主要依靠网络爬虫技术，采用 python3.6 语言开发，程序运行在 Windows10 系统上，整个网络爬虫的基本框架由爬虫调度器、URL 管理器、网页下载器、网页解析器和数据存储器五大部分组成，并基于单机多进程的方式并行采集以提高爬取效率。基于微信公众号的爬取方案需要先

使用抓包工具将微信中的公众号相应请求进行统一抓取，分部将所需要的微信公众号网页进行全部获取。在此基础上将获取的全部微信公众号网页下载到本机，最后对于所获取的本地网页进行页面内容提取，提取出所需要的文本内容。

2.3. 研究方法

2.3.1. 样本选取

本文选取“北京农业”近 5 年内的数据进行研究的总体分析样本。因本文使用文本挖掘的研究方法，所以这里不需针对样本总体进行抽样，以全样本为分析单位。

2.3.2. 数据获取

研究所采取的文本数据为“北京农业”中从 2018 年 3 月到 2022 年 12 月的发文信息，共 5103 篇文档。以 TXT 格式存储在数据库中。

2.3.3. 数据处理

1) 中文分词

中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。分词是自然语言处理的基础，分词准确度直接决定了后面的词性标注、句法分析、词向量以及文本分析的质量。本研究采用的是开源的 jieba 分词工具，具有良好的分词准确度和较快速度。可以支持包括精确分词、全模式以及搜索引擎模式在内的三种分词模式。

2) 关键词提取

关键词提取就是从文本抽取文章中意义最相关的词语，目前在文献检索、自动文摘、文本聚类和文本分类等方面有着重要的应用。关键词提取算法一般分为有监督和无监督两类。本研究采用 TF-IDF 算法提取文本中的关键词语，TF-IDF (term frequency-inverse document frequency，词频－逆向文件频率)是一种用于信息检索(information retrieval)与文本挖掘(text mining)的常用加权技术。

2.4. 统计分析

对于一些数据的统计分析，选用 python 的 pandas 库进行统计分析。pandas 是一个强大的分析结构化数据的工具集，Pandas 是提供高性能易用数据类型和分析工具。它的使用基础是 Numpy(提供高性能的矩阵运算)；用于数据挖掘和数据分析，同时也提供数据清洗功能。常使用的导入方式 import pandas as pd。在具体应用中，对微信公众号筛选后语料的每篇文章前 30 个关键词进行分词抽取，抽取之后继续计算每个季度的频率，用 pandas 库抽取季度时间，统计一段时间内的词频。

2.5. 筛选原则

2.5.1. 突出农业行业特点

“北京农业”政务微信围绕北京市农业农村行业特点，以政策解读、农业技术、产业发展、品牌建设等为重点内容进行宣传，通过优化宣传栏目、优化业务手段、优化展现形式、优化资源整合等方式，创新微直播、微视频、图文、图说等形式，不断提升搜读三农工作宣传效果。

2.5.2. 突显休闲农业特色

落实北京市休闲农业“十百千万”畅游行动计划，突出“游京郊、品京品、享京韵”主题，结合休闲线路的时令自然景观和农事景象，综合运用新媒体手段，重点打造 10 余条京郊休闲农业精品线路，并

拍摄《探路归来！舞彩浅山，给你新鲜氧气》等多部微视频，通过农小哥出镜推荐京郊三农发展最新成果，带动农民致富增收，打造京郊休闲农业新风景线和金色招牌。

3. 结果

3.1. 北京农业公众号的描述性分析

3.1.1. 内容特征分析

微信作为一种社交媒体应用，具有及时性、分享性、参与性以及互动性的特点。用户通过使用微信拓宽了自身参与和互动的空间，能够主动交流和分享信息，同时借助强大的链接功能，微信可以整合多种媒体，帮助用户获得交叉信息。每个用户也都可以是信息的制造者和传播者，以前一对一、多对一的传播方式转变为多对一、一也可以对多，传播的结构不再单一化。尤其对于突发事件或重大新闻来说，用户更倾向于将社交媒体来作为获取这类信息的首选。微信一方面通过朋友圈来维系用户的个人社交关系，另一方面通过微信公众平台来打开用户与外界的联系，综合了小社交圈和大社交圈，使得交流结构更具开放性。微信这类社交媒体的出现，其访问热度已经远远超过了主流媒体平台，更为人性化的功能使得微信的活跃度和普及率不断上升[2] [3]。

“北京农业”微信公众平台作为专门传播农业科技知识的新媒体传播平台，该平台自 2018 年 2 月 24 日发布第一篇文章，至 2022 年 12 月，共发布 5100 余篇文章，发文频率为每日 3 至 4 篇。

“北京农业”微信公众号主要分为农业资讯、农业视界和京彩农业三大内容模块，其中，农业资讯包括政策法规和最新资讯两个主题；农业视界包括微农科普、微农科技、微农课堂等三个主题；京彩农业包括品牌农业、优农佳品和休闲观光三个主题。

对微信公众号发文量按类型进行统计，除未归类的 1092 篇文章外，发文量累计前三的主题为：“最新资讯”786 篇，“微农科普”760 篇。“政策法规”664 篇，结合微信公众号后台用户各类型文章的点击率数据，可根据用户兴趣，合理调配每日发文类型组合及数量，增强公众号的用户粘合性。

3.1.2. 峰值发文比较分析

对微信公众号 2018~2022 年的每日发文量分别进行按月度、按日度统计，可以看出，“北京农业”公众号发文量最多的月份主要集中在四至七月，一月份为一年中发文量最低的月份。结合大数据对用户手机等移动设备使用时间的习惯分析，着重加强寒暑假、小长假等假期时的发文推送，以获得更高的用户关注度。从日度发文量的变化趋势，可以直观获得公众号的发文规律，掌握全年发文态势，为优化公众号运维模式提供参考。

3.2. 基于 LDA 主题模型的公众号热点主题建模分析

3.2.1. “农业资讯”主题建模分析

(一) 政策法规

“政策法规”是农民了解国家农业政策、三农政策和农业补贴政策等相关惠农政策的窗口。农民群体可以通过及时的关注了解最新农业农补等惠农政策，及时把握商机和获取更多农业补贴。“北京农业”公众号以政策解读的视角向广大农民群体，以及关注农业政策相关的群体传播最新政策法规，以通俗易懂的语言让用户群体知法懂法，切实保障自身权益。通过对平台该主题下发文的主题词抽取及频次统计，构建政策法规主题词云图(见图 1)。利用 TF-IDF 算法[1]计算主题词的重要性，得到重要性排名 TOP20 主题词“农村、乡村、野生动物、农产品、振兴”体现了近年来农业政策的关注重点。

[1]Term frequency-inverse document frequency 用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。

Figure 1. Policy and regulation word cloud map
图 1. 政策法规词云图

(二) 最新资讯

“最新资讯”传达了北京市农业的最新工作动态，以北京市农业农村局为主体，用户群体可获得北京现代农业的发展布局、高精尖农业产业发展、北京乡村振兴的战略规划等最新资讯。通过对平台该主题下发文的主题词抽取及频次统计，构建最新资讯主题词云图(见图 2)。利用 TF-IDF 算法计算主题词的重要性，得到重要性排名 TOP20 主题词“蔬菜、农产品、农机、技术、种植、质量”是北京现代农业发展的关注重点。

Figure 2. Latest news word cloud map
图 2. 最新资讯词云图

3.2.2. “农业视界”主题建模分析

(一) 微农科普

“微农科普”是对农业种植与生产的知识科普，平台发文随着作物物候期同步推出相关农事提醒与种植养护知识，具有较好的时效性，对农民群体有较高指导意义。通过对平台该主题下发文的主题词抽取及频次统计，构建微农科普主题词云图(见图 3)。利用 TF-IDF 算法计算主题词的重要性，得到重要性排名 TOP20 主题词“蔬菜、土壤、防治、植株、施肥”是平台对农业生产知识科普的热点。

Figure 3. Micro-agriculture word cloud map
图 3. 微农科普词云图

(二) 微农科技

“微农科技”是对农业种植与农产品生产等领域的技术科普，涵盖病虫害防治、水肥调控、新品种培育、农业绿色发展等诸多与生产息息相关的农业技术。通过对平台该主题下发文的主题词抽取及频次统计，构建微农科技主题词云图(见图 4)。利用 TF-IDF 算法计算主题词的重要性，得到重要性排名 TOP20 主题词“蔬菜、技术、种植、种质、秸秆”是微农科技知识传播的热点关键词。

Figure 4. Micro-agriculture science and technology word cloud map
图 4. 微农科技词云图

3.2.3. “北京农业”主题建模分析

(一) 品牌农业

“品牌农业”以北京农产品品牌推介为主，助力消费扶贫，推动京郊农业产业发展。通过对平台该主题下发文的主题词抽取及频次统计，构建品牌农业主题词云图(见图 5)。利用 TF-IDF 算法计算主题词的重要性，得到重要性排名 TOP20 主题词“采摘”是当前品牌农业的热点，“密云”、“延庆”是品牌农业较为聚集的两个地点。

Figure 5. Brand agricultural word cloud map
图 5. 品牌农业词云图

(二) 休闲农业

“休闲农业”是以农业园区观光体验、果蔬采摘为主题的特色农业，平台通过推送城市居民应季采摘休闲观光的“小广告”，并以诙谐有趣的语言，引人注目的实景照片及小视频，得到广大城市居民的关注，通过农旅融合，有效推进休闲农业的发展。通过对平台该主题下发文的主题词抽取及频次统计，构建休闲农业主题词云图(见图 6)。利用 TF-IDF 算法计算主题词的重要性，得到重要性排名 TOP20 主题词。

Figure 6. Leisure agriculture word cloud map
图 6. 休闲农业词云图

3.3. 基于知识共现图谱的公众号农业知识传播特征分析

微信公众号平台的农业知识聚类为五大类：一是，农业管理(绿色)；二是，北京特色农业(紫色)；三是，农业政策(蓝色)；四是，市场监管(红色)；五是，农业执法(墨绿色)。由关键词热度图(见图 7)可以看出不同季度关键词的热度变化，“生产”、“蔬菜”、“管理”、“技术”是春播季节的高频词，对于

指导农民春播农种具有重要意义，按生长季度推送用户所需的农业知识，更有利于增强用户黏性，获得用户更高的关注度。

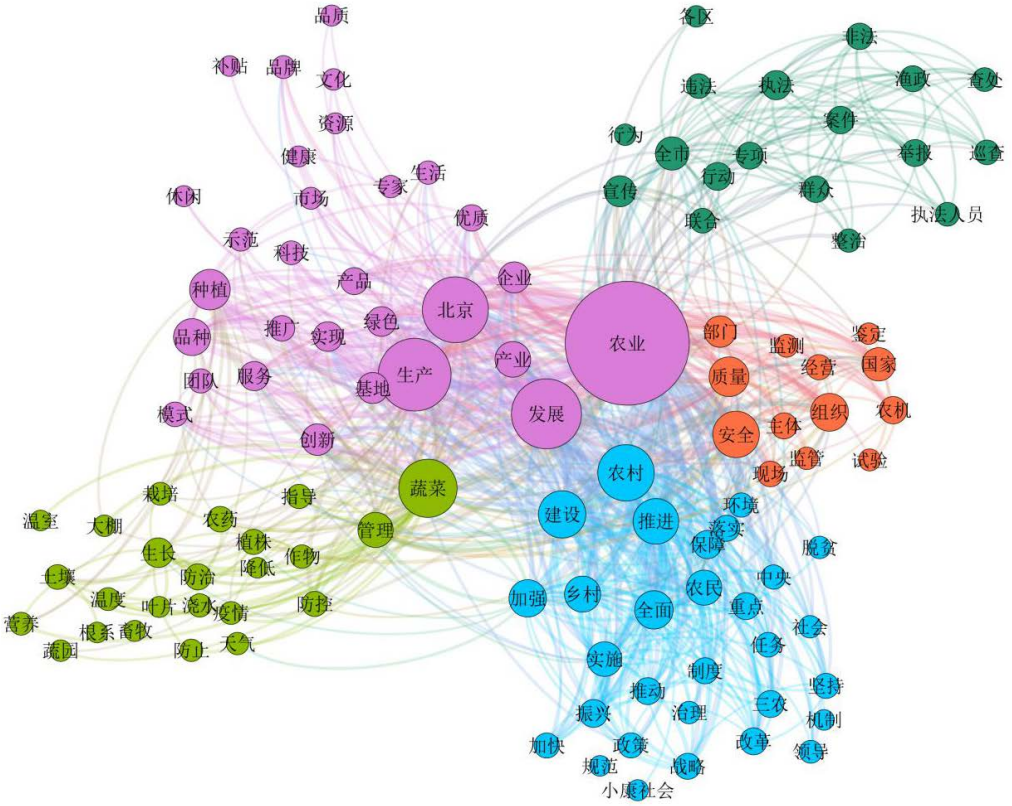

Figure 7. “Beijing Agricultural” WeChat official account agricultural knowledge co-occurrence map
图 7. “北京农业”微信公众号－农业知识共现图谱

3.3.1. 主题词季节演变特征分析

“生产”、“蔬菜”、“技术”等词在每个春季出现次数都较多，众所周知春季是播种的关键季节，也是影响整个农业发展的关键环节，所以在春季中公众号文章中“生产”有关的内容占有较大的比重，农业中对日常生活占有很大比重的“蔬菜”也在春季比较受关注，而对影响农业“生产”的“技术”也被人们所关注。在农作物生长的整个过程中，“技术”都是相对来讲比较重要的部分，因此在夏季、秋季以及冬季等也相对于其他词汇比较受关注。对于农作物“生长”在一年四季都比较受关注。纵向对比各个词汇在各个季节中的出现热度，温度在秋冬季节被提及较多，可能是在较为寒冷时更需要注意农作物的温度部分。

3.3.2. 热点主题词热度趋势特征分析

对热点主题词的季节性热度趋势进行分析，从中选取了农业知识聚类中的词频最高的三个词：“蔬菜”、“技术”、“农村”。综合各个热词与时间的热度趋势，可以总结发现，在春季是容易引起对前一年农业情况总结的一个季度，也是对今后的一年做出展望的一个季度。所以相比较而言，各个词在春季出现的热度都较大，在冬天的热度较低。在同一年之中对于相关词的报道在春季最大，随着夏季、秋季的季度变化逐渐降低，到冬天为最低。总体而言，在春季是四季中对农业生产最关心的季节，媒体也会在春天发布较多关于农业的报道以及相关类型的文章。

4. 讨论

新媒体的出现对传统媒体形成了巨大挑战，同时也给予传统媒体更广阔的融合和发展平台。通过与新媒体融合，传统媒体科普能够利用自身科普资源丰富、科普人才素质高的优势，并借助新媒体在科普工作中的传播优势，扩大科普知识的传播和共享范围，建立立体式的科普传播构架[4]。

北京农业微信公众号在近五年的运行中已逐渐成长为农业知识传播领域中较为成熟的专业平台，平台关注用户量持续上涨，并形成了良好的转发、评论、点赞等互动圈。通过微信公众号，为传统媒体提供新的宣传路径，扩大宣传广度，同时也为自身建立一个实体和网络结合覆盖的立体科普传播结构[5]。北京农业微信公众号在农业知识传播的内容组织方面已较为全面，在深刻剖析传播特征的基础上，应着力从需求挖掘、创造热点等方面让推文更有热度。微信公众号运维者要做到兼顾微信推文内容全面以保证阅读量，同时尽力做到推文分享的热度，以保证点赞量。无论微信推文传递什么样的内容，其本质不能脱离的都是面对人这样一种服务对象。每个个体都有其独特的需求，微信推文针对服务对象的实际需求去创作，往往能取得意想不到的传播效果。

参考文献

[1] 潘伟. 基于文本挖掘技术的微信公众号关系网络研究[D]: [硕士学位论文]. 南京: 东南大学, 2018.

[2] 周易军. 论微信公众平台的统计功能对企业微信营销的意义[J]. 视听, 2015(1): 117-119.

[3] 何镕飞. 基于文本挖掘的微信公众号文本内容传播效果研究[D]: [硕士学位论文]. 太原: 山西大学, 2017.

[4] 陈乐遥, 洪磊, 陈杨, 等. 基于文本挖掘的公安院校公众号主题类型挖掘研究[J]. 计算机时代, 2020(8): 6-9.

[5] 王磊, 吕鹏辉. 基于微信推文内容视域下图书馆微信服务现状研究及建议[J]. 情报杂志, 2017, 36(9): 202-207+191.