

基于机器学习算法和社会网络分析方法的议程设置验证

——以新冠疫情初期的“中国援助”为例

万鑫[#], 唐润发[#], 熊令纯, 李裕梅

北京工商大学数学与统计学院, 北京

收稿日期: 2022年9月13日; 录用日期: 2022年10月13日; 发布日期: 2022年10月24日

摘要

本文通过使用机器学习算法和社会网络分析方法, 以新冠疫情初期的“中国援助”为例验证了议程设置的三个层次。首先, 通过2020年3月1日~2020年4月30日的媒体报道量和百度指数, 初步分析媒体报道量与公众关注程度的简单相关关系, 得到两个结论: 一是公众关于“中国援助”的关注度与媒体的报道量高度相关; 二是当代媒介环境下, 媒体影响公众的最佳滞期由7天缩短到了当天。其次, 采用支持向量机分类方法, 将每篇媒体新闻报道与群众评论所蕴含的思想态度分成积极正向和消极负向两方面, 进而分析媒体新闻报道对群众评论的主观思想的影响, 通过分析认为媒体对疫情下“中国援助”偏向积极地报道带动了公众偏向积极地看待中国援助国外。最后, 通过TextRank算法和社会网络分析方法, 构建媒体属性矩阵和公众属性矩阵, 并分析媒体报道和公众评论的相关性, 发现媒体的报道推动了公众对“中国援助”事件的认知网络的构建。

关键词

议程设置验证, 新冠疫情的“中国援助”, 媒体报道, 公众评论, 机器学习算法, 社会网络分析方法

Verification on Agenda Setting Based on Machine Learning Algorithms and Social Network Analysis Methods

—A Case Study of “Chinese Assistance” in the Early Stage of COVID-19

Xin Wan[#], Runfa Tang[#], Lingchun Xiong, Yumei Li

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing

Received: Sep. 13th, 2022; accepted: Oct. 13th, 2022; published: Oct. 24th, 2022

[#]万鑫、唐润发同为第一作者。

Abstract

Machine learning algorithms and social network Analysis methods are used to verify the three levels of agenda setting in the early stage of “Chinese assistance” in COVID-19. Firstly, through the media coverage and Baidu Index from March 1, 2020 to April 30, 2020, this paper preliminarily analyzes the simple correlation between media coverage and public attention, and two conclusions are obtained: the public’s attention to “Chinese assistance” is highly related to the amount of media coverage, thus the first level of agenda setting is verified; in the contemporary media environment, the best time lag period for media to affect the public has been shortened from 7 days to the same day. Secondly, by using SVM model, the thoughts and attitudes contained in each media news report and public comments are divided into positive and negative aspects, and then the impact of media coverage on the subjective thoughts of public comments is analyzed. Through the analysis, it is considered that the positive media coverage of “Chinese assistance” has led the public to view Chinese assistance to foreign countries positively, thus the second level of agenda setting is verified. Finally, through TextRank algorithm and social network analysis methods, the media attribute matrix and public attribute matrix are constructed, and the correlation between media coverage and public comments is analyzed. It is concluded that media coverage affects the public’s cognitive network construction about the “Chinese assistance” event, thus the third level of agenda setting is verified.

Keywords

Agenda Setting Verification, “Chinese Assistance” in COVID-19, Media Coverage, Public Comments, Machine Learning Algorithms, Social Network Analysis Methods

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



1. 引言

1968年麦克斯维尔·麦库姆斯与唐纳德·肖一起提出“议程设置”，该理论在众多学者的实证研究中不断发展，内涵得到了扩充，已成为新闻传播领域中的经典理论。20世纪初期，伴随着媒介的发展，“大众传播威力论”、“子弹论”盛行，直到20世纪40年代，霍夫兰、拉扎斯菲尔德等人通过研究提出“有限效果论”，表示大众媒介的传播效果是有限的。在这一背景下，议程设置理论提出，通过媒介议程和媒体议程的比较，发现媒体的传播效果显著[1][2]。现如今，网络媒体盛行、媒介融合不断深化，传统媒体的议程设置效果亟待考察。议程设置理论发展至今，根据显著性考察的不同，可概括三个层次：客体议程设置、属性议程设置和网络议程设置。众多学者通过实证研究检验了“议程设置”理论的存在[3][4]，但是相关研究的方法较为单一、研究层次不够清晰，数据处理方面缺乏科学性。为了更加严谨客观的论证，本文主要采用机器学习法和社会网络分析法来弥补人的主观影响，并以“中国援助”相关媒体报道和公众讨论的文本为研究对象，来验证“议程设置”的三个层次：1) 公众对某议题的关注程度与他们所接触的媒体对这些议题的关注度是否高度相关？2) 大众媒体是否也影响到公众对议题属性的判断？3) 媒体能否决定公众如何将不同的信息碎片联系起来，构建出对议题的认知网络？于是我们采用相关性分析以及实质性分析来验证第一层次，用机器学习方法来验证第二层次[5][6][7][8]，用社会网络分

析方法来验证第三层次, 综合以上方法我们更客观的探究了媒介传播效果, 验证了“议程设置”的三个层次[9] [10]。

2. 基本知识

文章中涉及到一些机器学习的算法比较复杂, 在本章将给出这些算法的思想, 以及算法的基本步骤。

2.1. Word2vec 模型

Word2vec 模型是托马斯·米科洛夫(Tomas Mikolov)在 2013 年提出的[11], 作者的目标是从海量的文档中训练得到低维稠密词向量, 且词与词之间的语义关联程度可以通过向量的余弦表示。通过 Word2vec 得到的词向量通常作为其他神经网络的输入应用在一些自然语言处理的任务中。本文将用 Word2vec 模型对媒体报道和公众评论的文字进行数值向量化。

2.2. 支持向量机

支持向量机(Support Vector Machine)是 Cortes 和 Vapnik 于 1995 年提出的[12], 它在解决小样本、非线性及高维模式识别中表现出许多特有的优势, 并能够推广应用到函数拟合等其他机器学习问题中, 使得 SVM 适用分类范围越来越广[13] [14]。SVM, 从线性可分模型到非线性可分的软边界模型, 再从低维线性不可分映射到高维线性可分的核函数思想等过程组成。本文将用 SVM 对媒体报道和公众评论进行分类分析。

2.3. TextRank 算法

TextRank 算法是一种用于文本的基于图的排序算法[15]。其基本思想来源于谷歌的 PageRank 算法, PageRank 设计之初是用于 Google 的网页排名的, 以该公司创办人拉里·佩奇(Larry Page)之姓来命名, 它可以用来体现网页的相关性和重要性。在利用 TextRank 算法选出文章的关键词时, 最终会得到一系列关键词及其对应的权重。本文将用 TextRank 算法选出媒体报道和公众评论的关键词并得到相应的权重等。

2.4. 网络可视化及相似度检验

2.4.1. NetDraw 可视化

为了更直观体现在“中国援助”这个议题每个属性之间的联系, 本文使用 NetDraw 工具可视化媒体和公众的网状图。该工具由美国肯塔基州立大学 Gatton 商学与经济管理学院管理系 Steve Borgatti 教授开发, 目前已被广泛应用于社会网络分析研究[16] [17]。

本文, 在同一篇报道或同一公众评论中任意两个词语出现, 我们称这两个属性共现, 并认定为这两个属性之间存在关联。NetDraw 网状图是由节点和连线组成, 用节点大小表示属性的重要性, 节点越大表示该词出现的越多, 在网络中越重要, 用连接线表示两个属性的共线, 用线的粗细来直观表现属性之间的关联程度。

2.4.2. QAP 相关性分析

社会网络分析中, 另有一种方法用来研究两个方阵的相关性, 这种方法是 QAP 二次指派程序[18]。QAP 的计算原理是以矩阵置换为基础, 其针对两个 n 维的对称矩阵 A 和 B 的 QAP 计算过程见文献[18], 我们将用该方法分析媒体属性矩阵和公众属性矩阵之间的相关性, 从而分析媒体报道对于公众的影响。

3. 议程设置验证

本节将通过简单相关分析、SVM 模型、社会网络分析方法验证议程设置的三个层次, 并给出实验中数据处理的具体步骤。

3.1. 实验准备

3.1.1. Word2vec 模型训练

本文基于 Word2vec 模型将文本向量转换为词向量，在转换之前需要先训练 Word2vec 模型。该模型的训练数据，是在搜狗实验室下载的 1.5G 的全网新闻报道，训练由 Spyder 平台完成，Python 版本为 3.7。

3.1.2. 第一层次验证的数据收集

在慧科数据库，以“中国援助”为关键词，搜索条件设置为包含“中国援助国外”、“中国物资”、“中国专家”和“中国口罩”这些词中任意一个，时间段设置为 2020 年 3 月 1 日至 2020 年 4 月 30 日，数据库显示这段时间内关于“中国援助”的新闻报道共 31,293 篇。其中每一天的新闻报道篇数依次是：9、22、18、34、66、44、98、25、62、91、97、121、345、680、1105、1251、463、648、252、370、255、1331、1597、785、964、1570、1283、1065、1134、857、600、564、589、490、215、308、1532、1582、600、743、525、525、136、331、254、125、404、170、222、777、797、287、482、196、154、67、84、266、416、576、634。

百度作为全球最大的中文搜索引擎，以其快捷方便特点满足了中国国内民众的搜索习惯。百度指数 (Baidu Index) 是以百度海量网民的行为数据为基础的数据分析平台，某个关键词在百度平台的搜索规模大小，都会体现在百度指数的涨跌态势上，所以百度指数一段时间内的涨跌态势可以反映国内民众对这个舆论的重视程度的变化。为了通过大数据体现民众对中国援助的话题的关注度，本次研究选用了百度指数作为平台。

在百度指数，以“中国援助”为关键词，统计了 2020 年 3 月 1 日到 2020 年 4 月 30 日，有关“中国援助”的百度指数走势，61 天每天的指数数据依次为：173、218、198、174、207、176、225、296、319、354、426、591、1316、1220、1285、1679、2004、1450、1601、1471、1247、1523、1426、1212、1287、1589、1450、1268、1199、1067、1015、834、687、708、1101、614、510、441、426、385、395、399、338、326、324、370、364、339、332、336、285、238、234、294、255、315、295、261、229。

3.1.3. 第二三层次验证的数据搜集和处理

在第二层次和第三层次的证明中，若使用 2020 年 3 月 1 日至 2020 年 4 月 30 日这段时间内关于“中国援助”的所有数据，则不仅增加研究样本的噪声，而且在训练网络时浪费大量资源。故本次研究从中分别抽取 300 篇媒体报道和公众评论做分析。为了让数据具有代表性，需要保证每天都能抽到数据，下面设每天要抽取 X 篇，则 X 计算公式如下：

$$\frac{\text{当天报道天数}}{\text{总篇数}} = \frac{X}{300}$$

当 $X < 1$ 时， X 取 1；当 $X \geq 1$ 时，四舍五入取整数。这段时间内媒体报道的总篇数为 31,293 篇，比如第 1 天报道篇数为 9 篇，则第 1 天抽取的篇数 X 为 $(9/31,293) \times 300 \approx 0.086 < 1$ ，那么第 1 天就在 9 篇里随机抽取 1 篇；再如第 16 天报道篇数为 1251 篇，则第 16 天抽取的篇数 X 为 $(1251/31,293) \times 300 \approx 11.99 > 1$ ，那么第 16 天在就应当在 1251 篇里随机抽取为 12 篇。随机抽选是用计算机先在区间(0, 当天报道篇数)内生成 X 个随机数，再根据这 X 个随机数在当天的总篇数里挑选出对应的篇数。

抽取后的数据在输入 SVM 模型之前，还需要进行一系列处理，下面以媒体数据为例。

①标签设置：先把 300 篇新闻报道分成正面报道和负面报道，以数字“1”为正向标签，数字“0”为负向标签，300 篇媒体报道中有 238 篇正面报道，62 篇负面报道(公众有 194 篇正面评论，106 篇负面评论)。

②文本清洗：把 300 篇文章整合到一个文件，去除如空格、数字和字母等各种无用的符号。

③结巴分词：jieba 库是一款优秀的 Python 第三方中文分词库，多应用于自然语言处理领域，它可以将中文句子分解成词语。如“据央视新闻报道，国内新冠肺炎现有确诊病例降至 10,000 以下了！”这句话分词后的结果为“据央视新闻报道国内新冠肺炎现有确诊病例降至以下了”。

④去停用词处理：这项工作是为了降低数据噪声，避免把没有实际意义的词输入模型训练。本文选定中科院计算所中文自然语言处理开放平台发布的中文停用词表，其中包含了 1208 个停用词，结合本次实验的数据特征，在原有词典的基础上，添加了如“来源、人民日报、新华社、慧科讯业有限公司、文章”等 20 多个适用于新闻数据的特殊停用词，增强了去停用词的效果。

⑤文本向量化：中文文本是不能直接输入到模型的，需要用到 Word2vec 工具把中文文本转化为词向量形式。公众评论数据的处理与上面媒体数据的处理步骤类似。

3.2. 第一层次验证

3.2.1. 相关性分析

议程设置第一层次考察的是客体显著性的转移，可以理解为公众对话题的讨论趋势与媒体的趋势具有一致性，在数据上的表现就是在一段时间内，公众对某个议题的讨论的量应该与媒体对该话题报道的量成正比。

通过计算两组数据的相关性，发现相关系数达到 0.561，呈强相关。在 2020 年 3 月 1 日至 2020 年 4 月 30 日期间，民众对“中国援助”话题的关注度如下图 1 所示。

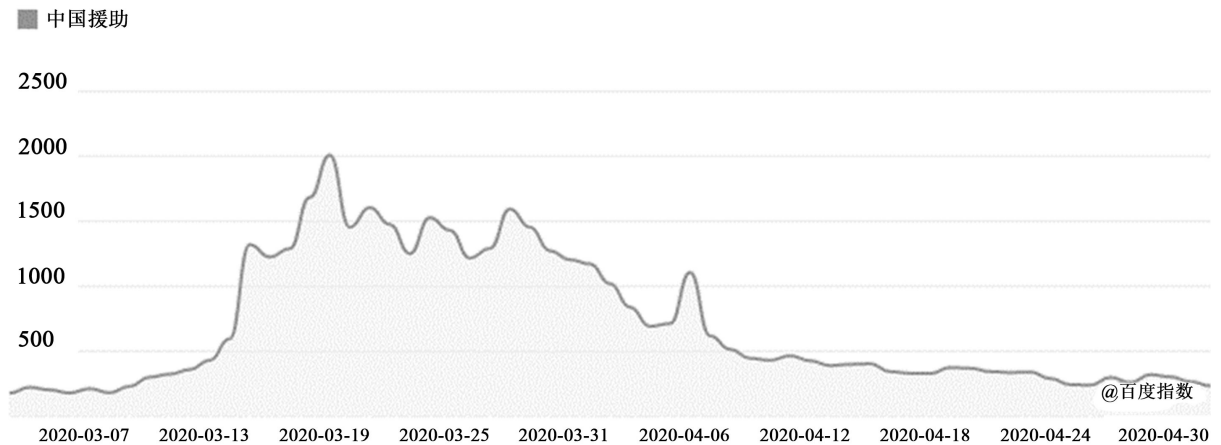


Figure 1. Public attention trend to “China assistance”

图 1. 民众对“中国援助”的关注趋势

由图 1 可知，公众对中国援助的讨论在 3 月 1 日到 3 月 9 日处于缓慢上升阶段(试探期)，3 月 10 号到 3 月 13 号呈现快速拉升趋势(升级期)，3 月 14 日到 3 月 30 日讨论热度居高不下(激化期)，4 月 1 日到 4 月 30 日热度下降回到平缓趋势(消退期)，其中公众讨论的高峰期出现在 3 月 14 日到 30 日期间，与媒体报道量对比，发现二者的趋势具有一致性。

媒体通过对“中国援助”这一议题的大量报道，引发了公众对该议题的大量讨论，第一层次议程设置效果显著。

3.2.2. 时滞期分析

在议程设置的研究中，时滞期一直是一个研究要点，时滞期是指议程的一方影响另一方需要的时间

[19]。Roberts 等人研究表明：传统媒体影响公众讨论的时滞期应该相对较短，传统媒体的新闻报道影响 EBB 上的网络讨论所需的时间为 1 到 7 天，并且当时间滞后为 7 天时，媒介议题与公众讨论议题的相关性最强[20]。利用上面的两组数据，本文对当代媒体的时滞期进行了有趣的研究。发现当把两组数据错开几天时，也就是把当天的新闻报道量，与过几天的百度指数相对应，相关系数变化结果如下表：

Table 1. Correlation coefficients between media coverage and Baidu Index

表 1. 媒体报道量与百度指数的相关系数

错开天数	0	1	3	7
相关系数	0.561	0.539	0.416	0.002

由上表 1 可以看出，随着两组数据错开天数的增加，相关系数直线下降，到错开 7 天时，两组数据变得完全不相关，当错开天数为 0 天时，即当天的媒体报道量对应当天的公众讨论，两组数据的相关性最强。

互联网使得万物互联、万物互通。在改变人们生活方式的同时，也极大的改变了传播的方式，给传统媒体的时效性带来了极大的挑战。通过相关性的比较发现，在当代媒介环境下，议程设置的最佳时滞期在变短，已经从 7 天缩短到了当天，媒体对公众的影响愈加及时。

3.3. 第二层次验证

本节将通过 SVM 模型验证议程设置的第二层次。首先在 2020 年 3 月 1 日至 2020 年 4 月 30 日这段时间按比例选出 300 篇新闻报道和 300 篇公众评论，作为测试数据，再将数据经过一系列的预处理、降维，最后将数据输入到 SVM 模型进行分类，并根据分类结果分析媒体和公众的情感倾向。

3.3.1. SVM 模型训练

SVM 是一个有监督学习的分类器，相当于一个人类的大脑，想要达到自主分类的功能，必须先训练“大脑”。训练数据是人工标记的，来自于慧科数据库的媒体报道数据，其中包含了正面和负面的评论各 300 条，输入 SVM 模型进行训练，并把训练好的模型保存好，在接下来的分类中使用。

3.3.2. 基于 PCA 降维的数据优化

新闻数据有两个特性，一是新闻报道篇幅过长且包含很多无用的信息，二是一篇报道里面会出现既褒又贬的观点，如赞成对一些国家援助的同时又反对对部分国家提供援助。考虑到这些因素，本文对数据使用了 PCA (Principal Component Analysis) 降维技术，PCA 可以将词向量矩阵里线性相关的项去掉，保留线性无关的特征，从而降低了数据维度，达到去噪的效果[21]。将数据进行 PCA 降维结果如图 2、图 3 所示。

由上图 2 和图 3 可以看出，媒体报道数据中，前 50 维数据对整个数据几乎作了全部贡献，即可以使用前 50 维数据来替代所有数据去进行分类，而公众评论数据中，前 20 维数据几乎作了全部贡献。因此，在 SVM 分类时，媒体数据选取前 50 维，公众数据选取 20 维。

3.3.3. SVM 分类

在机器学习领域，评价模型分类结果的好坏，常用到 ROC (receiver operating characteristic curve) 曲线。ROC 曲线位于平面坐标系中，其中横轴 X 轴为假阳性率，也称为误报率，X 轴越接近 0 准确率越高；其中的纵轴 Y 轴称为真阳性率，也称为敏感度，Y 轴则越接近 1 代表准确率越好(Hanley 和 Mcneil, 1982)。

如图 4 和图 5 所示，虚线把整个图划成了两部分，曲线下方的面积称 AUC (Area Under Curve)，用来

比较预测的准确性, AUC 值越高, 则曲线下方面积越大, 就说明预测准确率越高, 图 4 的 AUC 值为 0.903、图 5 的 AUC 值为 0.888。

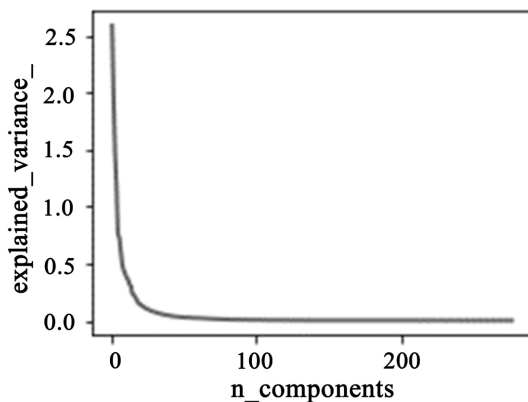


Figure 2. PCA dimensionality reduction results of media coverage data

图 2. 媒体报道数据 PCA 降维结果

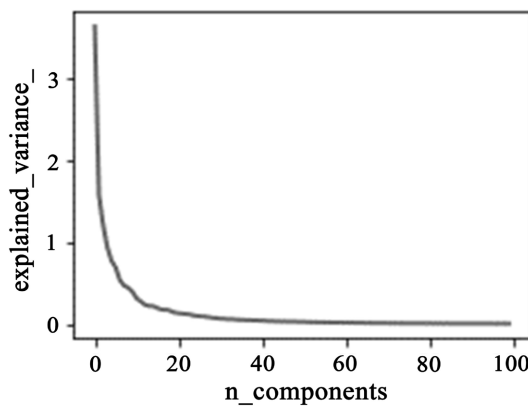


Figure 3. PCA dimensionality reduction results of public comment data

图 3. 公众评论数据 PCA 降维结果

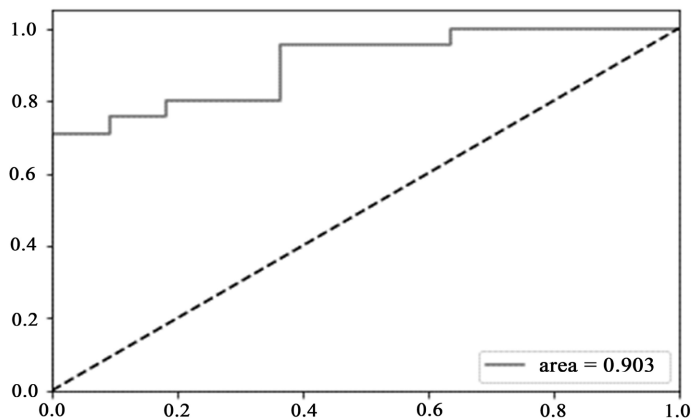


Figure 4. ROC curve of SVM classification in media coverage

图 4. 媒体报道的 SVM 分类的 ROC 曲线

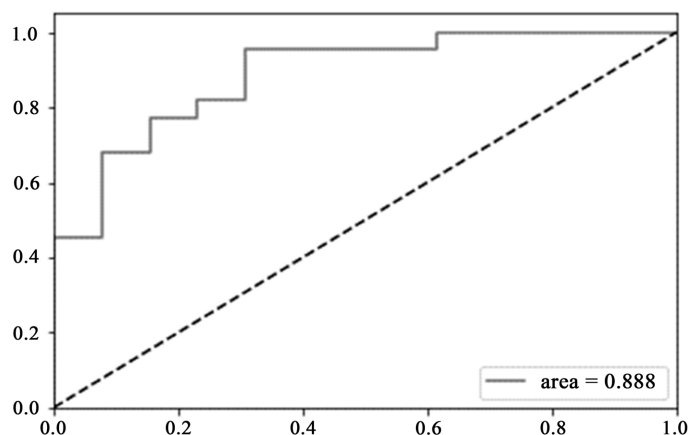


Figure 5. ROC curve of SVM classification of public comments

图 5. 公众评论的 SVM 分类的 ROC 曲线

3.3.4. 分类结果分析

经过 PCA 数据降维后，得到了媒体报道和公众评论的分类正确率如下表 2。

随机抽取的 300 篇新闻报道中，有 238 篇正面报道，62 篇负面报道，在 SVM 模型正确率为 87.5% 情况下，SVM 结果认为媒体对新冠肺炎下关于“中国援助”的报道是偏向积极的。随机抽取的 300 篇公众评论中，有 194 篇正面评论，106 篇负面评论，在正确率为 82.9% 的情况下，SVM 结果认为公众对新冠肺炎下关于“中国援助”的态度也是积极的。因此，在“中国援助”这个议题上，媒体的态度与公众的态度是一致的，媒体不仅影响了公众讨论什么，还影响了公众看待“中国援助”的态度，属性发生了转移，议程设置第二层次得到了验证。

Table 2. Classification and prediction results of media coverage and public comments

表 2. 媒体报道与公众评论分类预测结果

分类项目	媒体报道	公众评论
正确率(%)	87.5	82.9

3.4. 第三层次验证

本节是通过 TextRank 算法和社会网络分析方法验证议程设置的第三层次。首先会用 TextRank 算法选取出媒体报道和公众评论的 15 个属性词，再构建属性矩阵，进而通过 NetDraw 进行可视化，最后通过 QAP 分析两个属性矩阵的相关性，根据相关系数判定媒体网络对公众网络的影响。

3.4.1. 基于 TextRank 算法选取属性词

第三层次的关键是在媒体报道和公众评论这 600 篇文章中概括出属性词。之前的研究在筛选属性词过程中大多加入了主观情绪[22]，本文为了使选出来的属性词更客观也更具有代表性，使用了自然语言处理领域常用的提取关键词工具 TextRank 算法，它能够从一个给定的文本中提取出该文本的关键词、关键词组[15]，TextRank 算法会计算每个词的重要性。表 3 和表 4 是基于 TextRank 算法抽取的重要性靠前的关键词表格。

依据 TextRank 算法选取的重要程度靠前的关键词，去除一些无意义词和同义词(国家、没有、世界……)，最终选定了贡献度最高 15 个属性词：中国、意大利、美国、口罩、援助、物资、病毒、武汉、欧盟、俄罗斯、特朗普、人道主义、医疗队、巴基斯坦、武契奇。

Table 3. Importance of key words in media**表 3.** 媒体关键用词重要性程度

属性词	中国	意大利	疫情	美国	国家	口罩	援助	……
重要度	0.0161	0.0054	0.0053	0.0053	0.0046	0.0035	0.0034	……

Table 4. Importance of key words in public**表 4.** 公众关键用词重要性程度

属性词	中国	意大利	疫情	美国	国家	口罩	援助	……
重要度	0.0286	0.0107	0.0106	0.0095	0.0091	0.0054	0.0049	……

3.4.2. 构建属性矩阵

新闻报道或公众评论中相关属性的共现次数被计算出来用以测量任意两个属性间的关联强度。两个属性词在同一篇报道或同一个公众评论中共同出现的频率越高，这两个属性之间的关联就越强[11]。本文分别统计出媒体报道和公众评论中属性词的共现次数，共统计 600 次，获得媒体和公众的共现矩阵。

从表 5 和表 6 可以看出，属性矩阵是对称矩阵，对角线上是无意义数据，其他表格中的数据表示两个属性词在同一篇文章中同时出现的次数，如在媒体矩阵中，中国和意大利对应的表格显示为 2734，这表示在媒体矩阵的 300 篇文章中，中国和意大利同时出现在某一篇文章的次数和 2734。

Table 5. Media attribute matrix**表 5.** 媒体属性矩阵

公众矩阵	中国	意大利	美国	口罩	援助	物资	病毒	武汉	俄罗斯	欧盟	特朗普	医疗队	人道主义	巴基斯坦	武契奇
中国	—	2734	1272	1692	2436	2346	1063	299	179	361	162	324	109	353	354
意大利	2734	—	557	906	1340	1301	615	217	112	299	106	230	81	272	156
美国	1272	557	—	601	735	660	570	183	124	191	167	90	66	97	75
口罩	1692	906	601	—	1151	1187	573	213	102	203	108	170	85	238	178
援助	2436	1340	735	1151	—	1804	786	257	125	292	130	273	109	309	294
物资	2346	1301	660	1187	1804	—	694	237	122	274	113	243	104	312	275
病毒	1063	615	570	573	786	694	—	206	109	204	146	151	71	165	135
武汉	299	217	183	213	257	237	206	—	51	81	47	76	42	62	57
俄罗斯	179	112	124	102	125	122	109	51	—	41	31	27	17	38	30
欧盟	361	299	191	203	292	274	204	81	41	—	60	78	41	58	75
特朗普	162	106	167	108	130	113	146	47	31	60	—	37	33	32	24
医疗队	324	230	90	170	273	243	151	76	27	78	37	—	39	91	74
人道主义	109	81	66	85	109	104	71	42	17	41	33	39	—	47	21
巴基斯坦	353	272	97	238	309	312	165	62	38	58	32	91	47	—	63
武契奇	354	156	75	178	294	275	135	57	30	75	24	74	21	63	—

3.4.3. 网络可视化

将属性矩阵通过 excel 表导入 NetDraw，得到媒体属性矩阵如下图 6 所示。

Table 6. Public attribute matrix
表 6. 公众属性矩阵

公众矩阵	中国	意大利	美国	口罩	援助	物资	病毒	武汉	俄罗斯	欧盟	特朗普	医疗队	人道主义	巴基斯坦	武契奇
中国	—	147	276	108	485	236	97	28	33	40	17	36	26	23	21
意大利	147	—	43	23	79	51	23	6	14	12	6	18	9	4	1
美国	276	43	—	32	153	47	38	12	7	10	19	14	7	1	1
口罩	108	23	32	—	86	75	24	2	5	13	3	8	7	9	2
援助	485	79	153	86	—	172	76	24	17	32	14	28	23	12	19
物资	236	51	47	75	172	—	42	10	15	19	8	20	14	4	6
病毒	97	23	38	24	76	42	—	11	4	9	4	8	7	0	3
武汉	28	6	12	2	24	10	11	—	2	0	3	3	1	0	0
俄罗斯	33	14	7	5	17	15	4	2	—	5	1	2	1	2	0
欧盟	40	12	10	13	32	19	9	0	5	—	1	4	2	1	9
特朗普	17	6	19	3	14	8	4	3	1	1	—	2	0	0	0
医疗队	36	18	14	8	28	20	8	3	2	4	2	—	6	3	2
人道主义	26	9	7	7	23	14	7	1	1	2	0	6	—	1	0
巴基斯坦	23	4	1	9	12	4	0	0	2	1	0	3	1	—	0
武契奇	21	1	1	2	19	6	3	0	0	9	0	2	0	0	—

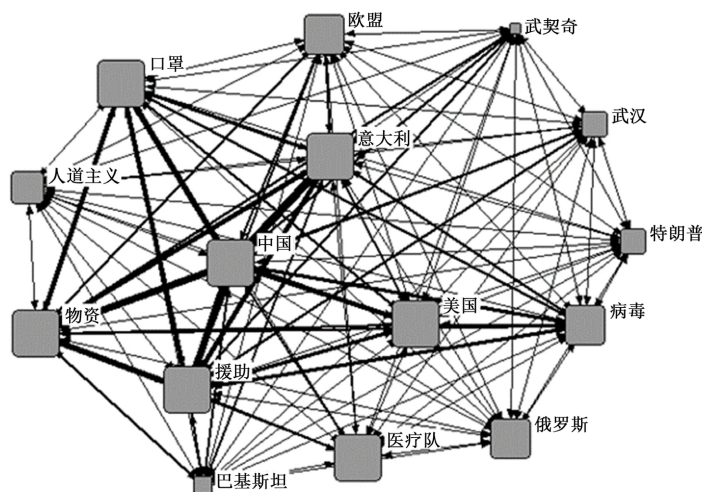


Figure 6. Media attribute mesh plot
图 6. 媒体属性网状图

本文通过网络可视化来分析新冠疫情下中国援助的媒体议程和公众议程各属性之间的联系。通过 NetDraw 对媒体矩阵进行可视化，得到了媒体议程网络。通过 UCINET 软件进行点度中心性分析发现，媒体议程网络各属性词根据数值排序依次为：中国(点度中心度 13,684)、援助(点度中心度 10,041)、物资(点度中心度 9672)、意大利(点度中心度 8926)、口罩(点度中心度 7407)、病毒(点度中心度 5488)、美国(点度中心度 5388)、病毒(点度中心度 5488)、欧盟(点度中心度 2258)、巴基斯坦(点度中心度 2137)武汉(点度中心度 2028)、医疗队(点度中心度 1938)、武契奇(点度中心度 1811)、特朗普(点度中心度 1196)、俄罗斯

(点度中心度 1108)、人道主义(点度中心度 865), 其中度是与一个节点相关联的边的条数, 有向图中每个节点都有一个出度和入度, 分别是指入边和出边的条数, 度被指为结点的“关联性”。

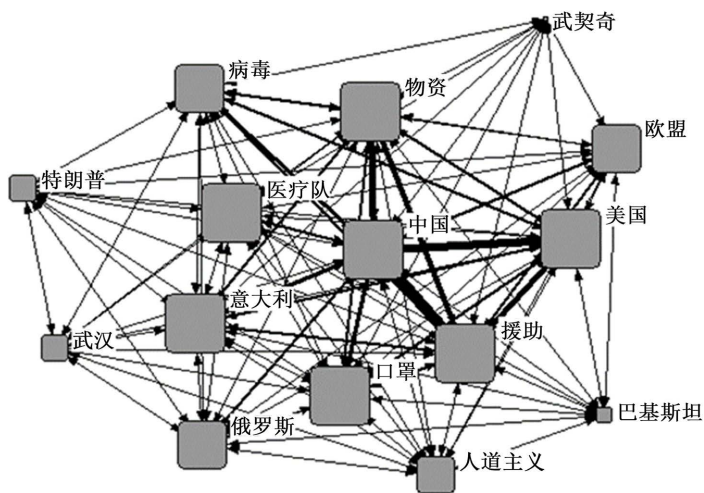


Figure 7. Public attribute mesh plot
图 7. 公众属性网状图

在议程网络中, 不同的属性词由不同大小的蓝色方块组成, 方块的大小代表了每个属性的点度中心度, 点度中心度表示与该点直接相连的点的个数。节点的中心度越大, 就意味着这个节点的中心度越高, 该节点在网络中就越重要。因此, 在媒体议程中中国、援助、物资等是强调的重点内容。度中心性越高, 与网络的其他属性联系就更多, 连接线的粗细可以代表关系紧密程度, 线越粗联系越紧密。在媒体议程设置中, 中国与意大利、口罩、物资、援助、美国之间有着紧密的联系。

同样, 将公众矩阵输入到 NetDraw, 得到公众矩阵如图 7 所示。

在 NetDraw 显示的图像中, 每一个节点都代表一个属性, 本研究以中心度结果显示节点的大小, 节点越大表明它的度中心性越高, 与其他节点联系更紧密, 在网络中也越处于重要的位置。公众议程中, 各属性次根据点度中心度数值排序依次为: 中国(点度中心度 1573)、援助(点度中心度 1220)、物资(点度中心度 719)、美国(点度中心度 660)、意大利(点度中心度 436)、口罩(点度中心度 397)、病毒(点度中心度 346)、欧盟(点度中心度 157)、医疗队(点度中心度 154)、俄罗斯(点度中心度 108)、人道主义(点度中心度 104)、武汉(点度中心度 102)、特朗普(点度中心度 78)、武契奇(点度中心度 64)、巴基斯坦(点度中心度 60)。

与媒体议程相同, 公众也对中国、援助、物资、美国、意大利、口罩、病毒、欧盟给予了重点关注。同时在公众的脑海中, 中国与物资、美国、援助之间存在着紧密联系。值得注意的是, 根据点度中性的数值和排序发现, 与媒体议程更关注“意大利”相比, 公众对“美国”的关注度最高, 并对巴基斯坦的关注度最低。

3.4.4. 基于 QAP 的相关性分析

QAP 分析(Quadratic Assignment Procedure)程序可以对媒体共现矩和议程共现矩进行相关性分析和回归分析, 两个矩阵相关系数越大, 证明矩阵间的相关性越大。通过数据显示, 媒体矩阵与公众矩阵的相关系数达到了 0.831, 媒体议程与公众议程正向显著相关, 媒体第三层次议程设置效果显著。议程网络实现了将公众脑海中的具体景象描绘了出来, 为媒体在新媒介下改善议程设置, 发现公众议程的关注重点和思维逻辑提供了重要参考依据。在“中国援助”这一事件中, 媒体不仅影响公众“想什么”和“怎么想”, 也成功影响了公众的认知网络。

4. 结论

本文基于机器学习算法和网络分析方法,研究了在新冠肺炎环境下“中国援助”的议题,客观的检验了议程设置的三个层次。与之前的学者不同,本文在第一层次的研究中,通过把媒体数据和公众数据错开天数作分析,得出在当代媒介环境下,议程设置的最佳时滞期从7天缩短到了当天,媒体可以及时影响公众;在第二层次的研究中,本文不再依据阅读媒体文章和公众评论自主归纳出媒体和公众态度,而是通过数据训练一个神经网络“大脑”,让机器的“大脑”去判断,极大提高了研究的客观性,结果更让人信服;在验证议程设置的第三层次时,属性词的选取一直是之前研究的难点,不论是通过研究者自己归纳还是依据问卷让公众提供,都存在较大误差,本文使用的TextRank算法结合上下文,选出文章中贡献度高的词语,为选取属性词提供了新思路,并让第三层次得到了客观和合理的验证。

通过本文的分析,还发现,在新媒介环境下,媒体可以通过以下路径进一步优化“中国援助”的议程设置。1)媒体需进一步弘扬人道主义缓解公众消极情绪;2)媒体应进一步提升国际影响力,改善国际舆论场;3)媒体可加强细节报道引发公众关注。

基金项目

北京工商大学教育教学改革研究项目(jg215203)。

参考文献

- [1] 陈贵梧. “无组织的有序”: 社会化媒体何以影响议程的设置?——以滴滴顺风车安全事件为例[J]. 电子政务, 2021(9): 2-12.
- [2] 李青青. 新媒体时代媒介议程设置理论嬗变与发展[J]. 中国出版, 2021(16): 28-31.
- [3] Neuman, W.R., Guggenheim, L., Jang, S.M. and Bae, S.Y. (2014) The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. *Journal of Communication*, **64**, 193-214. <https://doi.org/10.1111/jcom.12088>
- [4] Wanta, W., Golan, G. and Lee, C. (2004) Agenda Setting and International News: Media Influence on Public Perceptions of Foreign Nations. *Journalism & Mass Communication Quarterly*, **81**, 364-377. <https://doi.org/10.1177/107769900408100209>
- [5] 杨立公, 朱俭, 汤世平. 文本情感分析综述[J]. 计算机应用, 2013, 33(6): 1574-1578+1607.
- [6] 吴杰胜, 陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件, 2019, 36(9): 93-99.
- [7] 王晗啸, 李成名, 于德山, 巴志超. 基于上下文语义的网络议程设置研究——以红黄蓝事件为例[J]. 国际新闻界, 2020, 42(4): 77-97.
- [8] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Vol. 10, 79-86. <https://doi.org/10.3115/1118693.1118704>
- [9] 林聚任. 社会网络分析: 理论、方法与应用[M]. 北京: 北京师范大学出版社, 2009.
- [10] 郝帅, 罗晨. 第三层次议程设置检验——“李文星事件”中的媒体与公众议程网络分析[J]. 教育传媒研究, 2019, 19(2): 61-69.
- [11] Tomas, M., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. *Computation and Language*, arXiv: 1301.3781.
- [12] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. <https://doi.org/10.1007/BF00994018>
- [13] Mountrakis, G., Im, J. and Ogole, C. (2011) Support Vector Machines in Remote Sensing: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, **66**, 247-259. <https://doi.org/10.1016/j.isprsjprs.2010.11.001>
- [14] Schuldt, C., Laptev, I. and Caputo, B. (2004) Recognizing Human Actions: A Local SVM Approach. *International Conference on Pattern Recognition*, Cambridge, 26 August 2004, 32-36. <https://doi.org/10.1109/ICPR.2004.1334462>
- [15] Mihalcea, R. and Tarau, P. (2004) TextRank: Bringing Order into Texts. *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, 25-26 July 2004, 404-411.

- [16] 王运锋, 夏德宏, 颜尧妹. 社会网络分析与可视化工具 NetDraw 的应用案例分析[J]. 现代教育技术, 2008, 18(4): 85-89.
- [17] Creswick, N. and Westbrook, J.I. (2010) Social Network Analysis of Medication Advice-Seeking Interactions among Staff in an Australian Hospital. *International Journal of Medical Informatics*, **79**, e116-e125. <https://doi.org/10.1016/j.ijmedinf.2008.08.005>
- [18] 刘军. QAP: 测量“关系”之间关系的一种方法[J]. 社会, 2007, 27(4):164-174.
- [19] 窦慧彬. 我国传统媒体与社交媒体关于“二孩”政策的议程影响研究[D]: [硕士学位论文]. 上海: 上海外国语大学, 2017.
- [20] Roberts, M., Wanta, W. and Dzwo, T.H. (2002) Agenda Setting and Issue Salience Online. *Communication Research*, **29**, 452-465. <https://doi.org/10.1177/0093650202029004004>
- [21] Kim, C. and Klabjan, D. (2020) A Simple and Fast Algorithm for L1-Norm Kernel PCA. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **42**, 1842-1855. <https://doi.org/10.1109/TPAMI.2019.2903505>
- [22] Kweon, S.H., Go, T., Kang, B.Y., Cha, M.-K., Kim, S.-J., Kweon, H.-J., *et al.* (2019) Study on Agenda-Setting Structure between SNS and News: Focusing on Application of Network Agenda-Setting. *International Journal of Contents*, **15**, 10-24. <https://doi.org/10.5392/IJoC.2019.15.1.010>