

全国各省市的媒体镜像

—— 基于网络新闻大数据

QUANGUO GESHENSHI DE MEITI JINGXIANG

黄合水 彭丽霞 著



厦门大学出版社
XIAMEN UNIVERSITY PRESS

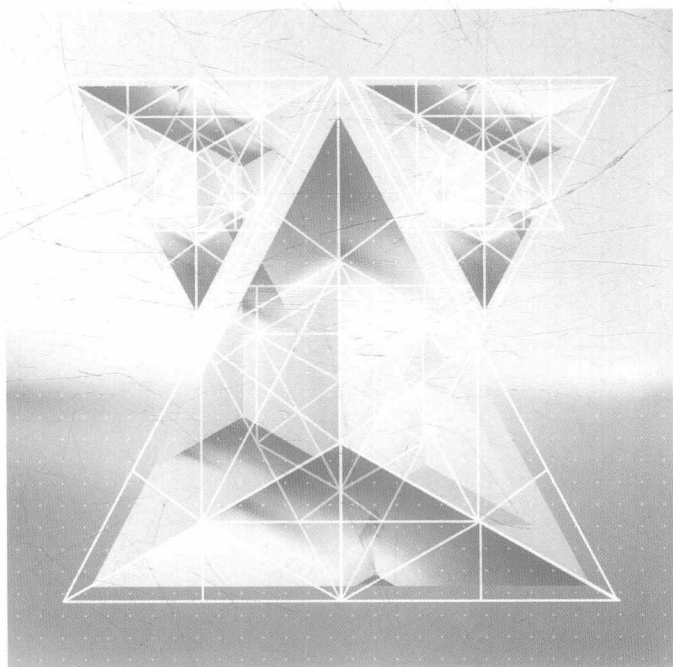
国家一级出版社
全国百佳图书出版单位

全国各省市的媒体镜像

—— 基于网络新闻大数据

QUANGUO GESHENGSHI DE MEITI JINGXIANG

黄合水 彭丽霞 著



厦门大学出版社
XIAMEN UNIVERSITY PRESS

国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

全国各省市的媒体镜像:基于网络新闻大数据/黄合水,彭丽霞著. —厦门:厦门大学出版社,2018.4

ISBN 978-7-5615-6821-7

I. ①全… II. ①黄…②彭… III. ①互联网络-新闻报道-研究 IV. ①G210.7

中国版本图书馆 CIP 数据核字(2017)第 322853 号

出 版 人 郑文礼

责任编辑 刘 璐

封面设计 李嘉彬

技术编辑 朱 楷

出版发行 厦门大学出版社

社 址 厦门市软件园二期望海路 39 号

邮政编码 361008

总 编 办 0592-2182177 0592-2181406(传真)

营销中心 0592-2184458 0592-2181365

网 址 <http://www.xmupress.com>

邮 箱 xmupress@126.com

印 刷 虎彩印艺股份有限公司

开本 787 mm×1 092 mm 1/16

印张 10.75

插页 2

字数 249 千字

版次 2018 年 4 月第 1 版

印次 2018 年 4 月第 1 次印刷

定价 69.00 元

本书如有印装质量问题请直接寄承印厂调换



厦门大学出版社
微信二维码



厦门大学出版社
微博二维码

前言

近些年,大数据研究在学术界变成显学,人们争相涉足。虽然我们不敢宣称自己是做大数据研究的,但反思一下过去二十多年的教学科研工作,却发现自己也慢慢地、不知不觉地从“小数据”走进了“大数据”。

从研究方法上讲,采用调查法、实验法进行的研究通常是小数据研究;采用内容分析法进行的研究更像是大数据研究。当然,当对固定样本持续不断地进行调查,积累了大量数据时,其研究未必不是大数据研究,反之,当分析的内容有限时,其研究也未必就是大数据研究。

在过去二十多年的教学科研中,我们先是采用调查法、实验法进行研究,这样的研究要找人接受调查、参与实验,需要付出大量的人力、财力,而且常常被诟病为样本太少或实验对象都是大学生,研究结果不具备普遍意义。后来,我们发现媒体(特别是平面媒体)中已经采集了大量的资料,分析起来比较容易,结果也很有意义,于是开始进行大量的内容分析研究。特别是当互联网媒体逐渐超越传统媒体成为人们获取信息的主要渠道之后,我们也慢慢地领悟到,其实我们的研究已经进入了大数据领域。

可归属于大数据研究范畴的比较直接的研究是2010年进行的“基于网络的明星代言价值测量”,该研究的出发点是评估明星代言价值,但在方法上具有大数据特点,直接从网络世界中采集明星的信息加以分析处理。2014年,我们开始了对品牌健康的连续监测研究,及对如何处理网络文本资料的方法研究。就这样,我们不知不觉地进入大数据领域。

网络世界里留存着大量痕迹,这些就是所谓的大数据。透过这些大数据不仅可以分析留下痕迹的人(主体)的特点,而且可以分析被描述对象(客体)在他人心目中的形象,还可以分析主体、客体及不同主体、不同客体之间的相互关系。

2015年,因部校共建,总觉得作为学者也应该为此做点什么,于是进行了自己认为有意义、有价值的关于“福建省在网络中的形象”的研究。以该项研究为基础,进一步拓展到全国,于是就有了本书。

从形态上看,本书有点像连环画。这是作者有意而为之,目的是满足大数据可视化要求,让读者能够比较轻松地阅读,同时留下较深的印象。

本研究数据采集客观,数据处理客观,但结果不一定符合读者已有的印象。因大数据动态变化快,课题完成时间又不长,所以难免存在谬误之处,欢迎批评指正。

本书的全国镜像地图系由本院实验室实验员张雨歌和硕士生金毓慧协助制作,特此致谢!

黄合水

2016年7月30日

于厦门大学新闻传播学院

目 录

导 读	1
一、研究目的	2
二、研究方法	4
1.操作定义	4
2.数据采集	6
3.数据编码	6
4.数据处理	10
5.数据表达	11
三、研究结果——单维镜像	12
1.全国月份	12
2.各省(直辖市、自治区)	17
3.各省辖行政区	40
四、研究结果——多维镜像	68
1.全国月份	68
2.各省(直辖市、自治区)	72
3.省城和单列市	80
4.其他地级行政区	89
五、研究总结	165

导 读

作为一个普通读者,你想知道以下问题的答案吗?

1.我国天灾哪个月份最严重呢?风调雨顺的省(直辖市、自治区)或城市有哪些?风灾、震灾等自然灾害最严重的又是哪些省(直辖市、自治区)或城市?

2.哪些省(直辖市、自治区)或城市堪称平安?

3.交通事故问题最严重的省(直辖市、自治区)或城市是哪些?

4.哪些省(直辖市、自治区)或城市的生产最安全、事故最少?

5.食品安全问题最多和最少的省(直辖市、自治区)或城市分别是哪些?

6.环境污染最严重、最轻的分别是哪些省(直辖市、自治区)或城市?

7.疫病多发的省(直辖市、自治区)或城市有哪些?

8.欺骗、违规、不诚信问题比较严重的省(直辖市、自治区)或城市有哪些?

9.腐败问题较重的省(直辖市、自治区)或城市是哪些?

10.社会治安问题比较严重的省(直辖市、自治区)或城市是哪些?暴力问题最少的省(直辖市、自治区)或城市又是哪些?黄、赌、毒在哪些省(直辖市、自治区)或城市严重一些?偷盗问题最严重的城市是哪个?

11.在各种天灾人祸中,你所在的城市或你的家乡什么问题最突出?什么问题最少?

12.你对各个省(直辖市、自治区)或城市的刻板印象和媒体镜像相吻合吗?

如果你是一个媒体工作人员,你还想了解以下问题吗?

1.媒体形成的关于某个省(直辖市、自治区)或城市的镜像,跟你的主观判断一样吗?

2.你或你所在的媒体想塑造的关于某个省(直辖市、自治区)或城市的形象,结果实现了吗?

如果你是政府官员,你还想了解以下问题吗?

1.各省(直辖市、自治区)或城市最突出的问题是什么?

2.媒体对你所在省(直辖市、自治区)或城市的报道符合实际情况吗?

3.你认为该省(直辖市、自治区)或城市的主要问题得到了恰当的媒体监督吗?

4.媒体对不同省(直辖市、自治区)或同一省辖的不同城市的报道真实、可靠吗?

下面我们将就如何透过对网络新闻大数据的分析研究来回答上述问题做详细介绍,希望各类读者能够从中读出自己的问题的答案。

一、研究目的

镜像,就是客体在镜子中的形象。平面的镜子,能够真实地反映客体的原貌;凸面镜、凹面镜以及其他形状镜面的镜子,所反映的是扭曲程度不同的客体。

人们对客体的认识、了解和印象,在许多情况下就是通过镜像实现的。照片实际上就是人的镜像,是一个人了解另一个人的载体。

媒体,在某种程度上也是社会的一面镜子,它不仅会形成关于特定客体的镜像^①,而且能将客体的镜像记录并保存下来,以便人们过后去看、去了解。

媒体中的客体镜像,是由零碎的信息构成的,十分庞杂,不经过一定的分析处理,人们很难认识客体的特征。有幸的是,现代网络大数据包括新闻大数据为人们分析描绘有关客体的媒体镜像提供了方便。

中国的新闻媒体,其主要功能一方面是宣传党的路线方针政策,起到党的喉舌的作用;另一方面是揭示社会问题,起到社会监督的作用^②。后一类新闻,通常也是人们比较关心、关注的,对人们了解社会、形成对社会的印象有重要的作用。

网络新闻中关于一个省(直辖市、自治区)的新闻信息浩如烟海,要利用所有信息全面地描绘省(直辖市、自治区)的镜像,十分困难。即使描绘出来了,其重要特征也不够突出。因此本研究从社会关心的热点以及研究的可行性角度出发,着重从以下几个方面对各省(直辖市、自治区)的镜像加以分析比较:(1)天灾(包括水灾、寒灾、热灾、震灾、风灾和其他自然灾害);(2)疫病;(3)食品安全;(4)交通事故;(5)治安(包括涉黄、赌彩、涉毒、偷盗和暴力);(6)生产事故;(7)环境污染;(8)违约(包括涉骗和违规);(9)涉腐;(10)伤亡。前9个方面是相对独立的,而最后一个方面是相对综合性的,与天灾、疫病、食品安全、交通事故、治安、生产事故都有交叉,甚至连“涉腐”也常常因自杀、谋杀等出现伤亡。这些方面都是人们日常比较关注的,所以我们对它们一一加以分析研究。

上述10个方面的新闻报道通常都包括如下三种类型:第一种是对事件进行披露,如揭示腐败事件;第二种是报道对问题的处治,如发现环境污染,采取了哪些措施等;第三种是关于现象的防范措施,如制定了哪些措施以避免色情传播等。第一种新闻是负面的,第二种和第三种新闻往往被认为是正面的,但是从一个省(直辖市、自治区)的角度来看,不管是正面还是负面的报道,只要有关于这些方面的报道,就说明该省(直辖市、自治区)存在相关问题或存在出现该类问题的风险。尽管没有报道并不等于没有问题,但不会形成

① 刘建明.当代新闻学原理[M].北京:清华大学出版社,2003:131.

② 李良荣.新闻学概论[M].上海:复旦大学出版社,2014:169-170.

镜像;报道了则证明问题存在,而且会形成媒体镜像,进而影响人们对该省(直辖市、自治区)的印象。

从以上 10 个方面来描绘各省(直辖市、自治区)的镜像,肯定是不全面、不完整的,但却是十分有意义的,也是十分必要的。研究结果不仅可为省(直辖市、自治区)政府的政策制定、决策管理提供依据,为省(直辖市、自治区)政府的效益考评、绩效管理提供参考,为媒体关于省(直辖市、自治区)新闻报道的评价和管理提供借鉴,而且有助于普通百姓了解各省(直辖市、自治区)之间的差别。

因此,本研究的目的在于完整、严密的理论表达,而在于探索一套利用新闻大数据描述省(直辖市、自治区)镜像的方法,同时试图将各省(直辖市、自治区)上述 10 个方面的新闻镜像科学地展示出来。

二、研究方法

1. 操作定义

本研究涉及的概念及操作定义如下：

(1)“省”：指中央政府直接管辖的行政区，包括省、直辖市和自治区。具体地说，本研究中的“省”只包括大陆的 22 个省、4 个直辖市和 5 个自治区，即安徽、北京、重庆、福建、甘肃、广东、广西、贵州、海南、河北、河南、黑龙江、湖北、湖南、吉林、江苏、江西、辽宁、内蒙古、宁夏、青海、山东、山西、陕西、上海、四川、天津、西藏、新疆、云南、浙江。香港、澳门和台湾暂未列入本研究范畴，特此说明。

(2)“市”：指“省”直接管辖的行政区，包括自治州、地区、盟和地级市等，全国共有 334 个地级行政区^①，此外，对于海南的 5 个县级市、4 个省辖县和 6 个自治县，河南的 1 个副地级市，湖北的 3 个副地级市和 1 个林区，也进行了分析。

(3)某省(市)新闻量：指特定时间里含有某省(市)名称的新闻标题数，用 NN_k 表示，单位是“条”。

(4)特征：指新闻标题中所含有的某特定信息，本研究中特指表 2-1 中的 22 个维度。

表 2-1 特征及其操作定义

序号	特征	操作定义
①	天灾	标题涉及非人为的包括水灾、寒灾、热灾、震灾、风灾和其他灾害，以及旨在控制这些灾害的行为或举措
	水灾	标题涉及由水直接造成或引发的灾害，以及旨在控制此类灾害的行为或举措
	寒灾	标题涉及由低温引发的灾害，以及旨在控制此类灾害的行为或举措
	热灾	标题涉及由高温引发的灾害，以及旨在控制此类灾害的行为或举措
	震灾	标题涉及由地震引发的灾害，以及旨在控制此类灾害的行为或举措
	风灾	标题涉及由大风引发的灾害，以及旨在控制此类灾害的行为或举措
②	疫病	标题涉及已经或将要发生瘟疫或疾病的现象，以及旨在消除瘟疫或疾病的行为或举措

① 中华人民共和国民政部全国行政区划信息查询平台：<http://202.108.98.30/map>。

续表

序号	特征	操作定义
③	食品安全	标题涉及可能或已经造成不良后果的食品质量问题,以及旨在消除此类问题的行为或举措
④	交通事故	标题涉及交通运输无法如期运行、发生人员伤亡或者财产损失的现象,以及旨在消除此类问题的行为或举措
⑤	治安	标题包括涉黄、赌彩、涉毒、偷盗、暴力等现象,以及旨在消除这些现象的行为或举措
	涉黄	标题涉及与“性”有关的行为和信息传播现象,以及旨在消除此类现象的行为或举措
	赌彩	标题涉及与赌博、彩票有关的现象,以及旨在消除赌博、管控博彩的行为或举措
	涉毒	标题涉及生产、运输、销售、吸食、藏匿毒品的现象,以及旨在消除此类现象的行为或举措
	偷盗	标题涉及偷窃、盗窃现象,以及旨在消除此类现象的行为或举措
	暴力	标题涉及利用武力伤害他人或损害公共财物的现象,以及旨在消除此类现象的行为或举措
⑥	生产事故	标题涉及机构在生产经营过程中发生的造成人身伤亡和财产损失的事故,以及旨在消除此类事故的行为或举措
⑦	违约	即违背契约,指标题中包含欺诈、不守信用、违反规章制度的现象,以及旨在消除此类现象的行为或举措,具体包括涉骗和违规
	涉骗	标题涉及欺骗、不遵守不成文的道德或道义契约的现象,以及旨在消除此类现象的行为或举措
	违规	指标题涉及明确指出不遵守明文规定的各种法律和规章制度的现象,以及旨在消除此类现象的行为或举措
⑧	环境污染	标题涉及已经或将要发生对环境造成不良影响的现象,以及旨在消除此类现象的行为或举措
⑨	涉腐	标题涉及揭露公职人员利用手中掌握的公权力为自己或他人谋取非法利益的行为及其所带来的后果,以及旨在防控此类现象的行为或举措
⑩	伤亡	标题涉及已存在或可能存在的受伤或死亡现象,以及旨在消除此类现象的行为或举措

(5)特征新闻量:指特定时间里含有某特征的新闻标题数,用 NF_{ik} 表示,其中 i 表示特征, k 表示“省”或“市”。

(6)千万人均特征新闻量:指每一千万人平均拥有特征新闻的数量, $TMPF_{ik}$ 表示。采用此指标的原因是人均特征新闻量数值太小,理解和表达都不便。

(7)特征关注度:指在特定时间里某特征新闻量占有新闻报道的比重,用 AF_{ik} 表示。同样因其数值较小,不便于处理和分析,而将其比重乘以 1000 来表示。

(8)特征指数:指用来衡量省(直辖市、自治区)与某特征相关程度的指标,是以特征关注度为加权系数加权后的千万人均特征新闻量,用 DF_{ik} 表示。

(9)特征指数对数:指用来衡量省(直辖市、自治区)与某特征相关程度的直观化数字,是将特征指数进行以 10 为底的对数转换而得,用 R_{ik} 表示。当特征指数值小于 1 时,令其对数值为 0。

(10)特征级别:指用来衡量省(直辖市、自治区)与某特征相关程度的等级化数字,是将特征指数对数进行级别转换而得,分为 0~5 共六个等级。

2. 数据采集

(1)采集工具:研究者自行开发的新闻标题采集软件——TCP。

(2)限定新闻源:七大新闻网站,包括人民网、新华网、新浪网、腾讯网、网易网、搜狐网和凤凰网。取七大网站而不是全国所有新闻网站的原因有以下两方面:一是小网站生生灭灭不稳定,不利于长期的观测研究;二是如果包括所有网站,数据量、数量级增多,会增加数据采集和分析的难度。

(3)采集过程:打开 TCP,设定采集关键词、时间范围和新闻源。关键词即省(直辖市、自治区)名称,如北京、新疆、福建等;时间范围,如2015年1月1日至2015年12月31日;新闻源,如 people.com.cn(人民网)或 ifeng.com(凤凰网)等。点击“百度一下”,并根据结果页左上角的新闻条数,选择点击800以上或800以下数据下载键,然后输入数据生成的文件名字,软件就会自动采集并形成 excel 格式的数据文件。下载结束后,返回到搜索页面,重复同一过程,直至采集到需要的所有信息。

(4)采集时间:2016年3月1日至2016年4月30日。

(5)获得数据:本研究采集到的所有数据包括标题中含有各市且时间为2015年的七大网站中的所有新闻标题。全部数据的新闻标题一共是2 991 508条。

3. 数据编码

(1)编码关键词

本研究的每一个特征都采用关键词进行编码。这些关键词是通过阅读大量的标题文本并通过不断地测试获得的。有些特征只需要一组关键词,有些特征则需要两组关键词组合;有些只需要一组或两组关键词进行一次编码即可,有的则需要至少两组关键词进行两次或多次编码。下面一一介绍各种特征的关键词编码及检索方法。需要说明的是,本研究中的天灾、治安和违约,都分别包含子维度,对于子维度的编码结果求和即可,故没有单独进行编码;其他灾害则纯粹是为了天灾而进行的编码,不做单独处理。

①天灾:包括水灾、寒灾、热灾、震灾、风灾以及其他灾害的关键词,共79个。其中水灾关键词包括“暴雨、强降水、强降雨、水灾、洪灾、大雨、洪水、山洪、防洪、海啸、泥石流、山体滑坡、雷雨、雹”14个;寒灾关键词包括“冻、冷空气、寒冷、寒潮、大雪、暴雪、风雪、雪崩、雪灾、强降雪、低温”11个;热灾关键词包括“旱灾、旱情、高温、雷击、雷电、雷劈、森林火灾、山火、森林大火、森林火险”10个;震灾关键词包括“地震、震中、余震、抗震、震灾、防震、震区”7个;风灾关键词包括“大风、强风、狂风、龙卷风、风灾、沙尘、台风、米克拉、海高斯、巴威、美莎克、海神、红霞、白海豚、鲸鱼、灿鸿、莲花、浪卡、苏迪罗、莫拉菲、天鹅、艾莎尼、艾涛、环高、科罗旺、杜鹃”26个;其他灾害关键词包括“蝗虫、虫害、赤潮、防灾、减灾、赈灾、救灾、灾区、灾害、受灾、大雾”11个。

②疫病:包括“病、疾、癌、瘟、疫、症、发烧、菌、埃博拉、登革热、发热、流感、麻疹、患者、N7N9、H1N1、H5N1、MERS、H3N2”19个关键词。

③食品安全:共49个关键词,分为两组:第一组是用于区分食品与非食品的,包括“茶、吃、饭、喝、假酒、粮、麦、米、奶、食、乳、饮用水、矿泉水、自来水、汤、饮、油、肉、菜、猪、牛、羊、鸡、鸭、鹅、药、农产品”27个;第二组是用于识别是否存在问题的,包括“假、安全、变质、不合格、超标、吃出、地沟、仿冒、腹泻、喝出、黑名单、呕吐、上吐、山寨、死猪肉、危害、问题、异物、下架、下泄、中毒、检疫”22个。编码时,组内是并集,组间是交集。只要新闻标题中既包含第一组的任一关键词,又包含第二组的任一关键词,则该标题被编码为与食品安全有关。

④交通事故:共46个关键词,分为两组:第一组是用于识别交通的,包括“BRT、车、船、的士、地铁、渡口、飞机、高速、高铁、公路、国道、航班、航线、货轮、机场、交通、快轨、码头、轻轨、铁路、轨道、轨交”22个;第二组是用于识别事故的,包括“沉、翻、祸、碰、伤、失、死、撞、坠、冲出、恢复、故障、起火、倾覆、事故、延误、遇难、着火、失火、火灾、追尾、取消、备降、盲降”24个。将两组字词交叉检索,即可识别出与交通事故有关的新闻标题。

⑤治安:包括涉黄、赌彩、涉毒、偷盗和暴力现象的关键词,共120个。其中涉黄关键词包括“娼、奸、裸、嫖、淫、黄赌毒、黄流、扫黄、妓女、牛郎、三陪、色情、色诱、涉黄、猥亵、脱衣、性爱、性病、性传播、性关系、性侵、性行为”22个;赌彩关键词包括“赌、博彩、中奖、中彩、六合彩、福彩、体彩”7个;涉毒关键词包括“冰毒、藏毒、查毒、大麻、毒巢、赌毒、毒贩、毒犯、毒驾、毒窟、毒品、毒情、毒窝、毒泉、毒瘾、毒资、杜冷丁*、贩毒、古柯、海洛因、缴毒、缉毒、戒毒、禁毒、K粉、咖啡因、可卡因、吗啡、买毒、卖毒、涉毒、三唑仑、扫毒、吸毒、鸦片、摇头丸、运毒、心毒、制毒”39个;偷盗关键词包括“偷、盗、窃、丢失、贼”5个;暴力关键词包括“刀、匪、捅、挨打、暗杀、绑架、暴打、暴恐、暴力、暴徒、被害、被劫、被砍、被抢、被杀、打劫、打人、打伤、打死、打砸、盗抢、黑帮、黑社会、奸杀、劫持、砍人、砍杀、砍伤、砍死、勒索、谋杀、抢夺、枪杀、强奸、抢劫、殴打、斗殴、杀害、杀人、凶手、遇害、遭抢、砸坏、砸破、杀夫、杀妻、家暴”47个。

⑥生产事故:共28个关键词,分为两组:第一组是用于识别生产的,包括“厂、矿、车间、仓库、工地、公司、集团、农场、企业、施工、店、商铺、大厅、营业厅”14个;第二组是用于识别事故的,包括“着火、失火、起火、火灾、漏、伤、死、损、塌、坍、爆炸、被困、事故、透水”14个。两组关键词交叉判别,即可识别出与生产事故有关的新闻标题。

⑦违约:包括涉骗和违规现象的关键词,共40个。其中涉骗关键词有37个,分为两组:第一组是用于识别涉骗的,包括“谎、假、瞒、冒、骗、欺、谣、诈、承诺、承认、诚信、兑现、坑害、蒙人、上当、守信、投诉、信用、质疑、作弊、虚报”21个;第二组是用来排除第一组检索中带入的非涉骗标题的,包括“假日、寒假、暑假、假期、休假、长假、冒烟、冒火、感冒、冒雨、度假、婚假、病假、产假、放假、年假”16个。编码时,两组关键词分别进行编码,结果是从第一组编码结果中剔除第二组编码结果。违规关键词包括“违、非法、不法”3个。

⑧环境污染:包括“PM、臭、霾、香烟、烟花、烟雾、烟囱、控烟、肮脏、脏乱、赤潮、高能耗、环保、环境灾害、垃圾、排污、受污、酸雨、污染、污水、治污、低碳、碳排放”23个关键词。

* 即盐酸哌替啶。由于新闻报道标题的通俗性,多用“杜冷丁”一词,本书予以保留。

⑨涉腐:共 109 个关键词,分为三组:第一组可以直接查出与腐败有关的新闻标题,包括“八项规定、表叔、反腐、反贪、房婶、房叔、腐败、纪委、巨贪、挪用、受贿、首虎、双规、双开、四风、贪腐、贪官、贪污”18 个;第二组是用于识别与腐败有关的人员,包括“厅官、高官、官员、组长、所长、村长*、科长、乡长、镇长、社长、院长、县长、区长、市长、署长、厅长、局长、司长、省长、部长、委员长、秘书长、校长、教务长、处长、军长、师长、团长、营长、连长、排长、班长、董事长、行长、检察长、机长、秘书长、常委、城管、城监人员、村委、调研员、干部、公安人员、公务员、公职、监察人员、检察人员、将军、警察、经理、人大代表、司令、书记、委员、巡视员、政委、支书、主管、主任、主席”61 个;第三组是用于识别腐败的,包括“被抓、被免、被曝、捕、查、渎、贿、挪、涉、贪、违、刑、亡、罪、罢免、撤职、处分、公车、公费、公款、降级、举报、开除、落马、跑路、情妇、通奸、停职、外逃、自杀”30 个。后两组字词交叉检索,可以查出第一组未能查出的与腐败有关的新闻标题。编码时先按第一组关键词进行检索,然后再将第二和第三组关键词交叉检索,将两次检索结果合并,即可以识别出绝大多数与腐败有关的新闻标题。

⑩伤亡:包括“伤、死、暗杀、被害、被杀、毒杀、奸杀、砍杀、罹难、命案、谋杀、枪杀、身亡、轻生、杀害、杀人、失踪、尸体、丧命、丧生、牺牲、遗体、遇难、遇害、自尽、自杀”共 26 个关键词。

(2) 编码信度

为了检验关键词编码的信度,本研究从原始标题数据中系统抽取 1461 条标题,并先后进行人工编码和关键词编码,表 2-2 是人工编码和关键词编码的结果,其中 A 列是人工编码的统计值;B 列是关键词编码的统计值;C 列是人工编码和关键词编码的交集,即 $C=A \cap B$;

D 列是关键词编码的正确命中率,即 $D=\frac{C}{A} \times 100\%$;E 列是采用关键词编码发生错误的比

率,即 $E=\frac{B-C}{B} \times 100\%$;F 列是关键词编码的信度,计算式是 $F=\frac{1461-B+C}{1461} \times 100\%$ 。

通过对两种编码结果统计分析,我们发现:第一,关键词编码信度均非常高;第二,命中率也很高,23 个维度中有 22 个能够命中人工编码结果的 80% 以上,只有暴力维度的命中率低于 80%,为 71.43%,这主要是因为“打”“杀”等暴力字的用法太灵活所致;第三,12 个维度的错误率都在 10% 以内,8 个维度的错误率大于 10% 且小于等于 20%,只有赌彩维度的错误率高达 50%。可见,采用关键词编码是可行的,但是个别维度的错误率还有降低空间。

表 2-2 关键词编码的信度检验

灾害	人工编码 结果(A)	关键词编码 结果(B)	两种编码 结果交集(C)	关键词 命中率(D)	关键词编码 错误率(E)	关键词编码 信度(F)
天灾	35	40	35	100	12.5	12.5
水灾	6	6	6	100	0	100
寒灾	9	11	9	100	18.2	99.9
热灾	4	5	4	100	20	99.9

* 即村主任。由于新闻报道的通俗性,多用“村长”一词,本书予以保留。

续表

灾害	人工编码 结果(A)	关键词编码 结果(B)	两种编码 结果交集(C)	关键词 命中率(D)	关键词编码 错误率(E)	关键词编码 信度(F)
震灾	1	1	1	100	0	100
风灾	7	8	7	100	12.5	99.9
其他灾害	8	9	8	100	11.1	99.9
疫病	18	21	18	100	14.3	99.8
食品安全	10	10	9	90	10	99.9
交通事故	28	31	28	100	9.7	99.8
治安	45	44	39	86.7	11.4	99.7
涉黄	8	10	8	100	20	99.9
赌彩	1	2	1	100	50	99.9
涉毒	7	7	7	100	0	100
偷盗	8	9	8	100	11.1	99.9
暴力	21	16	15	71.43	6.3	99.9
生产事故	13	11	11	84.62	0	100
违约	64	65	62	96.9	4.6	99.8
涉骗	35	36	33	94.3	8.3	99.8
违规	29	29	29	100	0	100
环境污染	33	34	33	100	2.9	99.9
涉腐	9	8	8	88.9	0	100
伤亡	36	34	32	88.9	5.9	99.9

(3) 编码过程

采用研究者自行开发的关键词编码软件——KWCP(界面见图 2-1)。编码过程中,先打开 KWCP 软件,点击文件,导入需要编码的文件,键入所要编码的关键词或关键词组



图 2-1 KWCP 软件的操作界面

合,并确定所要编码的是 excel 的哪一列以及编码赋值,单一文件点击“编码”,多个文件点击“全部编码”,软件就会自动在后台进行编码,编码结果直接写入编码的文件。

4. 数据处理

编码之后,将七家媒体的编码结果汇总,就可以统计出以月份为时间单位的各市特征新闻量(NF_{ik})以及总新闻量(NN_k),在特征新闻量和总新闻量的基础上,本研究依次进行如下处理:

第一,将某市某月的特征新闻量(NF_{ik})除以该市人口数(P_k),即可获得该市某月的人均特征新闻量。鉴于人均新闻量数值太小,不易理解和表达,为此,我们用千万人均特征新闻量(用 $TMPF_{ik}$ 表示,单位是条/千万人,取值范围为 $0 \sim \infty$)代替人均特征新闻量,计算公式如下:

$$TMPF_{ik} = \frac{NF_{ik}}{P_k} \times 10^7 \quad (1)$$

年千万人均特征新闻量,可通过求各月千万人均特征新闻量的均值而得。

统计处理时使用的各市人口数资料来自国家统计局及各省(直辖市、自治区)统计局网站公布的全国第六次人口普查资料^①。

第二,将某市特征新闻量(NF_{ik})除以该市总新闻量,即可获得该市的特征关注度(用 AF_{ik} 表示,同样因其数值较小,不便于处理和分析,而乘以 1000 来表示,取值范围为 $0 \sim 1000$),计算公式如下:

$$AF_{ik} = \frac{NF_{ik}}{NN_k} \times 1000 \quad (2)$$

第三,将某市特征关注度作为加权系数,乘以该市的千万人均特征新闻量,即可获得该市的特征指数(用 DF_{ik} 表示,取值范围为 $0 \sim \infty$),计算公式如下:

$$DF_{ik} = \frac{NF_{ik}^2}{P_k \times NN_k} \times 10^{10} \quad (3)$$

通过特征指数,能够比较客观地衡量各市与特征的相关程度。此指标可用于跨市、跨媒体、跨时间甚至跨特征之间的比较。但由于数据取值范围较大,不容易做直观的比较,所以需要进一步转化为指数对数。

第四,将某市的特征指数取以 10 为底的对数值,即可获得该市的特征指数对数,计算公式如下:

$$R_{ik} = \log_{10} DF_{ik} \quad (4)$$

^① 2010 年第六次全国人口普查主要数据公报: http://www.stats.gov.cn/tjsj/zxfb/201104/t20110429_12706.html, http://www.stats.gov.cn/was5/web/search?page=1&channelid=288041&orderby=-DOCRELTIME&was_custom_expr=like%282010%20年第六次全国人口普查主要数据公报%29%2Fsen&perpage=10&outlinepage=10。

通过式(4)得到结果之后,令特征指数小于1的任何数的对数值为0。由此得到的对数值最小为0,最大值不限,但一般不超过6。

省级数据是对各市的特征新闻量、总新闻量、各市人口数等数据分别求和,再根据式(1)至(4)计算特征指数对数。同理,全国数据、月份数据也是对相应的参数求和,再通过公式计算。

特征指数对数(R_{ik})的大小表面上反映的是该市与某特征的相关程度,但实质上反映了一个市该特征问题的严重程度。举个例子来说,如果一个市的涉腐特征指数或指数对数值高,说明该市涉及腐败的报道多。这些报道不管是揭露腐败现象,还是关于反腐的行为举措,都说明腐败问题严重,否则就不会有关于腐败现象的报道,也无须反腐。因此, R_{ik} 值大小反映的是特征问题的严重程度,即值越大,问题越严重。

5.数据表达

为了让读者能够比较直观地了解研究结果及其含义,我们主要用镜像蜘蛛图和镜像彩色地图及一种彩色表格将数据结果呈现出来。镜像蜘蛛图的制作直接采用 R_{ik} 值,而镜像彩色地图和彩色表格的制作则需进一步进行数据转化,即将 R_{ik} 值转化为等级值。具体见表2-3:

表 2-3 镜像解读说明

特征指数对数	级别	等级颜色	情况描述
$R_{ik} < 0.5$	0级	绿色	没有问题
$0.5 \leq R_{ik} < 1.5$	1级	蓝色	问题轻微
$1.5 \leq R_{ik} < 2.5$	2级	黄色	问题轻度
$2.5 \leq R_{ik} < 3.5$	3级	橙色	问题中度
$3.5 \leq R_{ik} < 4.5$	4级	红色	问题严重
$R_{ik} \geq 4.5$	5级	黑色	问题很严重

在解读镜像时,我们结合现实情况认为,达到3级或以上就构成问题并值得重视,3级以下则看作正常现象。

依据上述研究过程,我们进行了大数据的采集和分析,得到各省(直辖市、自治区)2015年在天灾等10个方面22个维度的媒体镜像。下面我们分两个部分来加以描述:第一部分是单维镜像,即从10个方面22个维度来比较全国不同月份的差异;比较各省(直辖市、自治区)的差异,并根据数值大小列出各维度中前三名和后三名;比较每个省(直辖市、自治区)中各直辖行政区的差异(限于篇幅,天灾、治安和违约之下的12个子维度不做比较),并列出各维度中前十名和后十名的地级行政区。第二部分是多维镜像,即从10个方面22个维度来描述各省(直辖市、自治区)以及单一行政区(包括省份、地级行政区)的镜像,以比较单一行政区在22个维度上存在问题的相对严重程度。