网络舆情研究新路径：大数据技术辅助网络内容挖掘与分析

张荣显 曹文鸳

摘　要：国内的舆情分析研究文献显示，舆情主要涵盖分析社会的现实和变动的状况，包括引发的事件本身及相关舆论生成的因素推论。针对当前网络舆情研究缺乏对舆情本质的理解和系统的分析框架，分析结果亦只依据描述性统计来作出等缺憾，提出一种全新的网络舆情研究路径，以覆盖度、测量和解释为网络舆情挖掘三大要素，搭建以人机结合的网络舆情大数据分析平台，即利用机器学习和网络挖掘技术初步概览舆情面貌，再以人工在线内容分析方法深度挖掘和解释舆情事件。将以具体案例说明此路径的实用性和可操作性。

关键词：网络舆情；大数据技术；网络挖掘；机器学习；内容分析

中图分类号：G20　　**文献标识码：**A

一、前言

依据中国互联网络信息中心最新的统计报告（2016），截至 2016 年 6 月，中国大陆互联网普及率达 51.6%，网民数量为 7 亿人，已经形成具有庞大规模的网民体量，网络成为重要的舆论平台。

随着网络舆情研究进入大数据时代，网络挖掘和机器学习等新技术使得快速甚至是即时搜集和处理大量网络数据成为现实，但是，大数据技术并非万能，在研究和探索舆情本质的过程中，依然需要人工判断作为主要的分析和解释手段，我们以此尝试解决当前舆情研究中缺乏整合性和系统性不足，甚至是被技术导向主宰的问题。

本研究针对网络舆情研究之现况及需要，提出一个全新的网络舆情研究路径，以覆盖度、测量和解释为网络舆情挖掘要素，搭建以人机结合的网络舆情大数据分析平台。在实践方面，将整个分析框架和机制，集合于一实时数据挖掘平台上，透过一体化的舆情监测和分析流程，力图达到高效、准确、广度和深度并重，以及随时跟踪舆情事件发展动向之目的。

张荣显　亚太区互联网研究联盟主席，澳门易研网络研究实验室总裁，博士

曹文鸳　珠海横琴博易数据技术有限公司资深研究顾问，硕士

二、网络舆情的概念及舆情数据的特点

闵大洪（2016）总结过去对舆情理论的研究成果，尤其是对概念的定义方面，汇整并形成总结性的概念定义：“舆情，系指社会的现

实和变动的状况，包括各种原因引发的事件本身及相关舆论的生成。”舆情监测既有苗头性又有全局性，苗头性即是需要在事件未形成舆论之前及早察觉和监测；全局性则是指需要对社会不同阶层、政经势力、利益相关或某一专门领域状况的整体呈现。

目前舆论阵地已大幅度延伸至网络环境中，喻国明（2010）、谢耕耘（2011）及尹培培（2013）等学者对网络舆情进行了全面论述，即是认为网络舆情是指民众通过互联网针对自己所关心或自身权益紧密相关的公共事件、社会现象等作出的主观反映，是多重态度、意见等交互的综合表现。网络舆情特点包括自由、情绪化、分散、即时、多变等，影响力强。网络舆情监测的总体目标是能够在最短的时间内发现所需要监测的舆情信息，寻找到首发的信息源，接着监测范围扩大至所有涉及信息来源，并分析传播的趋势和范围，时刻跟踪事态发展及带来的新情况（闵大洪，2016）。

相比传统媒体信息，网络数据内容更新快速（Velocity），数据形式多样（Variety），不仅限于传统内容的图文形式，更具有视频、动画等内容形式，网络舆论趋势不确定性高（Veracity），数据体量巨大（Volume），内容复杂（Complexity）和数据的非结构化（Unstructured）特征明显，蕴含无法忽视的高价值（Value）属性。从数据结构上的特点来说，如果数据简单、规律、重复性高，那么运用传统分析手段或简单的数据挖掘方法就能进行归类分析，然而，正是因为当前网络舆情数据包括大量的社交媒体和移动互联网数据在内，数据间存在关联性，同时呈现明显的非结构化特征（胥琳佳，2013），使其分析难度加大。

从事件特征上来说，在网络舆情的环境下，传统新闻叙事上的 5W1H 较难以辨认，不再有明确的事件发生地点（Where），取而代之是多样的来源；无固定的内容发布时间（When），即时更新成为常态；人物（Who）身份模糊、隐蔽；事件（What）本身焦点模糊；叙事（How）散乱；欲对事件原因（Why）的挖掘，则变成了难于理解事件的背后故事；难以测量理论；更难以发现形态。

三、网络舆情研究的现况与不足

因网络舆情具有前述特点，加大了舆情研究工作的难度，加上舆情监测行业发展年份尚浅，在当前网络舆情监测和分析领域中，存在诸多问题和不足。

目前国内网络舆情监测服务机构主要区分为三类，分别是：（1）依托人民网、新华网等主流媒体建立的舆情监测平台，以服务政府有关部门为主；（2）由高校或学术机构创办的舆情研究所，具有学术传统；（3）由软件公司或其与传统的市场调查公司联合成立的舆情监测软件企业，抓取互联网舆情数据能力较强。不同的网络舆情监测机构由于背景不同，在产学研等方面各有其优势及不足，整体而言，相关产业存在不同程度发展产品单一，同质化严重或缺乏产业内融合机制等问题（于新扬，2015）。

多位研究者在汇总和整理当前网络舆情研究文献及行业发展现状后，总结认为大数据时代下的网络舆情研究研究学科视角单一，缺乏跨学科的有关研究，未能进行动态化、立体化、全局化的综合探讨，为研究而研究，研究结果难以转化为实际应用系统。整体而言，存在系统性不足的问题（林源，2015）。由于当前网络舆情研究缺乏对数据的整合，未能有效地结合网络舆情数据与相关外部数据，导致数据割裂及解读片面；研究偏于平面和孤立，未能精到地解析舆情事件或话题背后的深层原因（燕道成和姜超，2015；上海交通大学舆情研究实验室，2014）。

更进一步，有研究者指出，当前网络舆情研究出现了技术导向的研究特点，即是过于围绕大数据展开网络舆情研究，缺少对社会舆情生成、发展、演化和衰退的内在机理来研究社会舆情信息的获取与识别、监测分析与预警、导控等治理决策方案（蔡立辉和杨欣翥，2015）。

四、新的舆情研究路径——大数据技术辅助网络内容挖掘与分析

本研究提出一种新的舆情研究路径——大数据技术辅助网络内容挖掘与分析，是以人机结合基本理念的舆情研究机制，有别于当前主流的网络舆情研究手段，以改善网络舆情研究遇到的方法论问题，具有挖掘广度、深度及监测结果更为全面和准确的特点。

（一）新舆情研究路径的理论框架

大数据时代，网络技术手段已可以支持以普查方式覆盖处理海量的网络数据，不再如传统舆情信息需要抽样以代表母体的处理方式，也由此得出了“数据足够大的时候，就可以自己说话结论”的论断。然而，虽然不再担心抽样偏差，却产生新的忧虑，即是需要考虑数据源本身的偏差。由于整体数据可能含有噪音，如不排除，则容易高估算法的精确度。同时，大部分的数据是孤岛状态，在整合处理时，无法准确地忽略和重合数据，也易导致数据结果偏差。可见，让数据“自己说话结论”是危险的论断，其中需要对数据源的清理，才能避免潜在误差。

本研究指出处理网络舆情数据面临的挑战，并提出以社会科学逻辑和业务思考为基础的解决方式，包括覆盖度（Coverage）、测量（Measurement）和解释（Explanation）三大要素。

1. 覆盖度（Coverage）

覆盖度即是解决数据是否齐全、代表性及数据质量的问题。舆情研究中，不论是传统媒体条件下还是大数据时代，相比全部数据来源，数据信息是否具有代表性更为重要。数据的过度覆盖易引入过多的含有歧义或无关的信息，会影响算法的精确度。同时应高度关注关键字搜索的设计和操作。由于自然语言使用灵活和含义丰富，简单的关键字设置搜索出的数据结果，与实际所需要的数据库结果可能存在较大偏差，从而导致误差存在。

不少学者也曾经以“谷歌流感趋势预测”（Google Flu Trend，GFT）为例，来说明这个问题。谷歌发现某些搜索关键词能够很好地标示流感疫情的现状，因此，谷歌使用经过汇总的谷歌搜索数据来预测流感疫情，并将其预测结果与美国疾病预防控制中心（Centers for Disease Control and Prevention，CDC）的监测报告作对比。然而在 2009 年，谷歌依据 2008 年前的资料建立起的数据模型所预测出来的结果远低于 2009 年实际所发生。而后，修正模型后，在 2013 年，其数据再次出现高估的问题，至此，谷歌关闭了 GFT 的功能，并且未再更新资料（https://www.google.org/flutrends/about/）。一项发表在《科学》杂志的研究指出，出现这种结果的两个重要原因是“大数据傲慢（Big Data Hubris）”和算法变化。“大数据傲慢”即认为大数据可以完全取代传统的数据收集方法，而这种观点最大的问题在于，绝大多数大数据与经过严谨科学试验得到的数据之间存在很大的差异，因为其忽略了最基本的有关测量、概念的信度与效度及数据之间的依赖性。另一方面，算法本身会经过调整和改进，算法的改变合并用户的搜索行为或是媒体的报道，均可能会影响 GFT 的预测，即是数据持续更新，算法无法做到随时调整，由此带来其结果的误差（Lazer et al，2014）。

因此，为掌控研究质量，需认识到数据过度覆盖和数据来源不足同样易造成数据质量不佳的情况，我们提出，舆情研究需要考虑合理的数据范围，可利用搜索关键词的逻辑设置，将舆论话题概念化，并利用可人工二度判断的手段来解决数据覆盖度的问题。

2. 测量（Measurement）

测量即是解决可以挖掘什么的问题。在大数据技术的协助下，机器已经能够完成许多自动化的测量工作，如网民行为（点赞数、阅读数、分享数、来源、路径、发展趋势、评论声量等）及文本的情感测量，当前舆情监测工作较为重视对行为的测量，准确度高，但是对于态度的测量仅以正负面的标尺为主，对舆情本质，如态度或意见的强度、有条件式的立场或意向等方面的测量较为欠缺，无法分析在什么情况下的“支持”或“反对”的意向，也难以辨别不同利益相关者之间的态度差异。

再者，往往对网民的意见数据测量存在缺

乏理论概念、甚至偷换概念的情况，如以声量代替影响力的测量、以正负面的情感来代替满意度和支持度等情况，因此导致测量效度不确定。另一方面，以中文语义技术为手段的情感分析，准确度尚不理想，与传统民调结果难以相提并论。以语料匹配方式所能达到的分析准确度少于60%，即便使用有优良的训练集的机器学习方式，在理想的场景下，可将准确度提高至80%（祝建华，2012），但此结果依然难以满足需求。因此，需要在适合舆情研究的理论框架和依据的支持下，建立具有科学性和系统性的测量标准，才能正确地进行舆情的深度挖掘。

3. 解释（Explanation）

解释即是解决如何分析和解释发现的问题。网络舆情的解释度视乎分析的深度，而当前主流的机器自动化分析，绝大多数基于描述性分析，即是以单变量分析为主，如各种排名榜单，分析单薄，解释性不强，提炼洞察困难。因此，需要在掌握单变量的数据信息基础之上，关注变量之间的差异和关系，以回答有意义和有深度的研究问题或检验假设。

（二）大数据技术辅助网络内容挖掘与分析研究机制

基于上述对舆情研究路径的理论框架的探讨，大数据技术辅助网络内容挖掘与分析研究机制设计使用人机结合的理念，力图避免当前网络舆情研究的误区和偏差。该路径的执行流程为，先采集网络上的海量信息，再结构化处理，随后利用网络挖掘和机器学习技术，结合人工在线内容分析，充分考虑分析结果的准确度，深度挖掘舆情事件，最终获得有价值的洞察。

1. 网络挖掘与机器学习

网络挖掘（Web Mining）是指互联网中普遍使用的数据挖掘方式。以研究目的区分，网络挖掘区分为三种类型：（1）内容挖掘（Content Mining）：以单个文件或网页为分析单位，以文本分析为主，用于分析半结构化或结构化处理后的信息；（2）结构挖掘（Structure Mining）：分析网页的节点和结构，包括从网页超链接中提取规则，或是挖掘文本结构；（3）使用行为挖掘（Usage Mining）：挖掘网页访问者的使用记录（Herrouz et al., 2013）。

机器学习（Machine Learning）定义为“机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能（Langley，1996）”，是借助数据或以往的经验，以此优化计算机程序的性能标准的方法（Alpaydin，2004）。

本研究综合运用网络挖掘与机器学习技术，结合技术专长与社会科学研究知识的积累，设定合适舆情分析的网络挖掘研究框架，具体为通过机器算法、语意分析技术和自动化关键字匹配等技术，快速挖掘网络舆情信息，以描述和挖掘舆情事件或现象的面貌。在网络挖掘的研究框架下，当前可透过机器挖掘自动化分析的主要面向和指标包括（不限于）：

	分析面向	分析指标
1	传播来源	网络数据来自具体的媒体来源，如社交网站（如Facebook、微博）、新闻网站、博客、论坛等；
2	传播量度	网络舆情或口碑的声量，以描绘事件的发展趋势；词云图以字体在图中的大小来表示声量大小或关注点等；
3	传播内容	网络舆情所涉及的话题、人物、机构、品牌等；
4	传播特征	以数量来描绘舆情话题的走势、事件发生的路径等，以解释传播过程和特征；
5	传播力度	点赞量、跟帖量、分享量、阅读量、排行榜等，还有参与度、曝光量、KOL等，以多项参数来综合解释舆情的传播力度；
6	传播效果	以情感分析作为尺度，衡量传播效果。

以下分别以两个发生在澳门的案例来说明上述的机器自动化分析结果。

例1：“台风“妮妲”袭澳事件舆论分析

本部分以台风“妮妲”袭澳事件为例，透

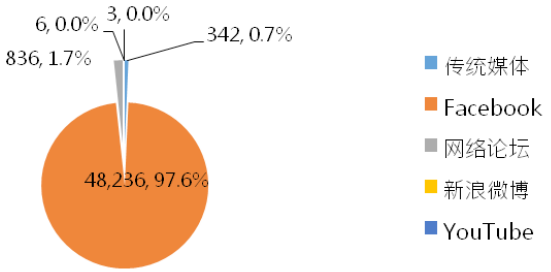
过网络挖掘和机器分析结果，综合说明舆情事件的传播来源、传播量度、传播内容、传播特征、传播力度、传播效果以及不同阶段的态度差异和变迁。

背景：2016 年 8 月，台风“妮妲”来袭，澳门于 1 日下午开始悬挂 3 号风球。香港天文台 1 日晚 8 点 40 分悬挂八号西北烈风或暴风信号（8 号风球），澳门未有像香港悬挂 8 号风球，引发社会质疑。

观察期内网络舆论（包括 Facebook、网络论坛、新浪微博和 YouTube）信息量明显高于传统媒体，其中以 Facebook 信息量最多，占 97.6%，明显高于其他传播来源。

台风“妮妲”袭澳事件信息量分布

检测日期：2016年8月1日至8月5日



进一步观察信息量最大的 Facebook 社交媒体，三个现场直播气象局发布会的 Facebook 专页获得较高的点赞数、评论数和转发数，三条直播主帖短时间内共获得 3.5 万回帖，占 Facebook 总帖数 74.1%，引起网民极大回响。其中，以 Facebook 专页「Lotus TV」直播发布会传播力度最为显着，共计获得点赞数 2,073 个，27,910 条回帖，转发次数达 2,491 次。

8月2日三个Facebook专页现场直播气象局发布会

作者	来自	信息	类型	心	💬	↻
Macau Cable TV 2016-08-02 15:25:56	Macau Cable TV	【现場直播】地球物理暨氣象局就於颱風"妮妲"的說明之記者會	video	447	5115	381
Lotus TV 2016-08-02 15:17:02	Lotus TV	直播氣象局情況	video	2073	27910	2491
澳門日報-Macao Daily News 2016-08-02 15:28:45	澳門日報-Macao Daily News	直播 氣象局發佈會	video	208	2721	173

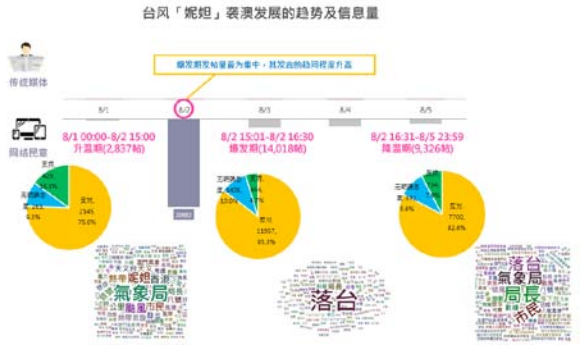
三条直播主帖及其回帖谈及“落台”、“局长”、“下地狱”等词最多，负面表达较为强烈，对局长的不满意见明显，要求其落台呼声较大。

三条直播主帖及其回帖词云图

注：词云图以相关文本数据源为分析基础，其结果的繁简体表达受文本数据自身字体限制，下同。

观察整个事件的发展趋势及信息量，可明显区分事件发展阶段，以 8 月 1 日零时至次日 15:00 为升温期，网络舆情内容有 2,837 条民意帖，该阶段的反对态度达到 75.6%，词云图显示关键词为“气象局”、“妮妲”、“台风”等，可见民意讨论集中于台风形势本身；第二阶段，8 月 2 日 15:01 至 16:30，短短一个半小时内网络民意猛烈增加，相关民意帖达到 14,018 帖，事件发展至爆发期，该阶段舆论反对态度扩散至 85.3%，“落台”为最明显的关键词，信息量远超其他关键词，说明该意见占据主流舆论；随后 8 月 2 日 16:31 至 8 月 5 日 23:59，对该事件的探讨明显下降，网络民意有 9,326 帖，讨论进入降温期，反对意见稍微减少，为 82.6%，但是依然高于升温期，关键词为“局长”、“气象局”、“落台”等。

结合事件发展的趋势及信息量，可以发现，在舆论爆发期发帖量最为集中，同时发言的趋同性升高，由词云图反映出，舆论走向由对台风天气本身的关注，转向对政府相关部门失职的问责，表达出强烈的反对态度。



注：其中各阶段态度取向分析为人工内容分析结果，在此引用仅为了更好配合网络挖掘数据进行解释。

例 2：“2015 年澳门施政报告”四象限词云图

2015 年 11 月 17 日下午，澳门特区政府行政长官崔世安在立法会发布 2016 年施政报告。报告范围包括持续提升民生素质、促进经济适度多元、建设宜居城市、深化公共行政改革等四方面。观察施政报告发布前后整一个月的相关舆情内容，利用分词、机器学习及数据可视化等一系列的技术手段，制作出四象限的词云图，以初步解释舆情的度量、内容和传播效果。在对词云图解读方面，关键字的字体大小表示讨论声量的多少，横坐标为正面及负面，纵坐标为讨论什么和怎么讨论。

如下图，在涉及“2016 年澳门施政报告”的舆情中，在词云图左侧“正面”情感方向上，网民讨论“施政”和“政府”最多（右上角），其次为“行政”和“博彩”，与施政报告内容和热点相关议题密切相关；在如何讨论（右下角）中，以“直播”为涉及最多，是对信息传播方式的讨论，再次为“保障”，说明其为网民提及施政报告时较为关心的方面。

词云图右侧“负面”情感方向上，左上角网民讨论什么中，最明显的关键词为“政府”和“施政”，与政府发布施政报告议题高度相关，其余较为明显且较为相关的关键词为“公屋”、“土地”、“承建商”和“房屋”，说明住房问题为网民讨论中较为负面的议题；右下角网民如何讨论话题以“填海”为最关键词，说明该议题的动作是讨论较多负面的内容。

2. 人工在线实时内容分析

舆情研究的本质是为在尽可能短的时间内于茫茫信息中获得所需要监测的舆情内容，并且深入解析舆情事件的方方面面，包括前因后果，即时动态，未来可能的发展趋势。因此，仅掌握机器挖掘出的数据并不足够。机器可帮助掌握包括传播来源、传播量度等单变量数据，然而在舆情研究中多变量之间的关系和差异是社会实证研究的核心问题，因此必须引入人工分析处理机制，以寻找更为深层次的解释。本研究主要从差异和关系两方面为挖掘重点。差异可使用交叉和聚类的方式获得，如交叉可了解议题之间的态度差异、强度差异，聚类能够帮助细分利益相关者等。对关系方面的挖掘，以相关关系、因果关系和预测及解释力为主。

1) 内容分析法

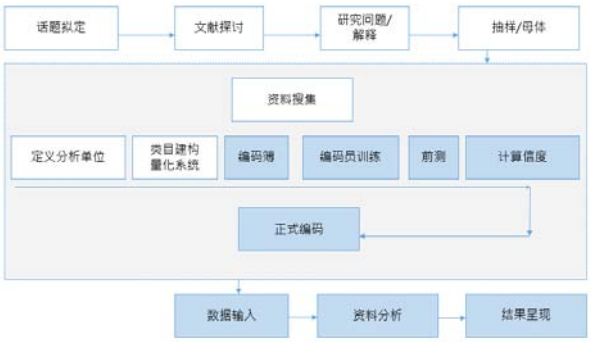
本研究中人工在线实时内容分析以传统内容分析法（Content Analysis）为研究方法理论依据，在此基础上改进操作，以改善运作效率、提高数据及编码的质量、产出更有信度和效度的分析结果。

内容分析法是社会科学研究方法中的一种对文本内容进行编码、分类、语义判断及形成可供统计分析之用的量化分析方法。它是指一种以系统、客观与量化的方式，来研究与分析传播内容，藉以测量及解读内容的研究方法（Kerlinger，1973）。在方法设计和执行方面，强调：（1）系统的方法，需采用随机样本、系统的类目建构和编码程序；（2）客观的程序，

需遵守明确的标准与规则；（3）量化的分析，需为所有的变量下操作性定义，确定测量标尺，进行统计分析。

传统的内容分析流程包含多个程序，设计以保证理论和操作的合规和准确性。整个流程以话题（研究题目）拟定为始，进而进行文献探讨，以确定研究问题及解释；在对分析对象范畴的确定时，可考虑对母体进行分析，或者采取抽样的方式，确定研究对象；通过资料搜集建立样本集；定义分析单位后，建构类目量化系统，制作编码簿，在正式编码之前，对编码员进行训练，进行前测编码，计算信度，当编码员间信度达至一定水平时，可开始正式编码，此后输入数据，分析资料，最终获得结果呈现。

一个传统的内容分析流程图



内容分析可以支持多种资料类型作为研究范畴，如采访稿、焦点小组结果、教材、新闻、论文、杂志、文章、政治演讲、小说、广告、社交媒体内容等，呈现的格式包括文字、图片、音频、视频等。内容分析方法可灵活应用于多种研究目的及不同领域，其中最为知名和经典的案例之一为 Harold Lasswell 在一次世界大战中的研究。Lasswell 在其著作《世界大战中的宣传技巧》（*Propaganda Technique in World War I*）中以宣传信息所使用的符号为分析对象，包括报纸、宣传手册、传单、书籍、海报、电影、图片等，发展出内容分析法以研究宣传运动中的技巧。还有其他的研究领域包括有研究者利用该方法确定文章作者的著作权的比例，例如 Mosteller 和 Wallace（1963）采用基于词频的贝叶斯技术，解决了《联邦主义者》（*The Federalist*）文章中的原作者的分布问题。商业领域中，有研究者使用内容分析法评估食品行业的发展趋势，例如，1998 年有一项研究钙摄入和减肥之间的关系，研究范围是青少年和女性杂志上的广告、文章和专栏内容（Kondracki, Wellman, Amundson , 2002）。社会服务方面，美国农业部森林服务局（United States Department of Agriculture Forest Service）利用内容分析法监测社会环境对国家森林管理措施的评价意见（West , 2001）。

2) 编码员之间信度（Inter-coder Reliability）

在内容分析中，需要多于一个的编码员来进行编码工作，这些独立的编码员对一段信息 / 记录内容的特征（也就是记录单位）作出判断，并且达成一致的结论。这种一致性以量化方式呈现，称之为编码者间的信度。不同的编码员应该对每一个分析的对象给予相同的评分（对等距或者等比标尺而言，即使不是完全相同的数值，也应该是相近的值），这种实质的同意程度是检验“编码者间的信度”的基础（Tinsley & Weiss, 2000）。

通常我们研究的信息有明显的内容（manifest content）和隐藏的内容（latent content）。对于明显的内容，例如文章字数、消息来源、人物或单位名称等，很容易以客观的判断来达成高度一致性。但是，对于隐藏的内容来说，例如报导态度或者价值观，编码员必须根据他们自己的思维系统作出主观的诠释。这样的话，编码员之间的相互主观判断变得更加重要，因为当这些主观判断由所有编码员共享的时候，也就是它们更有可能让读者产生相同的意义（Potter and Levine-Donnerstein,1999）。

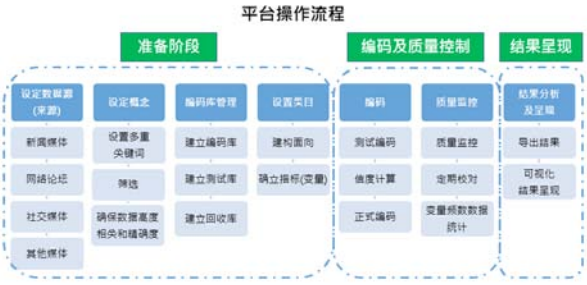
编码员间信度评估流程由编码指引开始，需要依据编码簿制作编码指引，帮助编码员准确理解编码类目，帮助编码员熟悉议题，理解编码类目；之后选取少量样本，各编码员需要独立进行编码，不可相互讨论或指导，计算信度系数以观察不同编码员是否已经达到可接

受的认知一致性水平，如未能达到理想的信度水平，则需要对编码员再次进行培训和指导，以确保编码员达到理想信度水平，可开始正式编码。学术上常用的编码员间信度有 Holsti 的信度系数（Holsti's Coefficient Reliability）及 Krippendorff 的 alpha 值（Lombard, Snyder-Duch 和 Bracken, 2002）。

3) 人工在线实时内容分析流程

在参考传统内容分析法的理论和操作方法基础上，本研究建立了由大数据技术辅助人工在线实时内容分析机制及平台—博易数据挖掘平台（DataMiner），整个流程包括准备阶段、编码及质量控制和结果呈现三大部分。

在完成前期文献搜索、确定研究问题等预备工作，可于平台上进行准备阶段的设定数据源、通过设置多重关键词以设定概念，在该过程中，可通过筛选工作以确保数据高度相关和精确度，并且完成编码库管理和设置类目的工作；进入编码及质量控制阶段，该部分尤为重要，正式编码前需要进行前测编码，以确保编码员间信度达到可接受的理想水平，在正式编码过程中，透过平台随时监控编码结果，并可定期校对以保证编码质量；完成上述过程后，可对结果进行分析和可视化呈现。



该操作流程设置多种质量保证机制，可随时监管编码员效率、编码准确度，以确保最终的工作结果可真正为舆情研究提供价值。

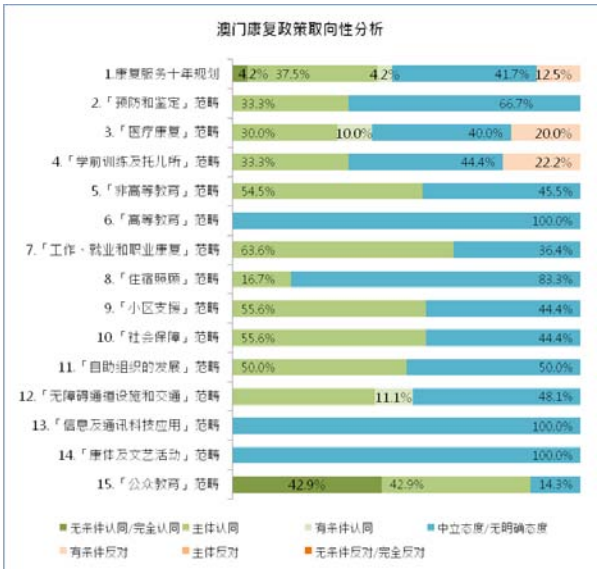
下图为线上内容分析机制页面，支持即时编码、即时检验、即时监控和即时结果。

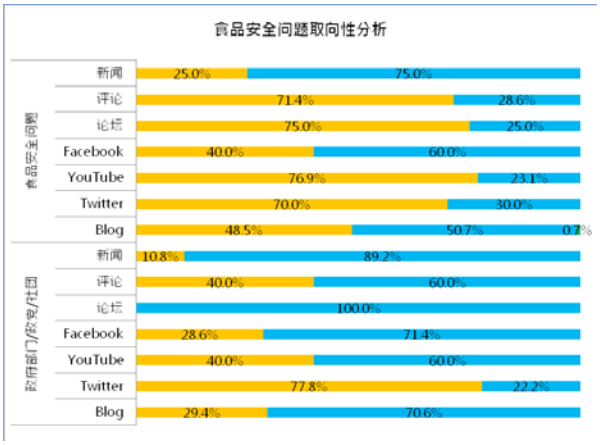
以下分别以发生在澳门的两个案例来进一步说明通过这种方法可以做到的分析结果。

例 3：“澳门康复政策”的网民态度分析

在某些舆情事件中，涉及的话题面向多向且复杂，需要人工处理和区分，在此基础上，才能得以进一步解析细分议题之间的态度差异及强度差异。

下图着重于对澳门的一项康复政策不同范畴的态度差异的解读。康复政策为总体政策类型，下属多个细分政策范畴，情况较为复杂，必须使用人工判断的方式予以分类和归整。结果显示，针对康复政策，除整体性的“康复服务十年规划”，其余区分分类共 14 个细分范畴，进而需要判断这 14 个范畴的态度如何。对态度的测量以七个层次划分观察网络舆情，区分为是无条件认同 / 完全认同、主体认同、有条件认同、中立态度 / 无明确态度、有条件反对、主体反对和无条件反对 / 完全反对。观察分析结果（模拟数据），以对“公众教育”范畴的认同程度较高，有 42.9% 为“无条件认同 / 完全认同”，42.9% 为“主体认同”；观察另一个方向的认同程度，以“学前训练及托儿所”和“医疗康复”两个方面的反对态度最为明显，分别有 22.2% 和 20.0% 表示了“有条件反对”态度。





例 4：“ 食品安全问题 ”的网民意向分析

另一个案例是有关食品安全的舆情分析，探讨关于不同利益相关者对于食品安全所持的立场差异。下图为关于食品安全问题的分析结果，采用机器学习和人工编码结合的方式得出。观察不同媒体来源中对食品安全问题的整体立场，可以看到新闻评论、论坛、YouTube 和 Twitter 上反对的声音较多（71.4%、75.0%、76.9%、70.0%)。不同媒体来源对于不同利益相关者（包括政府单位、政党和社团）的态度存在较明显的分布差异，以 Twitter 上的反对声音最多（77.8%），在论坛和新闻中表达出来的舆论声音以中立为主（100%，89.2%）。该案例说明，不同媒体渠道上所收集到的舆论声音可能存在差异，对事件的看法和立场会不一样。如单独使用网络挖掘，仅仅能看到整体的声量，无法解读到不同层次的内容，因此可见，仅仅看传播量等内容是远不足够的。

3. 大数据技术与人工在线实时内容分析的互动和促进关系

在本研究的网络舆情研究新机制中，以人机结合为核心理念，大数据技术框架下的网络挖掘和机器学习可执行广度的自动化分析和快速挖掘舆情信息，人工在线内容分析则能完成深度挖掘和解释舆情间差异和关系的任务。从舆情分析和解读的整体角度出发，以网络挖掘及机器学习为代表的大数据技术与人工在线内容分析两个体系是相辅相成的关系，构成一个良性循环，且存在彼此优化的特点，具体体现在三个方面：

（1）机器技术改善人工编码流程。通过网络和计算机技术的辅助，提供编码文本关键词高亮设置，相似主题文本优先派发等算法支持，得以提高人工编码效率；另一方面，平台提供快速简单的前测编码和即时质量监督功能，解决了传统内容分析中编码质量难以控制和校正延迟的难题。

（2）词云图帮助制作编码类目。利用词云分析技术，能够快速发现和掌握舆情事件主要面向，从某种程度上，以关键词的形式，表达了相关内容的热度情况。在传统的人工内容分析法中，制作编码类目前需大量检视相关内容文本，以获得对研究问题的大致了解。有词云的帮助，其快速挖掘的关键词能提供编码类目的线索，大大改善制作编码类目的效率及效度。

（3）人工内容分析结果帮助改良机器学习的准确度。经过信度检验的人工内容分析所累积的大量人工编码结果，可以作为优质的机器学习的训练集，用于改善自动化分析结果，如情感分析，以此方式不断提升机器分析的准确度。

五、结语

回顾当前网络舆情研究发展现况，网络舆情监测和分析的工作难度大，面对复杂的舆论场景，单靠机器或人工方式无法解答我们的研究问题及现实需求。为此，本研究提出新的舆情研究路径——大数据技术辅助网络内容挖掘与分析，并通过博易数据技术公司的“博易数据挖掘平台 -DataMinder”来实现。该路径综合汇总多年舆情研究经验，以社会科学实证研究的核心要素——覆盖度、测量和解释作为网络舆情挖掘的理论框架，配合人工在线实时内容分析方法，探索舆情事件变量间的差异与关系。其中所建立的分析机制及流程，乃将研究视角落实至研究舆情的本质，以回应学术界、政府和业界期望了解舆情的真正意涵及价值。

本研究着重于提供一种舆情研究的思路与方法论，不限于特定舆情研究的目的和用途，适用于实务应用，亦可用于学术研究；可用来掌握舆论形势，又可用于深度挖掘某一个话题，以解决实际问题为目的。

以上作为网络舆情分析路径上的初步尝试，乃经过一段时间的实践，并已取得一定的成效。然而作为新的探索，需要持续优化，尤其是理论上需要强化和补充，在实践上需要改善和提升。其中需要考虑是否能够应对各种舆情研究类型和情况，例如，当需要处理的数据量特别大的时候，运用人工内容分析时人力部分的压力过大，时效性会大打折扣，可考虑按照一定规则抽样处理，如对文本内容采用系统抽样或分层随机抽样方式，形成可供操作的编码样本库，这也是笔者提出作为未来研究和探讨的一个方向。

另一个值得关注的方向是，将质化与量化结果的相互结合解读的研究方法论。在对舆情的研究实践中，网络挖掘和机器学习是研究舆情的第一步，可快速获得初步的量化结果；第二步是使用人工编码和分析将文本内容做量化处理，即是质化文本材料的量化过程；第三步是量化和质化内容的相互补充，即是以原文文本补充和解读量化结果。以此完成由质化内容得出量化结果，再次回到质化内容，量化结果与文本之间相互补充和解释的循环方法论，未来或可进一步实践和探索该方法对舆情或其他类型研究的解释度和操作性。

参考文献：

[1] 于新扬 . 中国网络舆情监测发展现状及不足。《传媒观察》，2015（1），8-9 页。

[2] 上海交通大学舆情研究实验室 . 大数据与社会舆情研究综述。《新媒体与社会》，第十一辑。

[3] 中国互联网络信息中心 .《中国互联网络发展状况统计报 告（2016 年 7 月 ）》。http://www.cnnic.cn/gywm/xwzx/rdxw/2016/201608/W020160803204144417902.pdf

[4] 尹培培 . 大数据时代的网络舆情分析系统。《广播与电视技术》，2013（07）。

[5] 闵大洪 . 闵大洪：对中国网络舆情监测工作的观察与思考。《网络空间研究学刊》，2016 年 10 月 16 日。

[6] 林源 . 网络舆情研究综述。《科技情报开发与经济》，第 25 卷，146-150 页。

[7] 祝建华 . 一个文科教授眼中的大数据。中关村大数据日，2012 年 12 月 13 日北京。

[8] 胥琳佳 . 大数据对于传播学研究内容和方法的影响——基于社交媒体和移动互联网的思考。中国出版，2013（18）。

[9] 喻国明 .《中国社会舆情年度报告》。人民日报出版社。

[10] 谢耘耕主编.《中国社会舆情与危机管理报告》。社科文献出版社。

[11] 蔡立辉和杨欣翥 . 大数据在社会舆情监测与决策制定中的应用研究。《行政论坛》，第 128 期，1-10 页。

[12] 燕道成和姜超 . 大数据时代网络舆情研究综述。《视听》，2015（9），133-146 页。

[13] Alpaydin, E. (2004). Introduction to Machine Learning; MIT Press: Cambridge, MA, USA, 2004.

[14] Herrouz, A, Khentout, C, & Djoudi, M. (2013). Overview of Web Content Mining Tools. The International Journal of Engineering and Science. 2(6).

[15] Kerlinger, F.N. (1973). Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart & Winston.

[16] Kondracki, N. L., Wellman, N. S., Amundson, D.R. (2002). Content Analysis: Review of Methods and Their Applications in Nutrition Education, 2002(34), 224-230.

[17] Langley, P. (1996). Elements of Machine Learning. Morgan Kaufmann, San Francisco, CA.
[18] Lasswell. H. D.. (1971). Propaganda Technique in World War I. Mit Press.
[19] Lazer, D., Kennedy, R., King, G., Vespignani, A.. The Parable of Google Flu: Traps in Big Data Analysis Science 14 March 2014: Vol. 343 no. 6176 pp. 1203-1205.
[20] Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. Human Communication Research, 28, 587-604.
[21] Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. Journal of the American Statistical Association, 58(302), 275-309.
[22] Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. Journal of Applied Communication Research, 27(3), 258.
[23] Tinsley, H. E. A. & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley &S. D. Brown, Eds., Handbook of Applied Multivariate Statistics and Mathematical Modeling, pp. 95-124. San Diego, CA: Academic Press.
[24] West, M. D. (2001). Applications of Computer Content Analysis. Ablex Publishing Corporation.

New approaches to online public opinion research: Online content mining and analysis using big data technology

ZHANG Rong-xian CAO Wen-yuan

Abstract: Previous domestic research shows that public opinion mainly covers the social events and changes in society including the deductive factors triggering these events and public opinions related. Nowadays, most of online public opinion research lacks the understanding of the nature of public opinion and the systematically analytical framework is rarely adopted. Descriptive statistics are widely used to draw conclusion. Considering the above limitations of the current public opinion research, this paper presents a novel approach for online public opinion research which takes three major elements into accounts: coverage, measurement and explanation and is achieved by the combination of an online big data analytics and human judgment methodology. It first gives the overview of public opinion with the help of the machine learning and web mining technology built on the platform; then it mines deeply and explains events via a manual online content analysis method. Some cases will be elaborated in this paper to show the practicability and operability of this approach.

Keywords: online public opinion; big data technology; web mining; machine learning; content analysis

(责任编辑：李晓晖)