中山大学资讯管理丛书
中山大学资讯管理系创办30周年
暨 资讯管理学院成立 纪念特辑

ZIXUNGUANLIYANJIUJINZHAN

资讯管理研究进展

■ 曹树金 主编

中山大学出版社

中山大学资讯管理丛书
中山大学资讯管理系创办30周年
暨 资讯管理学院成立 纪念特辑

ZIXUNGUANLIYANJIN

资讯管理研究进展

曹树金 主编

中山大学出版社
·广州·

图书在版编目（CIP）数据

资讯管理研究进展/曹树金主编. —广州：中山大学出版社，2010.12
ISBN 978-7-306-03800-5

Ⅰ. 资…　Ⅱ. 曹…　Ⅲ. 信息管理—文集　Ⅳ. G203-53

中国版本图书馆 CIP 数据核字（2010）第 233452 号

出 版 人：祁　军
策划编辑：王俊辉
责任编辑：王　辉
封面设计：贾　萌
责任校对：杨文泉
责任技编：何雅涛
出版发行：中山大学出版社
电　　话：编辑部 020-84111996，84113349，84111997，84110779
　　发行部 020-84111998，84111981，84111160
地　　址：广州市新港西路 135 号
邮　　编：510275　　传　真：020-84036565
网　　址：http://www.zsup.com.cn　　E-mail:zdcbs@mail.sysu.edu.cn
印 刷 者：广州中大印刷有限公司
规　　格：787 mm×960 mm　1/16　23.5 印张　340 千字
版次印次：2010 年 12 月第 1 版　2010 年 12 月第 1 次印刷
印　　数：1~1000 册
定　　价：38.00 元

总　　序

我们的社会不仅进入了资讯时代，而且还迎来了数字狂潮。随着资讯效用的日益显著和资讯技术的广泛渗透，一方面，资讯管理已经成为各行各业都要完成的关键任务；另一方面，资讯管理的困难性和复杂性在不断上升。这样的环境对资讯管理学科和资讯管理人才的要求，既需要我们努力探明，更需要我们为之奋斗。

我国政府已经将信息资源的开发利用作为国家信息化的核心内容，资讯管理的核心任务就是开发利用信息资源，从而提高个人、组织、国家的竞争力。近些年来，资讯技术的迅猛发展正不可思议地改变着人们的学习、工作和生活方式，给资讯管理职业带来了严峻的挑战和前所未有的机遇。同时，建设和谐社会、贯彻科学发展观要求我们以资讯技术为手段，在资讯管理中落实以人为本，实现人文关怀。这就意味着资讯管理是一个跨度很大的领域，涉及技术、社会、人文的许多方面。

中山大学资讯管理系是我国培养资讯管理人才的重地之一，28 年来，先后培养了大量的图书馆学、档案学、信息管理与信息系统、情报学等专业人才。2005 年新组建为学校直属系，建立了涵盖图书馆学、情报学、信息管理与信息系统、档案学、文献学的资讯管理学科群研究和教学队伍，形成了从本科生到博士研究生的资讯管理人才培养体系。目前，每年培养本科生 260 多名，硕士和博士研究生 70 多名。

工欲善其事，必先利其器。为了提高资讯管理人才的培养质量，我们鼓励教师既对资讯管理类学科的教育规律和教学方法进行深入的探讨和实践，又对资讯管理领域的学术问题尤其是前沿课题开展细致的研究。基于此，我们决定编辑出版“中山大学资讯管理丛书”，比较集中地展示我们教师的研究成果，求教于广大同仁。

中山大学资讯管理系主任　曹树金

目　录

网络舆情信息监测研究进展

曹树金　陈少驰　陈珏静

【摘　要】随着互联网的快速发展，网络信息传播的迅捷性和交互性日益彰显，网络舆情已经进入主流社会视野，成为影响和决定社会生态的重要力量。论文对当前网络舆情监测的研究热点和研究方向进行了分析，总结了近几年来国内外关于网络舆情监测在理论、技术、应用产品三方面的研究情况，并在此基础上指出存在的不足和未来发展方向。

【关键词】网络舆情；舆情监测；研究进展

一、导　言

社会舆情与其发展历来是政府关注的重点。信息技术的发展和网络用户的不断增多，舆情逐渐网络化，网络舆情演变成为社会舆情的重要组成部分，并逐渐引起了政府与学术界的关注。2008 年 6 月 20 日，胡锦涛总书记在人民日报社考察工作时指出："互联网已经成为思想文化信息的集散地和社会舆论的放大器，我们要充分认识以互联网为代表的新兴媒体的社会影响力"。作为"观点的集散地"，"民生的集散地"，网络舆论已经进入中国的主流社会视野[1]。网络舆论是公开民意和意见的表达，侧重信息传播；网络舆情则表达了民众对社会事件产生的社会性政治态度，侧重情感表达。网络舆情是网络舆论的基础，网络舆论往往是先从个体的网络舆情演变而来。要更好地把握舆论，就必须对舆情开展研究[2]。针对网络舆情的研究虽然起始不久，但近几年呈现日益增多的趋势。在 CNKI 论文库查询显示，国内第一篇有关网络舆情的论文诞生于 2005 年 2 月的《新闻记者》。经过近几年的研究，网络舆情在国内的研究主要体现在两方面：一是理论方面的研

究，主要包括政策和机制研究，用户行为特征的研究，网络舆情传播模式的研究，其要点是试图通过政策管理和约束对网络舆情进行管理；另一个方面主要是网络舆情的监控和预警系统的技术研究，包括算法、模型及关键技术的研究。目前，类似的网络舆情预警系统或产品已经出现，但数量并不多，有商业软件公司开发的系统，如方正的舆情预警系统，卡塔尔的 Cymfony 网络舆情系统，也有科研机构、学校与相关政府部门及单位（如公安部门）合作开发的系统，如深圳公安局和北京交通大学合作开发的网络舆情信息挖掘系统。但是，这些系统由于单独开发或面向的对象特殊等原因，都存在比较大的局限性。总体而言，网络舆情挖掘与预警的研究还处于初级阶段，缺乏全面、成熟的研究，也没有功能全面实用性较强的系统。

本文对近几年国内外网络舆情监测的研究方向和研究热点进行初步的归纳和分析，主要从以下三个方面展开：①网络舆情相关理论；②网络舆情监测主要相关技术；③国内外已有网络舆情监测系统。

二、网络舆情相关的理论研究

（一）网络舆论的要素、特征研究

随着网络的迅猛发展和上网人数的不断增加，网络逐渐成为人们获取信息的主要渠道之一，网络舆论和传统的社会舆论一样融入主流视野并彰显了强大影响力。网络舆情是由于各种事件的刺激而产生的，通过互联网传播的人们对于该事件的所有认知、态度、情感和行为倾向的集合[3]。从社会学的角度辨析可得出网络舆情的双重含义：一方面是大众意见的高度共识，另一方面又是非理性的意见妥协——源自离散的个体，没有经历沟通过程，缺乏政府引导和调查论证[4]。

目前，绝大多数有关网络舆论的研究主要是从社会学的角度讨论网络舆论与社会舆论之间关系，网络舆情的要素及特征，网络舆情相关政策和机制等等。例如，截至 2010 年 9 月，有关网络舆情特点的研究论文就有 13 篇，《2007 年 BBS 跟帖凸显六大网络舆情特点》、《略论网络舆情的概念、特点、表达与传播》、《“两会”期间网络舆情呈现特点分析》、《网民的网络舆情主体特征研究》和《网络舆情的基本特点》等。

也有国外学者如 Wojcieszak 不是从理论分析角度，而是采取抽样调查网

民行为的方法，来分析网络舆情中个体和群体的相互影响关系[5]。相对应的，国内学者如曹效阳等则从网络舆情的结构及其网络特征来量化网络舆情，其研究成果对网络舆情监测的总体构思、技术线路产生了影响[6]。

一般来说，网络舆论和传统的社会舆论一样也包含以下7个要素：舆论的主体，舆论的客体，舆论的自身，舆论的数量，舆论的强烈程度，舆论的持续时间及舆论的功能表现。网络舆论却又不同于社会舆论，它的个性特征主要有：丰富性、多元性、透明性、盲目性、离散型和聚合性、愤青化等[7]。另外，曹劲松认为，网络舆情的基本特征主要有：传播爆炸性、主体隐蔽性、信源模糊性、网民动员性、意向指向性、影响显著性[8]。基于社会科学的角度探讨网络舆情，虽然不能直接对网络舆情系统的设计和实现产生作用，但是确实可以为认识网络舆情提供理论依据。这是实现网络舆情挖掘和监测的根本——只有将这些特征数量化才有意义，才能被计算机所识别和处理。

（二）网络舆情模式分析

对网络舆情进行科学解疑释惑的关键是要有科学的分析工具、模式和判别依据。在网络环境下，海量信息呈现离散分布，依靠先进的信息技术对网络舆情进行定量研究以实现自动监测逐渐成为发展趋势。但是，目前对网络舆情模式的研究是少之又少。

从定量和数理统计的角度研究网络舆情模式的很少，到目前为止只有上海交大的学者曾从统计学的角度对网络舆情内容及深度模式进行探讨，构建了互联网内容与舆情的热点（热度）、重点（重度）、焦点（焦度）、敏点（敏度）、频点（频度）、拐点（拐度）、难点（难度）、疑点（疑度）、粘点（粘度）和散点（散度）等10个分析模式[9]。

近年来对网络舆情的模式研究的方法，开始从单一的元素因子分析向网络舆情模型构建过渡。中国科学院的方薇等人，采用元胞自动机理论为基础，构建了网络舆情元胞自动机传播模型[10]。中国科学院计算研究所的戴媛等人，把系统的多元分析和模糊数学相结合，也在网络舆情安全评估模型上有所创新[11]。国外的Craemer分析了舆论观点的形成机理，进而提出一个无参数变量模型，来描述网络舆情和公众文化的演变[12]。从学科发展和进步的角度考量，该类研究动态地透视互联网内容及舆情形成和发展的基本特点，虽然这些分析模式并没有得到大量样本的有力论证，但是，它提供了一条很有价值的研究思路。它利用数值变化描绘网络舆情从产生到灭亡的整个生命周期的特征变化，量化定性指标，有助于采用数学方法（如函数）

对网络舆情进行数量上统计并预警，这是网络舆情实现自动化监测和预警的必要前提。

网络舆情模式分析从定性研究转向定量研究的过渡，是实现网络舆情监测与预警的一个很重要的转变，仍需要在此基础之上进行进一步研究。

（三）网络舆情新兴媒介和主体的研究

网络舆情的新兴媒介层出不穷，如 BLOG，FACEBOOK，TWITTER 和 YOUTUBE 等新媒体，在互联网上扮演着信息的时空穿梭机，也给广大网民提供了一个更多元化的平台。针对此类的研究也有不少，如 Greysen 论述了新信息环境下网络舆情媒介的三大挑战[13]。Ashlin 则在调查的基础上，得出“WEB－BLOG 是一个伴随社会政策和社会交流产生的全球化现象”的结论[14]。相关的研究都从社会学和行为学的角度进行探讨，围绕信息环境和公共政策展开，对现实有一定的指导意义。但是不管如何变迁，网络舆情媒介仍然不能逃离“媒介”的本质属性，因此将“新”媒介与传统媒介进行比较研究，从新闻和媒体的视角来研究网络新媒介，也就更加迫切和有意义。

网络舆情的主体是网民，他们经常性地以互联网作为传播和交流媒介，通过上网获取信息并参与网络互动，通过发表个人见解表达情绪和态度[15]。对网络舆情主体的研究是做好网络舆情监测的基础，只有充分了解其主体特征才能对症下药，即所谓的“知己知彼百战不殆”。可是目前对网络舆情主体的研究是一个空缺，国内外与这方面相关的论文极其少。根据中国互联网络信息中心（CNNIC）发布的第 26 次《中国互联网络发展状况统计报告》显示，到 2010 年 6 月底，中国网民规模达到 4.2 亿，宽带用户达到 3.6381 亿户，占总网民数的 86.62%。使用手机上网的用户达到 2.77 亿[16]。

要对如此庞大的群体的舆情进行监测，首先要了解他们的特征。一般来说，作为新兴媒体网络的使用者的网民都是有一定文化背景的，而且比较年轻化，他们有主见又敢于表态，更有些网民思想偏激。但是，他们中往往有一个“意见领袖”。所谓的“意见领袖”是指在信息传递和人际互动过程中少数具有影响力、活动力，既非选举产生又无名号的人[17]。他们的观点往往能够左右网民的判断并最终引导网络舆论的走向。网络舆情监测可以从这些“意见领袖”下手，密切监测他们的言行，并用恰当的方式与网络“意见领袖”进行沟通，引导他们发表有利于事态良性发展的舆论，他们的拥护与追随者也会继续跟进，进一步推动舆论的良性发展。这样通过网民引导网民，用网民自己的声音引导、感染网民，实现网民自我教育、自我引导，

往往能够达到事半功倍的效果。因此，对网络舆情主体的研究不容忽视。

（四）网络舆情监测在各行业中的应用研究

在互联网影响力日益增大的今天，各级党政机关、企事业单位和学术、媒体机构等都越来越重视互联网舆情的监测、预警和引导。综观最近五年的研究情况，后两年的研究主要针对各行业对网络舆情的应用。主要有：

◆各级政府运用网络舆情监测加强网络舆情公关

网络的便捷性、隐蔽性等特征使它成为社会大众参政议政和管理国家事务的主要阵地。网民对官员腐败、渎职等廉政建设问题的揭露有利于政府拨乱反正；同时，网上也时常出现有损政府形象的负面信息，这些信息通过网络渲染、传播，极易造成大范围乃至全国的动乱。有些网络谣言与虚假报道采取隐蔽、专业化的方式传播，极具迷惑性，混淆视听，滋生事端，往往造成群众的盲从与冲动[18]。对于网络舆情突发事件的应对研究是该领域的热点和重点。各种针对网络舆情突发事件的案例分析屡见不鲜，尤其是为政府如何更好地利用和引导网络舆情的参考咨询类研究占据了绝大部分。这方面的研究主要集中在宏观方面，如上海交通大学的《网络舆情对政府形象的影响及应对策略研究》，中国矿业大学的《论网络舆情对领导决策的挑战》，北京邮电大学的《非常规突发事件中网络舆情的作用分析》，等等。对于网络舆情的有力有效引导，越来越成为公共管理部门的新兴职能，也成为学者们关注的热点。

◆公检机关通过网络舆情监测提高执法公信度

涉及公检机关的职权范围内的敏感事件一直是网络的热点，我们可以看到2007年“跳楼砸死人”事件到现在网上还时有评论，“这属于故意杀人罪，要判死刑”、“这是意外伤害罪，要负一定的民事责任”、“被砸死的人也应该负一定责任，因为他未经他人同意，妨碍了他人的生死抉择权”等等，还有认为梁丽无罪，胡斌应重判，邓玉娇属正当防卫，罗彩霞的受教育权应该得到维护的帖子随处可见，之后还有汹涌般的跟帖，如此泛滥的舆论形成了大型的“网络审判庭”，使公检机关的言行被淹没在噪音中，严重削减了公检机关的执法公信度。公检机关不能坐以待毙，他们应该联手组建舆情部门，专门负责对涉及本机关的网络舆情进行日常监测和突发事件监测，通过实时巡查的方式随时掌握舆情动态，在发现不良舆论苗头时，即时召开新闻发布会或者在主流媒体上公布案件的真相、处理过程及结果。对于网民和媒体的质疑或者诉求，应该坚持全面、平衡、客观的立场，不掺杂任何意见或偏见，不带任何主观倾向，避免道听途说，偏听偏信[19]。在媒体和公

众面前塑造一个高效而负责任的形象。

◆高校重视网络舆情监测完善学生管理

大学生作为校园舆情的主体由于自身年龄层次、心理、思维特征以及他们在社会中所处的特殊地位，使他们在意见、态度、情绪表达上呈现出特有的方式[20]。他们发表的言论往往集中在两个方面，一是校内涉及大学生学习、生活利益，且受到大学生群体关注的事件，如学校的某项规章制度；二是发生在校外，且容易引起大学生群体刺激的各类社会政治、经济、道德伦理事件，如网络上传得满天飞的“躲猫猫”、“范跑跑”、“家乐福”等[21]。他们的网络舆论阵地主要是校园 BBS、留言板、博客、QQ 群等，这就要发挥学生会代表、教授、导师、班主任、辅导员等舆情监测、引导的角色作用。他们都可以以普通网民的身份进入到舆论空间，参与大学生的舆情过程，如果学生发表的舆论是建议性的，可以向相关部门传达意见，尽可能满足合理建议，对不合理建议则应该以学生比较可以接受的方式说明原因，避免他们因要求得不到满足而制造恶意舆论。如果舆论是对学校或师生不满的发泄性的，引导者应该直接与学生对话，安抚情绪，抑制不良舆论爆炸。当学生在舆论空间上讨论校内校外热点问题时，相关人员可以参与讨论，了解学生的思维，公开事实真相，争做“意见领袖”，引导学生往良性方向讨论。这对于完善高校管理意义重大，应该引起相关部门、人员的重视。

◆主流媒体借助网络舆情监测提高影响力

调查发现，由于主流媒体本身所具有的公信力、权威性和可靠性，网民通过网络获得信息时，尤其是出现突发事件的时候，网络上的观点多如繁星，令不知真相的一般受众十分迷茫，他们往往会在主流媒体上求证信息的真实性。当他们在众多的热点问题言论中不知拥护哪一个派别时，也常常在主流媒体上寻求答案。主流媒体就应该充分利用好这份优势，想方设法地监测网络动态，把国内外重要新闻事件、网络舆情热点整合到自己的领域范围内，开辟网络评论专栏，并培养一支专职撰写网络评论的“写手”队伍[22]，与网民倾情互动，解答网民的各种诉求，并对舆论进行正确引导，形成正面舆论的强势，树立网络舆论的中心地位，提高自己在网络中的影响力。

◆企业利用网络舆情监测化解经营危机

在个人媒体崛起的网络时代，网民以个人为中心，网络话语权平等，客户可以因一次不愉快的产品体验或者没有接收到预期的服务而在言论自由的网络上抱怨，甚至是竞争对手散布恶意谣言诋毁中伤。要是对这些负面信息处理不及时就可能酿成不可挽回的恶性影响，危机就这样爆发了：轻则赶跑

企业的潜在客户、降低产品品牌价值，重则导致企业被逐出市场。因此，企业开始关注并建立起网络舆情监测机制，对主流搜索引擎中与企业相关的关键词，一些知名度高、与企业相关性高的社区、门户网站、BBS 和近期网络中的热点问题及其帖子、回帖密切跟踪监控[23]，如发现相关的负面信息要即时采取有效措施将其扼杀在萌芽阶段进行控制，如发现对企业有利的信息则因势利导，借助舆论的力量提升企业的公共信誉和品牌价值。

此外，还有对涉农网络舆情的研究，如《涉农网络舆情走势与应对》，以 2008 年发生在四川广元旺苍县的“蛆柑”事件为例，揭示了涉农网络舆情对经济社会发展的影响力，呼吁政府应该积极应对，维护占我国绝大部分人口的农民群体的利益[24]。《对做好铁路网络舆情信息工作的几点认识》在大量事实的基础上，总结出铁路网络舆情的特点，并提出了正确研判和处理铁路网络舆情的方法和进一步做好铁路舆情信息工作的有关建议[25]。《兰州石化公司网络舆情监控系统的设计与实现》以兰州石化公司为代表，指出该公司网络发展现状及存在问题，在此基础上探讨了舆情监控技术、系统设计及其实现情况，为兰州石化公司网络信息监管提供辅助决策[26]。在严峻的医患关系大环境下，医疗纠纷小则增加医患矛盾，大则影响医院的声誉。浙江省临海市第一人民医院的王超超等发表的《医疗机构网络舆情的应对策略》为医疗机构提供了突发事件的预防、舆论指导及应对等决策，帮助医疗机构化解舆论危机[27]。但是，目前对这几类机构的网络舆情研究严重不足，而这几类机构都与我们生活息息相关，他们的健康发展对国家和人民意义非同小可，这需要引起我们的广泛关注以及进一步研究。

（五）网络舆情热点研究

自从网络舆情这一概念提出以来，社会各界对网络舆情热点的研究长盛不衰，规模比较大的成果是 2009 年 3 月，由中国人民大学舆论研究所、中央电视台《对话》栏目、京报集团、《北京晚报》共同推出的“舆情分析报告”，该报告对 2009 年 1～2 月份网络舆情进行统计，提出网络舆情热点事件主要集中以下八大方面：①个别政府官员的违法乱纪行为；②涉及司法系统、城管队伍等；③涉及部分政府部门、央企；④衣食住行等全国性民生问题（如房价激涨）；⑤涉及社会分配、贫富分化；⑥涉及国家利益、民族自豪感（如家乐福事件、台独、藏独事件）；⑦重要或敏感的国家地区突发性事件（如汶川大地震等）；⑧影响力较大的热点明星的火爆事件（如“艳照门”、“快男超女”）[28]。作者认为，学校是绝大部分网民的聚集地，发生在

学校的敏感事件渐渐地也成了舆论的热点，如“马加爵”事件、广州大学城逾百人中毒事件等，都广泛地引起了社会的关注和评论。还有对社会弱势群体的声援也越来越高涨，更多的人肯站出来为弱势群体们发言，尤其是关乎尊老爱幼优良传统的反叛事件，人人闻而觑之，舆论铺天盖地。以此看来，网络舆情热点涉及社会的方方面面，如不加以监测、预警和引导控制，其后果不堪设想。网络舆情监测关注的不应该仅仅是社会生态的某一个方面，而是应该广泛搜罗各种信息，及时予以应对。

（六）网络舆情研判研究

在当今时代，我们可以强烈地感受到，网络舆论正以惊人的速度渗透到社会生活的方方面面，网络舆情监测工作因此变得格外艰难。全面监控所有网上发布的信息是不太可能的，既浪费大量人力物力，又会“捡了芝麻，丢了西瓜”，失去监控的重点。这就需要监测与研判互动，各部门监测人员集中精力研究与自身密切相关的舆论并进行跟踪。同时准确判断度量非常恶性、恶性、中性、良性舆论，决定哪些先处理，哪些后处理，哪些不用处理，哪些采取遏制措施，哪些采用引导的方式，哪些应该积极鼓励，以做到“逐个击破”。研判的方法则有很多，如：系统研究方法，网络舆情往往同时涉及好几个领域内的问题，舆情监控者不能站在个人立场偏颇某一个领域，而是要站在社会角度全面系统地判断。内容分析方法，也可以说是定性的方法，这是最一般的舆情研判方法，通过舆论的内容了解事件严重性。定量分析方法，它是定性分析法的补充，两者常常交叉使用。通过统计各种性质言论的数量，判断大众对这件事的关注度和响应倾向，为下一步决策提供辅助。对比分析法，这是用得比较广泛效果也不错的方法，对某些复杂的舆情，一时间无法正确判断或者不适合采用单一的方法，就可以与以往处理过的类似案例对比来研判[29]。目前，有关网络舆情研判的成果主要是网络舆情方法和机制方面，实证研究有但不多，这是一个值得继续探讨的问题。

三、网络舆情监测的相关技术

（一）网络舆情信息采集技术

理论上讲，网络舆情信息采集应该主要针对占整个网络的50%的动态

网页，以便更好地完成对于动态页面主体内容的获取工作。

有学者曾提出基于 Rhino 实现 JavaScript 动态页面解析的解决方案，脚本引擎 Rhino 可以对输入的 JavaScript 脚本片断进行逐行解析，并以不同形式分别输出 JavaScript 动态页面中的超链接网络地址和页面主体内容[30]。但是，因为缺少超链接网络地址过滤模块来抛弃与动态页面内核心超链接无关的网络地址信息，所以处理时间长，存贮空间浪费多。再者，该方法只实现对 JavaScript 脚本语言的解析，对其他的例如 VBScript，PHP 等是不支持的，不具有普遍的适应性。通用搜索引擎所应用的通用网络爬虫的目标，就是尽可能多地采集信息页面。而在这一过程中它并不太在意页面采集的顺序，或者被采集页面的主题相关性。这消耗了非常多的系统资源和网络带宽，但并没有换来采集页面的较高利用率。

也有学者提出采用正则表达式匹配和 MD5 加密技术来解决上述问题，这种方法只适合主题网络爬虫[31]。

目前对网络舆情信息采集技术研究并不多见，而直接采用现有的网络爬虫也是不明智的，采集技术是网络舆情挖掘与预警系统开发的瓶颈之一。在现有的系统中，网络舆情信息的采集环节一般支持自定义 URL 的数据抓取，但是仍不能很好实现对网络上所存在的各种各样类型的文件采集和分析，乃至整个网络的信息采集和分析。作为网络舆情挖掘与预警的前提——信息采集，同样是网络舆情系统开发的研究重点。

（二）网络舆情信息过滤技术

采集到的 Web 页面并不是都能直接使用。为了提高网络舆情处理的效率和准确率，系统需要在网络舆情预处理之前判断采集到的页面是否有冗余信息，不相关的 Web 页面被过滤掉，只存储有相关性的 Web 页面。网络舆情信息过滤的方法很多，从过滤的手段来看，可以分为基于内容的过滤、基于网址的过滤和混合过滤三种。基于内容的过滤是通过文本分析、图像识别等方法阻挡不适应的信息；基于网址的过滤是对认为有问题的网址进行控制，不允许访问其信息；混合过滤是将内容过滤和网址过滤结合起来，控制不适应信息的传播。从是否对网络信息进行预处理来看，信息过滤可分为主动过滤和被动过滤两种。主动过滤是预先对网络信息进行处理，如对网页或网站预先分级、建立允许或禁止访问的地址列表等，在过滤时可以根据分级或地址列表决定能否访问。被动过滤是不对网络信息进行预处理，过滤时才分析地址、文本或图像等信息，确定是否过滤[32]。这类系统可以称之为网

络舆情监测系统的最初雏形，从功能上看，基本实现网络不良信息的过滤，但是无法实现网络舆情系统的热点发现和主题追踪等功能。

目前，判断页面内容与主题的相关性的方法仍然是基于关键词的模型匹配方法。为了便于计算机计算，目前进行信息主题过滤和聚合主要采用布尔模型和向量空间模型建立用户索引，然后进行语义信息匹配度计算[33]。

（三）网络舆情信息分词技术

在网络舆情中，抓取页面信息后进行分词是主题分析的基础，其结果在舆情监测中有重要影响。目前，比较流行的分词方法有：机械分词法、语义分析法和人工智能分词法。除此之外，目前已开发出许多实用的分词系统，如哈工大统计分词系统，清华大学 SEG 分词系统，中国科学院计算技术研究所研制的分词系统 ICTCLAS，北大计算语言研究所的分词系统，欧姆龙分词系统等。

但是，现有的分词技术和分词系统都是不尽如人意的。主要的缺陷在于，网络舆情的分词系统技术仅仅借用了中文分词技术的广义方法，没有针对网络舆情自身的特点来开发设计。目前相关的研究数量不多。郑魁，疏学明在中国科学院 ICTCLAS 分词系统的基础上，提出了网络舆情热点发现分词法，就是一个有针对性的应用创新[34]。

网络舆情分词还要考虑网络新词的发现和更新问题。分词是否合理直接影响到网络舆情信息处理分析的准确性，一个好的分析系统或者分词算法，不仅能发现已有的词语，而且还能实现简单的构词。构词对于网络环境下的信息处理是十分重要。网络上流行着大量新兴的词汇，这些词汇要不断补充到现有的词表当中。在社会热点事件之中往往会伴生许多新的网络词汇，这些新兴词汇有着向现实世界渗透的巨大影响力[35]。网络舆情离不开对新词的监控。它直接影响后期的主题监测识别。分词系统或者算法的选择是至关重要的，甚至还要对现有的技术和算法进行进一步的改进，以适应舆情监测的需要。

（四）网络舆情话题（主题）识别技术

主流的话题发现算法都采用文本聚类技术来实现，该类算法的主要问题就是准确率低、大类现象比较严重。在早期的网络话题相关研究中，为了简化问题，一般假定所有的话题没有层次之分，而且一个文档只能与一个话题相关。但随着研究的深入，从 2003 年开始，层次化话题发现作为话题发现

与跟踪领域一个全新的研究问题被提了出来，它突破了传统的话题组织忽略话题多粒度现象的不合理之处，采用层次化的结构对话题进行组织[36]。

比较典型的网络舆情主题识别是采用了话题检测与追踪（TDT）技术。话题识别与跟踪的基本思想源于1996年，当时美国国防高级研究计划委员会（DARPA）提出需要一种能自动确定新闻信息流中话题结构的技术。随后，该技术引起了广泛关注，经过很长一段时间反复实践，最后形成了完整的理论体系和测评体系。话题检测与追踪研究主要包括以下五个任务：新闻报道的切分、新事件的检测、报道关系检测、话题检测及话题跟踪。虽然TDT评测中有一些很好的话题发现模型，但由于本文中的热点话题发现处理的是动态的数据流，语料的规模更大，并且除了发现的话题要准确，也看重话题发现的效率和复杂度，所以需要寻求一种更好更适用的话题发现模型。例如，时达明综合考虑了评论数、评论内容、话题内容等要素提出话题热点度的计算公式，根据计算出的热点度HotDegree进行排序，得到热点话题的排序，从而进行热点话题的发现[37]。

基于话题特征统计排序从而得到热点话题的发现方法，将文本聚类问题的实现转换为话题特征聚类问题，也大大提高了话题的准确率和可读性。该方法分为文本预处理和话题发现与分析两个步骤。在文本预处理阶段，首先对文本进行分词，建立索引，将文本存入自建索引库中。在分词的基础上，提取文本中的关键词列表，用一定数量的关键词来表示该文本信息。需要注意的是，文本的关键词是文本的特征，但不是话题的特征[38]。

可以看出，目前网络舆情的主题识别技术，正在从传统的线性文本聚类分析，向更注重内容特征的话题标引统计识别技术过渡。在适应计算机模拟和计算要求上，新兴的话题识别技术，由于缺乏相应的模型支撑，仍然处于探索和待实证阶段。

（五）网络舆情信息聚类技术

在网络舆情分析系统中，聚类分析研究是比较普遍的网络舆情分析方法，它的应用面非常的广，既被用于热点话题的发现，也被应用于倾向性分析。聚类分析方法主要有文本聚类，主题聚类，自动聚类等，这些分析方法主要采用的算法都是常用的文本聚类算法，有决策树（Decision Tree，DT）、朴素贝叶斯（Naive Bayes，NB）、类中心点（Rocchio）、K最邻近（k-Nearest Neighbors，KNN）、支持向量机（Support Vector Machines，SVM）、Boosting、线性最小平方拟合（LinearLeast Squares Fit，LLSF）以及神经网络

(Neural Network，NN) 等[39]。综合考查和分析种种聚类方法的适用范围和特点，龚海军提出了增量多层聚类算法来发现话题，首次聚类采用 KNN 算法，二次聚类采用凝聚聚类[40]，罗晖霞则对 K - MEANS 聚类算法进行改进，使之更加适应网络舆情聚类的要求和有效度[41]。但是这些算法的提出，只是基于其他算法的缺陷采用排除法，同样也没有对各算法进行比较研究。

网络舆情信息聚类技术需要改进和努力的方向，就是要降低算法的复杂性，提高运算效率，排除噪声数据干扰，并尽量使得聚类结果可以使用。各种聚类算法在聚类效果上都是各有优劣，目前没有一种聚类方法得到广泛认可成为主流，很多系统在采用这些算法的时候，并没有对这些算法的效果做一个比较或者说只是做了个简单的合理性评价就给予采用了。在网络舆情的特定环境中，聚类算法的审慎选择和效果验证是极其必要的。

(六) 网络舆情信息挖掘技术

将数据挖掘技术应用到网络舆情系统中的案例是屈指可数的，其中北京交通大学的薛玮在这个方面做过研究。他开发的系统由用户管理模块、挖掘策略管理模块、挖掘模块、故障管理模块构成，主要完成文本分析、文本分类、概念聚类、语义索引、自动文摘、事件处理等任务。系统采用了基于余弦相似度分析的 Cosine Similarity PageRank (CSP) 算法来挖掘舆情信息，并通过文本分析技术进行特征提取，生成舆情信息文摘，接着采用文本分类技术来进行文档自动分类，将分类后的文档进行概念聚类，产生概念空间，然后采用神经网络算法建立具有联想功能的语义索引，最后为用户提供基于概念的检索查询接口，并通过事件处理提供舆情事件的发展过程即来龙去脉[42]。数据挖掘技术能帮助实现从提供无序低价值信息向提供高质高价值信息方向的转变，但是数据挖掘要依赖于大量的网络信息，这就需要相应发达的数据采集技术予以支撑，否则难为“无米之炊”。数据挖掘在网络舆情监测中的应用目前仍然处于初级阶段，还有待今后的研究。

Web 信息挖掘由传统数据库领域的数据挖掘技术演变而来。随着互联网的蓬勃发展，数据挖掘技术被运用到网络上，并根据网络信息的特点发展出新的理论与方法，演变成网络信息挖掘技术。Web 数据挖掘技术分为三大类：Web 内容挖掘、Web 结构挖掘和 Web 访问信息挖掘。利用 Web 信息挖掘不仅可以发现网络舆情，分析网络舆情的起源，发现网络舆情受众及其特点，还可以研究舆情在网络上的传播、扩散模式，甚至评估舆情影响效果。网络舆情是由特定的网络信息所体现的，利用 Web 信息挖掘技术对网