于卫红 著

R语言与网络舆情处理

清华大学出版社

于卫红 著

R语言与网络舆情处理

清华大学出版社
北京

内 容 简 介

进入互联网时代后，网络舆情形成迅速，影响着社会生活的方方面面，如何高效全面地采集舆情数据并利用数据挖掘算法及数据分析工具将舆情文本中有价值的信息挖掘出来，对于舆情监管、舆情研判、舆情引导至关重要。本书以R语言作为舆情分析的工具，在阐述相关原理的基础上，介绍了网络舆情信息采集、舆情信息预处理、舆情文本分类、舆情文本聚类、舆情数据关联规则挖掘、舆情相关指标预测等舆情分析环节，所有分析都使用R语言进行实现，给出了完整的过程和代码。本书可以作为舆情处理、数据分析等教学或科研的技术参考书，适于本科生、研究生、数据分析爱好者、舆情分析工作者及研究人员等阅读参考。

图书在版编目(CIP)数据

R语言与网络舆情处理/于卫红著. —北京：清华大学出版社，2018
ISBN 978-7-302-48257-4

Ⅰ. ①R… Ⅱ. ①于… Ⅲ. ①程序语言－程序设计 ②互联网络－舆论－研究 Ⅳ. ①TP312 ②G219

中国版本图书馆CIP数据核字(2017)第209829号

责任编辑：闫红梅　薛　阳
封面设计：常雪影
责任校对：梁　毅
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京泽宇印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10　　**字　　数**：244千字
版　　次：2018年1月第1版　　**印　　次**：2018年1月第1次印刷
印　　数：1～1000
定　　价：39.00元

产品编号：075768-01

前言

2013年8月19日和20日，习近平总书记出席全国宣传思想工作会议并发表重要讲话；2014年10月15日，习近平总书记主持召开文艺工作座谈会并发表重要讲话；2015年12月25日，习近平总书记视察解放军报社并发表重要讲话；2016年2月19日，习近平总书记到人民日报社、新华社、中央电视台三家中央新闻单位进行了实地调研后，主持召开党的新闻舆论工作座谈会并发表重要讲话。从这些讲话中，我们可以深刻地领会到：党中央高度重视舆论宣传工作，根据形势发展的需要，更是把网络舆情监督和引导当作重中之重来抓。

在当前的互联网及大数据的时代背景下，网络舆情形成迅速，影响着社会生活的方方面面，如何高效全面地采集舆情数据并利用数据挖掘算法及数据分析工具将舆情文本中有价值的信息挖掘出来，对于舆情监管、舆情研判、舆情引导至关重要。网络信息的不断膨胀给舆情工作提出了新的挑战，为了更好地进行舆情收集、舆情研判、加快构建舆情引导新格局，舆情工作方式、舆情管理思维、舆情数据分析技术等都需要不断创新。

作者在本书的写作过程中阅读了大量的相关文献。文献研究表明，目前，越来越多的学者加入到了网络舆情的基础理论、支撑技术和演化机制等的研究中，网络舆情的研究视角日益多样化，研究内容也越来越深入。从网络舆情分析的视角来看，其核心技术主要包括自然语言处理、文本分类、文本聚类、关联分析、智能预测等，相应的理论、算法等也日臻成熟。但是，在实际的网络舆情分析各个环节中，舆情信息如何有效地采集、舆情分析算法如何高效地实现、舆情分析结果如何可视化展示等问题仍然困扰着很多研究者和舆情分析人员。

基于上述考虑，本书以R语言作为舆情分析工具，在阐述相关原理的基础上，介绍了网络舆情信息采集、舆情信息预处理、舆情文本分类、舆情文本聚类、舆情数据关联规则挖掘、舆情预测等的技术和方法。作为数据分析的利器，与其他流行的统计分析软件（如Excel、Matlab、SAS、SPSS等）相比，R语言的优势主要体现在开源免费、易于扩展、数据包丰富、可视化功能强大、可运行于多种平台。

本书力求简明扼要、提供有价值的知识，以最浅显的语言、详尽的R语言实现代码向读者循序渐进地展现网络舆情分析的完整过程。本书共8章，具体章节结构如下。

第1章　网络舆情与舆情分析概述：主要介绍了网络舆情的定义及特征，并对网络舆情的研究热点及相关技术做了概述。

第2章　R语言基础：为了帮助不熟悉R语言的读者尽快入门，本章主要从数据读写、基本语法、绘图三方面对R语言的使用做了言简意赅的介绍。

第3章　网络舆情信息采集及R爬虫的实现：介绍了网络舆情信息采集的基本原理、八爪鱼数据采集器的使用，并通过示例讲解了如何使用R语言开发一个简单的信息采集爬虫。

第4章　基于R语言的舆情信息预处理：介绍舆情信息预处理中分词、去停用词、词频

统计、文本向量化等的基本原理以及R语言实现方法。

第5章　基于R语言的网络舆情分类：从分类的基本原理入手，介绍了决策树分类算法、网络舆情分类的基本原理，并通过“微信公众号文章分类”这一示例讲解了使用R语言进行网络舆情分类的方法和步骤。

第6章　基于R语言的网络舆情热点话题聚类：介绍了聚类的基本原理、经典的聚类算法、聚类算法在舆情分析中的应用，并通过“电商顾客评论热点话题聚类”这一商务舆情分析示例讲解了使用R语言进行网络舆情聚类的方法和步骤。

第7章　基于R语言的网络舆情关联规则挖掘：介绍了关联规则挖掘的基本原理、常用的关联规则挖掘算法、关联规则在舆情分析中的应用，并通过“雾霾舆情热点词关联模式挖掘”这一示例讲解了使用R语言进行网络舆情关联分析的方法和步骤。

第8章　基于R语言与BP神经网络的网络舆情分析：介绍了BP神经网络的算法原理、BP神经网络在舆情分析中的应用，并通过“微博转发数与评论数预测”这一示例讲解了使用R语言与神经网络进行网络舆情相关指标预测的方法和步骤。

本书系2015年度教育部人文社会科学研究规划基金项目“微信环境下基于大数据的高校舆情监管机制研究”(项目编号：15YJAZH102)研究成果之一。本书内容浅显易懂、代码详尽，希望能对舆情工作者及研究人员有所裨益。由于作者学识有限，书中难免有所疏漏，在此表示歉意，并请读者朋友们不吝赐教。最后感谢清华大学出版社为本书的出版所做的努力。

大连海事大学　于卫红

2017年3月

目 录

第1章 网络舆情与舆情分析概述

1.1 舆情与网络舆情的基本概念

1.1.1 舆情的起源及定义

作为舆情研究最基本的概念，舆情是一个充分体现中国历史文化传统的词语。“舆”字在古代指车。《说文解字·车部》：“舆，车舆也。”“舆人”即为造车工人。《周礼·考工记·舆人》：“舆人为车。”到春秋末期，“舆”逐渐演化为轿子，“舆人”也被赋予抬轿子的人的意思，并逐渐涵盖了车夫、差役、小官吏和随车士卒等下层的普通大众的意思。到了汉代，历史文献中的“舆人”，与“刍荛”“庶人”一样，成为普通百姓的代名词。“舆人”之后又出现了“舆人之诵”“舆人之议”等词语，表示一般百姓的意见、言论。

据查，“舆情”一词最早出现在《旧唐书》中，唐昭宗在乾宁四年（公元 897 年）的一封诏书中称：“朕采于群议，询彼舆情，有冀小康，遂登大用。”意思是说皇帝采纳群臣的意见，了解老百姓的看法，不仅有益于国家安康，更将对进谏者委以重用。显然，在中国古代皇帝布告臣民的专用文书中出现“舆情”字样，一来说明对“舆情”的最初使用源自官方，而非民间。再者，诏书中将“舆情”与“群议”两个词对用，充分强调了“舆情”特指普通老百姓的看法，而不是统治阶层的意见。

简言之，舆情是“舆论情况”的简称，是指在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，作为主体的民众对作为客体的社会管理者、企业、个人及其他各类组织及其政治、社会、道德等方面的取向产生和持有的社会态度。它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等表现的总和。

从传统的社会学理论上讲，舆情本身是民意理论中的一个概念，它是民意的一种综合反映。但是，从现代舆情理论的严格意义上讲，舆情本身并不是对民意规律的简单概括，而是对“民意及其作用于执政者及其政治取向规律”的一种描述，舆情是舆情因变事项发生、发展和变化过程中，民众所持有的社会态度。正确理解舆情概念，必须把握以下 4 层含义。

（1）舆情是民意集合的反映。换句话说，民意是形成舆情的始源，没有民意，就没有舆情。

（2）舆情所要反映的民意，是那些对执政者决策行为能够产生影响的“民意”，而非民意的全部。

（3）舆情因变事项是舆情产生的基础，研究、分析舆情，首先要深入研究、分析舆情因变

事项的发生、发展和变化的规律。

(4) 舆情空间对舆情传播及其对执政者决策行为的影响有重要作用。舆情定义中的"民众社会政治态度",是指民众对执政者及其所持有的政治取向的看法、意见和态度。民众的这种社会政治态度说到底是对自身利益需求的一种诉求和表达,它不仅包括民众对国家政治的看法、意见和态度,对社会政治的看法、意见和态度,同时还包括民众对社会事物的看法、意见和态度。

1.1.2 网络舆情

网络舆情是指在一定的社会空间内,通过网络围绕中介性社会事件的发生、发展和变化,民众对公共问题和社会管理者产生和持有的社会政治态度、信念和价值观。它是较多民众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等表现的总和。网络舆情形成迅速,对社会影响巨大。随着因特网在全球范围内的飞速发展,网络媒体已被公认为是继报纸、广播、电视之后的"第4媒体",网络成为反映社会舆情的主要载体之一。

网络舆情与传统意义上的社会舆情既有联系又有区别。二者的联系具体表现在以下几个方面。

(1) 网络舆情和社会舆情都是社会存在和发展状况的反映。网络舆情和社会舆情都不可能是超时代超社会的,它们都具有社会历史性。

(2) 网络舆情和社会舆情都是公开表达和传播的态度、意见和看法。不公开表达和传播的态度、意见和看法不能形成舆情,网络舆情和社会舆情都是人们面对客观现象和现实问题所公开表达出来的内心态度和意见,并通过公开传播来吸引和影响广大公众。

(3) 网络舆情和社会舆情往往相互影响相互作用。网络舆情和社会舆情的关联性很强,现实社会中人们关于某一社会现象或社会问题的议论很容易传到网上,而网上关于某一社会现象或社会问题所形成的议论也会很快向社会扩散。

网络舆情与社会舆情相互区别,主要表现在以下几个方面。

(1) 网络舆情和社会舆情的传播方式不同。社会舆情往往是通过人们的街谈巷议、口传心授,并以一定的意见、情绪、态度甚至行动倾向表现出来,而网络舆情的产生、形成并发挥作用的载体是网络,即网民的情绪、态度和意见等都是在网络中进行表达。

(2) 网络舆情与社会舆情在社情民意的反映面上不尽相同。作为网络舆情主体的网民只是社会人群的一部分,因此,网络舆情不能等同于社会整体的意见与情绪,它只是反映以网民为主的某些社会群体或阶层的意愿。

(3) 网络舆情与社会舆情的存在形式不同。与社会舆情主要通过人们的街谈巷议或行为举动等方式表现不同,网络舆情则是通过新闻跟帖、论坛、博客、播客、即时通信工具、搜索聚合等途径表达出来。

1.2 网络舆情的特征及表现形式

网络舆情的表现形式主要为新闻评论、BBS论坛、博客、播客、聚合新闻、新闻跟帖及转帖等。近年来,随着网络技术的推陈出新,除网络新闻、网络论坛等传统应用外,又出现了微博、维基Wiki、微信等新形态的信息交互模式。分析现有的研究成果,学者们普遍认为:网

络舆情是由主体、客体、载体、本体和受体等要素构成的过程整体。所谓主体，就是网民，即通过互联网络关注社会事件并发表自己观点和意见的普通民众。所谓客体，就是网络舆论的对象，即网民所普遍关注的，且在现实利益、社会关系、社会观念等方面相互关联的社会事件。所谓本体，就是网民的共同意见，即经过网民互动与竞争所形成的某种为网民群体普遍赞同，且能在心理上产生共鸣的一致性意见。所谓载体，就是网络舆论的传播媒介，即由两台或两台以上的计算机，通过信息技术互相联系而构成的传播网络。所谓受体，就是接受网络舆论影响的网民，只是这里的网民与作为主体的网民略有不同，他们往往会具体化为具有一定经济、政治、思想文化执政的政府部门或成员。

范围广、交互性强、更新速度快的互联网传播从根本上改变了传播者与受传者之间的关系，是对传统新闻媒介的传播模式的解构和颠覆。在网络这个人人共同拥有的信息平台上，传播者和受传者处于完全平等的地位，共同享有根据自己的需要选择信息的自由和发表意见和观点的权利。网络舆情对政治生活秩序和社会稳定的影响与日俱增，一些重大的网络舆情事件使人们开始认识到网络对社会监督起到的巨大作用。同时，网络舆情突发事件如果处理不当，极有可能诱发民众的不良情绪，引发群众的违规和过激行为，进而对社会稳定造成严重威胁。

研究网络舆情，需要掌握网络舆情的特点。网络舆情表达快捷、信息多元、方式互动，具备传统媒体无法比拟的优势。网络的开放性和虚拟性，决定了网络舆情具有以下特点。

（1）直接性。通过网络媒介，网民可以立即发表意见，下情直接上达，民意表达更加畅通；网络舆情还具有无限次即时快速传播的可能性。在网络上，只要复制粘贴，信息就得到重新传播。相比较传统媒体的若干次传播的有限性，网络舆情具有无限次传播的潜能。网络的这种特性使它可以轻易穿越封锁，令监管部门束手无策。

（2）随意性和多元化。网民可以随意发表言论，不受任何约束。网络舆情不同于传统媒体的另一特点是缺乏媒体“审核人”的角色。在网络上，任何一个人都能不经过审核直接发布信息。网民在网上或隐匿身份、或现身说法，纵谈国事，嘻怒笑骂，交流思想，关注民生，多元化的交流为民众提供了宣泄的空间，也为搜集真实舆情提供了素材。

（3）突发性。网络舆情的形成往往非常迅速，一个热点事件的存在加上一种情绪化的意见，就可以成为点燃一片舆论的导火索。

（4）隐蔽性。互联网是一个虚拟世界，由于发言者身份隐蔽，并且缺少规则限制和有效监督，网络自然成为一些网民发泄情绪的空间。在现实生活中遇到挫折，对社会问题的片面认识等，都会利用网络得以宣泄。因此在网络上更容易出现庸俗、灰色的言论。

（5）偏差性。互联网舆情是社情民意中最活跃、最尖锐的一部分，但网络舆情还不能等同于全民立场。随着互联网的普及，网民们有了空前的话语权，可以较为自由地表达自己的观点与感受。但由于网络空间中法律道德的约束较弱，如果网民缺乏自律，就会导致某些不负责任的言论，比如热衷于揭人隐私，谣言惑众，反社会倾向，偏激和非理性，群体盲从与冲动等。

1.3　网络舆情分析技术

1.3.1　网络舆情分析的研究热点

对于网络舆情的特点，舆情工作者应当了然于心，并能对现实中出现的各种网络舆论做

出及时反馈，防微杜渐，防患于未然。因此，必须利用现代信息技术对网络舆情予以分析，从而进行控制和引导。由于网上的信息量十分巨大，仅依靠人工的方法难以应对网上海量信息的收集和处理，需要加强相关信息技术的研究，形成一套自动化的网络舆情分析系统，及时应对网络舆情，由被动防堵，化为主动梳理、引导。

特别是在如今的大数据时代，网络舆情分析更要用数据说话，跟踪网络舆情的起源和演变，最终根据分析给出建议性结果，为政府、企业乃至个人应对舆情提供决策支持。网络舆情分析大致有两个工作重点，一是还原舆情发展过程，找到舆情产生的根源；二是预测，分析出网络舆情的未来走向，再根据预测结果提出应对方案。

在探讨网络舆情分析技术之前，我们先对舆情分析中几个常用的基础术语做一个统一的概念界定。

(1) 舆情：通常是指较多群众关于现实社会及社会中各种现象、问题所表达的信念、态度、意见和情绪表现的总和；简而言之就是社会舆论和民情。一个严格定义是：舆情是指在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，作为主体的民众对作为客体的国家管理者产生和持有的社会态度。

(2) 事件(Event)：在特定时间、特定地点发生的事情。

(3) 主题(Topic)：也称为话题，指一个种子事件或活动以及与它直接相关的事件和活动。

(4) 热点：也可称为热点主题。热点和主题的概念比较接近，但有所区别。其主要特点如下：热点通常是一个主题，包含种子事件及相关报道；热点和时间相关，通常指某段时间内的热点，例如当天热点、一周内热点；热点和主题某段时间内的文档数量相关。热点可以分为绝对热点和相对热点。其中，绝对热点为在某段时间内文档数量超过某个固定阈值的主题；相对热点为按照某种排序方式排名靠前的若干个主题。

目前，网络舆情分析的研究热点主要包括如下几方面。

1. 主题检测与跟踪

在目前信息爆炸的情况下，信息的来源已不再是问题，而如何快捷准确地获取感兴趣的信息才是人们关注的主要问题。目前的各种信息检索、过滤、提取技术都是围绕这个目的展开的。由于网络信息数量太大，与一个话题相关的信息往往孤立地分散在很多不同的地方并且出现在不同的时间，仅通过这些孤立的信息，人们对某些事件难以做到全面的把握。一般的检索工具都是基于关键词的，返回的信息冗余度过高，很多不相关的信息仅仅是因为含有指定的关键词就被作为结果返回了，因此人们迫切地希望拥有一种工具，能够自动把相关话题的信息汇总供人查阅。主题检测与跟踪(Topic Detection and Tracking，TDT)技术就是在这种情况下应运而生的。通过主题发现与跟踪，人们可以将这些分散的信息有效地汇集并组织起来，从而帮助用户发现事件的各种因素之间的相互关系，从整体上了解一个事件的全部细节以及该事件与其他事件之间的关系。简言之，主题检测与跟踪任务的主要工作是准确地检测话题并跟踪话题的动态演化过程。

与一般的信息检索或者信息过滤不同，TDT 所关心的话题不是一个大的领域(如美国的对华政策)或者某一类事件(如恐怖活动)，而是一个很具体的“事件(Event)”，如美国“9·11事件”、习近平主席访美等。与早期面向事件的检测与跟踪(Event Detection and Tracking，

EDT)也不同，TDT 检测与跟踪的对象从特定时间和地点发生的事件扩展为具备更多相关性外延的话题，相应的理论与应用研究也同时从传统对于事件的识别跨越到包含突发事件及其后续相关报道的话题检测与跟踪。

美国国家标准技术研究院为 TDT 研究设立了 5 项基础性的研究任务，包括面向新闻广播类报道的切分任务；面向已知话题的跟踪任务；面向未知话题的检测任务；对未知话题首次相关报道的检测任务和报道间相关性的检测任务。

1）报道切分任务

报道切分(Story Segmentation Task，SST)的主要任务是将原始数据流切分成具有完整结构和统一主题的报道。比如，一段新闻广播包括对股市行情、体育赛事和人物明星的分类报道，SST 要求系统能够模拟人对新闻报道的识别，将这段新闻广播切分成不同话题的报道。SST 面向的数据流主要是新闻广播，因此切分的方式可以分为两类：一类是直接针对音频信号进行切分；另一类则将音频信号翻录为文本形式的信息流进行切分。

2）话题跟踪任务

话题跟踪(Topic Tracking Task，TTT)的主要任务是跟踪已知话题的后续报道。其中，已知话题没有明确的描述，而是通过若干篇先验的相关报道隐含地给定。通常话题跟踪开始之前，为每一个待测话题提供 1～4 篇相关报道对其进行描述。同时还为话题提供了相应的训练语料，从而辅助跟踪系统训练和更新话题模型。在此基础上，TTT 逐一判断后续数据流中每一篇报道与话题的相关性并收集相关报道，从而实现跟踪功能。

3）话题检测任务

话题检测(Topic Detection Task，TD)的主要任务是检测和组织系统预先未知的话题，TD 的特点在于系统欠缺话题的先验知识。因此，TD 系统必须在对所有话题毫不了解的情况下构造话题的检测模型，并且该模型不能独立于某一个话题特例。换言之，TD 系统必须预先设计一个善于检测和识别所有话题的检测模型，并根据这一模型检测陆续到达的报道流，从中鉴别最新的话题；同时还需要根据已经识别到的话题，收集后续与其相关的报道。

4）首次报道检测任务

在话题检测任务中，最新话题的识别都要从检测出该话题的第一篇报道开始，首次报道检测任务(First-Story Detection Task，FSD)就是面向这种应用产生的。FSD 的主要任务是从具有时间顺序的报道流中自动锁定未知话题出现的第一篇相关报道。大体上，FSD 与 TD 面向的问题基本类似，但是 FSD 输出的是一篇报道，而 TD 输出的是一类相关于某一话题的报道集合，此外，FSD 与早期 TDT Pilot 中的在线检测任务(Online Detection)也具备同样的共性。

5）关联检测任务

关联检测任务(Link Detection Task，LDT)的主要任务是裁决两篇报道是否论述同一个话题。与 TD 类似，对于每一篇报道，不具备事先经过验证的话题作为参照，每对参加关联检测的报道都没有先验知识辅助系统进行评判。因此，LDT 系统必须预先设计不独立于特定报道对的检测模型，在没有明确话题作为参照的情况下，自主地分析报道论述的话题，并通过对比报道对的话题模型裁决其相关性。LDT 研究可以广泛地作为 TDT 中其他各项任务的辅助研究，比如 TD 与 TT 等。

2. 舆情热点研究

热点自动发现任务也可叫作热点检测，就是如何从不断涌现的网上舆情中及时发现新发生的热点信息，并对其进行持续追踪。热点检测任务可以在主题检测任务的基础之上，加入时间和数量两个因素的分析来解决热点发现的问题。

热点分析任务在热点自动发现任务的基础上，对自动发现的热点进行深入分析，从多方面、多角度综合分析和展现当前的舆情热点。研究内容包括舆情热点的关键词和摘要提取、情感分析、传播分析、趋势分析和关联分析等任务。

3. 情感倾向性分析

指通过计算机技术自动分析文本信息所包含的情感因素，例如喜欢或讨厌、正面或负面、快乐或悲伤、愤怒和恐惧等。在不同的文献中，情感分析也被称作情感分类、褒贬分类、观点提取、观点摘要、情绪分析、情感识别、情感计算等。同时，情感是一个很广泛的词汇，在不同场合研究者往往采用不同的词汇来表达，比如观点(Opinion)、情感(Sentiment)、情绪(Emotion/Affect)等。

对舆情文本进行倾向性分析，实际上就是试图用计算机实现根据文本的内容提炼出网络传播者所蕴含的感情、态度、观点、立场、意图等主观反映。

目前，情感倾向分析的方法主要分为两类：一种是基于情感词典的方法；一种是基于机器学习的方法，如基于大规模语料库的机器学习。前者需要用到标注好的情感词典，英文的词典有很多，中文主要有知网整理的情感词典 HowNet 和台湾大学整理发布的 NTUSD 两个情感词典，还有哈工大信息检索研究室开源的《同义词词林》可以用于情感词典的扩充。基于机器学习的方法则需要大量的人工标注的语料作为训练集，通过提取文本特征，构建分类器来实现情感的分类。

4. 舆情趋势预测

舆情同其他事物一样，是一种客观存在，有其产生、发展、变化的规律。只要对其予以客观、全面、科学的考察，细致、认真、仔细的分析，就能大致预测它的发展方向。特别是当前我们已处于大数据时代，大数据使网络舆情预测成为现实。对已经出现的网络舆情予以监测，这是网络舆情引导的传统做法，也是以往网络舆情管理的起始。但是利用大数据技术，可以对网络舆情中具有关联的数据进行挖掘并加以分析，使敏感信息在网络上传播的初始阶段就被监测到。在此基础上通过模型对网络舆情变化趋势进行仿真，使网络舆情预测成为现实。实现网络舆情预测，至关重要的是对数据的相关性进行全面分析。而在传统的网络舆情引导中，由于数据库的缺乏和计算分析能力有限，往往难以全面分析网络舆情，得出的结论也有失偏颇。大数据环境下，对网络舆情的分析由静态化向动态化转变，由片面化向立体化转变，由单一化向全局化转变。利用大数据技术解构海量信息，并对这些信息加以重构，对网络数据的相关性进行深度挖掘，可以全面科学地分析并预测网络舆情的发展趋势。此外，大数据使网络舆情实现量化管理。使网络舆情得以量化，是利用大数据对网络舆情进行科学预测的前提。网络舆情信息量巨大，而被挖掘出来的网络舆情信息需要进行量化，在此基础上再建立数学模型对信息数据进行计算和分析。数据的量化指的是数据是可计算的，

一是在密切关注网民态度与情绪变化的同时对其采用量化指标加以标识，二是对网络言论所持某一观点的人群数量进行统计，三是透过网络信息文字内容来对网民互动的社会关系网络数量进行统计。另一方面，大数据使网络舆情相互关联。网络信息是网络背后的网民所传达出来的信息的集合，因而对网络数据进行研究，实质上是对由人所组成的社会网络进行研究。要实现网络舆情预测，离不开对网络舆情之间的关系进行关联这一尤为重要的大数据技术。在大数据时代，每个网络数据都被看作是一个节点，能够在舆情链上与其他关联数据不受限制地产生乘法效应，这种关联如同数据裂变，会扩大至全体网络数据，使舆情分析更为准确。

5. 舆情信息可视化

可视化是一个可以处理海量数据的可行工具之一，它能使科研人员发现数据内部隐藏的信息，从而进一步找出信息所反映的规律，提高对海量数据的认识。在网络舆情研究过程中，使用可视化分析技术能够克服传统数据收集、分析与呈现方法上存在的效率低下以及难以发现其中的关键信息与潜在特征的不足，通过化繁为简、化抽象为具象，能够使用易于理解的图形图像揭示网络舆情的分布、发展和演化规律，因而在网络舆情研究中具有非常显著的应用价值。

在具体的研究过程中，可视化从严格意义上来说是一种信息分析框架，原始信息、数据表格、可视化结构和最终呈现在用户面前的视图被这一框架有机地链接在一起。针对不同类型的舆情信息又有不同的研究方向，比如，对于文本信息，比较常用的可视化分析主要有基于关键词的网络舆情文本内容的可视化、时序性网络舆情文本信息的可视化等。对于具有层次结构的舆情信息，研究者们通常根据自己的关注点选择合适的层次信息可视化技术来呈现信息项之间错综复杂的层次关系，常用的可视化技术包括节点链接树、双曲树、径向树等，比如想要探究网络舆情信息扩散的路径就可以使用节点链接树的方法来发掘其中的关键节点。网状结构的舆情信息也是当前研究的热点，对于舆情的社会网络分析，研究者们提出了一些网络节点布局方法，如：按照力导向布局、地图布局、环状布局等。除此之外，一些常见的统计的图形，如饼图、折线图、直方图、总量图、趋势图等也常用于展现网络舆情信息的时间趋势、情感倾向、区域分布等特征以及舆情统计报表、报告的呈现。

1.3.2　网络舆情分析的步骤

舆情分析从数据采集到最终的分析报告发布主要包括 4 个步骤：舆情数据采集、数据预处理、舆情分析和舆情报告发布，如图 1-1 所示。

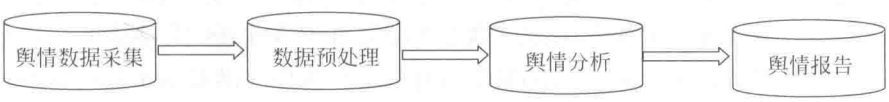

图 1-1　舆情分析的步骤

1. 舆情数据采集

互联网时代，要想达到舆情信息的快速准确采集，需要充分做到网上舆情采集和网下舆

情收集的互补，利用自动化的舆情监测工具，以网上舆情信息采集为主，保证舆情信息采集速度和数量。目前，国内许多软件公司开发出了多种网络舆情监测、采集、分析软件，这些系统利用爬虫技术，根据设定的监控关键词抓取重点媒体、论坛、博客、微博等网站里的舆情信息。其中，比较具有影响力的系统包括：乐思网络舆情监测系统、军犬网络舆情监控系统、Rank 舆情监测系统、谷尼舆情监测系统、红麦舆情监测系统等。一些免费甚至是开源的爬虫软件也可以用于舆情数据采集，这些软件主要包括集搜客网页抓取软件、八爪鱼爬虫、LoalaSam 网络爬虫等。研究者也可以根据实际情况使用 Java、Python、R 等语言有针对地开发满足研究需要的爬虫工具。

2. 数据预处理

信息预处理是对采集到的舆情进行初步的加工和处理，为后继舆情关键信息抽取和舆情内容分析奠定基础。

网络舆情数据大都是非结构化的文本数据，文本数据的预处理主要包括文本分词、去停用词（包括标点、数字、单字和其他一些无意义的词）、文本特征提取、词频统计、文本向量化等操作。

3. 舆情分析

舆情分析就是根据特定问题的需要，对针对这个问题的舆情进行深层次的思维加工和分析研究，得到相关结论的过程，可分为内容分析和实证分析。内容分析法是一种对信息内容做客观系统的定量分析的专门方法，其目的是弄清或测验信息中本质性的事实和趋势。提示信息所含有的隐性情报内容，对事物发展做情报预测。实证分析法是通过分析大量案例和相关数据后试图得出某些结论的一种常见研究方法。对舆情的分析要明确事件或话题本身所处的阶段，一般分为引发期、酝酿期、发生期、发展期、高潮期、处理期、平息期和反馈期等不同阶段。其次，应该在分析某一舆情热点之前对其进行科学的类型界定。热点事件一般主要分为突发自然灾害事件、生产安全事故、群体性事件、公共卫生事件、公权力形象、司法事件、经济民生事件、社会思潮、境外涉华突发事件等。

4. 舆情报告

根据舆情分析结果生成舆情分析报告。舆情报告是针对某个主题或者事件的舆论信息，以报告的形式展示主题情况，客观真实地展现某主题或事件在大众社会的看法和态度反馈，是调查报告的一种。一般舆情报告分为三个部分。第一部分对事件或主题进行概括式描述和简要介绍，交代事件的来龙去脉。第二部分是基于查找到的与主题有关的信息，以列表、绘图等方式来展现舆情发展。第三部分是对舆情分析的总结和对事件的客观评论，为领导决策做参考。

舆情报告不比新闻稿，它的时效性并不是十分快速，可能当报告出来时，事件早已平息，不再热门。这就是舆情报告的独特之处，它不在热门的时候画蛇添足，而是等人们的热情退去，给人们带来更深层次的理性的思考。