计算机网络基础研究书系

互联网信息监测系统研究

周 辉◎著

MONITORING THE INTERNET INFORMATION

知识产权出版社
全国百佳图书出版单位

本书的出版得到国家自然科学基金项目［61440019］，海南省自然科学基金项目［614228］，海南大学青年基金项目［qnjj1175］和海南大学科研启动基金项目［kyqd1232］的资助。

互联网信息监测系统研究

Monitoring the Internet Information

周　辉　著

知识产权出版社
全国百佳图书出版单位

图书在版编目（CIP）数据

互联网信息监测系统研究/周辉著. —北京：知识产权出版社，2015.9
ISBN 978-7-5130-2254-5

Ⅰ.①互… Ⅱ.①周… Ⅲ.①互联网络—信息管理—监测系统 Ⅳ.①TP393.4②G203

中国版本图书馆 CIP 数据核字（2013）第 213110 号

内容提要

本书循序渐进地介绍了互联网信息、信息监测的概念，手工监测方式及搜索引擎方式的不足，介绍了互联网信息监测系统存在的必要性和紧迫性，并尝试梳理了不同概念的内在差异和相互关系。在详细严谨的调研以及实际项目经验的基础上，归纳整理了企业用户在互联网信息监测上的功能性与非功能性需求，并以媒体监测、信息推送、咨询分析、监测规则管理为重点展开讲述了多个功能点。对从事互联网信息监测研究的学者及互联网信息监测的工作人员有指导借鉴意义。

责任编辑：国晓健　　**责任校对：**孙婷婷

封面设计：刘　伟　　**责任出版：**刘译文

互联网信息监测系统研究

周　辉　著

出版发行：	知识产权出版社有限责任公司	**网　　址：**	http://www.ipph.cn
社　　址：	北京市海淀区马甸南村 1 号	**天猫旗舰店：**	http://zscqcbs.tmall.com
责编电话：	010-82000860 转 8385	**责 编 邮 箱：**	guoxiaojian@cnipr.com
发行电话：	010-82000860 转 8101/8102	**发 行 传 真：**	010-82000893/82005070/82000270
印　　刷：	北京中献拓方科技发展有限公司	**经　　销：**	各大网上书店、新华书店及相关专业书店
开　　本：	787mm×1092mm　1/16	**印　　张：**	6.75
版　　次：	2015 年 9 月第 1 版	**印　　次：**	2015 年 9 月第 1 次印刷
字　　数：	95 千字	**定　　价：**	22.00 元

ISBN 978-7-5130-2254-5

前　言

当今，互联网信息的传播空前迅速，网民表达诉求的方式日益多元化。互联网信息是通过互联网传播的公众对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点，主要通过微博、论坛、博客、新闻评论、转帖、社交网站等实现并加以强化。社交网站、微博、微信等新媒体开始流行之后，信息在数小时之内足以传遍全国。及时掌握相关信息，无论对企业还是对国家机构，都有着至关重要的作用。

在互联网影响力日益增大的今天，各级党政机关，企业、事业单位和学术机构都越来越重视对网络信息的监测、研究和引导。如果引导不善，负面的网络信息将对社会公共安全形成较大威胁。对相关政府部门来说，如何加强对网络舆论的及时监测、有效引导，以及对网络舆论危机的积极化解，对维护社会稳定、促进国家发展具有重要的现实意义。

首先，本书循序渐进地介绍了互联网信息、信息监测的概念，手工监测方式及搜索引擎方式的不足，介绍了互联网信息监测系统存在的必要性和紧迫性，并尝试梳理了不同概念的内在差异和相互关系。其次，在详细严谨的调研及实际项目经验的基础上，本书归纳整理了企业用户在互联网信息监测上的功能性与非功能性需求，并以媒体监测、信息推送、咨询分析、监测规则管理为重点展开讲述了多个功能点。

在代表性需求的基础上，本书展开了需求分析和系统设计的工作，给出了具体的解决方案，并详细分析了互联网信息采集、海量信息管理与检

索，多维分析与机器挖掘等多个子系统的技术特点及设计要点。然后，从数据处理流程的角度，逐一剖析 HTML 网页信息的抓取，网页正文提取，结构化内容解析，中文、英文、维文等多语言支持，以及文本中的命名实体识别等技术环节。

在分析了系统结构、子系统的构成及数据处理过程中涉及的各个环节后，深入到互联网信息监测系统的细节中，探讨智能分词、文本聚类与分类，高可用性、访问代理等研究性较强的话题。

本书还列举了在互联网信息监测领域相关的厂商、代表性软件系统，以及与信息监测相关的知名的开源软件工具，从而帮助读者加深对相关行业及其动态的了解，为有志于开发互联网信息监测系统的企业和工程师提供参考思路。

笔者

2015 年 2 月

目　录

第1章 引 言

随着互联网的快速发展，网络媒体作为一种新的信息传播形式，已深入人们的日常生活。互联网已成为思想文化信息的集散地和社会舆论的放大器，网民言论前所未有的活跃，不论是国内还是国际重大事件，都能马上形成网上舆论。网民们通过网络来表达观点、传播思想，进而产生巨大的舆论效应，达到任何部门、机构、个人都无法忽视的地步。

1.1 互联网信息

互联网信息是通过互联网传播的公众对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点，主要通过微博、论坛、博客、新闻评论、视频、社交网站等实现并加以强化。因此在非常多的场合，互联网信息又被称为互联网舆论情况，简称网络舆情。

当今，信息传播空前迅猛，网络信息的表达诉求也日益多元。如果引导不善，负面的网络信息将对社会公共安全形成较大威胁。对相关政府部门来说，如何加强对网络舆论的及时监测，以及对网络舆论危机的积极引导，对维护社会稳定、促进国家发展具有重要的现实意义。

在互联网影响力日益增大的今天，各级党政机关、企业、事业单位和学术机构都越来越重视对网络信息的监测、研究和引导。2008 年 6 月 20 日，时任中共中央总书记胡锦涛在人民日报社考察工作时指出，“互联网

已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”。互联网已成为党和政府治国理政的重要新平台，网络舆论也越来越受到重视。

微博等新媒体开始流行之后，第一时间掌握相关信息和动态，无论对企业还是对国家机构，都有着至关重要的作用。最具有代表性的调查报告是2013年6月中国社会科学院新闻与传播研究所、社会科学文献出版社联合发布的新媒体蓝皮书《中国新媒体发展报告（2013）》。蓝皮书提出，2012年以来，中国新媒体用户持续增长、普及程度进一步提高，新媒体应用不断推陈出新、产业日趋活跃，新媒体的社会化水平日益提升、频频引发热点。新媒体的发展为人们的生产生活等带来了极大便利，但也出现了各种问题，如个人隐私泄露、不良信息传播、谣言层出不穷等，危害了网络环境，损害了人们对网络的信任。

蓝皮书提出，2012年，微博“国家队”异军突起，新华通讯社、人民日报、中央电视台等中央媒体齐齐发力，在微博舆论场尝试主导“微话语权”。截至2012年年底，新浪微博认证的媒体微博总数已经突破了11万个。中央媒体微博的崛起，改变了主流媒体应对网络热点时迟缓和失语的状态，提升了主流媒体的网络舆论引导能力。

中国微博用户整体呈现学历低、年纪轻、收入低、集中大中城市的特征。学生有9387万人，是微博用户的最大职业群体。月收入5000元以下用户占到总数的92.2%。其中，无收入群体人数最多，达到9183.5万人，这主要因为学生用户是微博最大的群体。

蓝皮书认为，不论是从相对数还是从绝对数上讲，新媒体都是最为主要的反腐倡廉事件的首次曝光媒介类型。从绝对数上说，反腐倡廉事件首次曝光于新媒体上的数量远大于首次曝光于传统媒体上的数量。其中，2010~2012年，反腐案件首次曝光于新媒体上的事件数量依次为67件、58件和31件，三年合计156件，是传统媒体的2倍。

蓝皮书认为，从微博意见领袖的信息来源来看，传统媒体仍是微博意

见领袖的重要新闻源。对意见领袖微博的信息来源进行统计，其中转载传统媒体信息的比例最高，占48%，而意见领袖的原创帖占44%，转载他人（非新闻机构和个人）仅占8%。目前，越来越多的媒体也加入到微博阵营中来。传统媒体的加入，特别是杂志和报纸在微博中的突出表现，证明了在碎片化阅读的新媒体时代，传统媒体仍能凭借自身优势在内容生产上胜出，达到引领舆论的效果。

虽然网络反腐借由现代信息技术而具有传统舆论监督无可比拟的优势，但由于其自发性、匿名性、虚拟性和开放性的特点，也使网络反腐中存在的问题十分突出。网络举报信息鱼龙混杂，真假难辨。网络赋予了网民言论自由，然而，一部分网民却滥用网络言论自由权，或为了赚取眼球，或为了窥探隐私，或出于发泄私愤，发布虚假信息，进行不实举报。不明真相的“围观”网民以讹传讹，虚假举报也获得很高的点击率和数量可观的跟帖和转发，使网络反腐陷入真假难辨的混乱无序状态。纪检及相关部门不得不花费大量人力物力对真假信息进行甄别查证，造成原本就很有限的反腐资源的浪费。例如，2012年11月30日，有微博曝光“四川达县县委书记有9名情妇”，一时引起网友热议。但经过官方大量的调查取证工作，证实所反映的问题严重失实。

而未经查证的“网络曝光”和“人肉搜索”，容易侵犯公民隐私，对被曝光主体构成网络暴力。一些网络举报人为了使自己的举报“言之凿凿”，不惜详细罗列与贪腐内容无关的所有个人信息，将被举报人的真实身份和个人信息暴露在亿万网民面前，甚至盗用他人照片佐证自己的观点或主张，使被举报人及家人的工作和生活受到许多困扰。例如，网帖所曝“拥有24套房产”的“房婶”，经纪委查实，只不过是一个普通工程师，而其6套房产也都是合法所得。虽然当事人的清白得以澄清，但其“家庭房产情况一览表”等个人隐私已被曝光。又如，厦门某女大学生的个人写真照片被盗用成雷政富情妇照片被大量转发和“人肉搜索”，对其个人名誉产生了极恶劣的影响。

网络反腐低俗化、娱乐化倾向严重，给网络舆论环境带来负面影响。专家表示，当前，在部分党员干部中的确存在奢靡浪费、生活腐化堕落等腐败现象，但这仅是腐败现象的主要表现之一。实际上，腐败的表现形式还有徇私枉法、以权谋私、权钱交易、买官卖官等。但从近几年的网络反腐案例看，无论是经查证属实的案例，还是虚假的网络曝光，大多与“情妇”“二奶”“包养”等字眼联系在一起。究其原因，这样的“桃色新闻”更具眼球效应，更易引发舆论关注，更易形成社会倒逼，因而受到网民的热捧。

在处置突然爆发的互联网事件中，一些地方政府官员往往因为发现不及时，对其前因后果掌握不全导致危机处置严重滞后，错过了最佳处置时间，导致危机爆发，严重影响政府部门的工作形象。

1.2 互联网信息监测

互联网信息监测，又叫网络舆情监测，是对互联网传播的公众对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点进行监视和预测的一种行为。具体讲，互联网信息监测是指整合互联网信息采集技术及信息智能处理技术后，通过对互联网海量信息自动抓取、自动分类聚类、主题检测、专题聚焦，实现用户的网络舆情监测和新闻专题追踪等信息需求，形成简报、报告、图表等分析结果，为客户全面掌握群众思想动态，做出正确舆论引导，提供分析依据。

提及互联网信息监测，就不得不澄清监测与监控的区别。通俗来讲，监控指的是对装备及系统的工作状态不间断地实时监测，并根据反馈信息自动对系统中异常部位实施相应措施的闭合自动控制作用。而监测，则指对装备、系统或其一部分的工作正常性进行实时监视而采取的任何在线测试手段。简单明了地说就是：监控比监测多了一个控制操作。对设备、系统我们常常强调监控手段，而对互联网舆论情况，系统能做的其实更多是

采集和分析，具有对互联网动态信息的知情权，而非控制广大群众的言论。

1.3 手工监测的不足

政府、网警、企事业等全国大大小小很多部门，相继成立了专门负责网络舆论监测的机构，由专人组成若干小组，24 小时不间断地对重点网站、重点论坛进行监测。此外，对网络热点较集中的知名论坛、贴吧、微博，通过聘用“网络调研助理”等方式，密切关注网站动态。通过各层次、各领域建立起来的组织机构，从物质、制度、资金、人力等方面，保证了互联网信息监测体系的日常运作。但对于互联网信息监测与分析，这还远远不够。

互联网信息监测，首先是一个从互联网上获取信息的任务，属于“互联网搜索”的技术范畴。搜索引擎是迄今为止最为成功的互联网技术，能够在极短的时间内，根据关键词返回相应的查询结果。但它并不能满足互联网信息监测和分析的要求。一个典型的搜索引擎，是从互联网上探测并下载原始的文档资料，然后对这些资料进行去重、抽取、分词、索引、存储，最后提供全文检索服务。

（1）搜索的前提是“已知关键词”，对于监测问题，这个前提往往并不具备。比如，在得知“天价烟”这件事情之前是想不到“天价烟”这个词的，等想到这个词时再搜索，应对的条件可能已经丧失了。

（2）监测相当于“大海捞针”，但搜索引擎返回的结果动辄上千万个，还是一个“信息海洋”，而不是“针”，因此无法从中获得有针对性的信息。

（3）由于搜索引擎只能收录开放网站，大量需要登录才能访问的网站被排除在外，特别是论坛、微博等在用户工作业务中重点关注的内容，往往是搜索引擎的盲区。

（4）搜索引擎的收录时间长短不一，知名网站收录速度快（1 小时内），但地区性网站则收录较慢（24 ~ 72 小时）。

（5）搜索引擎只能提供结果集合的一部分，比如 Google 最多只能翻 100 页，百度也不超过 76 页（760 个结果），因此无法保证检索结果的完整性。

（6）搜索引擎无法自动化，即用户每次都需要手工录入关键词，而与一个部门、企业相关的关键词数量是非常多的，其关键词的两两组合数量更是惊人。依靠用户每次手工录入后，肉眼查看结果，并在众多冗长的结果中翻页，设法找出与上一次搜索的不同，不仅劳民伤财，而且极不准确。

人工监测与系统监测的对比如表 1 - 1 所示。

表 1 - 1　人工监测与系统监测对比表

比较指标	人工监测	采用互联网信息监测系统
目标网站	几十个	几千个几万个
人力成本	需分别登录各个网站，手工查阅，还要手工复制粘贴，疲于奔命	网络信息的获取工作完全由软件自动进行，监测人员只需集中进行内容的浏览与分析，遇到紧急情况，系统将直接给用户推送短信
敏感信息识别	需要人工逐条查看确认	在自动判别的基础上再人工确认
信息保存	零碎，不可避免会出错	精确、全面，便于事后追踪
数据存储	Word 文件，分散，很难管理	统一存放在大型关系数据库中，集中管理
监测报告	基于手工统计和估计，数据支持不充分	基于自动化的统计分析，图文并茂，具有翔实统计数据支持，可以每日、每周、每月出报告
监测效果	覆盖片面，不及时，浪费人力资源	覆盖全面，及时，几分钟到几十分钟自动化、系统化

1.4　互联网信息监测系统

由于互联网具有虚拟性、隐蔽性、发散性、渗透性和随意性等特点，

越来越多的网民乐意通过网络渠道来表达观点、传播思想。当今，信息传播与意见交互空前迅捷，网络舆论的表达诉求也日益多元化。对相关政府部门、企事业单位、社会名人、事件相关人来说，如何快速获知与己相关的网络信息，进而加强对网络舆论的及时监测、有效引导，以及对网络舆论危机的积极化解，对维护社会稳定、促进国家发展具有极其重要的现实意义。

不同企业的网络信息监测流程可能不完全一样，但大致都可分为以下三个部分。

（1）制定危机预警方案。针对各种类型的危机事件，制定比较详尽的判断标准和预警方案，以做到有所准备，一旦危机出现便有章可循、对症下药。

（2）密切关注事态发展。保持对事态的第一时间获知权，加强监测力度。可以通过自动化的信息监测技术，在第一时间大量采集、汇总各种互联网上的信息。

（3）及时传递和沟通信息。与舆论热点所涉及的政府相关部门保持紧密沟通，建立和运用这种信息沟通机制，已经成为网络舆情管理部门的重要经验。以上海为例，无论是在地铁调价，还是在普陀城管打人等“网络热点”问题的处理上，都能做到各部门协同作战、相互配合、共同商议，判断危机走向，对预案进行适当修正和调整，以符合实际所需，这是危机应对的重要措施。

网络舆论的联动应急机制，指政府管理部门及其他相关职能机构，对网络舆情尤其是负面舆情的监测预警与控制，从而有效化解网络舆论危机。联动应急机制往往包括监测、预警、应对三个环节。在监测环节，有关人员和系统对网络舆情的内容、走向、价值观等方面进行密切关注，将最新情况及时反映到有关部门。在预警环节，对内容进行判断和归纳，对正在形成、有可能产生更大范围影响的舆论进行筛选，为接下来可能发生的网络舆情走向做好各种应对准备。在应对环节，当网络舆情变为现实的

网络舆论危机事件后，有关政府部门应采取具体行动。这三个环节有机组合，从整体上构成了网络舆论联动应急机制。

就技术保障而言，要监测互联网信息，少不了及时有效的信息搜集、信息处理、信息研判、信息反馈、信息决策系统。对互联网信息的监测与分析必须要浏览和查找海量的网络信息，包括网络新闻报道、相关评论、网络论坛等，从这些信息中提取与事件相关的内容，然后分析信息的时间与空间分布情况。随着互联网技术的不断更新，网络信息的监测和分析有必要通过与之相匹配的科技手段来进行。

互联网信息监测系统是针对互联网这一新兴媒体，通过对海量网络舆论信息进行实时的自动采集、分析、汇总、监视，并识别其中的关键信息，及时通知到相关人员，从而为第一时间应急响应敏感信息提供帮助的信息化系统。

1.5 全文组织结构

本章介绍了互联网信息、信息监测以及信息监测系统的概念，理清了不同概念的内在差异和相互关系。同时，比较了手工监测方法与使用软件系统自动化监测方法的差别。

接下来的第 2 章，根据实际项目经验，归纳整理了一般企业用户在互联网信息监测上的功能性与非功能性需求，并以媒体监测、信息推送、咨询分析、监测规则管理为重点展开叙述。

第 3 章在常见企业用户需求的基础上，开展需求分析和软件系统的设计，给出了具体的解决方案，并列举了互联网信息采集、海量信息管理与检索、多维分析与机器挖掘等多个子系统的技术特点及设计要点。

第 4 章从数据处理流程的角度，解释 HTML 网页信息的抓取，网页正文提取，结构化内容解析，中文、英文、维文等多语言支持，以及文本中的命名实体识别等技术环节。