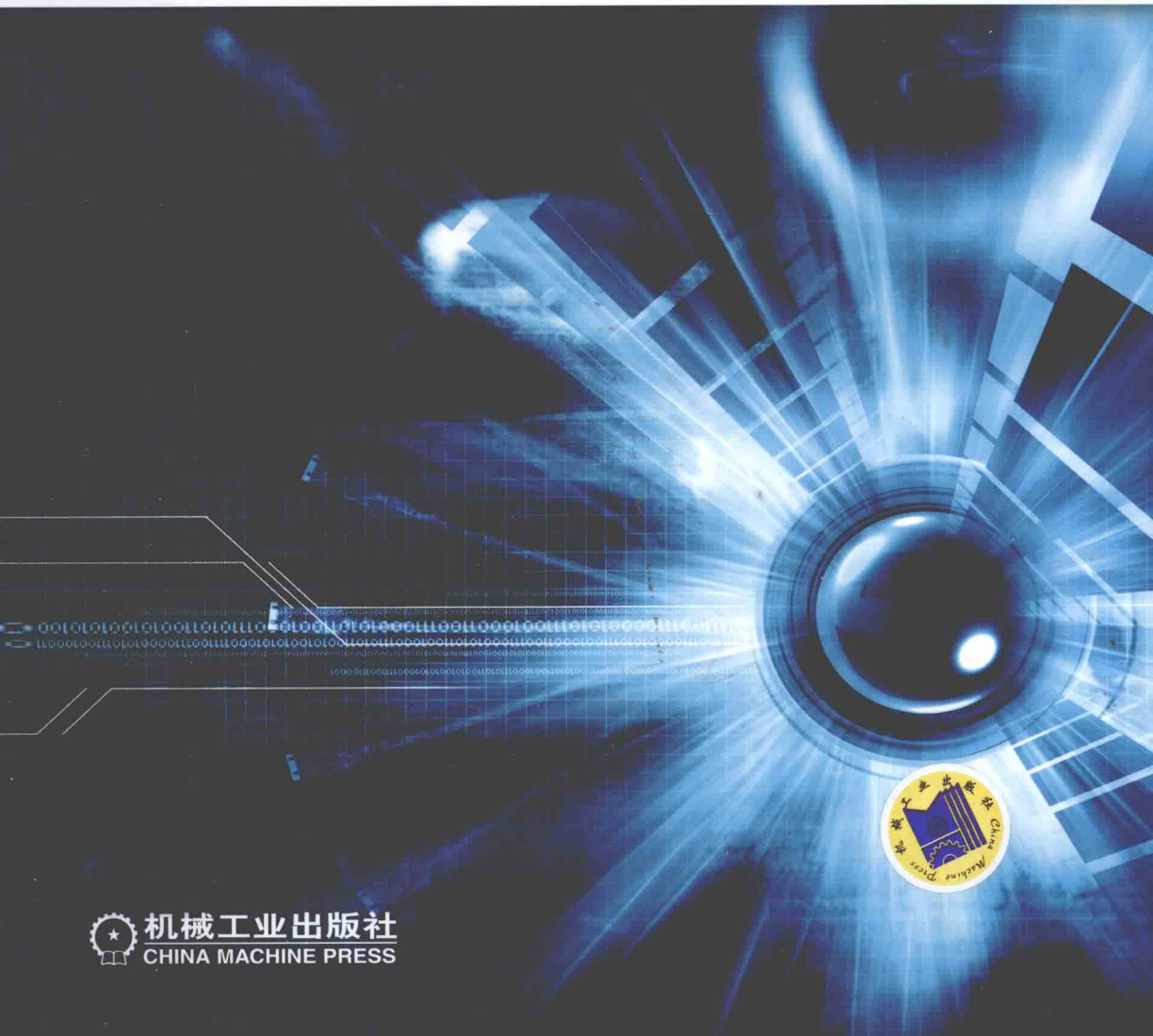
面向信息内容安全的新闻信息处理技术

杨伟杰 著



机械工业出版社
CHINA MACHINE PRESS

本书全面介绍了面向信息内容安全的网络新闻信息处理技术相关的基本概念、理论方法和最新研究进展。内容包括信息内容安全、新闻信息处理、自然语言处理、计算语义学、文本挖掘、信息过滤、话题检测与跟踪、社会网络分析、网络新闻评价、网络舆情分析、综合集成法等，既有对基础知识和理论模型的介绍，也有对相关问题的研究背景、实现方法和技术现状的详细阐述。

本书可作为高等院校计算机、信息技术等相关专业的高年级本科生的教材或参考书，也可供从事信息技术、数据挖掘、人工智能、管理科学、战略研究等相关领域研究的教师、研究生和科研工作者参考，借以提供思路和技术支撑。

图书在版编目（CIP）数据

面向信息内容安全的新闻信息处理技术/杨伟杰著. —北京：机械工业出版社，2011.3

ISBN 978-7-111-33166-7

Ⅰ.①面… Ⅱ.①杨… Ⅲ.①计算机网络－信息系统－安全技术 ②计算机网络－信息处理 Ⅳ.①TP393②G202

中国版本图书馆 CIP 数据核字（2011）第012179号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）

策划编辑：牛新国 责任编辑：牛新国

版式设计：张世琴 责任校对：肖 琳

封面设计：路恩中 责任印制：李 妍

北京诚信伟业印刷有限公司印刷

2011年4月第1版第1次印刷

169mm×239mm·13.5印张·261千字

0 001—2500册

标准书号：ISBN 978-7-111-33166-7

定价：39.80元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务 网络服务

社服务中心：（010）88361066 门户网：http://www.cmpbook.com

销 售 一 部：（010）68326294

销 售 二 部：（010）88379649 教材网：http://www.cmpedu.com

读者服务部：（010）88379203 **封面无防伪标均为盗版**

前　言

随着信息传播技术的迅猛发展，互联网成为不可忽视的舆论阵地，而互联网新闻作为一种重要的情报信息来源，成为网络舆论和社会舆论形成的主要源泉，因此，准确判断它的内容安全性，从而准确及时地把握社会舆论的动向变得尤为重要。但是由于新闻是一种典型的非结构化信息，同时互联网新闻具有无范围限制的特点，其内容安全性判断也变得相对复杂。本书针对这个问题，利用自然语言处理、数据挖掘等技术，对网络新闻进行分析，试图达到有效判断信息内容安全的目的。

本书主题是面向信息内容安全的网络新闻信息处理技术研究，包括网络新闻的分析、评价及其对网络舆情的影响，涉及自然语言处理、数据挖掘、网络安全、互联网管理等多个学科领域的交叉，处于学术领域研究的前沿，是当前研究的热点，并能够为以上领域的研究提供思路，为互联网管理提供科学有效的技术支撑。

本书内容共分为8章。

第1章　绪论。这一章首先介绍了信息内容安全的概念、产生背景及其对社会安全等方面的影响；然后介绍了网络新闻的特点及其内容安全的重要性；最后分析阐述了网络新闻信息内容安全的发展现状和应用前景。

第2章　网络新闻信息处理原理及相关技术。这一章首先提出了网络新闻信息处理的框架，然后针对框架中涉及的自然语言处理、计算语义学、文本挖掘等几个主要研究方向分别进行了详细介绍，包括经典理论、当前发展状况、最新提出的理论和技术，以及以后的发展方向和重点研究内容。

第3章　信息过滤。这一章介绍了信息过滤技术，包括信息过滤机制、信息过滤的模型和算法、信息过滤技术的实现，以及信息过滤在信息内容安全中的应用。

第4章　话题检测与跟踪。这一章介绍了新闻话题检测与跟踪的相关技术，首先提出了话题检测与跟踪的技术框架，然后介绍了当前在话题检测与跟踪方向提出的模型和算法，以及话题检测与跟踪技术的实现方法。

第5章　社会网络分析。这一章首先对社会网络相关的基础知识进行了介绍，然后提出了几种社会网络的构建方法和分析方法，最后讲述了社会网络在新闻信息内容安全方面的应用。

第6章　网络新闻信息的评价。这一章介绍了网络新闻信息的评价指标和几

种评价方法，以及新闻信息评价对新闻信息内容安全方面的作用。

第7章 网络舆情分析。这一章介绍了网络新闻与网络舆情分析的关系。首先阐述了网络舆情的概念与传播，然后归纳了网络舆情的搜索和收集方法，最后重点介绍了网络舆情分析的模型，网络舆情监控系统的体系结构，以及如何对网络舆情进行引导。

第8章 用综合集成法解决网络新闻分析系统的相关问题。这一章作为全书内容的一个总结和升华，根据复杂巨系统的定义，归纳了因特网（Internet）的系统学特性，提出了用系统学理论指导解决因特网的相关问题，并重点指出如何使用复杂巨系统理论为指导，从系统学角度考虑新闻分析系统相关问题的研究。

本书以国家863、973重点研究课题为背景，选题内容处于交叉学科的前沿，理论与实际相结合。本书的出版得到了北京工商大学“北京市属市管高校人才强校计划”项目的资助；作者在编写过程中，得到了中国科学院自动化研究所崔霞副研究员的指导与帮助，她对本书的篇章架构提出了很好的建议，在此表示衷心的感谢。由于本书内容处于交叉学科的前沿，很多问题没有取得共识，且涉及面广，作者水平有限，书中错漏之处在所难免，敬请广大读者批评指正。

作 者

2010年10月

目　录

第1章 绪 论

1.1 信息内容安全的概念和产生背景

不良信息传播与反传播的斗争一直伴随着人类文化发展的进程。近年来，互联网的飞速发展，尤其降低了信息发布和获取的门槛，使得这一斗争前所未有地凸显出来。网络以其前所未有的信息传播能力在给人们生活带来巨大便利的同时，也成为反动、色情、暴力等不良信息的载体。这些不良信息，尤其是有关国家安全的敏感信息借助于网络传播，成为一个危害极大的社会问题。从海量信息中迅速、有效地识别这类不良信息，进而阻止其非法传播，确保网上信息内容安全，已成为内容安全领域的重要研究课题，这对于维护社会稳定具有极其重要的意义。

1.1.1 我国互联网发展现状

根据中国互联网络信息中心（CNNIC）提供的数据[1]，截至2010年6月底，我国在网民人数与结构分布、互联网基础资源、上网条件以及网络应用等方面的情况都发生了巨大的变化，互联网已经凸显出其重要作用。

1. 网民规模与结构特征

我国网民数量增长迅速，截至2010年6月底已突破4亿关口，达到了4.2亿，较2009年底增加3600万人。互联网普及率攀升至31.8%，与2009年底相比提高了2.9%。互联网逐步向各层次的居民扩散。新增网民中，18岁以下的网民和30岁以上年龄较大的网民增长较快；初中及以下受教育程度的网民增长较快；低收入人群开始越来越多地接受互联网；农村上网人群增长较快。从接入方式上看，宽带网民数达到3.63亿人，手机网民数达到2.77亿人，这两种接入方式发展较快。除此以外，社会和政府还鼓励互联网往更广泛的群体渗透。从新增网民群体比重来看，互联网正逐步朝这一方向发展。

2. 互联网基础资源

互联网基础资源增长迅猛，资源结构有所调整。截至2010年6月，我国IPv4地址达到2.5亿，半年增幅7.7%。作为互联网上的“门牌号码”，IPv4地址资源正临近枯竭，互联网向IPv6网络的过渡势在必行。我国域名总数下降为1121万，其中“.cn”域名725万。“.cn”在域名总数中的占比从80%降至64.7%。与此同时，“.com”域名增加53.5万，比重从16.6%提升至29.6%。

网站数量下降到279万个，“.cn”下网站为205万个，占网站整体的73.7%。国际出口带宽达到998217Mbit/s，半年增长15.2%。

3. 上网条件

随着家庭使用电脑上网环境的不断改善，在家使用电脑上网的网民比例继续增加，达到88.4%，较2009年底提高了5.2%。在单位上网的比例上升到33.2%，在网吧上网的比例降至33.6%。手机上网方式也越来越多地被采用，所占比例攀升至65.9%。

4. 网络应用情况

平均上网时长继续增加，周平均上网时长达到19.8小时，增加1.1小时。上网时间延长，表明我国网民的网络使用深度在增加，网民对互联网有一定的依赖性。我国网民的互联网应用表现出商务化程度迅速提高、娱乐化倾向继续保持、沟通和信息工具价值加深的特点。2010年上半年，大部分网络应用在网民中更加普及，各类网络应用的用户规模持续扩大。其中，商务类应用表现尤其突出，网上支付、网络购物和网上银行半年用户增长率均在30%左右，远远超过其他类网络应用。社交网站、网络文学和搜索引擎用户数量增长也较快。电子商务应用的高速发展和娱乐社交类应用的较快增长，与我国互联网发展特点有关。我国电脑网民宽带普及率接近100%，青少年网民占整体网民数量的一半左右，中小企业电子商务应用呈普及化趋势。互联网作为全面的平台，成为了人们获取信息的常规来源、娱乐休闲的重要方式和商务交易的便捷渠道。现今我国各类网络使用所占比率和排名情况见表1-1。

表1-1 2009.12～2010.6 各类网络应用使用率及排名变化

类 型	应 用	2009.12 使用率	2010.06 使用率	2009.12 排名	2010.06 排名	排名变化
网络娱乐	网络音乐	83.5%	82.5%	1	1	→
信息获取	网络新闻	80.1%	78.5%	2	2	→
	搜索引擎	73.3%	76.3%	3	3	→
网络娱乐	网络游戏	68.9%	70.5%	5	5	→
	网络视频	62.6%	63.2%	6	6	→
	网络文学	42.3%	44.8%	10	10	→
交流沟通	电子邮件	56.8%	56.5%	8	7	↑
	即时通信	70.9%	72.4%	4	4	→
	博客应用	57.7%	55.1%	7	8	↓
	论坛/BBS	30.5%	31.5%	11	12	↓
	社交网站	45.8%	50.1%	9	9	→

（续）

类 型	应 用	2009.12 使用率	2010.06 使用率	2009.12 排名	2010.06 排名	排名变化
商务交易	网上支付	24.5%	30.5%	13	13	→
	网络购物	28.1%	33.8%	12	11	↑
	网上银行	24.5%	29.1%	14	14	→
	网络炒股	14.8%	15.0%	15	15	→
	旅行预订	7.9%	8.6%	16	16	→

5. 网络安全和可信度情况

据统计，半年内有59.2%的网民在使用互联网过程中遇到过病毒或木马攻击，遇到该类不安全事件的网民规模达到2.5亿人。

2010年上半年，有30.9%的网民账号或密码被盗过，网络安全的问题仍然制约着中国网民深层次的网络应用发展。

调查发现，89.2%的电子商务网站访问者担心访问假冒网站，而他们如果无法获得该网站进一步的确认信息，86.9%的人会选择退出交易。互联网向商务交易型应用的发展，急需建立更加可信、可靠的网络环境。

1.1.2 互联网上的不良信息问题

各种数据充分表明，互联网已经渗透到人类生活的方方面面，成为最主要的信息工具之一。然而，事物总是具有两面性的，当人们享受网络带来的种种便利的时候，也要看到网络巨大的信息传播能力潜在的危害性。反动、色情、暴力信息附身于网络，给人类正常文化生活造成了极坏的影响；垃圾信息充斥于网络，浪费了巨大的通信资源与用户时间；不法分子以网络为工具进行犯罪活动，严重影响了社会秩序；虚假信息、私密信息在网络上的泛滥，对社会公信力与社会道德规范产生了严重冲击；有关国家安全的敏感信息借助于网络进行传播，对国家与社会构成了极大的危害。2007年4月13日，公安部通过其网站公布了公安机关打击利用互联网违法犯罪活动的10个典型案例，主要涉及利用互联网传播淫秽物品、敲诈、赌博、诈骗、盗窃等方面。另有资料表明，非法团体组织活动的方式日益呈现多元化，除传统的传单、信件等方式外，电子、信息技术逐渐为这类组织所采用，其中互联网由于成本低、连接方便、覆盖范围广、信息发布门槛低等特点，成为首选方式之一。资料显示，全球计算机网络与信息安全问题十分突出，对各国政治、经济、国防、文化安全造成了很大威胁，大力加强计算机网络与信息安全工作，成为各国面临的紧迫任务。

对互联网上不良信息的监管和过滤就是在这个背景下产生的。该课题致力于

通过技术手段和政府监管相结合，制定符合我国国情的信息内容安全体系结构。所谓不良信息，是指违背社会主义精神文明建设要求、违背中华民族优良文化传统与习惯以及其他违背社会公德的各类信息，尤其包括违背《中华人民共和国宪法》和《全国人大常委会关于维护互联网安全的决定》、《互联网信息服务管理办法》所明文禁止的信息以及其他法律法规明文禁止传播的各类信息。

1.1.3 信息内容安全简述

信息内容安全是信息安全的一个重要分支。通常而言，信息安全以保障电子信息的有效性为目的，具体涉及信息的保密性（Confidentiality）、完整性（Integrity）、可用性（Availability）和可控性（Controllability）等方面。保密性是指对抗对手的被动攻击，保证信息不泄露给未经授权的人；完整性是指对抗对手主动攻击，防止信息被未经授权的篡改；可用性是指保证信息及信息系统确实为授权使用者所用；可控性是指对信息及信息系统实施安全监控。从应用的角度来讲，信息安全涉及信息传输的安全、信息存储的安全以及对网络传输信息内容的审计三方面。一切信息安全技术的最终目的是保障信息得以正常应用，保障网上的信息确实能够为文化、经济的发展以及社会稳定起到一定的促进作用，信息内容的审计成为一切信息安全技术得以发挥作用的最终保障。立足于这个视角，提出信息内容安全的概念是非常必要的。具体来讲，信息内容安全技术包括如下几个方面[2]：

（1）信息域的定义、划分，与不同信息域之间的信息隔离与安全交换技术，解决网络环境下不同信息域之间的信息隔离与安全交换等问题；

（2）信息内容截取与还原技术和信息阻断技术；

（3）信息内容的识别技术和智能信息内容分析方法，包括概念词典构造、语义关系与框架槽之间的映射关系、基于概念扩充的文本/模板匹配技术、基于语义分析的细选过滤技术等方面；

（4）信息内容监控技术，包括基于内容分级的标记与监管技术；

（5）信息隐藏技术，包括其他信息隐藏技术的实用化研究；

（6）安全浏览器技术，包括安全通信协议的深入剖析，实现高加密强度安全。

1.2 网络新闻的特点

新闻报道主要涉及每天发生的重要事件，其取材多样并且涉及面广，属于半结构化数据。新闻报道由于具有如下特征使得它的可利用价值远远超出了浏览与检索的范畴，综合考虑新闻报道，其主要特性如下[3]：

（1）作为一种公开的信息源，新闻报道，尤其是专题性新闻文本容易获取，并且具有报道及时、反应迅速和内容丰富等特点。

（2）新闻报道代表了不同国家、不同政治团体的政治立场和媒体呼声，能够反映其政治、外交和军事等不同领域的政策和态度。

（3）传播者创制新闻报道的目的在于为人们传播明确的事实信息，因而要求信息要置于明晰的编码之中，文本的意义不能过于依赖语境，也不能依赖于言外之意或字里行间的表达，以避免理解的多义和歧义。

（4）新闻报道从意义解释的角度讲基本上属于封闭性系统。新闻报道的文本是由一系列明确的事实判断语句构成的，从原则上排除意见和情感的主观表达，对开放性的理解形成了语义上的限制。

（5）文本形式的新闻报道结构相对简单，主要表现在以下几个方面：其一，新闻文本的结构类型相对单一，不像小说、散文等文本的具体结构形式可以丰富多样；其二，新闻文本的结构要素（主要是时间、地点、人物、机构等）稳定明确，它们支撑起新闻文本相对稳定的构架；其三，新闻文本的叙事结构也相对简单，大多数新闻文本的主体内容采用与新闻事实客观结构相一致的方式展开。新闻文本与新闻事实逻辑上的同构性，加上新闻传播主体再现新闻事实时的合理简化和必要提炼，会进一步增强新闻文本叙述结构的自然性和简明性。

（6）新闻报道的语言必须具有明确性，只有按照“准确、准确、再准确”的要求去做，才有可能使新闻事实的完整面貌得到准确的呈现。构成新闻报道的语言本质上具有传真性、写实性、再现性和记录性等诸多特征。

（7）新闻报道涉及面广、题材多，其内容更新快，新词的出现让人应接不暇，尤其在科技和财经领域。另外，人名、地名、机构名等新名词也层出不穷。

以上总结了新闻报道的主要特性，事实上，相对于传统新闻报道，网络新闻报道又有其自身的特点。

（1）互动性：将传统媒体与受众的单向传播关系转变为双向或多向互动的传播关系。网络新闻工作者可以通过新闻留言板、电子邮件、网络聊天、BBS 等方式实现双方信息共享，受众可以拥有更多搜寻、反馈的能力，一些网友将自己制作的 DV、图片在网上发布，甚至能影响到整个新闻事件的报道。

（2）网络新闻制作无范围限制，打破了传统新闻媒体在时空上的限制。

（3）网络新闻媒体的版面呈现方式不同，以图片、新闻标题与导航为主，并实现了多媒体整合运作。而且网络新闻没有篇幅和数量的限制。

（4）专题报道是网络新闻赢得受众的重要手段。而且网络媒体提供的数字化语言、文字、声音、图像信息和非线性互动传播，为深度报道的发展提供了更为广阔的天地。

从新闻的特点可以看出，作为一种信息传播的方式，新闻会对社会稳定产生

很大的影响。新闻舆论监督的勃兴，肇始于美国大法官斯特瓦特创设的“第四权力理论”，所谓的“第四权力”就是指新闻舆论。事实上，它不是国家权力，但随着新闻媒体在社会政治、经济、文化生活中的作用日益增强，它发挥着越来越重要的作用。同时，随着网络媒体“议程设置”功能的减弱和“沉默的螺旋”作用的不断增强，网络新闻作为网络舆论和社会舆论形成的主要源泉，因而准确判断新闻信息内容安全性对社会安全及其他相关方面具有重要意义[4]。

1.3 网络新闻信息内容安全分析技术的发展与现状

网络信息内容安全分析是一个综合性的概念，需要研究的技术相当多。目前还没有发展成为一个专门的学科，该项研究涉及的相关领域有信息安全、自然语言处理、网络理论、人工智能、机器学习、模式识别等。此外，基于文本挖掘、知识库、内容理解等方面的研究也非常多。迄今为止，在信息内容安全分析技术的研究及应用主要集中在不良信息监测、信息过滤，以及专门针对新闻信息的话题检测与跟踪等方面。

1.3.1 不良信息监测技术

根据中国互联网信息中心提供的数据，中国的互联网页面从内容上看，仍是文本居多，占到网页总数的85%以上，其次是图像，音频和视频网页数量相对比例仍旧不高。因此，不良信息监测相关的研究大多集中于文本信息的监测，本书中所指的不良信息也主要是文本形式的信息。归纳起来，不良信息监测技术研究主要分为以下几类：

1. 网络处理协议及体系结构研究

目前相关的研究大多集中在网关或用户端的信息过滤与自动屏蔽上，通常基于信息过滤技术。信息过滤系统中对信息源数据的获取往往采用网络监听的方法。网络底层信息监听可以采取两种方法：一是利用以太网的广播特性实现，二是通过设置路由器的监听端口实现。在这一方面，曲建华于2003年进行了Web上的信息过滤问题研究[5]；文自勇于2005年进行了分布式网络监听系统研究与实现[6]；郑海春于2003年进行了网络监听技术的研究与应用[7]。网络监听作为信息监测领域一个较成熟的手段，目前对于这方面的研究仍然占很大比重。

为进一步提高内容分析系统的处理能力和加快响应时间，谭建龙提出了扁平结构的网络内容分析模型[8]，其主要思想是把各个协议层的数据处理函数集中到一个层次中，从而减少内存访问次数，便于协议自动实现。对任何一个数据包，各个分析层尽可能地进行处理，包括尽可能早地执行关键词匹配，尽可能早地发现匹配规则，从而尽可能早地执行响应动作。

但是，采用网络底层的监听技术，需要对已有网络进行较大规模的改动。这种技术成本高、灵活性差，对监测点的选择提出了较高的要求，很难有效地应对不良信息传播者的“游击”策略。同时，该方法对于在网络用户端进行信息过滤有较大优势，而不适合本文所针对的应用需求。

2. 面向不良信息的文本分类研究

文本分类是实现不良信息监测的关键技术，目前在这方面的研究较多，是信息内容安全领域所关注的一个重点。

熊静娴、李生红在模糊集和语义网络的理论基础上，通过构建模糊值动态约束性概念网络，进行了面向不良文本信息监控的概念网技术研究，提出基于概念网络的文本分析算法[9]。黄海英、林士敏、严小卫也进行了基于概念空间的文本分类研究[10]，提出基于概念空间的文本分类机制，表现出明显的性能优势。郭莉、张吉、谭建龙提出一种基于后缀树的文本向量空间模型（VSM），并在此模型之上实现了文本分类系统[11]。对比基于词的 VSM，该模型利用后缀树的快速匹配，实时获得文本的向量表示，不需要对文本进行分词、特征抽取等复杂计算，同时还能保证训练集中文本更改能够对分类结果产生实时影响，具有较好的时间复杂度。分类过程和语种无关，是一种独立语种的分类方法。万中英、王明文、廖海波以提高分类精度为目的，提出一种基于投影寻踪（Projection Pursuit，PP）的中文网页分类算法[12]。他们首先利用遗传算法找到一个最好的投影方向，然后将已被表示成 n 维向量的网页投影到一维空间，最后采用 KNN 算法进行分类，能够有效解决“维数灾难”问题。林鸿飞、姚天顺提出基于示例的中文文本过滤模型[13]，首先对于用户提出的示例文本进行文本结构分析，采用文本层次的方法提取文本特征，形成主题词表示的用户模板，然后进行文本过滤；同时在用户反馈的基础上扩充示例文本数量，进而采用基于潜在语义标注的文本过滤方法，改进用户模板，提高过滤效率。樊兴华、孙茂松采用两步分类策略，提出一种高性能两类中文文本分类方法[14]，首先以词性为动词、名词、形容词或副词的词语为特征，然后将文本看做由词性为动词或名词的词语构成的序列，以该序列中相邻两个词语构成的二元词语串作为特征，以改进互信息公式来选择特征，以朴素贝叶斯分类器进行分类。该两步分类方法达到了较高的分类性能。卢军、卢显良、韩宏、任立勇针对网络信息的实时过滤问题，提出一种基于代理服务器的网络信息实时过滤机制[15]。为提高信息过滤的性能，还提出一种高效的关键词集合匹配方法（KPSMM），该方法可以实现关键词集合的高效过滤，其性能比传统的字符串过滤方法有较大提高。

此外，基于决策树（Decision Tree）、粗糙集（Rough Set）[16]、Ripper 方法[17]、Boosting 方法[18]以及 k 邻近（KNN）方法[19]、贝叶斯（Bayes）方法[20]、Rocchio 方法、支持向量机（SVM）[21]等的研究相当多。

3. 不良信息特征提取研究

文本特征的表示与特征提取是分类算法的基础与前提，在以上列举的文本分类算法中都有提及。但由于不同领域信息的形式特殊性，许多研究者也对特征提取进行了专门研究。

陈文亮、朱靖波、朱慕华、姚天顺以提高分类性能为目的，提出了一种结合机器学习和领域词典的文本特征表示方法[22]。他们利用了基于领域词典的文本特征表示方法增强文本特征的表示能力并降低文本特征空间维数；同时又提出一种自划分模型以解决领域词典存在覆盖度不足的问题，在特征数目较少的情况下，该方法表现出很好的分类性能。为解决分词给分类系统带来的消极影响，胡吉祥、许洪波、刘悦、程学旗提出了一种基于重复串的特征提取方法[23]，该方法无需分词便可以从文本中提取有意义的重复串作为特征，能降低特征空间维数，同时可有效改善传统以词为特征的聚类算法的性能。

1.3.2 面向信息内容安全的文本过滤技术

文本自动过滤技术是信息检索领域的重要研究课题，在大规模文本信息处理中具有很重要的意义。从信息处理的角度上看，文本过滤有如下几个应用领域[6]：

（1）提供选择性信息服务的企事业单位可以根据用户的信息需求过滤新闻信息，并且把用户可能感兴趣的内容发送给用户。这类似于图书馆和科技情报机构等提供的定题服务。

（2）在档案管理领域，文本过滤系统可自动地确定档案所属的类别。

（3）对终端用户而言，可以用具有文本过滤功能的代理程序来接收原始文本流（如 E - mail 和 Newsgroup），并从中选择用户可能感兴趣的内容。

（4）研究与开发具有自主版权的信息过滤系统，对于提高我国的网络和人工智能的研究和应用水平、保障国家信息安全、促进因特网技术在我国的健康发展也有着重要的意义。

文本过滤随着计算机应用的发展而从设想成为现实，并不断地完善自身的功能，经历了很长的发展时期，并在因特网日益普及的今天，在信息发掘方面发挥着越来越大的作用[24,25]。

1958 年，Luhn 提出了“商业智能机器”的设想，在这个概念框架中，图书馆工作人员建立了每个用户的需求模型，然后通过精确匹配的文本选择方法，为每个用户产生一个符合用户信息需求的新文本清单。这个设想为文本过滤的发展提供了有效的启发。

1969 年，美国信息科学协会进行了对 SDI（Selective Dissemination of Information，选择性信息分发系统）的研究。但是研究大都遵循 Luhn 模型，只有很少

的系统能够自动更新用户需求模型，其他大多数系统仍然依靠专门的技术人员或者由用户自己维护。SDI 兴起的两个主要的原因是实时电子文本的可用性和用户需求模型与文本匹配计算的可实现性。

1982 年，Denning 提出了“信息过滤”的概念，他的目的在于拓宽传统的信息生成与信息收集的讨论范围。他描述了一个信息过滤的需求的例子，对于实时的电子邮件，利用过滤机制，识别出紧急的邮件和一般例行邮件。他采用了一个“内容过滤器”来实现过滤。其中采用的主要技术有层次组织的邮箱、独立的私人邮箱、特殊的传输机制、阈值接收、资格验证等。

1987 年，Malone 等人发表较有影响的论文，并且研制了系统“Information Lens”。提出了三种信息选择模式，即认知、经济、社会。所谓的认知模式相当于 Denning 的“内容过滤器”，即基于内容的过滤（Content - based Filtering）；经济模式来自于 Denning 的“阈值接收”思想；社会模式是他最重要的贡献，目前也称为“合作过滤”。在社会过滤中，文本的表示是基于以前读者对于文本的标注，通过交换信息，自动识别具有共同兴趣的团体。

1989 年，信息过滤获得了大规模的政府赞助。由美国 DARPA 资助的“Message Understanding Conference”，极大地推动了信息过滤的发展。他将信息抽取技术用于信息的选择，在将自然语言处理技术引入文本过滤研究方面进行了积极的探索。1990 年，DARPA 建立了 TIPSTER 计划，目的在于利用统计技术进行消息预选，然后再进行复杂的自然语言处理，这个文本预选过程称之为“文本检测”。

20 世纪 90 年代以来，情况有了很大的改变，著名的文本检索会议（Text Retrieval Conference，TREC）和主题检测和跟踪会议（Topic Detection and Tracking，TDT）都把文本过滤作为主要研究内容之一，这就在很大程度上促进了文本过滤的发展。下面将着重介绍文本检索会议及其在文本过滤方面所做的工作。

文本检索会议，是由美国国家标准和技术局（National Institute of Standards and Technology，NIST）和国防部高级研究计划局（Defense Advanced Research Projects Agency，DARPA）组织召开的一年一度的国际会议，从 1992 年至今已经召开了 12 次，是文本检索领域最权威的国际会议之一，代表了当今世界文本检索领域的最高水平。

TREC 会议的宗旨主要有三条：通过提供规范的大规模语料（GB 级）和对文本检索系统性能的客观、公正的评测，来促进技术的交流、发展和产业化；促进政府部门、学术界、工业界间的交流和合作，加速技术的产业化；发展对文本检索系统的评测技术。

1.3.3 新闻话题检测与跟踪技术

新闻话题检测与跟踪又称为事件探测与跟踪（Topic Detection and Tracking，