JIYU YUYI KUOZHAN DE WANGLUO XINXI JIANGUAN

基于语义扩展的网络信息监管

张茂元 著

华中师范大学出版社

基于语义扩展的网络信息监管

张茂元　著

华中师范大学出版社

新出图证(鄂)字 10 号

图书在版编目(CIP)数据

基于语义扩展的网络信息监管/张茂元著. —武汉:华中师范大学出版社,2012.3

ISBN 978-7-5622-5269-6

Ⅰ.①基… Ⅱ.①张… Ⅲ.①计算机网络管理:信息管理—研究 Ⅳ.①TP393.07

中国版本图书馆 CIP 数据核字(2011)第 238312 号

基于语义扩展的网络信息监管

选题策划:第二编辑室

责任编辑:罗 挺 **责任校对**:刘 峥 **封面设计**:罗明波

编辑室:第二编辑室 **电话**:027-67867362

出版发行:华中师范大学出版社

社址:湖北省武汉市珞喻路 152 号

电话:027—67863426(发行部) 027—67861321(邮购)

传真:027—67863291

网址:http://www.ccnupress.com

电子信箱:hscbs@public.wh.hb.cn

印刷:武汉湖印印务有限责任公司 **督印**:章光琼

字数:110 千字

开本:880mm×1230mm 1/32 **印张**:5.25

版次:2012 年 3 月第 1 版 **印次**:2012 年 3 月第 1 次印刷

定价:18.00 元

欢迎上网查询、购书

前言

互联网迅猛发展，网络上的各种资源和信息异常丰富，互联网正在成为一种不可缺少的信息传播媒体。然而，不良信息、无用信息的传播和泛滥给传统的人工监督手段带来了难题，导致互联网信息内容的健康性问题变得日益突出。

网络信息资源呈现数量庞大、表现形式多样化、传播速度快、信息质量良莠不齐等特点。在表现形式方面，信息都是以字、词或者词组的形式来表示的，因而只有当查询词出现在文档中时，搜索引擎才有可能检索到包含该查询词的文档。

目前搜索引擎要求用户掌握一定的检索技巧，以词组或短语形式表达检索请求，这给普通用户使用搜索引擎造成一定障碍。一方面，即使文档中的词与输入的查询词表达相同概念，也会因为字面形式不匹配而无法被检索到，如“母亲”与“妈妈”；另一方面，如果文档中存在着错词，就会因为错词与正确词的形式不匹配而无法被检索到，如“适应症”与“适应正”。因此，在实现支持自然语言理解的信息检索之前，如何探索语义扩展检索技术来解决不拘泥于字面形式和不拘泥于错词形式的扩展检索问题，我们却知之甚少。

本书在查阅相关参考文献和分析国内外有关信息监管模

式及技术的基础上，围绕基于语义扩展的网络信息监管展开讨论，从网络信息监管机制、网页分类、概念扩展、纠错扩展、主动机制等方面进行研究。这些研究工作试图把信息学科的语义扩展技术有机地融入到社会学科的网络信息监管机制中，从概念扩展和纠错扩展的独特角度探索语义搜索技术，来解决不拘泥于字面形式和不拘泥于错词形式的扩展检索问题，以支持基于语义的网络信息监管机制。

本书的顺利出版，得到华中师范大学出版社的领导和编辑的热心支持，他们的辛勤工作有效地提升了该书的内容质量。同时，作者也对所有支持本书出版的人士表示由衷的感谢。

鉴于本书是对网络信息监管技术研究的一次尝试，而作者的专业水平和实践能力都十分有限，因此书中难免存在不妥之处，敬请广大读者批评指正。

本书可供从事网络信息管理、信息检索等领域的学者、科技人员阅读和参考，以及作为大专院校相关专业研究生教学参考用书。

作者

2011 年 10 月

目 录

1 绪 论

1.1 问题背景

与报纸、广播、电视等传统媒体有所不同，互联网是以文字、声音、图像、视频等形式来传播信息的一种数字化传播媒体，可为人们提供丰富的信息资源。与此同时，由于缺乏有效的网络监督手段，不良信息的传播和无用信息的泛滥导致互联网上信息内容的健康性问题变得日益突出[1]。例如，某网站发布广告称能生产具有防非典飞沫、防紫外线的防护眼镜，经浙江台州工商部门检查为虚假广告[2]，又如网络上充斥着有关色情、暴力等有害信息[3]。

网络信息监管的目的在于对网络信息的内容进行自动审查和过滤，防范不良信息和有害信息的随意传播。网络信息资源具有以下特点：

(1) 数量庞大，增长迅速

第27次中国互联网络发展状况统计报告指出，自2003年开始，中国的网页规模基本保持翻番增长，2010年网页数量达到600亿个，年增长率78.6%。由此可见，网络信息资源的增长速度是其他信息资源无法比拟的。

(2) 表现形式多样,覆盖面广

互联网资源多种多样,不仅有传统媒体的文字、图像信息资源,而且有音频、视频等多种类型的信息资源。这些资源涵盖了不同学科、不同领域,例如生活、科学、教育、健康、娱乐以及其他社会实践活动的各个方面。这些资源可以是单种语言描述的,如英语或汉语,也可以是跨语言描述的,如带有中文字幕的英语电影。

(3) 传播速度快,共享程度高

以光纤为骨干的大容量、高速数据传输系统,使网络信息资源传输速度变得极快。不仅如此,无线网络的带宽也比以前提高了许多,网络用户可以很方便地通过手机把信息快速地传到网络上。例如,2011 年李娜夺得法网冠军后,直播赛况的新浪体育第一时间将消息发布到微博上,这个微博在 5 分钟内转发超过 2 万次,10 分钟内转发超过 5 万次,刷新了单条微博转发速度纪录。

(4) 具有匿名性

网络的虚拟环境允许用户采用匿名方式发布和传播网络信息。工业和信息化部部长李毅中在 2010 年经贸形势报告会上作了"当前我国工业发展的若干重大问题"专题报告。该报告指出,目前,世界大多数国家的手机、网络都采用实名制,但在我国还缺乏法律依据。

(5) 信息质量良莠不齐

信息发布具有很大的自由性和任意性,缺少规范性,从而使网络信息的质量难以保证,呈现出良莠不齐的特点。

上述 5 个特点中:①特点(5)使网络信息中虚假信息、有害信息得以存在;②特点(3)使这些虚假信息、有害信息能快速地被传播,甚至引起多米诺骨牌效应;③特点(4)使

网络警察很难应用传统的侦破技术来定位网络信息的源头和终端；④特点(1)使监管部门靠人海战术来应付海量信息是不现实的；⑤特点(2)使目前尚不具有基于语义的检索功能的搜索引擎不能有效解决信息内容监管问题。

为了监管网络信息，政府出台了许多信息监管规定。例如，中国 2000 年出台了《互联网信息服务管理办法》(国务院令第 292 号)、2005 年出台了《非经营性互联网信息服务备案管理办法》(信息产业部令第 33 号)、《互联网新闻信息服务管理规定》。互联网协会也采取了相关措施，例如，让从业机构签署互联网行业治理公约。

尽管如此，面对着一个特殊的虚拟世界，这些规定和措施很难得以实施和生效。监管之难，难在相关措施的可操作性差。其原因主要是技术水平不足以及社会诚信体系不健全等因素，使监管部门无法针对网络信息的上述 5 个特点进行有效的监督和管理，从而给这些措施的“落地”带来很大困难。

要想有效地解决网络信息监管问题，监管技术就需要针对上述提到的网络信息特点来改进。①针对特点(3)和特点(5)，监管部门可以通过改进政府干预、行业自律、社会监督等机制来降低有害信息的数量和危害程度；②针对特点(4)，监管部门可以通过政府干预机制逐步实施网络实名制，并采用信息处理技术分析和跟踪网络信息的来源；③针对特点(1)，监管部门可以应用和改进网页分类等信息处理技术来处理海量信息；④针对特点(2)，鉴于目前的检索系统都是基于关键词匹配的检索，不能理解查询条件的语义而进行基于语义的搜索，监管部门非常有必要研究基于语义的信息搜索技术来实现网络信息内容的有效监管。

在当前几乎所有的检索系统中，信息都是以字、词或者

词组的形式来表示的，因此查询词出现在文档中时才有可能被检索到。一方面，在自然语言里同一个概念经常会有多种不同的表达方式，如“母亲”和“妈妈”，因而在相关文档中的词与输入的查询词表达相同概念，但因为在词形上不匹配而无法被检索到。另一方面，在网页、文本等数据源中存在错词，如“像棋”和“象棋”，这些文档会因为错词形式而无法被检索到。因此，讨论概念扩展方法来解决不拘泥于字面形式的信息检索，以及讨论纠错扩展方法来解决不拘泥于错词形式的信息检索，这对基于语义的信息搜索是非常有必要的。

因此，本书从概念扩展、纠错扩展等多个角度，讨论语义扩展技术，为解决不拘泥于字面形式和错词形式的信息检索做一些基础性工作，从而为给信息监管提供一个基于语义的信息搜索支撑技术做一些铺垫工作。

1.2 国内外研究现状

国内外现有的网络监管基本模式是：政府管理＋行业自律＋社会监督＋技术对抗。政府管理指通过立法，由相关行政管理部门来实施监管；行业自律指通过业内协会等组织共同订立守则来自我规范行为；社会监督指通过媒体、公众等社会力量的积极参与来维护互联网的信息传播秩序；技术对抗是指为了实现有效监管还需依靠“以技术对抗技术”，针对网络的高科技犯罪，采用信息搜索、反病毒技术、反入侵技术、信息过滤技术等来实施对网络违法行为的监控。

1.2.1 监管方式

(1) 美国

在互联网管理法规的数量上,美国以 130 多项法规居世界之首。美国政府于 1986 年通过《计算机欺诈和滥用法》,并相继于 1988 年、1994 年、1996 年、2001 年、2002 年和 2008 年进行过 6 次修订。该法案不仅惩治任何已经实施网络犯罪的人,也处罚"密谋策划网络犯罪"的人。

美国奥巴马政府从国家最高层面维护网络安全的战略角度,于 2009 年成立了"白宫网络安全办公室"、"全国通信与网络安全控制联合协调中心"。前者是直接对国家安全委员会和美国总统负责的网络安全机构,凌驾于军队和政府情报部门之上,负责统筹全国网络安全事务。后者协调和整合 6 大网络安全专职机构的信息,分析并上报全国网络空间的运行状况。

在行业自律方面,美国政府于 1998 年出台《2000 年数字著作权法》,提倡在联邦与州立法的总体框架内,由互联网运营企业自行制定自律性的监管规定。国际环球联合会要求世界各互联网信息发布机构、服务机构、监控机构对互联网上的相关信息进行分类标记,并推行采用互联网监控软件标记和审查网络信息。

(2) 欧洲

面对网络的有害信息,欧洲国家也毫不手软。欧盟委员会于 1996 年通过了《互联网有害和违法信息通信》、《在新的电子信息服务环境中保护未成年人和人的尊严》两本绿皮书,推动了欧洲国家的网络信息监管工作。

法国积极推广互联网以推动社会和经济的发展,并出

台相关法律。政府于1998年出台《未成年人保护法》,加重惩罚诱惑青少年网络犯罪的行为;于2006年颁布了《信息社会法案》,以加强对互联网的"共同调控";而且,从2004年起要求所有学校都在其网站上链接了涉及淫秽及种族歧视的"黑名单",来保护学生使其不受不良网站的侵害。

英国探索出一种有效的互联网管理的行业自律模式。1996年成立了"互联网监看基金会",其主要工作是处理各种不良信息报告,通知网络服务提供商删除非法内容,并将有严重情况的问题移交给执法机构。英国主管网络产业的国务大臣艾德·韦泽在互联网监看基金会的2010年年报发布仪式上表示,他和政府中一些官员非常欣赏这种行业自律模式。

俄罗斯通过立法清除网络违法信息,以加强保护未成年人。2010年底,俄罗斯立法通过了《保护青少年免受对其健康和发展有害的信息干扰法》草案,该草案规定所有的上网电脑必须设置内容分级系统,以防范青少年浏览色情、暴力等网页。

(3) 日本

日本是互联网高度普及的国家之一,其主要通过出台法律和法规来加大互联网监管的力度。政府相继颁布了《规范互联网服务商责任法》、《打击利用交友网站引诱未成年人法》、《青少年安全上网环境整备法》和《规范电子邮件法》等法律和法规。

(4) 中国

中国已成为全球经济发展的重要市场之一,其互联网发展同样非常迅速。中国在推动互联网发展的同时,也积极建构良好的互联网信息环境。2006年,中共中央办公

厅、国务院办公厅颁布了《2006—2020年国家信息化发展战略》。该战略指出，“坚持法律、经济、技术手段与必要的行政手段相结合，构建政府、企业、行业协会和公民相互配合、相互协作、权利与义务对等的治理机制，营造积极健康的互联网发展环境”。

中国政府先后出台了《中华人民共和国电信条例》、《非经营性互联网信息服务备案管理办法》、《互联网新闻信息服务管理规定》等法律和法规。2010年，中国全国人大常委会通过了修改后的《国家保密法》，首次把互联网纳入监管范围。

在行业自律方面，中国互联网协会于2001年成立，积极协助政府推动互联网的治理工作。2002年，协会制定并发布中国互联网行业治理公约，还开展自查互查活动，重点进行公约执行情况的检查。2004年协会公布了《中国互联网协会公共电子邮件服务规范》，该规范被国际电信联盟翻译并转载在其网站上。2008年协会组建了12321网络不良与垃圾信息举报受理中心，该中心为中国退出全球垃圾邮件发送国的“TOP 10”作了很大的贡献。

国务院新闻办网络局副局长刘正荣指出，中国政府对互联网的监管还处在不断摸索与完善的阶段。他说，“互联网一直在快速发展之中，网络技术的演变日新月异，政府只有不断总结经验、完善相关立法和治理措施，才能真正确保互联网的安全”。

1.2.2 监管技术

目前国内外用于互联网信息监管的技术主要有：

(1) 信息搜索技术