学校编码：10384　　分类号_____密级_____
学号：X2012230409　　UDC_____

厦门大学

工　程　硕　士　学　位　论　文

网络舆情监控系统的设计与实现

Design and Implementation of Network Public Opinion Monitoring System

乔　涓

指导教师姓名：　廖 明 宏　教 授
专 业 名 称：　软　件　工　程
论文提交日期：　2016 年 1 月
论文答辩日期：　2016 年 2 月
学位授予日期：　2016 年 6 月

指　导　教　师：_____
答辩委员会主席：_____

2016 年 1 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

当今信息技术发展迅速，网络已经成为人们发布、传递和获得信息的主要途径。网民可以通过论坛、博客、微博等交流平台发表自己的见解。网络给人们的生活带来了极大的便捷，甚至改变着人们的生活方式。与此同时，网络信息的快速传播也带来了很多负面影响，人们发表信息的随意性、信息没有经过核实就直接在网络中传播，对社会稳定产生影响。因此，监测复杂网络环境中的舆情，为相关部门提供充裕的应对时间,具有十分重要的理论和应用价值。

本文以在线社会网络为研究对象，通过采用网络爬虫实现对微博、博客、轻博客的数据获取。在此基础上分析在线社会网络节点特征与舆情传播规律之间的关系，得出重要用户节点对舆情传播有着极其重要的作用。以此作为理论基础筛选在线社会网络之中的重要用户节点，并对此进行监测，从而实现在线社会网络中舆情的监测。系统主要采用 B/S 架构体系，主要用 Java 语言进行开发，通过 XML 技术完成系统的配置，系统数据库主要采用 SqlSever 关系型数据库，HTTP 服务主要用于连接数据库，同时，为了提高系统性能，通过连接池、对象缓存等多种技术实现；通信协议通过 TCP/IP、HTTP 实现，并采用 Web Service 协议格式作为调用接口，系统通过扩展对象插件与内嵌浏览器技术实现接口集成，网页界面布局主要通过 HTML 一级 Jquery 框架集成技术实现。本论文的工作能够为用户行为模式分析、在线社会网络信息传播动力学、在线社会网络舆情引导以及异构在线社会网络信息传播模式等提供一定的理论基础和研究方法支持。

关键词：在线社会网络；网络舆情监测；网络爬虫

Abstract

Today's information technology is developing rapidly, the network has become the publishing, delivery and access to information, the main way. Internet users can express their opinions through forums, blog , microblogging platform. Network to people's lives has brought great convenience , and even change the way people live . At the same time , the rapid spread of information networks has also brought many negative effects , it is published arbitrary information on the dissemination of information has not been verified directly in the network, have an impact on social stability. Therefore , the complex network environment to monitor public opinion , providing ample time to deal with a very important theoretical and practical value for the relevant departments .

In this paper, an online social network for the study, through the use of web crawler data on the micro -blog, blog , blog access to light . Analyze the relationship between online social network node characteristics and propagation of public opinion between this basis, draw important user node spread of public opinion has an extremely important role. As a theoretical basis for screening online social networks among the major user node, and this monitoring , in order to achieve online social networks to monitor public opinion . Mainly adopts B/S architecture system, mainly in the Java language development, through the XML technology to complete system configuration, system database mainly adopts SqlSever relational database, the HTTP service is mainly used to connect to the database, at the same time, in order to improve the performance of the system, through the connection pool, object caching and other technical implementation; Communication protocol over TCP/IP, HTTP, and USES the Web Service protocol format as call interface, system by extending the object plug-in and embedded browser technology to realize the interface integration, primarily through the HTML Web interface layout level integration technology to realize the Jquery framework.

Key Words: Online Social Network; Network Public Opinion Monitoring; Web Crawlers

目 录

Contents

第一章 绪论

1.1 项目开发背景与意义

当今时代，计算机技术的发展越来越快，网络的发展也便捷了更多的人，人们的生活方式也发生了巨大的变化，智能手机的普及，网民随时随地的对社会事件进行评价，互联网企业也推出了越来越多的网络互动产品，提供评价事件的平台。许多社交网站应运而生，如：人人，各公司出品的微博、博客，猫扑等等，使得人们可以随时随地的了解新闻、娱乐圈，社会事件，并自由的评价与沟通。

Internet 的发展源于 20 世纪 60 年代，截止到目前，其发展规模、应用范围、自身功能对社会发展与经济建设都有十分重要的作用。当今社会的迅速发展离不开互联网技术的发展与进步，离不开计算机技术的支持，新媒体、自由社交网络的产生带来的是信息技术的改革，这场声势浩大的改革促进了人类社会发展的大步伐进步，在人类发展史上有里程碑的意义。

互联网的特性有很多，其中最重要的是其快速传播、高度粘贴的特点。用户对互联网的使用呈现较高的粘贴性，且有很深的依赖性，对互联网产品的忠诚度较高。用户对互联网产品的广泛使用，有较强的社交性、跟随性，互联网的聊天工具，评价平台等为顾客提供了很好的互动平台，对互联网产品的推广有很强的促进作用。互联网的客户群也越来越多，CNNIC 调查表明，2007 年有 2.1 亿人，到 2008 年我国网民发展为 2.98 亿人，占据中国 1/5 以上人口，网民比率迅速增长 41.9%。网民数量于 2008 年 6 月快速增长，一度超过美国网民占有率，在全球第一位。

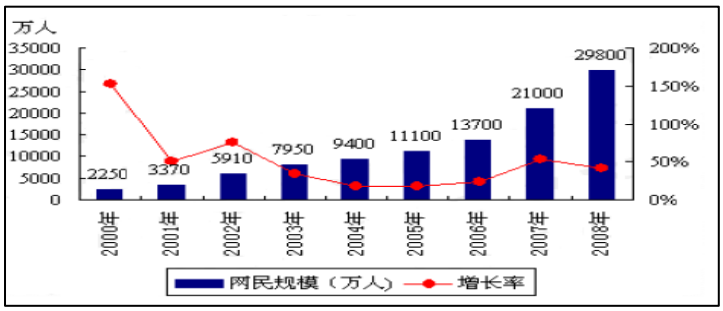

图 1-1　2000 年至 2008 年中国网民数量与增长情况

我国互联网产品中，博客产品的使用人群相对较多，且客户数量仍呈现较高增长，据 2008 年年底的统计数据，2.98 亿网民中超过 50%的网民使用博客产品，且使用者的绝对数在 1.6 亿以上。与此同时，博客产品顾客的活跃度也呈现较高程度的增加，例如：2008 年后半年博客新增顾客数量与 2007.12 相比上升 11%以上。随着智能手机、移动上网设备的推广，博客使用者可以随时随地登录其博客账户，与好友互动，进行社交活动。这一功能方便了广大的用户，成为当今社会人们社交不可或缺的重要组成部分。通过博客产品，以及移动电子设备的支持，提高了我国网民对政治的参与程度，国内外大事都可以第一时间了解并评价参与，成为有效的政治监督，对我国政治文明、精神文明都起到了很好的促进作用[2]。

在广大学者的研究中，舆情是民众对身边社会事件的产生、发展与评价的一种政治态度[17]。当某些社会事件发生时，广大群众各有其观点，其对社会事件表现出来的政治立场、政治态度、表述的意见、民众表现出来的政治信念、情绪等都是舆情的表现。通过对舆情理论的深入分析可得，舆情不能直接传播，其存在于传播都需求一定的渠道和载体，其中，这种渠道和载体既可以是杂志、书籍等以出版物为载体，也可以是网民的评价等以网络评价平台为载体，甚至可以通过公民的某些实际行为表现出来。互联网应用的日益普及、计算机技术的发展，使得网络成为民众评价的平台，舆情也与之一起产生。分析当今的舆情形式，传播最快、影响范围最广的为网络舆情[7]。网络舆情的主要平台为网络聊天工具、新闻门户网站、博客、微博等。通过上述网络舆情传播平台进行传播的舆情都包含了网民对社会发生的事件的观点、态度。

互联网是一个开放性的平台，同时也具备很强的虚拟性，互联网的这些特性使得网络舆情具有与普通舆情不同的特征：

（1）沟通无障碍。网民通过互联网社交评价平台，可以在第一时间传播自己对事件的观点，并与其他网民进行互动，在沟通上无任何障碍；

（2）舆情突发性。互联网是一个开放的平台，对于社会事件的反应迅速，网络舆情的传播也具有突发性，某些过激的言论和观点，在众多网民的传播下，在短时间就会变成网络舆情；

（3）舆情偏差性。网民借助于新闻媒体或者自己主观的看法了解事件的始

末，存在严重的信息不对称性，将直接导致其言行在态度和情绪上存在一定的主观性、偏差性。网民对事件的理解千差万别，传播的观点也有失偏颇，甚至有些网民认为网络的虚拟性，传播恶意谣言。这都造成了网络舆情的偏差性。

通过上述可以得到：互联网以及计算机技术的发展，使得互联网作为一个开放的言论平台，得到众多网民的青睐，网民数量一路攀升。博客等应用软件的广泛使用，使得网民对事件的反应快速，且具有很强的传播性。网民对事件的评论与传播，提高了网民对政治、社会、经济的参与，在一定程度上也对舆论的形成与传播造成压力。随着网络技术的迅速发展，网络舆情产生的影响也不得不被重视。正确的引导网民言论，限制恶意言论的传播，对网络言论的监督与控制已经迫在眉睫[11]。这就需要对网络舆情进行监测、并及时处理舆情信息。目前国内外学者对网络舆情的研究，依然还处在探索阶段，不能迅速的监测到舆情信息，就更谈不上及时控制网络舆情的传播。在网络舆情的监测上，需要相关部门采取有效的手段，及时监测出舆情信息，并采取措施进行控制与处理，以减少舆情对社会产生的不利影响[15]。而要想实现上述目标，就必须推进网络舆情监测系统的研究，为其提供强有力的技术保障。

1.2 国内外研究现状

近年来，国内外学者对网络舆情监控系统的设计开展了一些研究，研究的概述如下：

1.2.1 话题跟踪

通常，国内外学者对传统话题跟踪的研究一般从两个角度进行：知识角度、统计角度[13]。从知识角度对话题跟踪的研究主要是分析大量信息之间是否存在继承、关联等联系，通过将此领域的相关信息进行聚类，继而跟踪话题。从统计角度对话题跟踪的研究采用概率分析方法，对信息之间的相关度进行分析求解，从而跟踪话题。下面从这两个角度分析话题跟踪的研究现状。

从知识角度对话题进行跟踪的研究现状：Watanabe 通过研究日语新闻的广播信息，对话题跟踪进行分析，研发出了话题跟踪系统[10]。该系统的对“正如发生的”、“正如提到的”等话题分析其领域的知识，从而对话题进行检测与跟踪。通过实验与现实数据表明，从知识角度研发的话题跟踪系统可以有效的对特定知识领域的话题进行跟踪[1]。

从统计角度对话题进行跟踪的研究现状：对话题进行统计分析，主要方法是对话题信息进行过虑。信息过虑的步骤是：对动态信息进行识别、截取，从而得到可能是舆情的话题，该方法的前提是话题跟踪的需求是静态的。传统的话题跟踪通常基于先验话题模型，继而对后续的话题进行跟踪，从统计角度进行话题跟踪的方法主要是对话题进行分类。如：CMU 通过决策树算法和 K-最近邻算法对传统话题跟踪进行评价。决策树算法首先对话题材料进行分析，构造出决策树，其每一个节点代表一个发表相关话题条件的属性，其对应的分支为决策，其叶节点是对话题的分类。通过设置决策树，将需要检测的信息带入决策树，可以得到该被检测信息的类别。K-最近邻算法对话题相关度进行划分，对需要检测的信息进行分析，找到与之最类似的 K 最近邻，通过分析最近邻得到被检测话题的类别。

Umass 对 ODT 的研究进行分析总结，通过二元分类对话题进行跟踪，对需要检测的信息进行分类，若当前分类不能概括带检测信息，就将其分为其他的新话题，该方法将被检测话题划归为相关与不相关，并根据相关性的概率分布对分布器进行训练，依据判定式判定需要检测的话题[6]。

Michael 以及 James Allan 采用 Rocchio 算法对话题进行跟踪，该算法最重要的是构造话题经验策略。加重符合话题的报道在模型中的权重，并降低不符合话题的报道在模型中的权重。

对传统话题跟踪的相关研究中，重要的算法还有分析信息与话题相似度的算法。如：Dragon 综合分析信息的一元的语言模型与二项式文本匹配得到话题与信息的相关性；Carley 以及 Franz 采用聚类分析法设计并实现话题跟踪系统；Larkey 与 Yiming Yang 采用源语言模型与翻译模型对不同语种的信息进行话题跟踪[3]。

1.2.2 意见挖掘

对一些有主观性信息要进行意见挖掘，其中，主观性信息包括网络评价、评论、判断等信息[4]。

1997 年 McKeown 与 Hatzivassiloglou 在设计系统时通过学习词汇语义，对信息进行发表者态度、立场的判定。通过实验表明，其系统在学习了词汇语义之后，对信息中语义倾向进行判别时有较高的准确率，已达到 0.82，在系统中

加入形容词等有连续性的词汇之后，准确率升至 0.9。

2002 年 Pang 等通过机器学习的方法实现信息分类，该方法对向量机算法分类语义的准确度支持率达到 0.8，具有很强的有效性。

2003 年，Pottenger 与 Holzman 对网络中的聊天信息进行语义识别，通过对比多种算法，得到在对网络信息的语义识别方面 KNN 算法具有较高准确度，达到 0.9。

2004 年 Gamon 提取有舆情信息的特征，并进行分析，采用 SVM 判别信息的语义；Chambers 设计并实现了情感标注的分类语义倾向系统，该系统采用 Java 平台对信息进行分类；Zhongcha Fei 等采用短语的模式对信息中包含的语义进行分析，并综合识别句子长度，得到信息的分类。该方法是一种无监督的学习法，需要人为对短语和词汇的语义进行设定。

2005 年 Lee 与 Pang 等学者对识别信息语义之后的工作进行研究，提出对分类结果进一步细分，该步骤的依据是语义倾向程度。Lee 与 Pang 提出的进一步细分法在后续研究正，也被证实是可行的。Alm 等采用有监督的机器学习算法研究信息识别。Chao Wang 研究了网络评价的语义，在其研究中，采用了朴素贝叶斯与特征选择算法。Read 对计算机中语言的字符图示（如 ∩_∩）进行分析，降低识别过程中对机器学习方法的依赖。Qiangye 等通过支持向量机(SVM)、统计测度法识别评价，得到其精确度分别为 0.7887、0.7333。

2006 年 Hovy 与 Kim 对信息的特征进行总结，并通过最大熵(ME)学习法识别网络评论。

1.3 论文研究内容

本文在分析网络已经成为人们发布、传递和获得信息主要途径的背景的基础上。网民通过博客、论坛、微博等交流平台发表各种见解。网络给人们的生活带来了极大的便捷的同时，网络信息的快速传播也带来了很多负面影响，人们发表信息的随意性、信息没有经过核实就直接在网络中传播，对社会稳定产生影响。因此，监测复杂网络环境中的舆情，为相关部门提供充裕的应对时间具有十分重要的理论价值和应用价值。

本文以在线社会网络为研究对象，通过采用网络爬虫实现对微博、博客、轻博客的数据获取。在此基础上分析在线社会网络节点特征与舆情传播规律之