学校编码：10384　　分类号_____密级_____
学号：X2010230119　　UDC _____

硕　士　学　位　论　文

武警部队网络舆情系统的分析与设计

Analysis and Design of Armed Police Forces Network Public Opinion System

陈晨

指导教师姓名：　林坤辉　　教授
专　业　名　称：　软　件　工　程
论文提交日期：　2012 年　5　月
论文答辩日期：　2012 年　6　月
学位授予日期：　　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2012 年　5 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

在我国，不管是从硬件上看还是软件上看，网络技术都取得了巨大发展。通过 CNNIC 的调查报告，截至 2011 年 12 月底，中国网民数已增至 5.13 亿人。信息技术的飞速发展，使得武警部队官兵们也越来越多地借助网络表达自己的观点和看法，例如通过在论坛上发言、回帖，在博客上留言转载等等。因此网络已经彻底地融入官兵们的日常生活和工作中。

由于网络的自由等特性，使得灰色、偏激的言论在网络上能够得到广泛传播，不仅影响了官兵们的日常工作和生活，也对武警部队内部安全稳定与和谐造成很大的影响。对武警部队网络舆情必须加以监控，及时掌握当前的网络舆情走势，并采取适当措施进行控制和引导。因此建立一个武警部队网络舆情监控和分析系统就显得非常有必要。

本文首先分析了研究背景与意义，对当前国内外的相关技术进行综述。将武警部队网络舆情系统分成五大模块，并针对每个模块，阐述了各模块的核心技术。采用软件工程的理论和技术，对系统进行概要设计，明确系统的功能需求和整体结构。最后进行系统的详细设计，详细阐述了基于 Ontology 的主题网络爬虫技术、敏感和热点话题的检测技术以及基于情感 Ontology 的文本倾向性分析。

本文所设计的系统能对网上的海量网络舆情信息进行实时的监控和分析，可以实现对舆情的及时了解和掌握，为武警部队领导决策提供第一手资讯。

关键词：网络舆情；Ontology 主题爬虫；倾向性分析

Abstract

In our country, either at the hardware aspect or at the software aspect, network technology has made a great progress. Through the CNNIC survey report, as of the end of December 2011, the number of Chinese netizen has increased to 513 million. With the rapid development of information technology, armed police soldiers are increasingly using the network media to express their views and opinions, for example, to speak and reply at the Forum, leave and reprint message in the blog and so on. Network media has been completely integrated into soldiers' daily lives and work.

Because the free and other characteristics of the network, like gloom and incite statements can be spread widely, not only affecting soldiers' daily life and work, also causing a great impact on armed police forces stability and harmony. So it is necessary to monitor the network public opinion, grasp the tendency of current network public opinion and take appropriate measures to control and guide. Therefore it is very important to establish an armed police forces network public opinion system.

This dissertation first the analyzes the background and significance, related technologies at home and abroad were reviewed. This armed forces network public opinion system is divided into five modules, describes the core technology of each module. To make certain this system’s function needs and integral structure, it needs preliminary design. Finally, the detailed design of the system are made, detailed describe the ontology-based focused web crawler technology, sensitive and hot topic detection technology as well as emotional ontology-based text orientation analysis.

The system designed in this dissertation can real-time monitor and analysis mass Internet public opinion information, aiming to understand and master custom timely, to provide the first-hand information for armed police forces leadership.

Keywords: Internet Public Opinion; Ontology Focused Web Crawler; Orientation Analysis

目 录

CONTENTS

第一章 绪论

1.1 研究背景及意义

CNNIC 发布的《第 29 次中国互联网络发展状况统计报告》显示，截至 2011 年 12 月底，我国网民数已增至 5.13 亿人，互联网普及率达到 38.3%。即时通信用户规模达到 4.15 亿人;网络新闻的用户规模达 3.67 亿人;微博用户数量从 6311 万迅速增长到 2.5 亿，信息化趋势已不可阻挡。

担负着处突和维护社会稳定重要任务的武警部队，亦十分重视信息化建设，提出了“建设信息化武警，实现跨越式发展”的总目标。由于武警部队机构点遍布全国各地，每天的大事要情繁多，因此建设一个综合的信息化网络，实现高效率多触角管理十分必要。经过几年的建设，武警部队综合信息网已经逐步建立并从总部覆盖到总队、支队和中队四级单位。加上近年来官兵们的知识结构也在不断的发生变化，科技素养逐步提升，信息化已渗透进官兵们生活、工作、学习的点点滴滴。据统计，截至 2011 年，能熟练使用计算机、网络和各类软件的官兵占 80%以上。官兵们能够熟练运用信息网络即时了解国际国内和军外军内的大事要闻。当前社会正值转型发展期，部队内部也处于变革时期，各类矛盾交织凸显，因此，有许多负面的和涉及官兵切身利益的问题，较易引起官兵思潮浮动。由于部队内部管控较严，而网络因其虚拟、开放的特性，环境相对宽松，这就决定了官兵更倾向于选择这种方式，乐于通过论坛、博客（微博）、即时聊天工具和信箱等自由阐述、发表言论意见、宣泄情绪。不少重要的论坛，访问量高达几千，甚至上万人次。在交流传播的过程中，彼此的观点相互交织、演化，逐渐形成舆情。

具体分析可以看出，武警部队网络舆情具有以下五个显著特性：一是便捷性。随着网络的普及，使得官兵能够更加方便快捷地随时随地了解最新发生的事情，即时发表言论。而其具有特色的原创内容更是能够吸引更多人关注，并发表评论、表达观点，随时随地就能针对某事件发表自己的意见看法；二是快速性，官兵的观点通过网络论坛跟帖回复、博客及即时通讯软件的传发，使得往往在短时间内就迅速地传播扩散，涟漪效应明显；三是隐蔽性，特别是由于计算机和 IP 地址

较少，部分基层官兵往往共用一台自动分配 IP 地址的电脑上网。因此，隐蔽性较强，在某种程度上，更加助长了官兵们发表观点自不自觉都会带有一些自利性，或者可能不是出于正常目的的倾向，而有可能是基于某些潜在隐藏的目的；四是演化性，各种各类观点交织融会的网络舆情产生后，极易聚合发酵，舆情事件可能出现多个不同的发展趋势，或者与原有的动机产生偏离。五是语言鲜明性，由于上网的官兵们年龄普遍在 35 周岁以下，加之部队的用语比较倾向于直接化和命令式，褒贬鲜明，故官兵在上网时也不自觉的带入了这种表达方式。

因此，网络舆情的强大影响力不言而喻，这股如洪水般强大的舆论力量，如若引导不好，有时就会反作用于焦点、热点事件，一定程度上能影响武警部队思想政治建设，负面的网络舆情甚至会左右事态的进程，对部队内部稳定形成较大威胁。如河南假冒武警车牌案件，超女掌掴武警哨兵事件等产生的负面影响不言而喻；再比如一些社会热点问题，如房价问题、医疗、教育等和官兵切身利益相关的话题也容易引起讨论和关注。由于武警部队对网络舆情的监测分析起步较晚，尚处于理论研讨阶段，在网络技术监测、分析方面研究不够深入，因此无法及时借助技术手段找准切入点，进行有效引导疏导。所以，加强对武警部队网络舆论的及时监测、有效引导，形成一个健康的网络舆论导向，保持部队内部思想纯洁、高度集中统一，这对于维护武警部队稳定，提高官兵思想政治建设水平具有很重要的现实意义，也是创建和谐警营的应有内涵，必须给予高度重视。

鉴于武警部队网络舆情的特点，决策层要想做到了然于心，仅靠人工方法很难去收集并分析。因此，一个合理的网络舆情监测分析系统就显得非常重要，能及时、准确地从武警部队综合信息网中检测到隐藏的“热点”舆情信息，并分析舆情的倾向性，为领导提供科学化、全面化、直观化的数据支撑，确保对负面舆情信息进行适当和有效的干预并给予正确引导。这样才能及时地了解官兵心声，服务基层建设，防微杜渐，防患于未然。

1.2 国内外研究现状和技术实现状况

1.2.1 国内外研究现状

网络舆情系统一般主要包括信息采集模块、信息预处理模块、事件提取模块、舆情信息识别模板、舆情观点挖掘模块、信息决策模块等。其中在信息采集模块

主要需要用到主题爬虫技术，进行信息数据的收集；在信息预处理模块主要需要用网页去噪技术、网页排重技术等；在事件提取模块主要用到分词技术、事件主题提取技术等；在舆情信息识别模块主要需用到话题追踪与聚类等技术；在舆情观点挖掘模块中主要需用到文本倾向性分析等技术。由于本文主要针对武警部队网络特点，因此主要分析在爬虫技术与文本倾向性分析中引入 Ontology(本体)的相关概念，以及在舆情信息识别中考虑加入作弊探测，因此接下来主要介绍这三类技术的研究现状。

（1）有关“Ontology”

Ontology 的最早定义是在哲学中，指对客观存在事物的一个系统的解释和说明，抽象出客观现实的本质。随着人工智能的快速发展，人工智能界对 Ontology 也赋予了新的定义。1998 年 Studer 提出了具有代表性的定义：Ontology 是指共享概念模型的明确的形式化规范说明[2]。

Ontology 是以获取该领域的知识为目标，使得能共同理解该领域知识，据此该领域内共同认可的词汇能够被确定下来，而要明确定义这些词汇（术语）和词汇之间的相互关系，需要从不同层次的形式化模式上考虑。

目前，图书馆及信息系统中已经有 Ontology 的应用，用来帮助用户明确其信息需求，并指导用户浏览结构化的搜索结果。Ontology 可以在文本检索领域的语义层次上，帮助用户处理网络上的海量信息，提高查全率和查准率，因此受到越来越多的关注。近年来，Ontology 的应用已经陆续出现在人工智能、自然语言理解、软件工程、知识管理等领域。

（2）有关“主题爬虫”

早期的网络爬虫主要是利用图论知识，该类爬虫采集速度低，信息维护繁杂。主题爬虫应用在垂直搜索引擎中，能够尽可能多地获取与给定主题相关的网页信息。

国外对网络爬虫的研究比较早，主题爬虫雏形最早出现在 1994 年，即 Fish 系统。指出了主题爬虫系统的一个主要研究方向，即如何使用启发式策略来为 URL 评分、预测 URL 的相关性[3]。

2001 年 Mencze 对 BestFirst(利用网页与主题的相似度进行排序)、PageRank、InfoSpiders(利用查询向量与神经网络的思想进行排序)三种搜索策略进行比较，

实验发现在锁定主题方面，BestFirst 效果最优，PageRank 最差[4]。

刘金红等在文献[5]中将主题网络爬虫主要分为如下四类：一是基于文字内容的启发式方法，主要是基于网页文本内容、URL 字符串中等文字内容。主要包括 Best first search 方法、Fish search 方法和 Shark search 方法；二是基于 Web 超链图评价的方法，此类方法的爬行算法有以下两种：BackLink 和 PageRank。三是基于分类器预测的方法，基于分类器引导的主题网络爬虫，可以克服前两种方法的低效率，网页的主题相关性能够被更加准确地算出，而不是单单匹配关键字。四是其他主题爬行技术。

随着本体论在信息系统中的应用，Ehing 将本体的思想应用在主题爬虫中，在计算 URL 优先级的时候，引入本体相关知识，计算网页主题相关度[6]。

（3）有关“文本倾向性分析”

目前已有的中文词语语义倾向性分析方法主要有以下两类：基于 HowNet 的词汇语义倾向性分析法和基于同义词林的方法[7]。

王晓东等在文献中指出常见的文本倾向性分析方法有基于统计的文本倾向性分析方法和基于语义规则的文本倾向性研究方法[8]。基于统计的文本倾向性分析三种主要方法为朴素贝叶斯、最大熵及支持向量机，这些方法国内外得到了较为广泛的应用。针对后一种方法，作者提出了一种基于情感 Ontology 的文本倾向性分析方法，通过实验得出的结果比普通的基于语义规则的文本分析准确率提高了 10%左右，为文本倾向性分析提供了新的思路。

（4）有关“观点作弊探测”

作弊探测目前在国内外研究的不多，文献[9]指出作弊探测主要有三种方法。

1.以评论内容为中心的作弊探测，通过比较内容的相似度进行评判。

2.以评论者为中心的作弊探测，观察同一个评论者在同一网站上发表评论的时间及评论对象是否具有一定的规律。

3.以服务器为中心的作弊探测，若一个用户使用同一个 IP 地址在同一个网站上进行多次注册，并且对同一个产品发表了若干个评论，或使用多个 userid 对若干个产品发表了评论，那么该用户是作弊者的可能性就非常大。

1.2.2 技术实现状况

Goonie 互联网舆情监控系统是由谷尼国际软件开发的，系统结构如图 1-1

所示。通过自动获取互联网上的海量信息，进行聚类，主题检测与聚焦，实现网络舆情监测分析、话题追踪。形成的分析报告，为客户全面掌握舆情动态提供分析依据。该系统通过话题抽取识别，相似性去重等技术，可以获取网络中的热点敏感话题。根据统计等策略，分析不同主题在不同时间内被人们关注程度，预测网络事件的发展趋势。

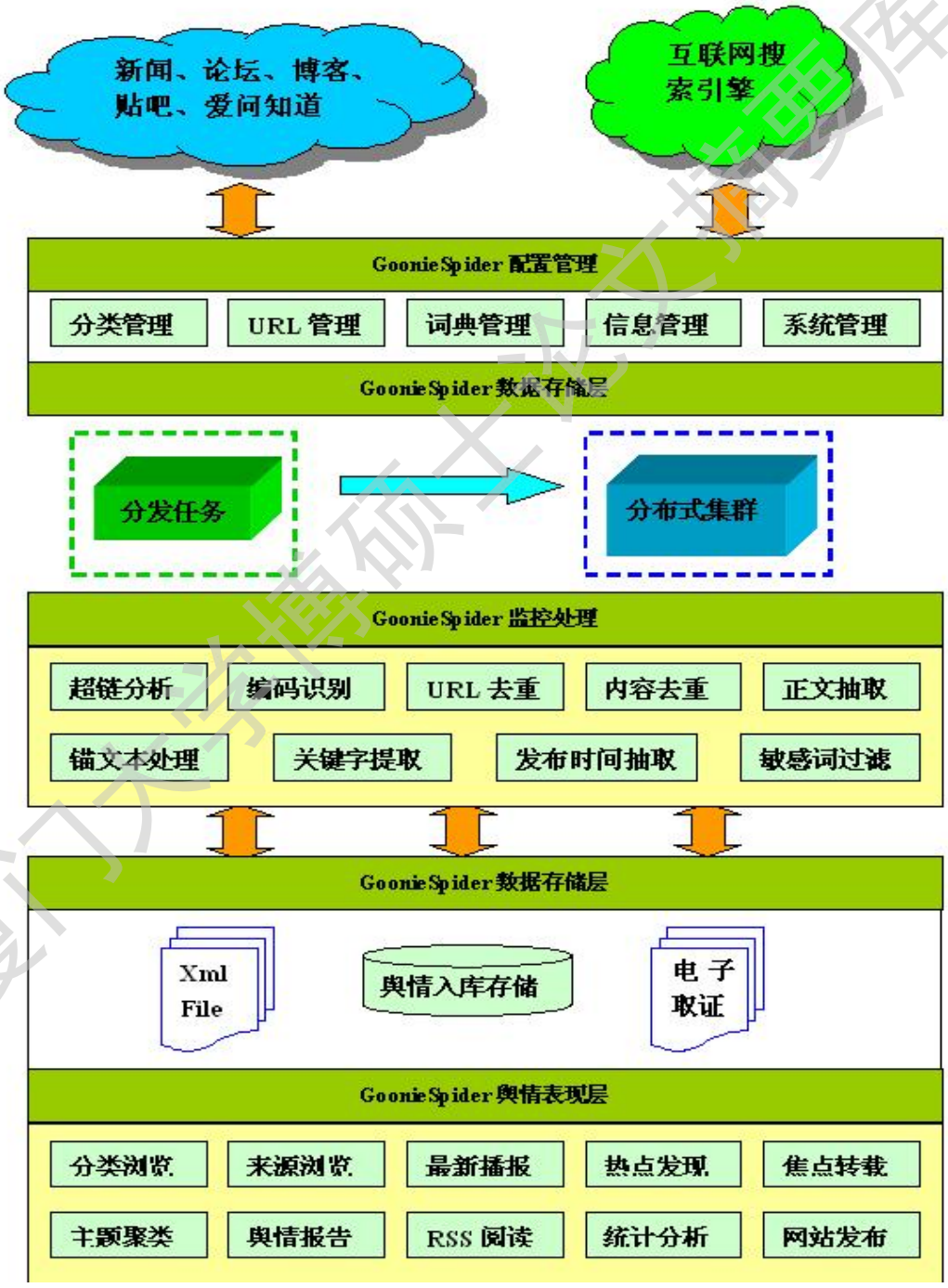

图 1-1 Goonie 互联网舆情监测系统架构图

中科点击开发的军犬网络舆情监控系统，由舆情采集工具（军犬网络信息采集系统）、舆情加工分析模块、舆情服务模块和舆情检索模块四部分组成。军犬网络舆情监控系统架构如图 1-2 所示。该系统能够分析出事件相关性，以及舆情是负面、正面的还是热点的，以及每条舆情信息的传播演化路径等，最后生成基于网络舆情平台中的数据、图表生成的简报专报。军犬网络舆情监控系统架构如图 1-2 所示。

图 1-2 军犬网络舆情监控系统架构图

TRS 网络舆情监控系统由北京拓尔思信息技术股份有限公司开发，是一款基于中文信息采集和处理的平台软件，将最新的文本检索、文本挖掘和内容管理等技术进行融合，自动地采集互联网上的各类信息，并智能地处理和分析这些海量信息，可以实现监控热点舆情、自动分析和统计、文本检索等。借助该系统，用户可快速构建一套网络舆情系统，该系统高效稳定，且捕捉各类舆情信息及时、准确，能较好的完成预定目标。TRS 网络舆情监控系统架构如图 1-3 所示。