分类号_____　　密　级

UDC _____　　学校代码 10497

# 武汉理工大学

# 学　位　论　文

题　目 网络舆情热点信息发现及其倾向性研究

英　文 The Research on the Discovery and Polarity of

题　目 Hot Topics of Public Opinion

研究生姓名 李 海 林

指导教师　姓名 聂规划　职称 教授

单位名称 经济学院　邮编 430070

申请学位级别 硕 士　学科专业名称 国际贸易学

论文提交日期 2010 年 11 月　论文答辩日期 2010 年 11 月

学位授予单位 武汉理工大学　学位授予日期_____

答辩委员会主席_____　评阅人_____

2010 年 11 月

# 独 创 性 声 明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得武汉理工大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：李海林 日 期：2010.11.10

# 关于学位论文使用授权的声明

本学位论文作者完全了解武汉理工大学有关保留、使用学位论文的规定。特授权武汉理工大学可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

研究生签名：李海林 导师签名：[illegible] 日 期：2010.11.10

# 摘 要

随着信息技术的发展和互联网的日益普及，网络已经成为广大民众获取信息的主要渠道，同时网络也成为人们发表评论、表达民意的重要平台。面对互联网上飞速增长的新闻话题以及人们的评论信息，如何从海量信息中采集到满足特定需求的信息，如何将互联网信息组织整理成有效的机器数据，如何从采集到的数据中区分有用信息和无用信息等等这些问题都是信息科技发展所面临的难题。网络舆情是指民众通过互联网对政府管理以及现实社会中各种现象、问题所表达的政治信念、态度、意见和情绪的总和。网络舆情与社会舆情相互作用、相互影响。两者不仅在内容表现形态方面具有一致性，同时网络舆情在一定程度上会影响社会舆情的发展趋势，对社会影响巨大。因此，政府部门对网络舆情信息必须具备一定的监控能力，能够及时掌握一定时期内民众所关注的热点问题，了解民众对热点事件的看法和态度，从而做出正确的决策，主动引导舆论走向。

本文在分析网络舆情热点信息发现和网络舆情热点信息倾向性研究现状的基础上，从舆情信息的来源入手，设计了详细的采集流程。针对大众和政府部门都比较关注的热点信息，本文根据热点信息的概念和特征建立了热点信息的判断标准，并将热点信息的特征定量化，构建数学模型，用算法来描述热点信息的发现和获取。针对热点信息的倾向性分析，本文首先手工构建了极性词典，并对极性词典进行了扩充和修正，将未登录词汇、否定词和强调副词对原始极性词的影响做了进一步分析，并提出相应的解决办法。对于普通的文本信息，用向量来进行表示，通过计算特征词的权重来选取文本的特征词条。由于中文句子以标点符号进行划分，本文对句子进行句法分析，解析出词语之间的依存关系，并对词语进行词性标注。本文建立了语义模板，通过语义模板的匹配来确定句子的语义模式，利用极性词典计算出词语的极性值，再结合句法分析和模式匹配得出其上下文极性。句子的倾向性由组成句子的主题词和极性词及其极性值决定，文本的倾向性由句子的倾向性和句子在整个文本中的权重计算得出。最后，本文对所做的研究工作进行了模拟实验，对实验结果进行了讨论与分析。

**关键词：** 网络舆情，热点信息，极性词，文本特征

# Abstract

With the development of information technology and the growing popularity of Internet, the network has become the main channel for general public people to get the information, as well as an important platform for expression of public opinion. At the face of rapid growth of news information and people's comments on the Internet, how can we get the information which meets the specific needs from the mass information? How to organize Internet information into an effective machine data? How to distinguish the useful information and useless information from the collected data? All these problems are difficult at the process of the development of information technology. Public opinion is the sum of political beliefs, attitudes, opinions and emotions about the government administration, as well as the variety of phenomena in the real world which are expressed by general people through the Internet. The Internet public opinion and the social public opinion are interaction and affect each other. The Internet public opinion and the social public opinion has a consistent on the content, The Internet public opinion to a certain extent, will affect the community development trends of social public opinion, and will have a huge impact on the community. Therefore, the Government needs to have some information on the network to monitor public opinion, and the ability to grasp a hot issue which the general people concern on the certain period of time, understand the attitudes and views of hot events in order to make the right decisions, and take the initiative to guide public opinion towards.

Based on the analysis of the discovery on public opinion hotspot information and the research on tendency analysis of public opinion, this paper designs a detailed collection process from the source of public opinion. For the hot information which is concerned by the general public and government departments, this paper has established criteria for judging hot information according to the concept and characteristics of hot spots, and quantitative characteristics of hot information to build mathematical model, using algorithms to describe the discovery and access of hot spot information. To the tendency analysis of hot information, first of all this paper hand-built the polarity dictionary, and the polarity dictionary was expanded and amended, then have the further analysis on the no-logged vocabulary, the negative words and stressed words to the impact of the polarity on the original word, and give the solutions. This paper uses vectors to carry out the ordinary text messages, and

selects the characteristics words of the text by calculating the weights. As the Chinese sentence is divided by punctuation, this paper carried a sentence parsing, parsed out the dependencies between words, and tagged the part of speech. This paper built the semantics template, and determined the sentence semantic model through the matching to semantic template, calculated the polarity value of the words using the polarity dictionary, got its context polarity by combining with syntactic analysis and pattern matching, the tendency of sentence is determined by the composition of the sentence and the polarity value of the words, the tendency of the text is calculated by the tendency of sentence and the weight of the sentence in the whole text. Finally, this paper made simulation experiments about the research work, discussed and analyzed the experimental results.

**Key words:** Internet public opinion, Hot topic, Polarity word, Text feature

# 目 录

# 第1章 绪论

## 1.1 研究目的及意义

随着信息技术和科学技术的飞速发展，互联网在人们的日常生活中越来越普及，甚至成为必不可少的一部分。2010 年 1 月 15 日，中国互联网络信息中心（CNNIC）在京发布了《第 25 次中国互联网络发展状况统计报告》。报告显示，截至 2009 年底，我国互联网普及率达到 28.9%，我国网民数达到 3.84 亿，宽带网民数达到 3.46 亿，域名数达 1682 万，三项指标继续稳居世界排名第一[1]。

根据以上数据显示，随着互联网融入生活的程度不断加深，在中国国内上网的人数在不断的攀升，已趋世界水平。与此同时互联网网页的数量也在不断增加，目前已有的网页数量已约有 160.8 亿个。由温家宝总理 2009 年 2 月 28 日与网友在线交流这一事件也可以看出，随着互联网络不断的发展深入，网络舆情将逐渐取代原来的纸质媒体、口口相传的传统舆情，影响力不可限量。

舆情，顾名思义即指社情民意，学者王来华在《对舆情、民意和舆论三概念异同的初步辨析》一文中将舆情定义为，在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，作为舆情主体的民众对国家管理者产生和持有的社会政治态度[2]。由舆情的定义可以得出网络舆情即指网民在互联网空间内，对于各种各样的社会问题所持有的态度和意见[3]。目前随着互联网普及率的不断提高，越来越多的人习惯于从网络上来获取信息，包括工作安排、日常生活、浏览新闻、参政议政等等各大方面。由于互联网的开放性，人们可以通过多种方式来表达自己的意愿和情感，比如：BBS、聊天室、博客、新闻评论等，同时互联网的虚拟性决定了人们在网络上发表言论的自由和隐蔽，由于缺乏有效的规则限制和监督机制，网络上极容易出现一些比较片面，比较灰色，甚至是反动的舆论信息。网络舆情与传统媒体相比，具有信息丰富、互动性强、及时快捷等优点。当然，我们在享受互联网给带来便利的同时，也应该认识到其中存在的舆论导向潜在安全问题。最近在网上议论较多的事件，如浙江杭州闹市飙车致死案、湖北巴东县邓玉娇案、湖北石首群体骚乱事件、云南晋宁县躲猫猫事件、四川成都市公交车燃烧事件等等，只要有足够的网民关注，某些事件很快就能成为舆论热点，然而在网络时代，舆论的导向具有很大的不确定性，如果这些热点信息被不法分子所利用，以此来控制事件的舆论走向，那么将会对社会稳定产生极大的威胁。因此，面对海量的网络信息，要充分发挥互联网的积极作用，重点关注网络舆情热点信息，为正确引导舆论导向提供决策参考。

一般来说，网络舆情热点信息具有两重功能：一方面，它是在特定历史条件

下所形成的某种群体观念，是在一定历史环境中，人们在思想上、愿望上和要求上的共同反映，这往往会成为当时社会气候的晴雨表和社会信息的显示器。社会热点的信息显示，有利于政府部门采取措施、解决问题、安定民心、稳定社会。这就是社会热点的积极作用。

另一方面，舆论热点信息往往呈现自发、松散状态，如不及时给予正确指引，便会因民间的传播、感染、认同而逐渐形成社会舆论合力，冲击人们情绪，不利于社会稳定。这便是社会热点的负面效应。而话题是由一些原因、条件引起，发生在特定时间、地点，并可能伴随某些必然结果的一个事件。它包括一个核心事件或活动以及所有与之直接相关的事件或活动。可能是一起事故、一起纠纷、一场争议，等等。它对时效性有很高的要求，有一定的存在时间，即话题在一定的时间内产生、发展，并随着时间的推移而消亡。

随着互联网在全球范围内的飞速发展，网络媒体已被公认为是继报纸、广播、电视之后的“第四媒体”，网络成为反映社会舆情的主要载体之一。舆情的收集比以往任何时期的政府部门都要容易得多，丰富得多。网络参与人数暴增导致舆情形成迅速，网络舆情与社会舆情相互作用、相互影响，网络舆情与社会舆情在内容表现形态方面具有一致性，网络舆情在一定程度上会影响社会舆情的发展趋势，对社会影响巨大，因此不仅需要各级党政干部密切关注，也需要社会各界高度重视。因此研究网络舆情热点信息的倾向性具有十分重要的意义：1)对于个人而言，它能使我们及时、方便地获取当前社会中比较重要的热点信息。2)对于企业而言，它能使企业更及时掌握相关领域的最新动态、热点技术咨询，增强企业竞争力。3)对于国家而言，它的意义更加重大，它能帮助政府有关部门及时了解当前社会重要事件、流行趋向、舆论方向，有利于相关部门迅速进行舆论引导，发扬积极、健康、正确舆论，抑制消极、错误的舆论，及时解决问题，促进经济社会健康、稳定发展，它为确保我国互联网络大众媒体的舆论导向的正确性起到一定的辅助作用。

## 1.2 国内外研究现状

随着互联网的迅速发展，网民数量的快速增长，网络已经成为人们获取信息和反映民意的新兴渠道。网络舆情也日益受到政府部门的重视和关注，学术界也对网络舆情做出了一定的研究和探讨。但截止到目前，国内学者关于网络舆情的研究还不够成熟，相应的研究成果也不多，笔者在中国期刊全文数据库中检索得到的相关文献资料较少。下面就对现有的资料作综述研究。

### 1.2.1 舆情基本理论

据资料记载，词语“舆情”最早出现在中国的唐朝时期，其用意在于老百姓在表达自己意愿的时候能出现正直廉明的大臣。“舆情”在文献中也被解释为“民众的意愿”。在18世纪的法国，卢梭首次提出了公众意见（Opinio publique）的概念。据《牛津英语大辞典》介绍，公众意见一次最早出现于1781年。

目前，我国对舆情的定义还没有一个统一的标准，除了上文中提到的王来华学者给出的定义外，还有张克生学者的《国家决策：机制与舆情》一书中认为舆情是国家决策主体在决策活动中必然设计的、关乎民众利益的民众生活、社会生产和民众中蕴含的知识和智力等社会客观情况，以及民众在认知、情感和意志基础上，对社会客观情况以及国家决策产生的主观社会政治态度。简单的说，就是社会客观情况与民众主观意愿，即社情民意[4]。学者丁柏铨认为，舆情就是民意情况，涉及对社会生活中各个方面的问题尤其是热点问题的公开意见或情绪反应[5]。毕竟认为舆情是指处于不同历史阶段的社会群体对某些社会现实和现象的主观反映，是群体性的意识、思想、意见和要求等的综合表现[6]。

对于舆情的内容分析，笔者也查阅了部分文献。天津社会科学院舆情研究所的张丽红把舆情空间划分为硬空间和软空间，即有形空间和无形空间，并提出舆情空间具有内含因素的多样性和互动性、变动性、相对界限性、相对层次性等特点[7]。谢海光等人提出构建互联网内容与舆情的十个分析模式，即热点、重点、焦点、敏点、频点、拐点、难点、疑点、粘点和散点，并借用校园公共安全危机管理案例对网络舆情热点进行了判据试释[8]。许鑫等人采用多学科融合交叉的方法，给出了互联网舆情分析的基本思路和方法，并构建了互联网舆情研判平台的想法[9]。吴绍忠等人通过设定网络舆情预警分为四个等级，并选取了11个指标构建预警体系，运用德尔菲法确定指标体系的权重，遵循网络预警的工作流程，就可以发现网络舆情，掌控其发展变化过程，进行深入挖掘与分析从而正确预警，并及时采取预控措施进行引导，以保障社会的安全与稳定[10]。梅中玲提出采用web信息挖掘技术来自动在网络上收集信息，由此分析出网络舆情信息的源头、受众及其特点，通过对大量历史数据的收集和学习，建立网络舆情传播的先验模式，再通过信息挖掘发现其传播特点和模式[11]。戴媛等人提出了网络舆情信息挖掘内容重要的“六个点”，即热点、焦点、兴奋点、波动点、重点和诱发点，针对网络舆情产生、阅览和转载三个阶段不同的特点提出了不同的信息挖掘方式，并选取舆情流通量、舆情要素、舆情状态趋势等三大指标来构建网络舆情安全评估体系，量化了评价舆情发展态势，为管理者提供预警和辅助决策的科学依据[12]。纪红等人指出搜集网络舆情要正确把握网络舆情的生成规律，清楚了解网络舆情的存在空间，分析网络舆情要通晓社会思潮和复杂形势，引导网络舆情要

做好信息发布等工作[13]。天津社会科学院舆情研究所的姜胜洪指出网上舆情主要通过网络新闻、电子邮件、网络论坛、新闻评论、博客等方式进行传播，具有直接性、突发性、丰富性、互动性、偏差性等特点，网络舆情给中国政治安全和文化安全带来了挑战，日益受到政府部门的高度重视[14]。刘鹏飞在《网络舆情抽样与分析方法》一文中论述了网络舆情选题与抽样的方法、原则以及网络舆情分析框架与办法[15]。吕洪波等人针对搜索引擎采集到的大量与主题无关的信息，提出将无关信息清理掉，并在系统中将信息清理和中文采词结合起来，提高系统运行效率和准确度[16]。黄晓斌等人首先介绍了网络舆情的特点（广泛性和匿名性、自由度与可控性、互动性与即时性、丰富性与多元性、影响范围和程度大）与作用（桥梁作用、耳目作用、决策依据、预警作用、导向作用），分析了文本挖掘技术的主要功能，构建了网络舆情信息文本挖掘的模型，给出了详细的实施步骤，分为准备阶段、处理阶段、分析阶段，最后进行实验验证[17]。杨频等人重点分析了网络舆情中的文本倾向性，提出结合使用 HowNet 中的语义相似度和基于人工评分的统计方法，对于两种方法获得的结果计算加权和得到词语的情感倾向度[18]。王来华等人认为舆情汇集分析机制可以划分为汇集机制、分析机制、报送机制、反馈机制、工作保障和激励机制，在运用舆情信息汇集和分析机制时，应注重服务对象、操作主体、基本环节和方法的科学性以及规章制度的建立和完善[19]。

### 1.2.2 网络舆情热点信息发现研究现状

网络舆情主要关注的是热点信息，在当前海量的互联网信息中，如何快速有效地主动发现网络舆情热点信息已经成为一项重要的研究课题。到目前为止，国内外学者对这个课题作了一定的研究。

对于热点信息发现的研究，国外较早出现在 TDT 领域，话题检测与跟踪（Topic Detection and tracking，简写为 TDT）是由美国国防高级研究计划局带头开展的一项研究，该项研究始于 1996 年，现已经取得一些研究成果[20][21]。这里的话题概念是指包括一个核心事件或活动，以及所有与之直接相关的事件或活动[22]。话题识别与跟踪研究的主要任务有 5 个，即对新闻报道的切分、对新事件的识别、对报道关系的识别、对话题的识别、对话题的跟踪[23]。由于 BBS 是出现热点信息的主要方式之一，针对 BBS 方面的热点信息发现，日本学者提出影响力传播模型 IDM，主要用来发现 BBS 上有一定影响力的网民和话题[24]。在影响力传播模型中，认为网民对某一事件的观点主要体现在帖子的关键词中，关键词在同一帖子中出现次数或传递次数的多少反映了其影响程度。该模型主要是通过分析关键词的影响力传递来发现热点话题。

在国内，在网络舆情分析领域做得较为成功的是北大方正技术研究院推出的方正智思舆情辅助决策支持系统。如图 1-1 所示，该系统以内容管理平台、知识管理平台、辅助决策支持平台来分别支撑舆情采集存储、舆情分析处理和舆情服务三层应用功能。其中网络信息的主题检测和追踪技术是该系统实现网络舆情热点信息自动发现的核心技术之一，并采用基于智能中文分词技术来发现预警不良信息，起到决策支持作用。方正智思舆情辅助决策支持系统的主要功能包括智能网络页面获取、智能检索、自动摘要、关联分析、自动聚类、自动分类等等[25]。

除了北大方正推出的舆情辅助决策支持系统外，也有许多国内学者从理论上对热点信息发现进行了研究。上海交通大学的黄宇栋等人针对网络热点信息主动发现系统中需要高速处理大规模数据的特点，提出把聚类特性引入传统 K-Means 算法，进一步融合基于密度的聚类思想，由此形成 DCFK 聚类算法，并且基于该算法构造中文聚类模型进行网络舆情热点信息的发现研究[26]。该模型主要包括文本分词、特征选择、向量表示和核心聚类等功能，并有效的解决了传统聚类算法中聚类类别数和初始中心点难以确定的问题。王林等人从用户兴趣角度出发，形成具有同一兴趣的论坛用户网络，将社区结构发现的理论与方法应用于论坛热点话题的主动发现，提出了极大社区的概念和反复挖掘极大社区的方法[27]。

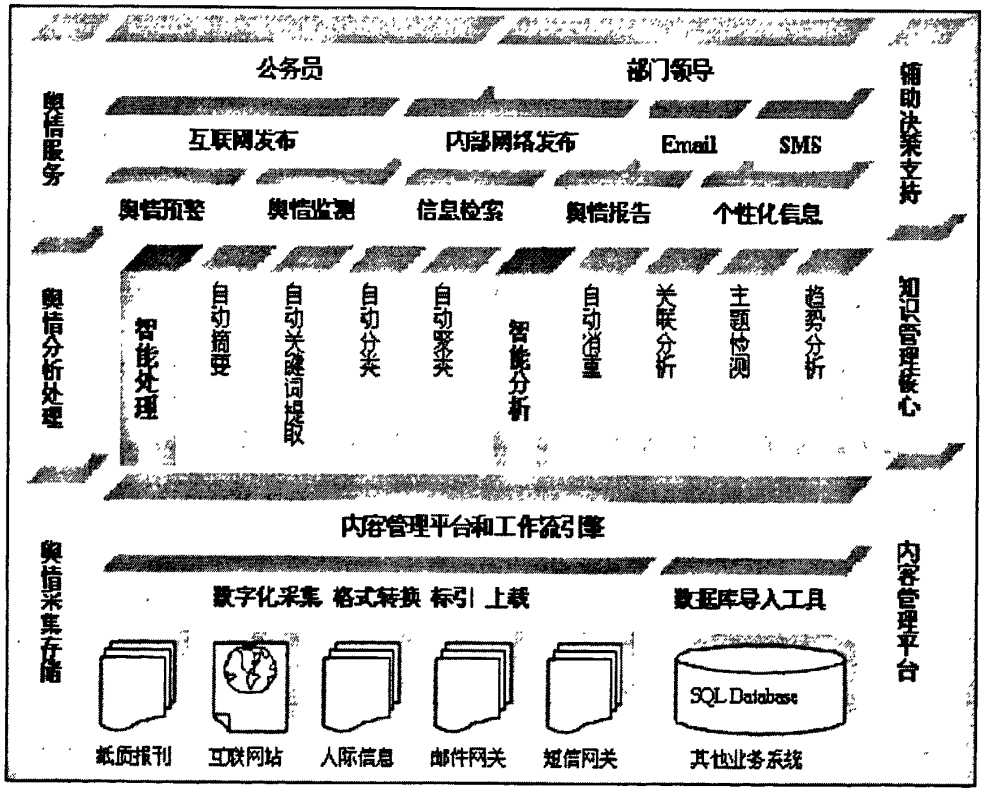


图 1-1 方正智思舆情辅助决策支持系统结构图

(来源：http://www.smelz.gov.cn/news/35897.htm)

鲁明宇等人认为一个帖子线索中可能会包含多个话题，提出采用模糊聚类方法对

BBS 话题进行识别，选取话题的帖子数、精华帖子数、平均单位时间浏览数、平均回复数等影响因素，对聚类得到的话题进行热点评分[28]。王义等人将字符串核函数运用到中文文本聚类算法中，该算法的工作步骤是信息获取、文本清洗、数据过滤及超文本抽取、聚类分析、热点新闻发布等[29]。周亚东等人将网络热点话题进行形式化描述，即用热点词语、核心标题、信息发布网站等向量来表示，认为热点话题可由多个热点词语来表示，由此提出一种流量内容热点词语相关度计算方法，该方法可以量化热点词语之间的相关度。采用 DBSCAN 聚类算法将具有较大相关度的热点词语聚合在一起，结合热点词语相关的网页标题和网站地址等信息得出网络热点话题的属性描述[30]。中国科学院的曾依灵等人认为网络热点信息应该具有简洁性、时效性、信息量大等特征，设计了网络热点信息发现算法，基于多级滤噪进行切分词拼接，利用特定的噪声库与多级滤噪策略严格控制拼接过程，挑选合理的收录策略，提取出能够准确反映海量网络数据中的热点信息串[31]。周启海等人针对信息在生活中受关注的程度，提出将同构化理论运用于信息的重要性与受关注程度及其度量，通过构建特殊的同构化映射来得出同构化信息温度的方法，从而使同构化信息温度用作测度信息重要性与受关注程度的新尺度，进而来发现网络信息中的热点[32]。刘兴兴等人设计了一个热点事件发现系统，该系统针对网络新闻预料具有数据规模大和时间特征明显两个特性，将语料按天分组，对每天的语料采用凝聚聚类得到微类，选取某段时间内的所有微类，再做聚类得到事件列表，利用事件热度计算公式，把候选事件按热度进行排序[33]。系统采用了两层聚类的策略，分为批处理和实时处理两个阶段，在事件热度计算中，选取了事件的报道频率（时间频率和数量频率）和事件的平均相似度作为特征量。

综合现有的研究方法，在网络舆情热点信息发现方面的研究主要有两个方向，一个是基于自然语言理解的处理技术，另外一个是基于数据挖掘的角度，涉及到的技术有中英文分词、词频统计、文本分类聚类等。目前专家学者在这两方面的研究都取得了一定的成果，但也还存在着一些不足之处，比如说对于热点信息的判断还没有形成一套成熟的指标体系，还没有很好的办法对热点信息的发展趋势进行分析等。

### 1.2.3 网络舆情热点信息倾向性研究现状

网络舆情分析的另一重点是对文本进行倾向性分析。文本倾向性分析就是指通过分析文本的感情色彩，从褒义、贬义的角度得出作者在文本信息中所持有的态度取向。目前国内外有不少学者在这方面作了相应的研究。

在国外，Wiebe 等人从词语的角度入手，在一些分散的词语中寻找线索，并

进行 N-grams 分析，识别句子的褒贬性，进而对文本进行倾向性分类[34]。Riloff 等人采用贝叶斯分类方法对情感名词、谈话特征和情感线索进行分类，其中用到算法来提取 1000 多个情感特征词建立情感分类器[35]。Turney 等人提出采用统计规则对词语进行倾向判断，根据含有褒贬意义的倾向信息来对文本进行分类。其研究对象主要是形容词、动词、名词、副词等等，采用观点来标注语料库，并根据相关的主观因素来判断上下文中所表达的情感[36]。除了这些以词语为基本单位的研究外，也有学者从短语或句子的角度出发来分析文本的倾向性。Casey 等人从文本中提取出具有强烈感情色彩的短语来作为特征评价组，再对其分析得出文本的倾向性[37]。Hatzivassiloglou 等人从较大的文档集中提取出连接词对，根据词对之间的含义生成词汇之间具有同义或反义的连接图，再采用聚类的思想将这些词汇聚集成褒义和贬义两类，由此来分析文本的主观倾向[38]。

在国内，朱嫣岚等人基于 HowNet 的基础上，选取具有强烈褒贬含义的基准词，通过计算单词与基准词之间的语义相似度和语义相关场来得出词语的语义倾向值[39]。熊德兰等人也是在基于 HowNet 计算词汇语义相似度的基础上，提出了基于语义距离和语法距离的句子倾向性计算方法[40]。李钝等人以短语为基本单位来分析文本情感倾向性，首先分析 HowNet 中词语间的语义倾向度，再分析短语中词汇的语法组成结构，对于不同类别的中心词采取不同的倾向值计算方法[41]。胡熠等人提出了一种基于语言建模的文本情感分类方法。该方法首先假设有褒义和贬义两个语言模型，通过设计一个距离函数来计算测试文本和两个假设语言模型之间的距离，进而得出文本的情感倾向[42]。李艳玲等人采用类别空间模型来描述词语对类别的倾向性，分析了词频、词的文本频、词的分布三种统计特征，提出首先采取组合特征提取办法来除去低频词和噪音词，再除去类别倾向性不明显的词[43]。徐琳宏等人先通过计算词汇与知网中已标注褒贬性的词汇间的相似度获取词汇的倾向性，选取具有强烈褒贬含义的词汇作为特征词，利用向量空间模型进行分类，采用否定规则来消除否定句对文本观点的影响，并按词汇与程度副词切分出的距离计算出一个观察窗口，由此来判断褒贬义词汇应增加的词频[44]。王素格等人选取具有情感色彩的词语，设计了 4 种类型的候选特征，计算每个候选特征的信息增益，选取信息增益最大的前 N 个特征作为最终特征；另一种方法是选取情感词汇表与文本中出现的词汇交集，除去信息增益中已选的特征词，再利用支持向量机进行文本倾向性分类。

总的来说，国内外在文本倾向性分析的领域研究上还并不是很成熟，尤其是在国内，中文领域的观点评论倾向性研究还只是刚刚的起步。汉语句子的组合特征使得中文文本的倾向性分析更加困难，在自然语言理解的处理上，人们不得不寻求一种能准确表述上下文语义的方法。随着计算机科技的不断向前发展，中文

领域的观点评论挖掘和倾向性分析依然是值得继续钻研的热点课题。

## 1.3 研究内容和组织结构

网络舆情热点信息的获取与分析是当前研究的重点问题之一，目前已有不少专家学者在该领域的研究取得了一定的成果。本文在前人研究的基础上，围绕网络舆情热点信息的发现及其倾向性分析这一主题作了一定的研究，其主要工作有以下几个方面：

（1）对当前国内外学者的研究现状进行了综述分析，对现有热点信息的发现及其倾向性分析方法作了具体的介绍。

（2）建立了网络舆情热点信息的判断标准，在此基础上，构建了网络舆情热点信息发现的模型，并对热点信息的发现和获取作了详细的算法分析。

（3）手工构建了一部极性词典，并对极性词典的扩充和修正提出了相应的解决办法，该方法能消除未登陆词、否定词和强调词的影响。

（4）设计了一种获取网络舆情信息倾向性的算法，该算法从句子的浅层语法分析出发，计算了词汇的上下文极性、句子的主题极性和整个文本的倾向性。

本文的组织结构如下：

第一章：绪论。主要介绍本文研究的背景和意义，并对舆情基本理论、网络舆情热点信息发现和网络舆情热点信息倾向性分析的研究现状进行综合分析。

第二章：网络舆情热点信息发现研究。主要分析网络舆情信息的采集来源、采集方式和信息预处理，建立热点信息的判断标准，并对热点信息的发现和获取作详细的算法分析。

第三章：网络舆情热点信息倾向性分析。主要给出舆情倾向性分析的工作流程体系，构建极性词典，并对其进行扩充和修正。重点设计网络舆情倾向性分析的算法。

第四章：实验结果与分析。主要是对网络舆情热点的发现和倾向性进行实验，并对实验结果进行讨论分析。

第五章：总结与展望。主要对本文的工作进行总结，并对下一步的研究方向作探讨。

# 第 2 章 网络舆情热点信息发现研究

## 2.1 网络舆情信息的采集

### 2.1.1 信息来源

只有掌握了网络舆情信息的来源，才能更好地对舆情信息进行采集。

（1）BBS。BBS 的全称是 Bulletin Board System，即电子公告板，是目前网民参与讨论、表达意见的最主要场所。它是一个有多人参与的讨论系统，当用户进入讨论区后，可以浏览其他用户发表的文章或话题，也可以发表自己的文章和见解，或者回复他人，展开讨论，这些行为又被称为“发帖”和“跟帖”。BBS 还设有版主，角色类似于传统媒体的“把关人”。目前，国内已经形成一些颇具影响力的 BBS，比如天涯社区、强国论坛等等。网民就一些社会热点、焦点问题发表的言论显示出强大的影响力，甚至影响到政府决策。

根据人民网所作的调查，强国论坛的网友以在科研、教育事业单位和党政管理机关工作的人最多，其中专业技术人员占 35.2%，国家行政机关管理人员占 18.9%，两者占到一半。从学历上看，大专以上学历占到 80.6%，其中本科比例又最高，达到 42.9%。由此可见，在像强国论坛这种有很大影响力的论坛里，人们多具有较高的文化素质，有较高的政治热情和政治眼光，对问题进行分析和判断的能力较强。网络论坛为他们提供了一个交流信息、表达意见的平台。在这种相对宽松的言论环境下，各种思想相互碰撞，必然会对社会政治、经济、文化等多个层面产生潜移默化的影响。

（2）博客。博客（Blog）是网络日志“Web log”的缩略语，是指以网上日志的形式，把个人每天的事件、意见和信息等发布到 Web 上，与他人分享、交流。它以其坦率、无保留、富于思想而奇怪的方式提供无拘无束的言论。博客被誉为“平民媒介”，它是互联网应用上赋予个人以力量的工具，它赋予个人前所未有的权力去影响世界。博客的魅力就来自于在浩瀚世界里构筑了一个个人表达和被倾听的空间，人们可以自由地表达自己、展现自己、经营自己，并对传统的传播力量进行反抗和消解，重构了人们的话语权。

目前，几乎各大门户网站都设有博客这一版块，比如新浪博客、搜狐博客、网易博客、凤凰博客、天涯博客等等，这些博客给广大网民提供了自由发表评论和抒发情感的空间。博主可以在博客上针对某一事件发表自己的看法，读者也可以针对作者的意见留下自己的评论。博客发出的各种声音中所蕴含的民间智慧放

射后所产生的效果表明，对于中国政府的管理层来说，不仅仅是多了一条倾听民间声音的渠道，他们已经开始以行动来回应这些声音，博客这一舆情表达通道必将会促进我国政府与公众之间的良性互动。

（3）新闻组。新闻组是网络媒体中基于电子邮件系统的一项大众化的信息服务方式，是网络用户就相互感兴趣的话题结成的世界范围的讨论小组。它就像一个可以离线浏览的 BBS，是个人向新闻服务器粘贴邮件的集合地。电脑在线时，我们可以通过新闻组浏览软件将新闻组里面的帖子全部下载到本地电脑中来阅读，当然，我们也可以自由地在新闻组服务器上粘贴信息。使用新闻组既可以节省大量上网时间，又可以阅读到大量资料。

每个新闻组都集中于特定的兴趣主题，主题也各式各样，无所不包。在新闻组里，用户可以阅读各类寄来的电子邮件，可以发表文章予以附和或反驳，也可以发表自己的文章到新闻组供他人讨论。在很多特征上，新闻组更像 E-mail 与 BBS 的结合。新闻组在国外的使用频度很高，人们通过新闻组来交流意见，发表看法。

（4）即时通讯。网络聊天是在互联网上实现一对一、一对多、多对多的直接文字交流方式， 是网络实时信息交流的典型体现，也是目前网络中最受使用者欢迎的一种网络服务。除了文字聊天外，音频和视频聊天也很流行。网络聊天主要是通过即时通讯工具或聊天室来实现。即时通讯工具是当前被广泛使用的聊天工具，种类很多，比如 QQ、ICQ、MSN 等等。这类软件的特点就是互动性强，具有网上实时交流信息的功能。腾讯 QQ 目前拥有庞大的用户群体，不少用户根据自己的兴趣爱好组成不同的 QQ 群，人们可以在群里就某一事件自由发表自己的想法和意见，往往会形成一股舆论倾向，甚至演变成用实际行动来表达这种感情倾向。

（5）新闻门户网站。新闻门户网站往往是发布中央重大决策，社会重大事件的发源地，并且现在很多的新闻网站上，都在新闻结尾处设有“参与新闻评论”的模块，这种评论设置给了网民很大的言论空间，任何人在浏览完新闻后，都可以对自己感兴趣的新闻发表自己的看法和见解，并且网民相互之间的留言是可见的，并会在相互之间产生一定的影响。如果某一新闻引起了大量网民关注，并产生了大量的留言信息，那么就可以从这些留言中分析得出网民的情感倾向。比如新浪的“我要评论”、雅虎的“我来说两句”、中国新闻网的“提交评论”等等。

### 2.1.2 采集方式

网络舆情信息比较特殊，主要是那些能够体现网民对某些公共事件的情绪、态度和意见的信息，它包括文字、图片、音频、视频等多种形式，还包括那些反

映网民上网行为的一系列信息。因此，要想全面、及时、准确地搜集到网络舆情信息也不是一件容易的事情。目前，我国相关部门往往采取人工手段来搜集互联网上的舆情信息，这一方式不仅效率低，而且也不一定能满足决策者的需求。因此，本文考虑利用搜索引擎来自动获取舆情信息。

一般来说搜索引擎都有搜索器、索引器、检索器和用户接口 4 个部分组成，现有的搜索引擎主要分为基于目录的搜索引擎、基于机器人的搜索引擎、基于客户的搜索引擎、元搜索引擎和分布式搜索引擎[45]。

网络舆情信息的采集流程如图 2-1 所示，其具体步骤如下：

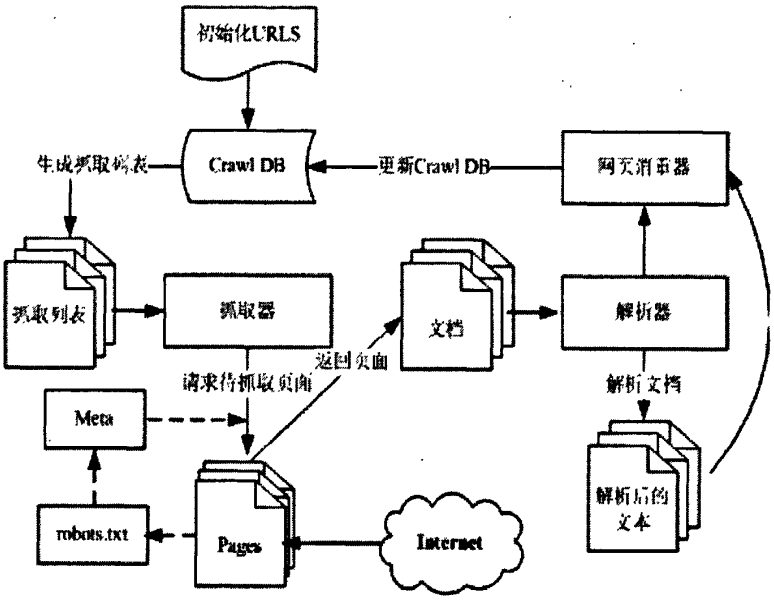


图 2-1 网络舆情信息采集工作流程图

（1）注入抓取 URL。因为搜索引擎的抓取程序要抓取网页，就必须给定一个或一些初始的 URL 入口，从而定位到某个或某些网页，在此基础上，搜索引擎按照广度优先或者深度优先的遍历策略进行抓取。在这一过程中，会构建一个 Crawl DB，将 URL 进行格式化和过滤，以消除部分不合法的 URL，并将已抓取过和未抓取的 URL 区分开来，避免重复抓取。

（2）生成抓取列表。搜索引擎抓去程序需要抓取到很多的网页，具体是哪些页面需要有互联网上的 URL 来指定。这一步骤的主要是对上一步提交的 URL 集合进行分析，确定抓取任务的详细信息。分析提交的 URL 集合之后，建立一个抓取任务列表，在以后的抓取工作中就可以根据预处理的列表进行抓取工作了。

（3）执行抓取，获取网页信息。这一步就开始根据前面生成的抓取任务列表中指定的 URL 对应的页面，这时候开始抓取工作了。需要将抓取到的这些页面文件存放到指定的位置，这些页面文件可以是经过简单预处理以后而被存储到文件系统中，也可以是原生的网页文件，以备后继流程基于这些文件来进一步处

理，比如分词，建立索引。

（4）文档解析。由于抓取的数据是结构和内容非常复杂的数据，而我们感兴趣的主要是文件的内容，通过 content parser 解析器，最终获取到的就是文本内容和其它一些可能需要用到的数据。有了这些可以识别的文本内容和数据，就可以基于此来建立索引库，而且需要将本次抓取任务的详细信息登录到 Crawl DB，为下次抓取任务提供有用的信息(比如：避免重复抓取相同的 URL 指定的页面)。

从图 2-1 可以看出，网络信息的采集是一个不断循环的过程，即由初始化 URL开始抓取，获取网页信息后对文档进行解析，得到新的URL存储到Crawl DB进行更新，重新开始抓取，不断循环直到设置的网页全部抓取完毕[46][47]。

## 2.2 网络舆情热点信息判断标准的建立

### 2.2.1 热点信息的概念和特征

舆论是民众对于公共事务公开表达的具有影响力的意见，舆情则是民众关于现实社会中各种现象、问题所表达的政治信念、态度、意见和情绪的总和；而网络舆论是民众对于公共事务通过信息网络公开表达的具有影响力的意见，网络舆情就是民众通过互联网对政府管理以及现实社会中各种现象、问题所表达的政治信念、态度、意见和情绪的总和；网络舆情热点则是网民思想情绪和群众利益诉求在网上的集中反映，是网民热切关注的聚焦点，是民众议论的集中点，反映出一个时期网民的所思所想。网络舆情热点紧扣社会舆情，往往是社会重大事件，或是与群众切身利益密切相关的问题，很容易在短时间内引起网民广泛关注，对现实社会产生深刻影响。

网络热点信息是从不断更新的海量网络信息中提取出来的，能呈现网络当前重要事件、关注焦点、舆论方向的，经过精简组织的相关信息。它应该具有如下一些特征：

（1）简洁性。简洁性是指用一个或多个精炼的词语或词组就可以表达出热点信息的主要内容。如“杭州飙车案”、“75 新疆暴乱”等。

（2）及时性。及时性是指网络信息必须能够及时体现当前发生的重大事件，因此，在文本信息中可能会出现比较生僻的词汇，如“欺实马”、“躲猫猫”等。

（3）信息量大。信息量大是指热点信息中的关键要素和因子，能够对信息起到代表性作用。如“十七大”、“911”等。

### 2.2.2 热点信息的判断标准

网络舆情热点是对互联网舆情信息进行深刻剖析的最基本的点，也是社情民意的重要体现。舆情热点按其影响范围来分可以分为地方热点和国家热点。地方热点事件是在一定范围内具有较大的影响，但是尚未受到全国广泛的关注；而国家热点是指受到全国人民普遍关注的事件，最为典型的就是 2008 年 5 月 12 日发生的汶川大地震。只要你登录互联网，不论是百度首页还是腾讯首页，头条新闻总是在报道有关灾区的最新情况。进行热点分析首先要明确以下几个问题：分析的热点事件是什么；关注哪些重要参数；网络论坛热点话题是什么。一般而言对于分析的重要参数，人们会关注总发文数（从第一篇与热点事件有关的文章开始一直到统计时点总共发表的关于这一主题的文章数），单位时间发文数（在统计期内，与分析主题相关的文章数），话题阅读总人数（从话题发表开始，到统计时点内所有阅读过该话题的人数），参与讨论总人数（从第一个发表文章的人开始到统计时点所有在该主题下发表文章的总人数），单位时间参与讨论人数（在统计期内，在该主题下发表文章的人数）。关注总发文的变动趋势，能够判断这一主题在当前是受到更多关注还是在逐渐被湮没；关注参与讨论总人数的变动趋势，可以对参与者的活动规律进行研究。对于热点话题的讨论初期，其关注者和参与者人数增长最快，网民发表意见次数最频繁、最活跃，从而也可能是网民情绪急速积聚时期，要予以特别关注和积极引导。通过舆情分析如果我们能够及时的发现舆情热点并对其进行追踪，这将有利于我们对舆情信息疏导与控制。

要制定判断舆情热点信息的指标体系，首先要分析影响舆情热点信息的主要因素。从热点信息的产生、发展、扩散到消亡，有很多影响因素，我们不可能将所有的影响因素都列入判断指标体系中来，因此，本文选定以下几个特征作为热点信息的判断标准：总发文数、单位时间发文数、话题的阅读人数、参与讨论总人数、单位时间参与讨论人数和热点信息的时间跨度。一般认为，对于某一话题，如果关于这个话题的报道数量不断增加，则认为这个话题比较重要；考虑时间因素，单位时间内某一话题的报道数量占这段时间内总发文数的比重如果较大，则认为这个话题比较主要；如果有很多人阅读这个话题，说明对这个话题感兴趣的人比较多；当某一话题发表后，如果参与讨论该话题的人数不断增加，则认为这个话题比较重要；在单位时间内人们对某一话题讨论得越多，则认为这个话题受人们关注的程度越大，这里可以用网民对话题的回帖数来表示参与话题的讨论人数；时间跨度是某一话题从发表到终止的时间间隔，如果某一话题的时间跨度越长，则认为这个话题受关注的程度较大。

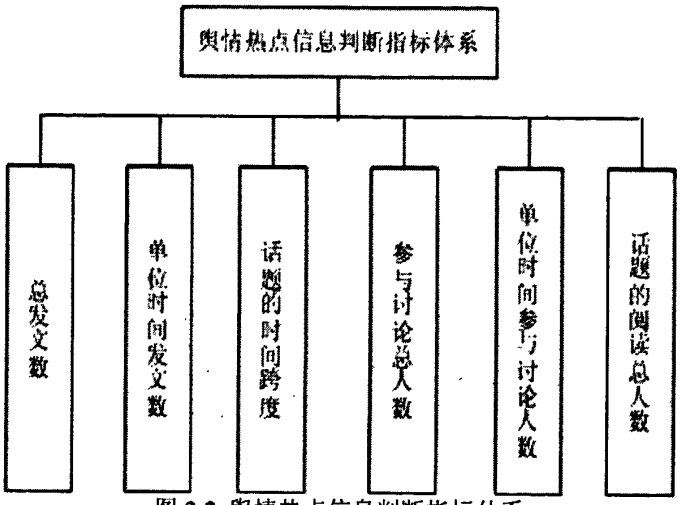


图 2-2 舆情热点信息判断指标体系

## 2.3 网络舆情热点信息发现算法的设计

传统的热点信息发现主要是通过网站的排名得到的，这一排名的依据主要是通过评论数量得出的，评判指标比较的单一，往往并没有考虑评论内容与信息内容之间的相关性。因此，本文按照前面制定的舆情热点信息判断指标体系，将该指标体系分为媒体关注度和用户关注度两部分，并把各个部分的指标进行定量化，在特征项量化的基础上得出热点信息关注度的计算公式。可以设定一个阈值，如果计算得出的关注度大于该阈值，则认为该信息是热点信息。

### 2.3.1 热点信息特征项的定量化

媒体关注度是指新闻信息收到新闻媒体关注的程度。对于互联网这一新兴媒体来说，就是指对于同样的新闻信息，有多少网站对其进行了报道，每个网站对于该新闻的宣传力度有多大。当然，如果某一新闻被报道得越多，则说明该新闻受到的媒体关注度越高；对于同一网站，报道同一新闻信息的相关新闻越多，也说明该新闻受到媒体的关注度越高。新闻媒体对热点信息的形成具有主导作用，只有新闻被网络媒体报道出来，才能够得到网民的关注，关注的多才能成为热点新闻。媒体关注度包括总发文数、单位时间发文数和话题的时间跨度，为了数学计算的方便，本文取某一时间段内的报道天数和报道频率来进行定量化，用参数分别表示为 rt（reported time）和 rf（reported frequency）。

用户关注度是指网民在浏览互联网信息时，对某一新闻或者信息表现出兴趣的关注程度。热点信息的形成同样离不开用户的作用，对于一些新闻门户网站，如新浪新闻，往往会根据用户的浏览量来给出新闻的排行榜，这样会使得新闻的关注度越来越高。对于 BBS，某一话题是否能够成为热点信息，主要就是靠网

民对该话题的回帖数，回帖数越高，该话题的关注度也越高。用户关注度的影响因素主要体现在用户的浏览行为方面，比如，用户如果对某一新闻比较感兴趣，他就会点击该新闻进行阅读，我们记录下用户对新闻的浏览行为就可得知新闻的浏览量。当然，并不说只要用户浏览了某一新闻就能肯定用户对新闻内容感兴趣，现在大部分网站的新闻都设置有评论的功能，如果用户阅读完新闻后留下相关的评论，并且评论越多，则说明该新闻受关注的程度越大。用户关注度主要包括：话题的阅读人数、参与讨论总人数和单位时间参与讨论人数，同样为了数学计算的方便，本文对这几个指标进行优化处理，取单位时间内的阅读人数和单位时间内的评论人数进行定量化，用参数分别表示为 rn（reader number）和 cn（comments number）。

### 2.3.2 热点信息发现算法分析

（1）基于媒体关注度的热点信息发现模型。媒体关注度的主要特征是 rt（报道天数）和 rf（报道频率），本文根据 TF*PDF 的思想来定量描述一段时间内某一信息受媒体关注的程度。TF*PDF 算法主要是用来判别一定时间内不同信息的特征项，其算法的思想基于以下前提条件：一条热点信息必然会被多个网站报道，并且关于该信息的报道数量和报道频率都较高[48]。对于不同新闻来源的信息本文都设定相同的权重，而描述热点信息的特征项必定会在每个新闻来源的多篇报道文档中频繁出现。所以，如果多个新闻来源都含有一个高权重的特征项，该特征项很有可能描述一条热点信息。

其具体的计算模型如下：

$$T_m(i,t) = |rf_i(t)| * \exp\left(\frac{rf_i(t)}{y(t)}\right) * rd(t) \qquad \text{（公式 2-1）}$$

$$|rf_i(t)| = rf_i(t) / \sqrt{\sum_{n=1}^{n=N} rf_i(t)^2} \qquad \text{（公式 2-2）}$$

其中，$T_m(i,t)$ 表示在时间段 $t$ 内，网站上关于信息 $i$ 的媒体关注度，$t$ 可以是任意的时间段，比如一天，一周，一月等等；$rf_i(t)$ 表示在时间段 $t$ 内关于信息 $i$ 的报道总数；$y(t)$ 表示网站上的所有报到总数；$N$ 表示网站上的信息总数；$rd(t)$ 表示在时间段 $t$ 内，关于信息 $i$ 的报道天数。

从公式 2-1 中可以看出，媒体关注度主要包括三个因素。第一个因素是 $|rf_i(t)|$，其含义表示网站上关于信息 $i$ 的标准报道频度；第二个因素是 $\exp\left(\frac{rf_i(t)}{y(t)}\right)$，描述信息在网站上的 PDF，即 proportional document frequency，说明某一信息的报道数量越多，该信息的 PDF 值就越大；第三个因素是 $rd(t)$，把时间因素考虑进去，说明某一信息在某一时间段内报道越集中，那么该信息的关

注度也越高。

（2）基于用户关注度的热点信息发现模型。从用户的角度来考虑热点信息的形成，主要有两个方面，一个是信息的阅读人数，即$rn$（reader number）；另一个是信息的评论人数$cn$（comments number）。对于不同的信息来源方式，两种因素在热点信息的形成方面要设定不同的权重，一般来说，如果用户对某一信息很感兴趣，他会在阅读之后留下自己的评论意见，所以，本文认为评论人数的权重要大于阅读人数的权重。

其具体的计算模型如下：

$$T_u(i,t) = \log_{0.5P_r+P_c+\lambda}(0.5P_r + P_c) \quad \text{（公式 2-3）}$$

$$P_r = rn/(rn+cn) \quad \text{（公式 2-4）}$$

$$P_c = cn/(rn+cn) \quad \text{（公式 2-5）}$$

其中，$T_u(i,t)$表示在时间段$t$内用户对信息$i$的关注程度，同样，$t$可以是一天，一周，一月等等；$P_r$表示关于信息$i$的阅读人数所占的比例；$P_c$表示关于信息$i$的评论人数所占的比例；$rn$表示时间段$t$内关于信息$i$阅读的人数；$cn$表示时间段$t$内关于信息$i$评论的人数（这里的阅读人数指只阅读不发表评论的人数）；$\lambda$在这里表示动态调整因子，用来平衡公式中相关因子对公式的影响。

本文综合考虑媒体关注度和用户关注度两个方面，将结合基于媒体关注度的热点信息发现计算公式和基于用户关注度的热点信息发现计算公式结合起来，构建以下模型来计算热点信息的发现。

$$T(i,t) = \alpha * |rf_i(t)| * \exp\left(\frac{rf_i(t)}{y(t)}\right) * rd(t) + \beta * \log_{0.5P_r+P_c+\lambda}(0.5P_r + P_c) \quad \text{（公式 2-6）}$$

对于参数$\alpha$和$\beta$，可以根据经验来设置其数值大小，并作适当的调整，其主要作用是用来调节“媒体关注度”和“用户关注度”的数值差异，以平衡各因子对整个公式的影响大小。

### 2.3.4 热点信息的获取

利用前面构建的热点信息发现模型，即公式 2-6，计算最近在搜狐网站上各个领域所发表的新闻信息的关注度，将计算结果按照从高到低的顺序排列后，得到一个初步的热点信息排名。考虑到每条热点信息都有从产生、发展到消亡的生命周期，本文借鉴文献[48]中提出的概念“话题指数”，其具体计算公式如下：

$$E_x = \frac{T_x}{T_1} * E_1 \quad \text{（公式 2-7）}$$

其中，$E_x$表示话题在第$x$天的话题指数，$T_x$表示话题第$x$天的关注度，$T_1$表

示话题出现第一天的关注度。利用公式 2-7 可以计算得出关于某一话题在一定时间段内的关注度变化情况，从而可以得到一条反映热点信息生命周期演变过程的曲线。

现在可以依据“话题指数”来得出初步热点信息的话题发展曲线，这样做的目的主要是为了剔除少部分虽然有高报道率高关注度但并不是热点的信息。比如 NBA，每年在各大媒体上都会有大量的报道，并且持续时间也较长，这样的信息即使有很高的报道率和关注度，也不能算是热点信息。但如果姚明在比赛中有什么突发性的状况，可能会引起广大球迷的深切关注，这一信息可以算是热点信息。

在这个基础上，可以根据经验设定某一阈值，当计算结果大于该阈值时，就认为该信息是热点信息，如果小于该阈值，则不认为是热点信息。

# 第 3 章 网络舆情热点信息倾向性分析

## 3.1 网络舆情倾向性分析的思路和方法

网络舆情的倾向性分析主要是判断文本信息中所包含的情感倾向。人们通常是通过评论来发表自己的观点和态度等，评论文本中往往含有丰富的感情色彩，如果能够计算出评论文本的褒贬倾向，也就能得出人们对某一事件的情感倾向。从情感倾向性角度来看，人们对某一事件的看法主要会有三种情况：支持、反对、中立。因此，本文将文本分类的标准也分为三类：

（1）支持：指人们对某一事件持肯定态度，得到民众的支持和赞扬等。其主要表现为评论文本中带有明显的称赞、支持、褒奖等含义的词汇。

（2）反对。指人们对某一事件持否定态度，对此事件进行抵触等。其主要表现为评论文本中带有明显的批评、责备、痛斥等含义的词汇。

（3）中立。指人们对某一时间持中立态度，既不支持也不反对。

对于文本信息的倾向性分析，本文主要考虑以下几个方面：(1) 词汇。词汇是文本信息最基本的组成单元，人们往往通过带有褒贬含义的词汇来表达自己的情感倾向，因此，判断词汇中含有的褒贬倾向是分析文本倾向性的基础， 如何构建一个比较完全的极性词典也是本文的主要工作之一。(2) 句子。句子作为文本信息的组成成分，首先是如何提取出含有褒贬倾向的主观句，并能够反映出作者的情感倾向，其次是要准确分析出句子的褒贬极性，即能够判断出句子的褒义、贬义或中性。(3) 段落。段落的情感倾向是由段落中的句子倾向组合而成，可以将若干句子组合成一个段落来进行分析，也可以取某一自然段落的首尾句子来分析整个段落的情感倾向，还可以计算段落中具有较强代表性的褒贬词汇倾向性或者是主观句的褒贬倾向来确定段落的倾向性。(4) 文章。文章的倾向性体现的是作者对某一事件的整体情感倾向，判断文章的倾向性可以利用机器学习对文本进行分类，也可以分析文章中的句法和词法结构，分析词汇在文章上下文中的极性来分析文章的倾向性。

网络舆情情感倾向性分析的主要流程如图 3-1 所示。对于采集到的网络文本信息，首先对其进行中文分词，提取出能够反映文本信息的特征词汇，计算特征词汇与极性词典中词汇的相似度来获得特征词汇的极性值，考虑到否定词和修饰词对词汇极性的影响，在这里引入否定词词典和修饰规则，按照算法来修正词语的极性值。在提取特征词汇的同时，对文本信息进行句法分析，采用哈工大开发

的 DeParser 工具获取词语之间的依存关系，通过事先建立好的语义模板来对句子进行语义模式的匹配，将其与词汇的极性值结合起来获取词语在上下文中的极性，将句子中词语的极性值累计相加得出句子的倾向性。最后计算评论文本中句子所占的权重，并结合句子的倾向性来获取文本的倾向性。

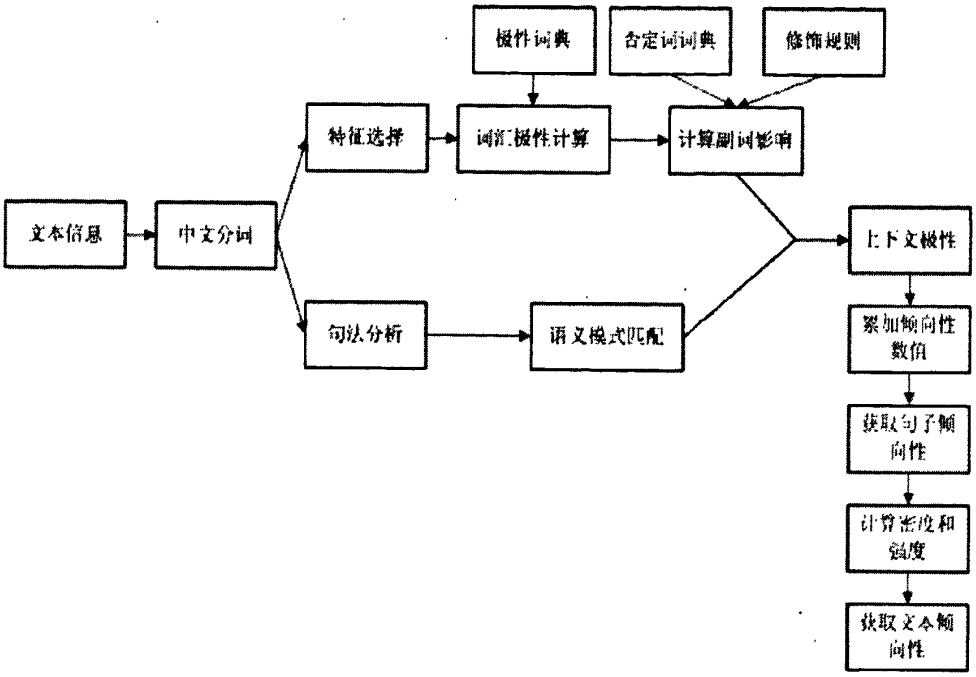


图 3-1 网络舆情倾向性分析流程图

## 3.2 极性词典的构建

在网络舆情信息中，人们往往会对比较热门的信息做出一些评论，通过这些评论来表达自己的情感，对于某一产品是喜爱还是厌恶，对于某一说法是赞成还是反对，这些情感的表达主要由某一些词语来体现。所以一般就把这类带有情感倾向的词语称为极性词。极性词所带有的情感倾向大致可以分为三类：褒义、贬义和中性。比如，褒义词有勤奋、高尚、优秀等，贬义词有懒惰、邪恶、愚蠢等，中性词有普通、运动、评论等。

一般来说，含有褒贬义的极性词基本上主要是形容词和副词，所以在进行词法分析后的形容词和副词都有可能是极性词。但这并不是说所有的形容词和副词都是人们对某一信息做出评论的情感表达，比如“经济快速的增长”中的“快速”就不是一个极性词，在选取极性词的时候就要设计一个合理的算法提取信息中的极性词。国外有学者认为出现在产品特征前后的形容词和副词就应该是能表达人们评论情感的极性词[50]，国内有学者通过对评论信息进行词法和句法分析来确定极性词[51]。本文对极性词的采集和标注都是在现有语料库的基础上，由人工

手动完成的，并构建了一个极性词典模型，如图 3-2 所示。

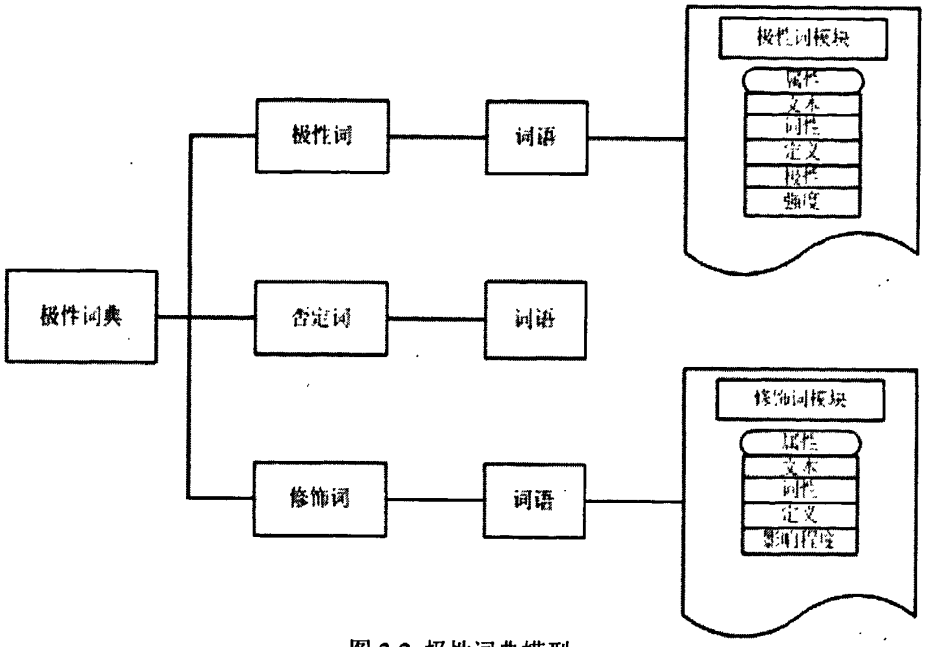


图 3-2 极性词典模型

### 3.2.1 极性词典的初始构建

极性词典就是用来存储词语极性的一种词典，其主要功能是用来快速查找和计算得到词语的极性。极性词典的构建是进行舆情热点信息倾向性分析的基础和关键。

本文设计的极性词典利用已有语料库构建中文极性词库，将现有的英文极性词典在 HowNet 的基础上进行转换，并进行人工修正来提高准确率。其具体过程是首先从英文极性词典中获取英文极性词，然后在 HowNet 中查找与每个英文极性词相对应的词条，获得对应的中文极性词，包括词性、极性和概念定义。接下来在已经生成的中文极性词典中查找，看对应的中文极性词是否已经包含在里面，如果没有，就将其添加进中文极性词典。

极性词典的初始构建工作主要包括词语的极性分类和极性强度的确定。词语的极性分类就是从语料库中挑选出其中的极性词，将其分为褒义、贬义和中性三类，再对这些极性词进行极性强度的标注，显示出词语的极性方向和极性强度。本文将极性词的极性强度分为 5 个等级：-2、-1、0、+1、+2，其中褒义词的极性方向为正，强度值是 1 或 2，贬义词的极性方向为负，强度值是 1 或 2，中性词的没有极性方向，强度值为 0。数值的绝对值越大则表示词语的极性强度越大。举例如下：

褒义词：优秀，+2；良好，+1。

贬义词：阴险，-2；差劲，-1。

中性词：普通，0。

对于某个特定领域的极性词汇，我们很难在现有的极性词典中查找到与之相对应的词语。由于本文是针对网络舆情热点信息进行倾向性分析，因此笔者采用手工方法在《中国网络语言词典》中挑选出具有较强代表性的部分褒贬义词语，并将其作为极性词代表添加到初始的极性词典中。

经过初始构建的极性词典，初步给定了部分词语的极性方向和极性强度，但这样的极性词典在实际应用中会存在着较多的问题。

（1）在汉语中往往一个词语会有多个含义，它们在不同的句子中所表达的情感倾向也不一样。比如，“这台笔记本的配置较低”；“这台笔记本的价格较低”。这两句话所表达的情感倾向就不一样，前面一句话是说“配置较低”，表达的是贬义倾向；后面一句是说“价格较低”，表达的是褒义倾向。

（2）有的词语随着修饰词的不同所具有的情感倾向强度也不同。比如，“这台笔记本电脑的性价比较高”；“这台笔记本的性价比非常高”。很明显，后面这句话所表达的情感强度就要比前一句话大得多，因为“非常高”与“较高”相比，修饰词所体现出的强调作用强化了原有的感情色彩。

（3）现有的极性词典中没有将一些带有褒贬义情感的成语收录进来，极性词的覆盖范围有一定的局限性。比如，“这台笔记本物有所值”，其中的“物有所值”就是带有褒义情感的词语，但像这样的词语在极性词典中并没有出现。

综上所述，初步建立的极性词典只能覆盖一部分极性词，并且由于汉语中普遍存在一词多义的现象，使得同一个词语在不同的句子中会表达出不同的极性。所以，现有的极性词典还不能满足进行文本倾向性分析的需要，还需要对其进行进一步的扩充和修正。

### 3.2.2 极性词典的扩充和修正

（1）未登录词汇的扩充

未登录词汇是指在现有极性词典中还没有收录的词汇，包括生僻词、低频词等。对于未登录词汇的极性计算，国内外已有学者对此作了一定的研究，朱嫣岚等人分别利用 HowNet 语义相似度和 HowNet 语义相关场计算单词的语义倾向值，实验结果表明，基于 HowNet 语义相似度的计算方法比基于语义相关场的计算方法准确率要高[39]。本文采用了基于语义相似度的计算方法。

首先从现有的极性词典中选取 40 对褒贬含义非常显著的词语，根据褒贬强度进行降序排列作为褒贬基准词。每对褒贬基准词都包括一个褒义词和一个贬义词，本文用符号 $x$ 褒义基准词，用符号 $y$ 表示贬义。对于未登录词语 $w$ 的极性值计算公式如下：

$$P(w)=\sum_{i}^{40}sim(x_i,w)-\sum_{j}^{40}sim(y_j,w) \qquad \text{（公式 3-1）}$$

其中，$P(w)$表示词语$w$的极性值，即褒贬倾向的强烈程度，$sim(x_i,w)$表示词语$w$和第$i$个褒义基准词之间的语义相似度，$sim(y_j,w)$表示词语$w$和第$j$个贬义基准词之间的语义相似度。对于词语$w$的褒贬倾向，可以根据$P(w)$值的大小来判断，如果$P(w)$大于 0，则认为词语$w$是褒义词；如果$P(w)$小于 0，则认为词语$w$是贬义词。$P(w)$值的大小就代表词语$w$的褒贬强度大小。

（2）否定词的影响

否定词是文本中经常出现的词汇，并且否定词能明显改变极性词的褒贬义强度，甚至是改变褒贬义偏离方向。比如，“ThinkPad 笔记本不便宜”，这句话中的“便宜”是个褒义词，但在前面加个否定词“不”，使得句子的所表达出的情感倾向就变成了贬义。因此，在判断句子极性的时候必须将否定词的作用考虑进来。否定词的收集是通过人工在现有语料库中挑选的，将一些具有典型否定意义的词语，如“不是”、“否”、“不会”等，抽取出来汇集成否定词集合。

本文采取否定规则匹配来消除否定句的影响，首先从复旦大学提供的语料库中提取出否定句，在这些大量的否定句中由人工提炼出高频的否定规则集合，作为对否定句的判断准则。然后将文本中的否定句与否定规则集合进行匹配，如果否定句中恰好有包含褒贬义倾向的极性词，则用与原有极性词相反褒贬含义的极性词替代，以此来消除否定词对句子中极性词情感倾向的影响。

（3）程度副词的极性计算

程度副词是用来对极性词起修饰作用的词语，程度副词分为相对程度副词和绝对程度分词两大类，无论是相对程度副词还是绝对程度副词都会对句子的极性强度产生很大的影响。比如：“这台笔记本有点便宜”；“这台笔记本很便宜”；“这台笔记本非常便宜”。这三个句子表达的意思都是指这台笔记本便宜，但很显然它们便宜的程度是逐渐递增的。因此，判断句子的极性还要考虑程度副词的修饰作用，为此，本文设计了一种简单的方法来计算长度副词对极性词的影响效果。

修饰词模块则是对极性词模块的进一步判断，如极性词“好”，其极性为褒义，按照本文的极性值设定办法该词的极性为 1，但是若在“好”之前加上修饰词“很”或“绝对”，其结果又将发生怎么样的变化。本文设计了一个简单的极性判定算法，其主要思想是：利用中文分词结果将语句中的极性词及其极性值进行初步定位，句法分析结果则对修饰词极性相应描述，然后对带有修饰词的极性词进行强弱程度计算。

计算方法：

A 若极性词 X 极性值为 1 或-1。

1)若修饰词强度值为 1，则极性词极性值*1；

2)若修饰词强度值为 2，则极性词极性值*2;

3)若修饰词强度值为-1，则极性词极性值*-1;

4)若修饰词强度值为-2，则极性词极性值*-2。

B 若极性词 X 极性值为 2 或-2。

1)若修饰词强度值为 1，则极性词极性值*1;

2)若修饰词强度值为 2，则极性词极性值*1;

3)若修饰词强度值为-1，则极性词极性值*-1;

4)若修饰词强度值为-2，则极性词极性值*-1。

## 3.3 网络舆情热点信息的文本表示

### 3.3.1 文本特征表示

网络舆情信息的主要内容是用中文文本形式描述的，这样的信息只有人经过自身的理解之后才能明白文本中所表达的意思，但对于计算机来说，它并不具备人类的智能，这样的信息它是无法理解的。所以必须将非结构化的数据信息转化成结构化的数据，这样的过程本文称之为文本表示。文本表示就是指根据文本表示的知识（词典）及统计信息将自然文本转换成可为计算机表达的数值向量或符号向量[52]。

中文与英文的表达方式有很大的区别，即英文词与词之间有空格隔开，可以很容易区分出关键词语。中文中的词语连在一起组成句子，句子由标点符号隔开，所以在进行文本处理之前必须对文本进行分词处理。目前的中文分词方法有很多，比如正向最佳匹配法、逆向最佳匹配法、逐词遍历法、切分标志法、特征词库法等。在本文中使用了中国科学院计算技术研究所自行研制的分词系统——“汉语词法分析系统 ICTCLAS（Institute of Computing Technology, Chinese Lexical Analysis System)”进行语料预处理，该系统基于多层隐马模型，具有的功能有：中文分词；词性标注；未登录词识别；新词识别；同时支持用户词典。根据最近的 973 专家组评测结果显示，分词正确率高达 97.58%，基于角色标注的未登录词识别能取得高于 90%召回率，其中中国人名的识别召回率接近 98%，分词和词性标注处理速度为 31.5KB/s。经过先后精心打造五年，内核升级 6 次，目前已经升级到了 ICTCLAS3.0。ICTCLAS3.0 分词速度单机 996KB/s，分词精度 98.45%，API 不超过 200KB，各种词典数据压缩后不到 3M，是当前世界上最好的汉语词法分析器。

在对文本信息进行预处理之后，就可以开始构造文本在计算机中的表示模型。目前主要的文本表示模型有布尔逻辑模型、概率模型和向量空间模型等。在

本文中将采用向量空间模型来表示文本特征。向量空间模型是由国外学者在上世纪60年代提出来的，近年来该模型得到了广泛的应用[53]。

空间向量模型的主要思想是：每一个文本都被映射成为一组规范化正交词条矢量所组成的空间向量中的一个点，即形式化为n维空间中的向量，其具体表现形式为$d_i=(T_{i1},W_{i1};T_{i2},W_{i2};...T_{in},W_{in})$，其中$d_i$表示文本$i$，$T_{i1}$表示能代表文本$i$的特征词条之一，$W_{i1}$表示特征词条$T_{i1}$的权重，$n$表示文本$i$中特征词条的个数。这该模型中，有两个比较重要的影响因素，一个就是特征词条的选择，另一个就是特征词条权重的设置。特征词条必须能够代表文本的主干思想，尤其是能正确表达出文本总所包含的情感倾向；特征权重代表了特征词条在文本中的重要程度，权重越大，则表示在该特征词条在文本中的代表性越强；权重越小，则表示该特征词条在文本中的代表性越差。

### 3.3.2 文本特征词条选取和权重设置

文本特征词条的选择是指从给定的文本中按照某一方法选择出具有一定代表意义的特征词汇，并将这些词汇组成一个特征集合。在向量空间模型中，文本特征词条的选取对文本情感倾向性分析起着决定性的作用，所以本文在选取文本特征词条时，应该把握以下几条原则：首先，特征词条要能够充分代表文本的主要思想，即特征词条在该文本中出现的频率要比较高；其次，特征词条要能将该文本与其他文本区别开来，体现出该文本独有的特性，即该词条在其他文本中出现的频率比较低；最后，特征词条要带有一定的情感倾向，即该词条能反映出文本中作者所表达的观点倾向。

目前国内外关于特征选择的研究已经取得了很大的进展，常用的计算方法有文档频率、信息增益、互信息、期望交叉熵、文本证据权等[54]。各种计算方法都有自己的优缺点和适用范围，本文考虑到网络舆情的文本信息量较大，采用文档频率算法并对其进行改进，来选取文本特征词条。

文档频率DF（Document Frequency）的核心是计算词条在指定文本中出现的次数，其主要思想是：设定一个阈值，如果某一词条的DF小于该阈值，则说明该词条是低频词，这样的词条含有较少的有效信息，应该将该词条从特征集合中能够删除。

本文首先计算词条在文本中的权重，其计算公式如下：

$$w(t_j,d_i)=\sqrt{tf(t_j,d_i)}\times \log\left(\frac{|D|}{df(t_j)}\right) \qquad \text{（公式3-2）}$$

将词条权重值进行归一化处理后，得到如下计算公

$$w'(t_j,d_i)=\sqrt{tf(t_j,d_i)}\times\log\left(\frac{|D|}{df(t_j)}\right)\Bigg/\sqrt{\sum_{t_j}\left(\sqrt{tf(t_j,d_i)}\times\log\left(\frac{|D|}{df(t_j)}\right)\right)^2} \quad \text{（公式 3-3）}$$

其中，$t_j$ 表示文本中第 $j$ 个特征词条，$d_i$ 表示第 $i$ 个文本，$w(t_j,d_i)$ 表示词条 $t_j$ 在文本 $d_i$ 中的权重值，$tf(t_j,d_i)$ 表示词条 $t_j$ 在文本 $d_i$ 中出现的次数，即词条的频率，$df(t_j)$ 表示在文本集合 $D$ 中出现过词条 $t_j$ 的文本数，$|D|$ 表示所有文本数的集合。

在计算得出词条 $t_j$ 在文本 $d_i$ 中的权重值的基础上，本文通过计算词条 $t_j$ 在文本集合 $D$ 中的权重值来进行特征选择。具体的计算思想是：以公式 3-1 的计算方式为基础，定义 $W(t_j,D)$ 为词条 $t_j$ 在文本集合 $D$ 中的权重值，事先给定某一阈值，如果该权重值大于阈值，则将该词条收录到特征词条集合中，否则予以删除。其计算公式如下：

$$W(t_j,D)=\sqrt{TF(t_j,D)}\times\log\left(\frac{|D|}{df(t_j)}\right)\Bigg/\sqrt{\sum_{t_j}\left(\sqrt{TF(t_j,D)}\times\log\left(\frac{|D|}{df(t_j)}\right)\right)^2} \quad \text{（公式 3-4）}$$

$$TF(t_j,D)=\sum_i tf(t_j,d_i) \quad \text{（公式 3-5）}$$

其中，公式 3-4 是经过归一化处理后的计算公式，这样使得词条的权重取值范围在 0—1 之间。

词语是文本中最基本的表示单元，不同的词语对文本的代表性不一样。一般来说，常用词语几乎在所有文本中都会出现很多次数，这样的词语在词频统计算法中往往被作为高频词汇，但实际上它们对文本的代表性很差，比如“的”、“是”、“把”等。当然，有一些在文本中出现次数较少的低频词也有较强的代表性，因为有的词语含有很强的褒贬含义，能够反映出文本的情感倾向。对于这样的特殊词条，本文做了简单的处理，针对出现频率很高的常用词，可以把这些词汇收集起来组成一个禁用词表，凡是出现在禁用词表里的词汇，在统计词频时全部过滤掉。针对有较强代表性的低频词，本文建立了极性词典，对于含有较强极性的词语，予以保留。其详细的流程如图 3-3 所示。

## 3.4 网络舆情倾向性分类算法分析

### 3.4.1 舆情文本信息的浅层句法分析

浅层语义分析是指能够识别句子中某些结构相对简单的成分，比如名词短语、动词短语等，近年来在自然语言处理中比较的流行[55][56]。浅层语义分析的主要工

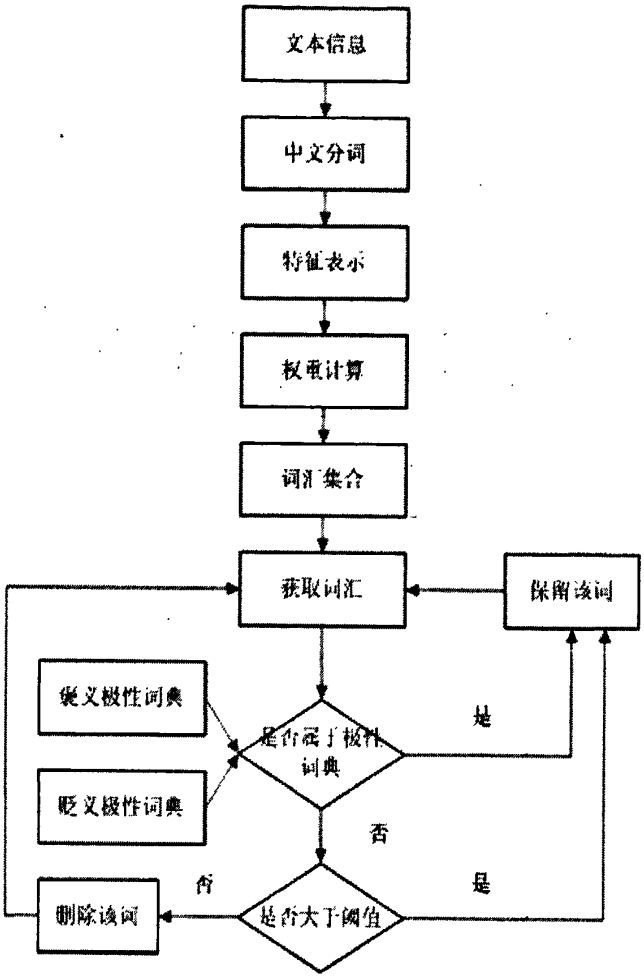


图 3-3 文本信息特征选择流程图

作就是句法分析，本文利用 DeParser 对句子进行句法分析，DeParser 是由哈尔滨工业大学信息检索实验室开发的汉语句法分析器。对于输入的一段中文信息，在其输出结果中有多种可选的显示方式，包括分词、词性标注、命名实体、语义消歧、句法分析和语义分析，用户根据自己的需要选择相应的结果。本文注重的是词性标注和句法分析，在对句子进行了分词和词性标注时，会在句子中每个词及词性的前面加上序号，句子的末尾增加一个句尾标志“<EOS>”，由其支配全句的核心词；在句子中词与词之间的依存关系中，每个关系以一个依存对表示，依存对中的第一个词是核心词，支配第二个词，如：“[2]笔记本_[3]配置(ATT)”这个依存对表示“配置”和“笔记本”之间存在依存关系 ATT，其中，“配置”是这个关系的核心成分，“笔记本”依存于“笔记本”。

例句：惠普笔记本配置比较好，价格也不贵，我觉得很不错！

词性标注的结果：

[1] 惠普/nz [2] 笔记本/n [3] 配置/v [4] 比较/d [5] 好/a [6]，/wp [7] 价格/n [8] 也/d [9] 不/d [10] 贵/a [11]，/wp [12] 我/r [13] 觉得/v [14]

很/d [15] 不错/vg[16] ！/wp [17] <EOS>/<EOS>

句法分析的结果如图 3-4 所示：

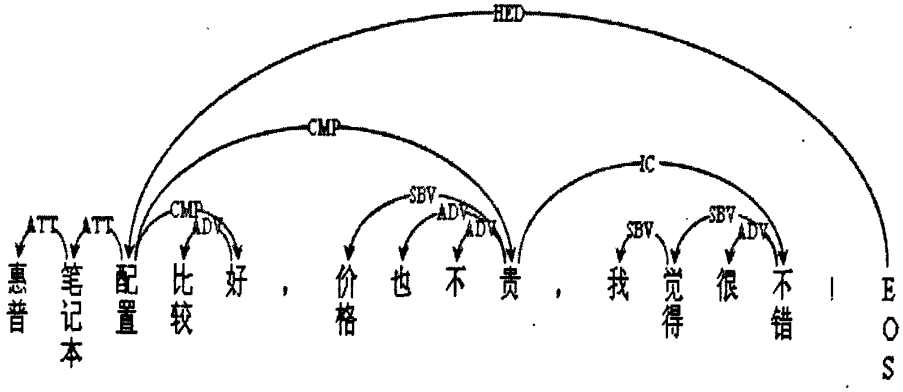


图 3-4 句法分析结果示意图

依存关系对：

[1]惠普_[2]笔记本(ATT)　[2]笔记本_[3]配置(ATT)

[3]配置_[5]好(CMP)　[4]比较_[5]好(ADV)

[7]价格_[10]贵(SBV)　[9]不_[10]贵(ADV)

[11]我_[13]觉得(SBV)　[13]觉得_[15]不错(SBV)

[14]很_[15]不错(ADV)　[17]<EOS> _[3]配置(HED)

括号中的符号表示两个词之间的修饰关系，详细定义见表 3-1。

表 3-1 Deparser 词性和依存关系标记注释

| 词性与依存关系 | 标记 | 词性与依存关系 | 标记 |
|---|---|---|---|
| 定中关系 | ATT | 连动结构 | VV |
| 数量关系 | QUN | 同位关系 | APP |
| 并列关系 | COO | 前附加关系 | LAD |
| 后附加关系 | RAD | 动宾关系 | VOB |
| 介宾关系 | POB | 主谓关系 | SBV |
| 比拟关系 | SIM | 核心 | HED |
| 关连接构 | CNJ | 语态结构 | MT |
| 独立结构 | IS | 状中结构 | ADV |
| 动补结构 | CMP | “的”字结构 | DE |
| “地”字结构 | DI | “得”字结构 | DEI |
| “把”字结构 | BA | “被”字结构 | BEI |
| 独立分句 | IC | 依存分句 | DC |

### 3.4.2 语义模板的建立

在对文本进行句法分析之后，本文可以得到词语与词语之间的语法依存关系，比如动宾关系、主谓关系等。为了更好地对文本进行语义分析，本文提出建立语义模板来抽象表示文本信息，主要是用来分析文本中可能出现的语义模式。该模板不仅仅是简单的包含特征词汇，并且要能体现出词汇之间的语义信息，根据汉语句子的描述特点，本文将句子抽象出以下几种模式：

（1）主体+行为+客体。比如：我喜欢 ThinkPad 笔记本！

（2）客体+被动词+主体+行为。比如：Thinkpad 笔记本被多数商务人士所喜爱！

（3）主体+行为。比如：Thinkpad 笔记本太贵了！

（4）客体+被动词+行为。比如：Thinkpad 笔记本被抢购一空！

其中主体是指句子中的主语，是“行为”的实施者；客体是指句子的宾语，是“行为”的承受者；行为是指句子的谓语，是句子中事件的描述。为了简化语义模板的建立，本文将模板中的主体和客体统一划分为个体，简化后的语义模板可以表示为：

（1）主谓宾模式：个体+行为+个体。

（2）主谓模式：个体+行为。

（3）动宾模式：行为+个体。

### 3.4.3 文本倾向分类的算法设计

本文设计的文本情感倾向分类的算法与传统的文本分类算法有所不同，由于本文的重点是分析文本中所带有的褒贬感情色彩，因此词语在上下文中的褒贬极性是关键影响因素之一。算法步骤描述如下：

输入：经过预处理的网页文本信息。

输出：褒贬倾向性分类后的文本信息。

第一步：对文本信息进行词法分析，提取出特征词汇，并用来对文本进行向量表示；

第二步：查找与给定语义模板中相匹配的语义模式，得到所有的匹配模式；

第三步：结合极性词典，计算每个匹配模式中特征词汇的褒贬倾向性；

第四步：计算词语在上下文中的极性，得出句子的褒贬倾向性；

第五步：综合所有句子的褒贬倾向值，得出整个文本的褒贬倾向性。

下面本文将对算法中涉及到关键技术进行详细的说明。其中的词法分析和特征词汇的提取在前文中已有详细的方法介绍，在这里就不再累述。

在第二步中，语义模式的匹配是整个算法的关键步骤之一。本文将文本信息

进行浅层句法分析之后，可以把文本分解成简单的短句，对每一个短句提取出它的模板，然后将提取出的句子模板与语义模式库中的所有模板进行匹配，得出匹配结果。

在第三步中，主要是计算词语在上下文中的极性值，这里就要用到本文之前建立的极性词典。在计算词汇极性值的时候，如果只是简单计算单个词汇的极性，而不考虑词语在上下文中的语境，往往会导致错误的计算结果。句子中所把包含的感情色彩，除了词语本身的极性以外，一些程度副词的修饰作用对词语的极性值也有很大的影响，否定词往往会时词语的极性方向变反，比如："不喜欢"；而强调词则会使词语的极性强度发生变化，比如："非常喜欢"。因此，词语的极性值计算需要重点考虑其在上下文中的极性。本文对于词语的上下文极性，先是对句子进行句法分析，得出句子中词语之间的依存关系，主要是确定句子中的极性词、否定词和修饰词。然后就对照极性词典，查找出极性词的极性值（包括褒义和贬义），如果出现修饰极性词的否定词，对照否定词典将极性词的极性取反；如果出现修饰极性词的强调词，则按照前文中设计的计算方法对极性词的极性值进行修正。最终综合计算的结果就是词语的极性值。

在第四步中，主要是获取句子的褒贬倾向。句子的褒贬倾向性是指句子中所描述的某一事件或产品的感情色彩，是褒义、贬义还是中立的。极性词的极性值在上面的步骤中已经计算得出，在这一步中需要判别出句子中的主题词语。根据之前对句子的句法分析可以得出词语之间的依存关系，一般来说，主谓结构（SBV）能够描述主语和谓语的修饰关系，其中主语部分可能是发表观点的作者，也可能是句子中的主题词汇，谓语部分可能是形容词，也可能是动词。按照句法依存关系，总能直接或者间接找出修饰主题词的词语极性，本文认为修饰主题词的词语极性就是主题词的极性，这样可以得出句子中所描述主题的极性值。

在第五步中，主要是获取整个文本的倾向性。整个文本的倾向性计算并不能简单的由各个句子的倾向值相加减得出，往往一段评论文本中谈论到的主题词会有多个，不同主题词之间的极性值直接加减的实际意义不大。因此，本文按照主题词的不同将文本中的句子分成不同的类别，对于含有同一相关主题词的句子归为一类，统计出描述该主题的极性词数量（包括褒义词和贬义词），将极性词的极性值相加，其和就是该主题的极性值。对于整个文本来说，本文结合句子的权值和倾向值来计算最终的倾向性。其计算公式如下：

$$C = \sum_{i=1}^{n} S_i * W_i \qquad \text{（公式 3-6）}$$

其中，$C$ 表示整个文本的倾向值，$S_i$ 表示文本中第 $i$ 个句子的倾向值，$W_i$ 表示文本中第 $i$ 个句子的权值，$n$ 表示文本中的句子总数。为简化计算，本文中句子的权值 $W_i$ 用主题词的权值来代替。

如果$C$大于0，则表示整个文本的倾向性为褒义，如果$C$小于0，则表示整个文本的倾向性为贬义。

# 第 4 章 实验与结果分析

## 4.1 网络舆情热点信息发现实验

本文的实验预料来自于各大新闻门户网站，提取了新华网、新浪网和搜狐网上在某一定时期内的时事要闻信息。每天信息采集工具会自动采集指定网站上的新闻信息，将采集到数据保存下来，并记录下新闻的浏览人数、评论人数等信息参数。本文列举了部分话题如下所示：

Topic 1：个税起征点年内可能提高 400 元

Topic 2：江苏盐城大面积停水，百姓抢购矿泉水

Topic 3：云南躲猫猫事件

Topic 4：全国两会召开

Topic 5：中俄欧盟等交替发声欲终结美元独霸

Topic 6：浦发银行高管薪酬有增无减涉嫌违规

Topic 7：中央文件要求县长及书记每月一天接待群众来访

Topic 8：浙江杭州飙车案肇事者被提请逮捕

Topic 9：陕西神木全民免费医疗

Topic 10：四川公交车燃烧事件

Topic 11：湖北邓玉娇刺死官员案

针对列举的部分话题，本文采用第二章中设计的算法公式来计算话题的媒体关注度和用户关注度，如表 4-1 所示，根据计算结果得出关注度较高的话题。

表 4-1 热点话题列表

| 序号 | 话题内容 | 关注度 |
|---|---|---|
| 1 | 云南躲猫猫事件 | 0.40253 |
| 2 | 四川公交车燃烧事件 | 0.38637 |
| 3 | 浙江杭州飙车致死事件 | 0.33581 |
| 4 | 全国两会召开 | 0.32695 |
| 5 | 湖北邓玉娇刺死官员案 | 0.31965 |

本文将该热点话题列表与人民网舆情监察室发布的 2009 年上半年地方应对网络舆情能力排行榜对比发现，关注度较高的话题大部分与人民网发布的一致。权威媒体在判断热点话题时主要是考虑宏观价值和潜在长远的影响，而本文的热点话题判断主要是考虑媒体和人们对话题的关注度，注重群众对话题的情感反

应。总的来说，本文设计的热点发现算法能在某种程度上反映出某一定时期内人们比较关注的话题

## 4.2 网络舆情倾向性分析实验

### 4.2.1 中文分词

利用中国科学院计算技术研究所自行研制的分词系统——“汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)”，对该文段进行中文分词。操作选项选取了“一级标注”，输出结果选取“973”标准。运行结果如图 4-1 所示：

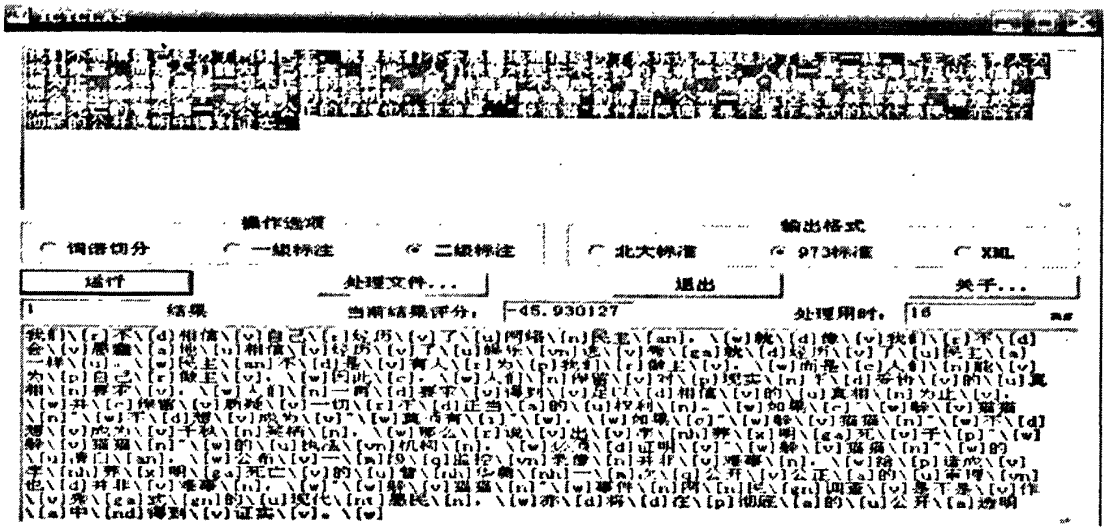


图 4-1 ICTCLAS 分词结果

从上面的运行结果可以看出，ICTCLAS 按照中文理解方式对文段进行了很好的切分， 并对所有切分词的词性进行了精准的标注。但是对于人名并没有做到精确地划分，如“李莽明”、“普华勇”。

### 4.2.2 句法分析

本文利用 DeParser 对句子进行句法分析。采用哈工大的“语言技术平台(LTP)2.0 测试版”进行实验。

（1）该测试版也带有分词及词性标注功能，其结果如下：

我们/r 不/d 相信/v 自己/r 经历/v 了/u 网络/n 民主/a ，/wp 就/d 像/v 我们/r 不/d 会/v 愚蠢/a 地/u 相信/v 经历/v 了/u 娱乐/v 选秀/n 就/d 经历/v 了/u 民主/a 一样/a 。/wp 民主/a 不/d 是/v 有人/r 为/p 我们/r 做主/v ，/wp 而是/c 人们/n 能/v

为/p 自己/r 做主/v ，/wp 因此/c ，/wp 人们/n 保留/v 对/p 现实/n 不/d 妥协/v 的/u 真相/n 要求/n ，/wp 人们/n 一再/d 要求/v 得到/v 足以/d 相信/v 的/u 真相/n 为止/v ，/wp 并/c 保留/v 质疑/v 一切/r 不/d 正当/a 的/u 权利/n 。/wp 如果/c “/wp 躲猫/n 猫/n ”/wp 不/d 想/v 成为/v “/wp 莫须有/i ”/wp ，/wp 如果/c “/wp 躲猫/n 猫/n ”/wp 不/d 想/v 成为/v 千秋/nt 笑柄/n ，/wp 那么/c 说/v 出/v 李荞明/nh 死/v 于/p “/wp 躲猫/n 猫/n ”/wp 的/u 执法/v 机构/n ，/wp 必须/d 证明/v “/wp 躲猫/v 猫/n ”/wp 的/u 清白/a ，/wp 公布/v 一/m 段/q 监控/v 录像/n 并非/v 难事/n ，/wp 给/p 造成/v 李荞明/nh 死亡/v 的/u 普华勇/nh 一/m 次/q 公开/a 公正/a 的/u 审理/v 也/d 并非/v 难事/n ，/wp “/wp 躲猫/n 猫/n ”/wp 事件/n 网民/nh 调查/v 是/v 不/d 是/v 做秀式/b 的/u 现代/nt 愚民/n ，/wp 亦/d 将/d 在/p 彻底/a 的/u 公开/a 透明/a 中/nd 得到/v 证实/v 。/wp

从上述分词结果可以看出，该系统对于姓名做到了很好的切分，但是对于“躲猫猫”并没有识别出来，这有待于舆情热点信息发现对未登录词的处理。

（2）句法分析

该系统将本文段分为三个句子，分别对其句法结构进行分析。

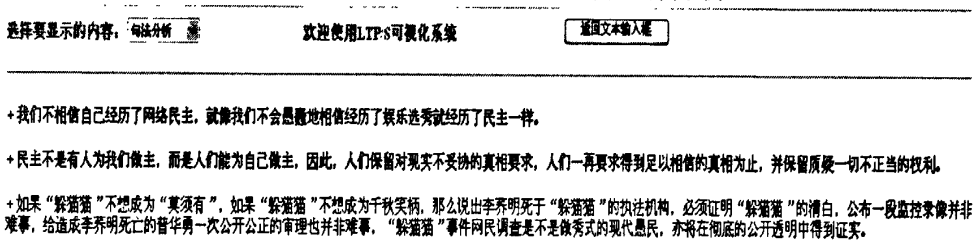


图 4-2 LTP 句法分析

本文选取第一个句子进行词语依存关系分析，得到的结果如图 4-3 所示。

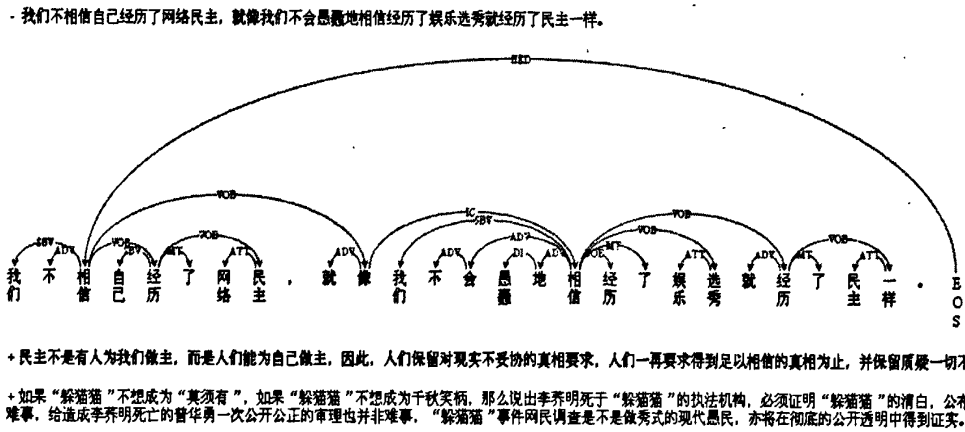


图 4-3 词语依存关系图

虽然截图本身并不完整，但是可以从上图很好地看出句子中词与词之间的关系。其中 SBV(subject-verb)为主谓关系，ADV(adverbial)为状中关系，VOB(verb-object)为动宾关系，DI 为“地”字关系，IC (independent-clause)为独立分句，HED(head)为核心，MT(mood-tense)为语态结构，ATT(attribute)为定中关系。根据此句法分析结果可以很好的确定修饰词与极性词的关系，给舆情热点信息倾向性分析奠定了良好的基础。

### 4.2.3 网络舆情倾向性判断

根据本文之前建立的极性词典，故在本实验中对于极性词及其性质的确定采用了人工选取的方式进行。结果如下：

相信 1 民主 1 愚蠢 -1 娱乐 1 妥协 -1 真相 1 质疑 -1 正当 1
权利 1 莫须有 -2 笑柄 -2 死 -1 清白 1 难事 -1 死亡 -1 公开 1
公正 1 做秀式 -1 愚民 -1 透明 1

词语之间的依存关系如下：

不相信(ADV) 不会相信(ADV) 不是(ADV) 是做主(VOB) 为做主(ADV)
能做主(ADV) 不妥协(ADV) 不正当(ADV) 不想 (ADV) 想莫须有(vv)
不想 (ADV) 想成为(CMP) 成为笑柄(VOB) 并非难事(ATT)

根据第三章中所设计的算法进行极性计算 (极性值*修饰强度值)：

不相信 1*-1=-1 民主 1 不会愚蠢-1*-1=1 不会相信 1*-1=-1
娱乐 1 民主 1 民主 1 不妥协-1*-1=1
真相 1 相信 1 真相 1 质疑-1
不正当-1*1=-1 权利 1 不想莫须有-1*-2=2 不想笑柄-1*-2=2
死-1 清白 1 并非难事-1*-1=1 死亡-1
公开 1 公正 1 并非难事-1*-1=1 不是做秀式-1*-2=2
愚民-1 公开 1 透明 1

统计结果如下：

褒义词数：20 贬义词数：7
褒义值：23 贬义值：-7

褒义 $\frac{23}{23+|-7|}\approx 77\%$ 贬义 $\frac{|-7|}{23+|-7|}\approx 23\%$

从上述结果可以看出，虽然分析过程较为粗糙，但还是可以得出舆情热点的倾向性的。本文段的倾向性偏正向。实际上从人的理解上看，本文段也是正向意味较浓重，并不是对“躲猫猫”事件进行简单粗暴的评价，而是引人朝积极方面思考的文字。

# 第5章 总结与展望

## 5.1 全文总结

### 5.1.1 全文主要内容

鉴于当前的互联网环境日益复杂多变，网民人数逐年增多，人们的生活越来越多地受到网络的影响，因此迅速地发现网络舆情热点信息对应对网络突发事件具有重要的现实意义，准确地进行网络舆情信息的倾向性分析，有利于政府部门及时做出精准的决策，正确地引导舆论走向。面对网络上海量的信息量，本文给出了一套舆情信息的采集方案，从舆情信息的来源入手，设计了详细的采集流程。针对大众和政府部门都比较关注的热点信息，本文根据热点信息的概念和特征建立了热点信息的判断标准，并将热点信息的特征定量化，构建数学模型，用算法来描述热点信息的发现和获取。在取得网络热点信息后，就要对这些信息进行倾向性分析，判断出大众对这一信息的情感倾向，是褒义还是贬义。本文首先手工构建了基础的极性词典，并对极性词典进行了扩充和修正，将未登录词汇、否定词和强调副词对原始极性词的影响进行了进一步分析，并提出相应的解决办法。对于普通的文本信息，要对其进行相应的处理，使得计算机能够识别，用向量来表示文本，通过计算特征词的权重来选取文本的特征词条。考虑到中文句子表达的特殊性，本文对句子进行句法分析，解析出词语之间的依存关系，并对词语进行词性标注。为了结合语境分析词语在上下文中的极性值，本文建立了语义模板，通过语义模板的匹配来确定句子的语义模式，由此，利用极性词典计算出词语的极性值，结合句法分析和模式匹配得出其上下文极性，句子的倾向性由组成句子的主题词和极性词及其极性值决定，文本的倾向性由句子的倾向性和句子在整个文本中的权重决定。

总的来说，网络舆情热点信息的获取和分析是目前一个比较热门的研究课题，本文研究工作在理论上具有一定的参考价值，距离实际应用还有一定的差距。相信在不久的将来，国内外的学者会在实际应用方面取得丰硕的研究成果。

### 5.1.2 主要创新点

本文研究工作的主要创新点如下：

（1）设计了基于媒体关注度和用户关注度的网络舆情热点信息发现算法。该算法结合了基于媒体关注度和基于用户关注度两个热点信息发现模型，分析热点信息的主要特征，并将这些特征定量化，用数学公式来表示计算模型，通过设

定阈值来实现热点信息的发现和获取。

（2）设计了基于词汇语义色彩的网络舆情情感倾向性分类算法。该算法通过构建极性词典来计算词语的极性值，结合句法分析和语义模板匹配来分析上下文极性，句子的倾向性由主题词和极性词及其极性值来表示，整个文本的倾向性由句子的极性值和句子在文本中的权值计算得到。

## 5.2 研究展望

网络舆情热点信息发现及其倾向性分析是一个充满挑战的研究领域，本文所做的工作仅仅是初步的研究与探讨。作者在本文的研究写作过程中，发现还有一些问题需要进一步的探索与研究，主要有以下几点：

（1）本文中给出的网络信息的采集方式存在着一定的缺陷。面对互联网上海量的数据信息，现有的采集方式在采集效率、信息存储等方面无法满足现实需要，因此，是否可将分布式搜索引擎的原理运用到信息采集方式上来，是一个值得探讨的问题。

（2）本文只设计了网络舆情热点信息的发现算法，但在现实中往往需要能够预测到话题未来的发展趋势，即能提前判断某一话题是否会发展为热点信息，这对于国家政府部门来说显得尤为重要。因此，如何将热点信息发现与热点信息预测结合起来将是下一步工作的重点研究课题。

（3）本文所设计的文本倾向性算法还有一些方面需要做出改进。首先文中极性词典的构建和语义模板的建立基本上都是手工完成，不仅工作量巨大，工作效率也不高，而且词语的极性值设置易受人的主观因素影响，因此，是否考虑构建一部权威的中文极性词典，并采取机器学习的方法来建立语义模板，值得进一步探讨。另外本文虽然对句子作了句法分析，但该工具对于长句的分析结果并不是很理想，使得句子的语义分析效果也受到影响，如何更加准确地发现句子中的语义关系，如何更加准确地确定语境中的上下文极性，比如中文一词多义问题，这些都对文本的倾向性分析有很大的影响。

# 参考文献

[1] 中国互联网信息中心. 第二十五次中国互联网络发展状况统计报告. 中国互联网统计报告, 2010(1)

[2] 王来华, 林竹, 毕宏音. 对舆情、民意和舆论三概念异同的初步辨析. 新视野, 2004(5): 64-66

[3] 张毅. 网络舆情管理及分析系统的构建. 湖北成人教育学院学报. 2009(5): 64-65

[4] 张克生. 国家决策：机制与舆情. 天津: 天津社会科学院出版社, 2004: 32

[5] 丁柏铨. 略论舆情——兼及它与舆论、新闻的关系. 新闻记者, 2007: 8-11

[6] 毕竟. 试论高技术传播时代的舆情预警. 新闻记者,2006(4): 39-31

[7] 张丽红. 论民众舆情形成、变化和发生作用的情境. 前沿, 2008(2): 140-142

[8] 谢海光, 陈中润. 互联网内容及舆情深度分析模式. 中国青年政治学院学报, 2006(3): 95-100

[9] 许鑫, 章志成. 互联网舆情分析及应用研究. 情报科学, 2008(8): 1194-1200

[10] 吴绍忠, 李淑华. 互联网络舆情预警机制研究. 中国人们公安大学学报, 2008(3): 38-42

[11] 梅中玲. 基于 Web 信息挖掘的网络舆情分析技术. 中国人民公安大学学报, 2007(4): 85-88

[12] 戴媛, 姚飞. 基于网络舆情安全的信息挖掘及评估指标体系研究. 情报理论与实践, 2008(6): 873-876

[13] 纪红, 马小洁. 论网络舆情的搜集、分析和引导. 华中科技大学学报, 2007(6) : 104-107

[14] 姜胜洪. 试论网上舆情的传播途径、特点及其现状. 社科纵横, 2008(1): 130-131

[15] 刘鹏飞. 网络舆情抽样与分析方法. 青年记者, 2009(3): 4-5

[16] 吕洪波, 姚锦峰. 网络舆情分析系统信息清理的研究. 硅谷, 2009(8): 70

[17] 黄晓斌, 赵超. 文本挖掘在网络舆情信息分析中的应用. 情报科学, 2009(1): 94-99

[18] 杨频, 李涛, 赵奎. 一种网络舆情的定量分析方法. 计算机应用研究, 2009(3): 1066-1068

[19] 王来华, 温淑春. 舆情信息汇集和分析机制刍议. 天津大学学报, 2007(9): 420-423

[20] Yiming Yang, Jaime G. Carbonell, Ralf D.Brown et al. Learning Approached for Detecting and Tracking News Events. IEEE Intelligent System, Intelligent Information Retrieval, 1999: 32-33

[21] kuan-Yu Chen, Luesukprasert, and Seng-cho T. Chou. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentense Modeling. IEEE TRANSCTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2007,19(8): 1016-1025

[22] Allan J, Carbonell J. Topic Detection and Tracking pilot study: final report. Proceedings of

the DAPPA Broadcast News Transctiption and Understanding Workshop, San Francisco: Kaufmann Publishers,1998: 194-218

[23] Wayne C. Multilingual Topic Detection and Tracking: successful research enabled by corpora and evaluation. Language Resources and Evaluation Conference, Greece, 2000: 1487-1494

[24] Matsumura, N. , Ohsawa, Y. , Ishizuka, M. Influence Diffusion Model in Text-Based Communication. Journal of the Japanese Society for Artificial Intelligence, 2002,13(3): 259-267

[25] 北大方正技术研究院. 以科技手段辅助网络舆情突发事件的监测分析——方正智思舆情辅助决策支持系统. 信息化建设, 2005: 50-52

[26] 黄宇栋, 李翔. 互联网媒体信息热点主动发现技术研究与应用. 计算机技术与发展, 2009(5): 1-4

[27] 王林, 戴冠中. 基于复杂网络社区结构的论坛热点主题发现. 计算机工程, 2008(6): 214-216

[28] 鲁明宇, 姚晓娜, 魏善岭. 基于模糊聚类的网络论坛热点话题挖掘. 大连海事大学学报, 2008(11): 52-55

[29] 王义, 张阳, 李书琴. 基于字符串核函数的热点新闻发现系统. 广西师范大学学报, 2007(12): 212-215

[30] 周亚东, 孙钦东, 管晓宏. 流量内容词语相关度的网络热点话题提取. 西安交通大学学报, 2007(10): 1142-1145

[31] 曾依灵, 许洪波. 网络热点信息发现研究. 通信学报, 2007(12): 141-146

[32] 周启海, 黄涛, 张元新. 同构化信息温度与热点发现应用初探. 计算机科学, 2007(11): 113-117

[33] 刘星星, 何婷婷, 龚海军. ; 网络热点事件发现系统的设计. 中文信息学报, 2008(11): 80-85

[34] Wiebe J, Wilson T, Bell M. Identifying collocations for recognizing opinions. In: Proc. ACl-01 Workshop on Collocation: Computational Extraction, Analysis and Exploitation, 2001

[35] Riloff E, Wiebe J, Wilson T. Learning Subjective Nouns using Extraction Pattern Boot strapping. In: Conf. on Natural Language Learning(CoNLL), 2003: 25-32

[36] Turney P, Littman M. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transations on Information Systems, 2003(4): 315-346

[37] Whitelaw C, Garg N, Argamon S. Using Appraisal Group for Sentiment Analysis. In: Proceedings of the 14th ACM international conference on information and knowledge management, Bermen, Germeny, 2005: 625-631

[38] Hatzivassiloglou V, Mckeown K R. Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computationl Linguistics(ACL97), 1997: 174-181

[39] 朱嫣岚, 闵锦, 周雅倩. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2005: 14-20

[40] 熊德兰, 程菊明, 田胜利. 基于 HowNet 的句子褒贬倾向性研究. 计算机工程与应用, 2008(22): 143-145

[41] 李钝, 曹付元, 曹元大. 基于短语模式的文本情感分类研究. 计算机科学, 2008: 132-134

[42] 胡熠, 陆汝占, 李学宁. 基于语言建模的文本情感分类研究. 计算机研究与发展, 2007(4): 1469-1475

[43] LI Yan-ling, DAI Guan-zhong, QIN Sen. A Rapid Method for Text Tendency Classification. 电子科技大学学报, 2007(6): 1232-1236

[44] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制. 中文信息学报, 2007(1): 96-100

[45] 刘晓红. 搜索引擎技术及其发展趋势. 广西医科大学学报, 2008(9): 109-110

[46] Shan-Hua Lin, Jan-Ming Ho. Discovering informative content block from Web documents. In: SIGKDD, 2002

[47] Soumen Chakrabarti, Mukul M.Joshi and Vivek B.Tawde. Enhanced topic distillation using text markup tags and hyperlinks. In: SIGIR, 2001

[48] Bun KK, Ishizuka M. Topic Extraction from News Archive Using TF*PDF Algorithm [A]. In: Proceedings of the 3rd International Conference on Web information Systems Engineering(SISE 2002), Singapore, 2002: 73-82

[49] Ellen Riloff, Janyce Wiebe, Theresa Wilson. Just how mad are you? Finding strong and weak opinion clauses. Proceedings of the 19th National Conference on Artificial Intelligence, 2004: 761-767

[50] Hu M, Liu B. Mining opinion features in customer reviews. In the Proceedings of AAAI (American Association for artificial intelligence), San Jose, California, 2004: 755-760

[51] 娄德成, 姚天昉. 汉语句子极性分析和观点抽取方法的研究. 计算机应用, 2006(11): 622-625

[52] 金晓鸥. 互联网舆情信息获取与分析研究[硕士论文]. 上海交通大学, 2008

[53] Salton G, Wong A and Yang C.S. A vector space model for automatic indexing. Communications of ACM Vol.18, No.11, P613-620, 1997

[54] 许高建, 路遥, 胡学钢. 一种改进的文本特征选择方法的研究与设计. 苏州大学学报, 2008(2): 18-22

[55] C. Wayne. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. Proc. of the Language Resources and Evaluation Conference. 2000: 1487-1494

[56] Abney Steven. Partial parsing via finite-state cascades. Proc. of the ESSLLI'96 Robust Parsing Workshop. Prague, Czech Republic, 1996: 23-40、

# 在读期间的科研成果

## 参与项目

[1] 邢台市桥西区悟思物流中心研究与规划。研究方向：物流系统设计。项目来源：桥西市人民政府，省级。参与物流数据分析。

## 发表论文

[1] 聂规划, 李海林. 基于产业结构的物流需求研究. 中国集体经济, 2009(03): 108-109

# 致 谢

在这篇硕士论文完稿之际，向所有给予我无私帮助、支持的老师、同学、朋友表示最衷心的感谢。

首先要感谢我的导师聂规划教授。本文是在他的悉心指导下完成的，从前期培养计划的制定到论文阶段的论文选题、内容安排、初稿撰写、论文修改直到最终定稿，都倾注了他大量的心血和关怀。在这两年多的课程学习和论文撰写期间，聂老师不仅在学业上给予我精心指导和深深教诲，而且还从精神上鼓励我奋进，在生活上给予了无私的帮助和支持，使我得以顺利完成学业。聂老师知识渊博、治学严谨、为人宽厚、诲人不倦，不仅在学识上为我引路，而且在为人上言传身教，使我深深感动，这宝贵的精神财富将使我终身受益。值此论文完成之际，向聂老师致以崇高的敬意和深深的祝福！

感谢电子商务研究所的陈冬林副教授，刘平峰副教授和其他各位老师，你们在论文设计构思、撰写和修改过程中提出了许多宝贵意见，对论文的完善起到了重要作用。在此真诚地对各位领导和教授的关怀和指导表示崇高的敬意。。

感谢武汉理工大学的王惠敏、曹洪江、付魁、杨爱明、申学武等老师，你们对我的帮助和指导，使我开阔了眼界，丰富了知识，获得了宝贵经验，所有一切都让我受益非浅。同时也感谢研究所的其他师兄弟和师妹们在日常生活和学习中给予的帮助和支持。

最后还要感谢关心过我学习和生活的朋友们！感谢家人对我生活和学习上的关怀与照顾，也感谢同事们在工作上给予的理解和支持！

# 网络舆情热点信息发现及其倾向性研究

作者：李海林
学位授予单位：武汉理工大学