基于twitter的国外网络舆情

分析方法和技术研究

# 摘要

目录

摘要 1

第1章 绪论 5

1.1 研究背景和意义 5

1.2 国内外相关研究 5

1.2.1 国内外网络舆情分析技术 5

1.2.2 国外对于社交网络舆情相关研究 6

1.2.3 国内对于社交网络舆情的相关研究 7

1.3 研究内容和技术路线 7

1.3.1 社交网络舆情主题爬虫技术 7

1.3.2 社交网络舆情文本挖掘技术 8

1.3.3 社交网络舆情社会网络分析技术 8

1.3.4 社交网络舆情数据可视化 8

第2章 社交网络舆情爬虫技术 9

2.1 网络爬虫相关技术 9

2.2 检测要素抓取－突发特征词 9

2.3 基于主题的网络爬虫 9

第3章 社交网络舆情文本挖掘技术 10

3.1 文本分词 10

3.1.1 常用分词方法 10

3.1.2 分词词典扩展 10

3.1.3 人名提取方法 10

3.1.4 事件关键词提取方法 10

3.1.5 热点词权重计算 10

3.2 文本聚类 11

3.2.1 常用聚类算法 11

3.2.2 时间向量相似性 11

3.2.3 人物地点相似性 11

3.2.4 话题聚类模型 11

3.3 主题提取 12

3.3.1 热点话题发现 12

3.3.2 敏感话题发现 12

3.3.3 话题走势 12

3.4 情感倾向性分析 13

3.4.1 倾向性分析与意见挖掘理论 13

3.4.2 词语情感倾向性 13

3.4.3 句子情感倾向性 13

3.4.4 段落情感倾向性 13

3.4.5 社交网络舆情情感倾向性分析模型 13

第4章 社交网络舆情社会网络分析 14

4.1 社会网络分析相关理论 14

4.2 构建社会网格 14

4.3 社交网络舆情行为分析 14

第5章 社交网络舆情数据可视化技术 15

5.1 事件主题河图 15

5.1.1 标签云 15

5.1.2 明细表格 15

5.1.3 网络图 15

5.1.4 分类图 15

5.2 事件发展趋势图 16

5.4 事件分布地形图 16

第6章 社交网络舆情语义网分析 17

6.1 语义网及本体相关技术 17

6.2 社交网络语义挖掘模型 17

参考文献 18

# 第1章 绪论

## 1.1 研究背景和意义

王来华的《舆情研究概论》对舆情的定义是在一定的社会空间内，受中介性社会事项的刺激，作为主体的民众对作为客体的国家管理者产生和持有的社会政治态度。它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和。南京社科院舆情研究所刘毅对网络舆情定义是通过互联网表达和传播的，公众对自己关心或与自身利益密切相关的各种公共事务所持有的多种情绪、态度和意见叫错的总和。

随着网络互联网爆炸式的发展，网络舆情在呈现“舆论多元化”特点的同时，也兼具了开放性和虚拟性，这些特性使网络舆情的形成发展异常迅速，一个现实中的热点事件在一些情绪化或者煽动性言论的带领下，就可以迅速演变成为一个事件的导火索.在目前网络虚拟身份无法实名的状态下，“网络暴力事件”接连发生，网络对社会秩序和社会稳定的影响日益加剧。

而网络中社交网络的迅猛发展对于网络舆情的发展起着举足轻重的作用。我国互联网信息中心(CNNIC)发布的中国互联网发展状况数据统计报告显示，我国国内微博的用户就超过了4亿。微博就已悄然而迅速地改变了社会大众的生活状态，同时被改变的还有传统的社会舆论的变化和微博给社会环境带来的改变，成为社交网络中的重要媒体。近年来，“郭美美事件”、“崔永元方舟子事件”和“360大战腾讯事件”等都是在微博的推波助澜下被曝光给广大社会公众，不仅在社会中引起了轩然大波。

因此如何利用先进的计算机信息处理技术，智能、自动地对网络舆情语言文字进行分析整理，加强监管力度，把握正确的舆情导向，及时通过先进的技术手段发现、防范、分析互联网舆情信息，是当今执政部门以及公安机关的一个重要任务，这也必然是会带动公安机关信息化的一场警务革命。

## 1.2 国内外相关研究

### 1.2.1 网络舆情分析技术

网络舆情分析方法和技术包括以下几个方面：

（1）基于搜索引擎的舆情分析。通过网页内容自动的收集处理信息、进行敏感词过滤、并进行自动聚类、分类、检测主题、聚焦专题、统计研判，完成系统对自己舆情管理的需求，形成最终的舆情研判分析报告，为最后决策提供全面舆情动态，提供引导方向及依据。代表产品是谷尼公司推出的Goonie网络舆情监控分析系统。

（2）基于文本挖掘的舆情分析。以图形图像处理和中文语言处理技术为基础，通过文本管理的方法，对海量的互联网信息进行自动抓取，并进行分类、聚类，同时检测主题和聚焦专题，实现使用者对互联网信息的监测研判需求，并得到最终的以表格、报表、简报形式呈现的分析报告，为使用者提供最合理有效地分析结果。代表产品为方正公司推出的方正智思舆情预警辅助决策支持系统，该系统可以有效地帮助用户解决在大数据环境时所面临的决策问题。

（3）基于语言语法研究的舆情分析。在国外，比较成熟的舆情分析方法是基于关键词统计分析，这也是对自然语言处理的一种研究。通过软件对语言中每个单词的语法成分分析，去除文章中无效的语言干扰，保留有效语法成分，从而确定语言指向和情绪倾向。代表产品是英国的科波拉软件公司的“感情色彩”分析软件。这款软件可以自动分析网络反馈，帮助政府机构对相应的新闻报道或特殊事件确定公众的反应。

（4）基于数学算法的舆情分析。国际方面的舆情分析技术比较成熟，这源于国外对舆情分析的历史和文化发展比较早，也因其有大量的专业机构和分析人员对舆情分析的进一步传承发展。目前，国外的舆情监控系统主要使用基于概念和模式匹配的数学算法，通过建立概率性建模，使舆情分析系统利用文字间的逻辑关系来理解搜索信息，从而在海量信息检索和自然语言检索中，可以实现上下文摘要、搜索结果分组、信息联展等操作。显然，迄今为止计算机还不能做到完全理解文字内容，但通过加强系统的概念识别技术，已经可以将错误率降到最低。主要产品有Autonomy公司为我国政府信息监测部门定做的互联网舆情监控分析系统。

但目前为止还未有针对大规模网络信息使用文本挖掘、分类聚类、趋势分析、社会网络分析和数据可视化等数据挖掘技术对网络舆情进行综合分析方法方面的研究。鉴于数据挖掘在各应用领域所取得的显著成绩，将数据挖掘技术应用于网络舆情分析将会得到十分满意的结果。

### 1.2.2 国外基于社交网络的舆情分析相关研究

最早是 2006 年出现在 blogger 的创始人埃文·威廉姆斯推出的叫“ TWITTER”的一项网络服务中，刚开始被翻译为推特。但时至今日短短的不到十年时间里，中国的推特——微博在全球以互联网为依托的到了广泛的发展。仅在中国，微博的用户已达4亿。随着微博在社会中的广泛使用，在微博上使用的热门词汇也迅速变成流行用语，“微博效应”正在快速形成。

有感于微博的迅速推广和广泛使用，国外与微博有关的学术研究也在近年逐步出现。理查德·威廉姆斯在发表的名为《喳喳，惆啾，咕咕，鸣叫：基于推特的政府机构的沟通模型分析》的学术论文中广大社会组织对微博的使用模式和策略进行了详尽的分析和说明。并认为社会中各单位组织仅仅利用微博的单向交流策略向社会公众发布信息，缺少对社会公众在微博上进行询问的回复和回应，无视了利用微博可以实现与社会公众进行沟通交流的互动式交流策略，建议社会单位和企业组织采用他的卓越关系理论，这个理论在于强调利用微博可以实现双向交流来使社会单位和企业组织可以最大限度的实现和达到与企业组织单位相关利益群体的的长期合作和关系的处理。便于社会单位和企业组织实现达到长期与相关利益群体和社会公众的交流沟通能力，以利于目标的实现。

著名学者勒庞关于研究网络舆论的形成和发展进程的著作《乌合之众》，就经常被西方研究社会舆论的学者所提及和引用，其中最著名的关于社会舆论对大众产生影响的一个专业词汇叫做“集体无意识现象”，这个词汇在对网络影响下的社会舆论或者称之为“网络舆论”指的是大批或者众多网民在网络上被影响和群体性行为，很多学者就借用集体无意识现象对网民的行为进行分析和研究。瑞士心理学家荣格在后来的研究论文中提出了集体无意识的定义和内涵。他把研究重点放在了集体无意识下的的非理性行为所带来的不好的影响和消极作用。这也就是现在普遍认为的集体无意识在社会和网络交流现象中的具体特性和缺陷，有关在网络群体和网络舆情中集体无意识现象更是网络突发事件爆发和激化的的重要原因。

随着时间的发展，国外关于微博网络舆情的研究也在不断的发展和深入，对微博网络舆情进行研究的角度也在不断的发生变化。西方很多学者在社会舆情和网络舆情中的的理论、观点以及他们提出的对待网络舆论的引导和管控手段对我国和本文所进行的微博网络舆论的研究都有重大的借鉴意义。

### 1.2.3 国内基于社交网络的舆情分析的相关研究

在我国社交网络主要包括，微博、微信、论坛、开心网等网络平台。其中以微博最为重要，已拥有超过4亿用户。在微博和网络的使用日益频繁的今天，针对微博舆情的网络舆情研究在国内还比较欠缺，对微博舆情的研究有待深入，研究层面主要还处于向规范研究的发展阶段，在实证研究方面的理论不够深入。对于微博舆情的对策研究和建议也停留在微博舆情问题的表面上，难以在微博舆情的实践中起到良好的指导作用。

虽然国内对微博舆情的研究开始阶段相对较晚，直到近几年才有关于微博舆情的相关研究成果见诸报端。学者杨树林在他发表的《新网络媒体对于社会管理的创新价值一一以当前微博的快速发展为例》研究论文中，就重点讲到了微博舆情对我国社会管理虽然提出了极大地难题和挑战，微博的大量使用对我国社会管理创新发展中的积极作用，最后阶段提出了微博舆情对我国各部门在舆情实践经验上的政策建议。

学者韩玺，顾萍的舆情研究小组在对微博舆情的相关问题进行研究后，在发表的《微博推动下政府网络舆情的传播学研究》一文中对我国目前政府在应对微博舆情时，应该对我国目前发生的微博舆情的现状和问题有一个深入了解，政府英爱加强对微博舆情事件的研究，并在传播模式和政府微博舆情事件的掌控能力和化解能力上有所针对性。专家陈显中主要把自己的研究方向定在了微博舆情的引导上，并把政务微博作为政府进行微博舆情引导的突破口和转变政府职能的推动力量加以重视，在他发表的《政务微博引导网络舆情的机制研究》一文就对我国政府政务微博的发展现状和存在问题作了深刻细致的分析，对我国政务微博的构建和引导提出了解决机制。

## 1.3 研究内容和技术路线

本研究课题将采用数据挖掘中的文本挖掘技术提取网络中的对分析主题相关文本数据进行抓取、去重和清洗，形成分析数据库；对于该分析数据库使用分类聚类、趋势分析和社交网络分析等传统的数据方法进行分析；最后通过数据可视化方法，将分析结果直观的展示出来。

数据挖掘是指从大量的数据中通过算法提取、挖掘未知的、有价值的模式或规律等知识的复杂过程。科学知识图谱的绘制使用了很多数据挖掘方法，常用的方法有聚类分析、数据可视化和社会网络分析等。

### 1.3.1 社交网络舆情主题爬虫技术

聚类分析是将物理或抽象的对象集合分成相似的对象类的过程。簇是数据对象的集合，同一个簇中的对象彼此相似，而不同的簇彼此相异。文献聚类分析是聚类分析技术在引文分析中的具体应用。处理方法是将文献通过分词、去停词等方法转化为词向量，并将每个词条赋予不同的权重，这样一篇文献就可以由词条权重值组成的特征向量来表示，所有文献将组成了特征向量空间模型，在该模型中使用聚类分析技术进行引文分析。

### 1.3.2 社交网络舆情文本挖掘技术

### 1.3.3 社交网络舆情社会网络分析技术

### 1.3.4 社交网络舆情数据可视化

数据可视化（信息可视化）是指将抽象数据用图形图像等可视化形式表示出来，以利于分析数据、发现规律和支持决策。常用的可视化算法有：

自组织特征映射网络SOM（Self-organizaing Feature Map）是一种基于神经网络的算法，它通过把高维数据映射到低维空间进行聚类，并保持一定的拓扑有序性。

寻径网络图谱PFNET（PathFinder Network）是对不同的概念或实体间联系的相似或差异程度进行评估，应用图论中原理和方法生成的一类特殊的网状模型。

三、实验设计及评测

# 第2章 twitter数据收集

## 2.1 传统数据收集方法－网络爬虫

## 2.2 基于twitter API的数据收集方法

2.2.1 twitter REST API

2.2.2 twitter Streaming API

2.2.3 JSON解析

2.2.4 JSON数据的存储

# 第3章 社交网络舆情文本挖掘技术

## 3.1 文本分词

### 3.1.1 常用分词方法

### 3.1.2 分词词典扩展

### 3.1.3 人名提取方法

### 3.1.4 事件关键词提取方法

### 3.1.5 热点词权重计算

## 3.2 文本聚类

### 3.2.1 常用聚类算法

### 3.2.2 时间向量相似性

### 3.2.3 人物地点相似性

### 3.2.4 话题聚类模型

## 3.3 主题提取

### 3.3.1 热点话题发现

### 3.3.2 敏感话题发现

### 3.3.3 话题走势

### 3.4 情感倾向性分析

### 3.4.1 倾向性分析与意见挖掘理论

### 3.4.2 词语情感倾向性

### 3.4.3 句子情感倾向性

### 3.4.4 段落情感倾向性

### 3.4.5 社交网络舆情情感倾向性分析模型

# 第4章 社交网络舆情社会网络分析

## 4.1 社会网络分析相关理论

## 4.2 构建社会网格

## 4.3 社交网络舆情行为分析

# 第5章 社交网络舆情数据可视化技术

## 5.1 事件主题河图

### 5.1.1 标签云

### 5.1.2 明细表格

### 5.1.3 网络图

### 5.1.4 分类图

## 5.2 事件发展趋势图

## 5.4 事件分布地形图

# 第6章 社交网络舆情语义网分析

## 6.1 语义网及本体相关技术

## 6.2 社交网络语义挖掘模型

结论

# 参考文献

[1] 人民网舆情监测室. 如何应对网络舆情-网络舆情分析师手册[M].(新华出版社, 2011).

[2] 韩家炜, 坎伯, 裴健等. 数据挖掘: 概念与技术[M]. (机械工业出版社, 2007. 2007).

[3] Cleveland, W.S. Visualizing data[M]. (Hobart Press, 1993. 1993).

[4] Yang, Y., Akers, L., Klose, T.等. Text mining and visualization tools–impressions of emerging capabilities[J]. World Patent Information, 2008, 30, (4): 280-293.

[5] Wasserman, S. Social network analysis: Methods and applications[M]. (Cambridge university press, 1994. 1994).

[6] Python

[7] Django

[8] Pandas

[9] NLTK

[10] NetworkX

[11] scikit-learn

[12] PyTables

[13] Bootrap

[14] D3