社会焦点透视镜系统——大数据视角下的舆情观测平台

赵妍妍[1]，秦兵[2]，刘挺[2]
1. 哈尔滨工业大学机电学院媒体技术与艺术系，黑龙江 哈尔滨 150001；
2. 哈尔滨工业大学计算机科学与技术学院社会计算与信息检索研究中心，黑龙江 哈尔滨 150001

摘要
Web2.0时代的开启和社会媒体的不断发展，使得互联网上的数据规模呈爆炸性增长。网络大数据不仅为社会治理领域带来了新的契机，也对数据处理技术提出了巨大的挑战。构建了一个社会焦点透视镜系统，结合新浪微博数据，不仅能够实时提供每日的焦点事件及其情感分布展示，供舆情分析部门进行检测，还能够深层剖析焦点事件的情感分布原因和人群分布，协助社会治理领域进行策略的提出和实施。以“9·3阅兵”为例，呈现社会焦点透视镜系统深度剖析的结果展示。

关键词
网络大数据；社会焦点透视镜；焦点事件抽取；情感分布

中图分类号：TP391.1 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016018

Social event sensor: a public opinion platform from the big data perspective

ZHAO Yanyan[1], QIN Bing[2], LIU Ting[2]
1. Department of Media Technology and Art, Harbin Institute of Technology, Harbin 150001
2. Research Center for Social Computing and Information Retrieval of Computer Science and Technology School, Harbin Institute of Technology, Harbin 150001

Abstract
The development of Web 2.0 and social media has led to the explosive growth of online user generated content. Big data brings a new opportunity for social governance, but also poses a great challenge for the data processing technology. A social event sensor system was constructed, which not only can automatically extract the daily hot events and their emotion distributions in real time for opinion monitoring, but also can deeply analyze the emotion distribution causations and the population distributions to help policy-making in social governance. Finally, one case study “9.3 Parade” was showed to show the deeply analysis of social event sensor system.

Key words
big Web data, social event sensor, hot event extraction, sentiment distribution

1 引言

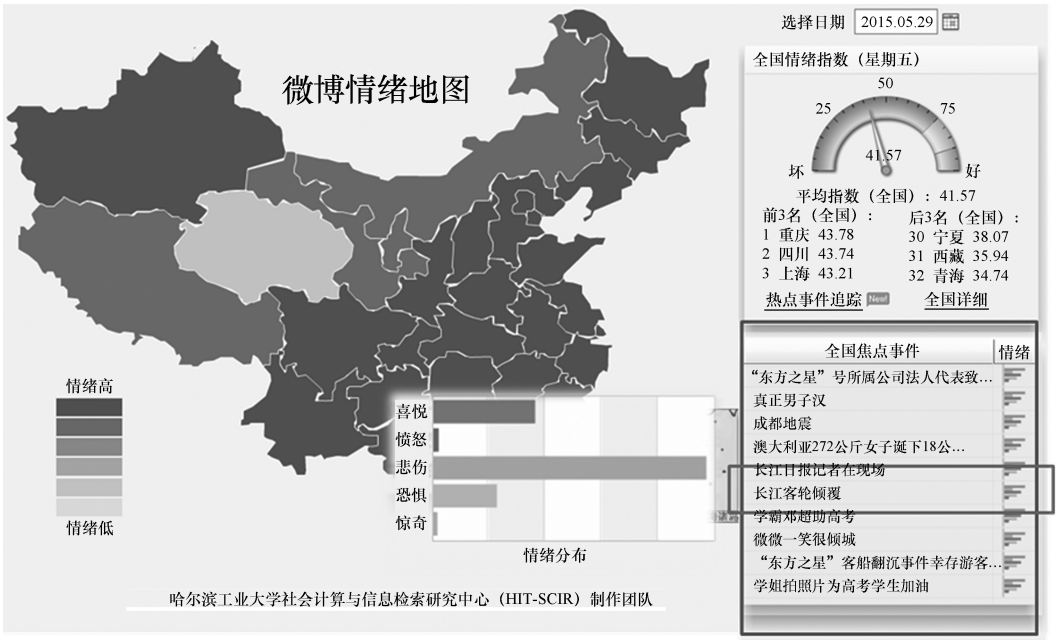
Web 2.0时代的开启和社会媒体（如微信、微博）的出现使得大量用户从被动地在网络上接收知识转变为海量网络数据的产生者。据统计，互联网上的数据每年将增长50%，每两年便翻一番，网络大数据应运而生。目前，大数据的研究和应用价值已在很多领域初见端倪。例如：在零售业，可以在大数据中挖掘出高消费者和高影响者两类有价值的客户，进行产品推荐和口碑宣传，与社交网络相结合创造出新的商品营销模式。此外，社交网络中的大数据也为很多政治选举提供了新的宣传手段，最典型的如在Facebook上开展的奥巴马的总统竞选运动。

随着大数据理念和相关技术的不断深入，大数据应用也在慢慢向社会治理领域渗透。2015年8月31日，国务院以国发〔2015〕50 号印发《促进大数据发展行动纲要》。大数据发展与“提升政府治理能力现代化”紧紧相连，成为全文亮点。大数据将如何助力政府治理，以改善百姓民生、社会服务成为大家最为关注的话题。众所周知，爆炸性增长的大数据蕴藏着巨大的价值，因此寻求有效的大数据处理技术、方法和手段成为基于大数据进行社会治理的最本质的需求。

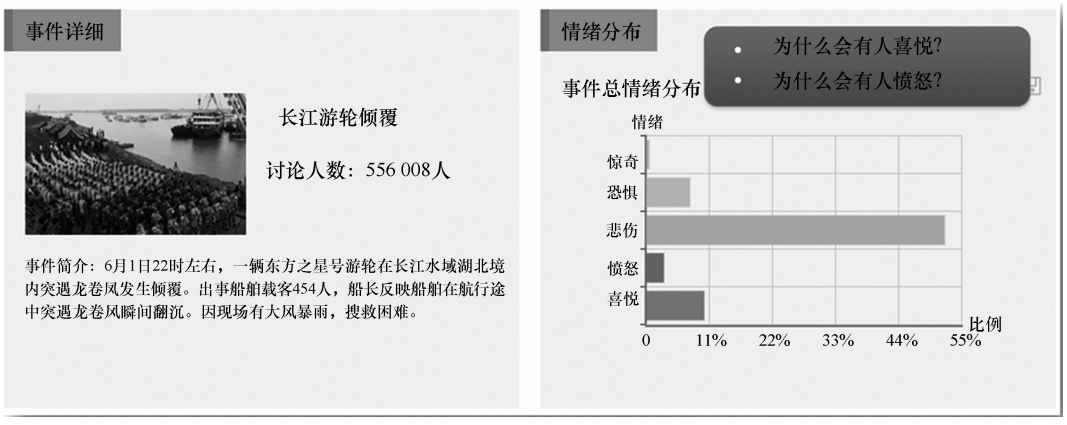
在众多的大数据形式中，社会媒体数据，如微博和微信数据，是很好的一种洞察民情、观测大众行为的数据形式。例如，当某一焦点事件发生时，大量民众在微博上发表自己的观点，可以通过观测相关的微博大数据来统计并获取民众对于该事件的情感分布趋势，继而协助相关部门进行社会治理策略的提出和实施。此外，微博大数据还可以挖掘出民众普遍关注的话题类型、暴露出民众的整体情绪趋势，供舆情部门监测。

目前国内外已经有多项借助微博或Twitter来进行浅层社会治理和分析的技术和系统。Zhao等人[1]构建了一个名为MoodLens的中文微博情感分析系统，将微博的情感分为愤怒、厌恶、高兴和低落4类，进行异常或突发事件的监测。Wang等人[2]构建了一个实时的预测2012年美国大选结果的系统，该系统通过统计Twitter上民众对于4位候选人的情感分布来进行结果预测。Ciot等人[3]研究了Twitter上进行用户性别预测的算法。Diao等人[4]研究了如何在Twitter上实时发现突发事件。Jennifer等人[5]研究了在Twitter上某个事件的发生时间预测算法。以上这些有代表性的系统和算法均是围绕微博或Twitter大数据中焦点事件抽取和情感分析这两大项任务进行的，属于浅层的大数据分析结果显示，存在的问题是缺乏事件和情感的深层分析和透视。这些传统的系统和研究往往只关注民众关心的焦点事件是什么，情绪走向是什么。如**图1（a）**所示，基于自然语言处理技术，可以对2015年5月29日的微博大数据进行分析，挖掘出全国十大焦点事件以及每个事件的民众情感分布，属于浅层分析，分析出的结果可以为相关部门提供一定的预警信号。

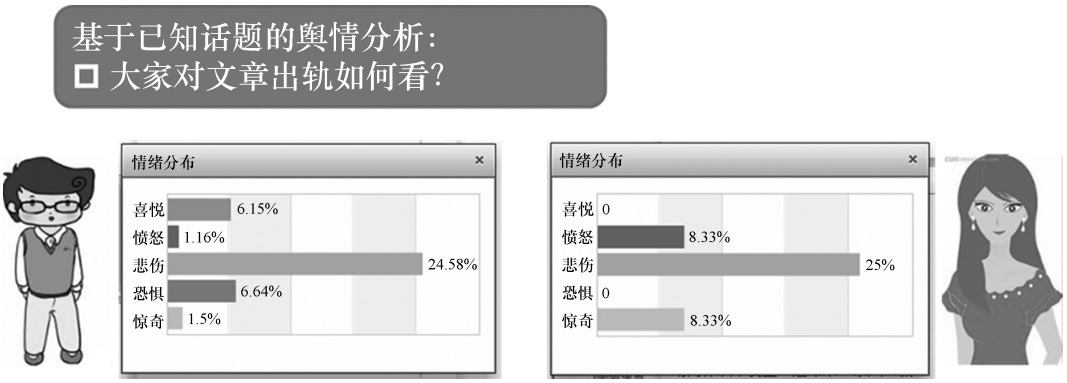
然而，对于社会治理而言，相关部门更关心的是为何某一事件的发生会产生异常情绪、什么样的人群会导致某些情绪的产生等深入的原因剖析，基于此来指导社会治理方案的制定。如**图1（b）**所示，看到民众对于“长江游轮倾覆”事件的情绪分布后，相关部门更想知道为何会有人喜悦、为何会有人愤怒等异常情绪的形成原因。又如**图1（c）**所示，相关部门还想知道针对同一焦点事件，不同的用户画像（性别、职业、年龄等）产生的情感分布的差别是什



（a）每日微博焦点事件发现与情感分布展示（浅层分析）



（b）焦点事件的情感分布原因剖析（深层透视）



（c）焦点事件的不同用户画像（性别）的情感分布展示（深层透视）

图1 面向焦点事件的情感浅层分析和深层透视

么，用以框定某一异常情绪的用户群体进行监测。相比微博大数据的浅层分析和呈现而言，深层透视能够更精准地聚焦原因和人群，显然对社会治理有更大的帮助。

基于此，本文将详细展示一个大数据视角下的舆情观测平台——社会焦点透视镜系统。该系统围绕微博大数据进行焦点事件及其情感分布的深层透视，旨在为新时代的社会治理提供创造性的思路。如前文所述，社会焦点透视镜系统包括两大部分的内容：焦点事件发现与情感分布展示；焦点事件情感分布原因和人群的深层透视。在第一部分内容中，本系统主要采用了事件抽取技术和情感分析技术；在第二部分内容中，主要采用情感原因分析技术和用户画像技术。本文以“9·3阅兵”为例，呈现社会焦点透视镜系统的浅层和深层分析展示结果。

2 社会焦点透视镜系统

社会焦点透视镜系统是一个实时的互联网大数据舆情监测平台。通过对微博海量数据的分析、挖掘和可视化，构建社会焦点事件的发现、追踪和挖掘的深层透视。**图2**展示了社会焦点透视镜系统的流程，共包括两部分内容：社会焦点事件的浅层分析和社会焦点事件的深层透视。

- 社会焦点透视镜的浅层分析：主要包括事件发现和情感分析两个模块。浅层分析可以每隔2 h实时更新当天的焦点事件，并实时对这些焦点事件进行民众情感的分析。如**图1(a)**右侧框中所示的焦点事件以及情感分布。此外，社会焦点透视镜的浅层分析还可以实时给出全国各省民众的整体情绪指数及各省民众关心的焦点事件。如**图1(a)**中显示的地图，从深至浅代表了情绪指数(喜悦情绪)由高至低。
- 社会焦点透视镜的深层透视：主要包括情感归因分析和基于用户画像的情感分析两个模块。深层透视是对某一段时期的某一个焦点事件的深层剖析。主要从两个角度入手，一个是导致某一种情绪的事件原因，另一个是导致某一种情绪的人群归类。

下面将详细介绍每个模块。

2.1 数据来源

选择新浪微博作为实时的数据来源。新浪微博汇集了有关焦点事件的民众的多角度评论以及民众每天的行为情绪动态。社会焦点透视镜系统每天的微博处理总量在1 600万条微博左右，每2 h更新一次。数据格式见**表1**。

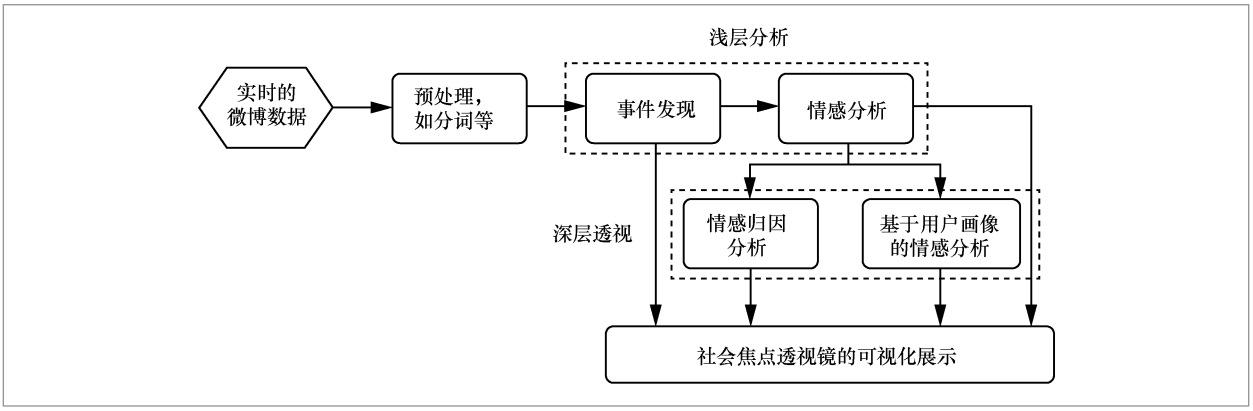


图 2 社会焦点透视镜的系统流程

表1 微博数据格式

数据项	示例
用户ID	2054436031
发布时间	Tue Apr 30 00:01:29 CST 2013
省份编号	11
所在地	北京 延庆县
性别	男
微博内容	失而复得的心情甭提有多开心了，哈哈

2.2 预处理

预处理主要包括两个部分：文本噪声预处理和文本分析预处理。

文本噪声预处理部分包括去广告、去水军和文本去重等步骤。考虑到微博可能存在的广告会对后续的事件发现和情感分类等步骤造成干扰，本文收集了400条广告标记短语，用于过滤带有广告词汇的微博；同时结合新浪微博数据中心的水军过滤算法，初步缓解部分话题水军泛滥的问题；考虑到每日系统需要处理海量数据的微博，其中部分微博存在表述重复的现象，针对性地对其进行去重处理。

文本分析预处理部分包括必要字符的替换、分词和词性标注等步骤。考虑到微博文本的特点，即用户信息（例如“@张三”）和短链接信息（例如“http://t.cn/Ryrc”）等会对后续的步骤造成干扰，所以本文对其进行必要的替换或屏蔽；后续步骤本文使用哈尔滨工业大学语言技术平台（language technology platform，LTP）[①]对文本进行精准的分词与词性标注。

① http://www.ltp-cloud.com/

2.3 事件发现

这里的事件具体是指微博焦点事件，即在短时间内被大量用户高度关注、讨论的话题。有些话题与社会事件密切相关，如“长江游轮倾覆”、“马航失联”等；有些话题与社会事件无关，但仍在短时间内获得了很高的关注度，如：“你最喜爱的男神”、“最美英语教师”等。微博焦点事件不同于官方媒体的头条新闻，微博植根于草根之中，是普通大众的心声或思想的网络直接反馈。基于此，挖掘浩瀚如海的微博中的焦点事件变得尤为重要。

在社会焦点透视镜系统中，设计了一个实时微博焦点事件抽取框架。该框架的核心是基于统计的思想，利用启发式规则和聚类算法。该框架主要包含3个组成部分，分别如下。

- 话题发现：候选热点话题发现。
- 话题聚类：候选热点话题聚类。
- 话题排序：对聚类得到的话题聚簇进行排序，排序靠前的即焦点事件。

具体的算法可见参考文献[6]。

2.4 情感分析

这里使用的情感分析技术具体是指面向焦点事件的情绪分类，最终显示为如**图1(a)**所示的焦点事件的情绪分布。其中的基础技术环节是，针对一条包含焦点事件的微博，判断它所表达的情绪是“喜悦”、“愤怒”、“悲伤”、“恐惧”还是“惊奇”。

情绪分类是情感分析领域研究得比较

深入的一项基础任务，主要有基于情感词和基于分类器两大类方法。其中基于SVM（support vector machine，支持向量机）和丰富特征的方法是最经典和快速的方法[7]。近年来，随着深度学习在自然语言处理的深入发展，深度学习技术在情感分类领域也取得了较好的效果[8]。因此，在社会焦点透视镜系统中，笔者采用了词向量和SVM经典特征相结合的方法[9]，取得了较好的性能。

这里值得一提的是，微博的口语化较为严重，充斥着隐式情感（如："满满的正能量"，"我给他打满分"）和反讽（如："你真是太给我长脸了！"）、隐喻（如："此人是垃圾"）等丰富的语言现象，这给情感分析技术提出了较大的挑战，这也是未来努力的目标。

2.5 情感归因分析

如前文所述，"事件发现"和"情感分析"模块属于社会焦点透视镜系统的浅层分析。用户更想探究的是为何会有某种情绪的产生、什么导致了某种情绪等更深层的透视。这也是本文的社会焦点透视镜系统不同于国内外其他现有系统的重要区别所在。在该系统中，第一层次的透视就是面向焦点事件的民众情绪的原因分析，具体体现为哪个子事件的发生导致了这种情绪。

本系统首次提出了情感归因分析任务，并使用自动抽取用户自然标注的Hashtag作为子事件的算法来解释焦点事件的原因分析。**图3**显示的是"长江游轮倾覆"事件的"喜悦"和"愤怒"两种情绪的原因分析。从**图3（a）**中可以看出，子事件"沉船内部有生命迹象"和"载客458人已救起8人"的情绪分布中"喜悦"的情绪占据了一定的比重，因此这两个子事件可以用来解释"长江游轮倾覆"事件所表露出的"喜悦"情绪。从**图3（b）**中可以看出，子事件"乘客家属收到诈骗短信"的情绪分布中"愤怒"的情绪占据了大部分的比重，因此该子事件可以用来解释"长江游轮倾覆"事件所表露出的"愤怒"情绪。具体的算法细节可见参考文献[10]。

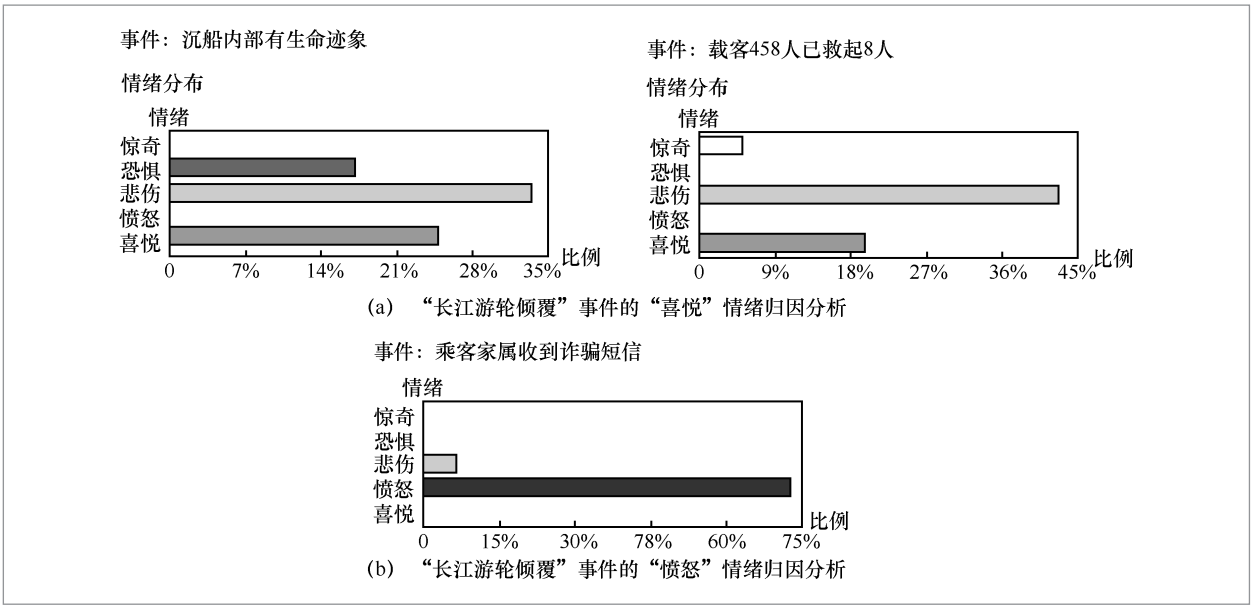


图3 "长江游轮倾覆"事件的"喜悦"和"愤怒"两种情绪归因分析

2.6 基于用户画像的情感分析

除了焦点事件的情绪归因分析之外，用户不同群体与情绪之间的对照也是社会焦点透视镜系统深层透视的重要组成部分。事实证明，不同的用户群体对同一事件的情绪反馈也不同。如**图1(c)**所示，不同性别的用户对“文章出轨事件”的情绪反馈是不同的。用户群体的特性除了用性别表示外，还有职业、年龄、地域等不同的用户画像角度，都可以从不同的侧面展示出不同的用户群体对同一事件的情绪反馈。如果能够将用户的各个画像角度与情绪分析相结合，无疑是从用户角度对焦点事件的深层次透视。

在目前的社会焦点透视镜系统中，仅仅针对用户的地域和性别两个维度的属性，对提及的微博数量进行了统计分析，**图4(a)**和**图4(b)**分别展示了针对“长江游轮倾覆”事件用户在省份和性别这两个维度上的微博数量。当然，将不同的用户属性与其情感分布进行对照是更深入的社会舆情透视，这也将是下一步的研究工作。

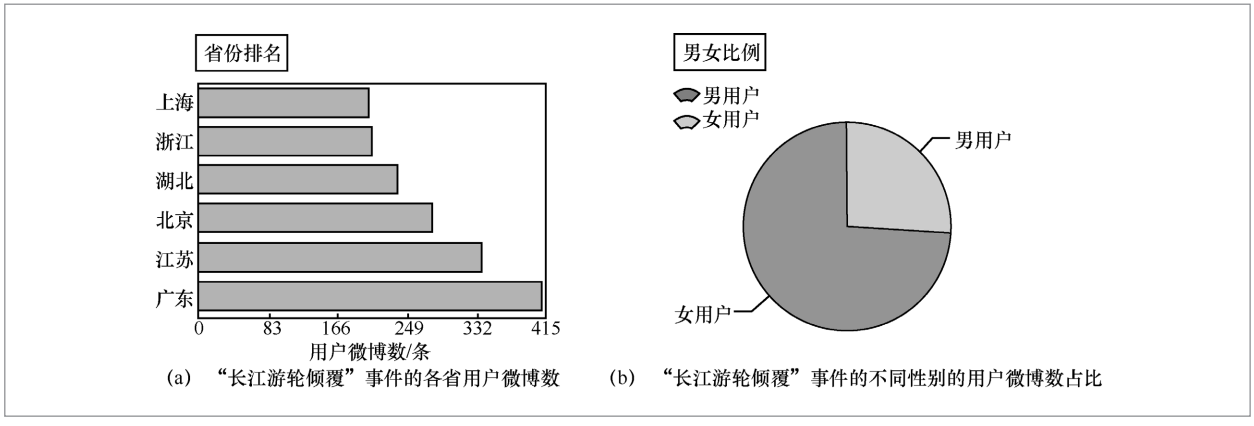
除了以上几个重要模块的展示外，社会焦点透视镜系统还有其他一些丰富的展示效果，详情请见http://qx.8wss.com。

3 社会焦点透视镜系统的应用实例——“9·3阅兵”

与人民网和新浪微博合作，笔者将社会焦点透视镜系统的关键技术用于了2015年的“9·3阅兵”话题中，推出了阅兵大数据“网民情绪展示”平台，该平台每隔15 min刷新一次数据，进行展示。

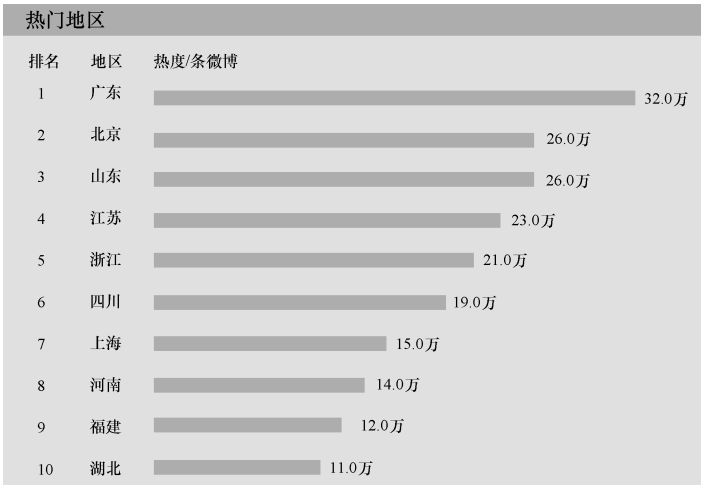
“网民情绪展示”平台的主要功能包括：网民实时关注热门地区排名、网民实时评论阅兵热点高频词分析、网民实时热点话题排行榜以及整个阅兵过程中的舆情走势等。该平台共采集了9月3日8：30-12:30这4 h内网民在新浪微博平台上的阅兵相关话题，并进行分析统计。大数据分析结果显示：原创微博及转发微博总帖量共计453万人次，网民参与发帖的峰值点出现在中午12:00，峰值数据为50万人次；热门地区被广东、北京和山东包揽前三甲；网民热议的高频词有“国泰民安、挺身而出、舍生忘死”等；“习近平宣布将裁军30万”成为网民最热议的话题。

图5是“网民情绪展示”平台的部分数据截图。

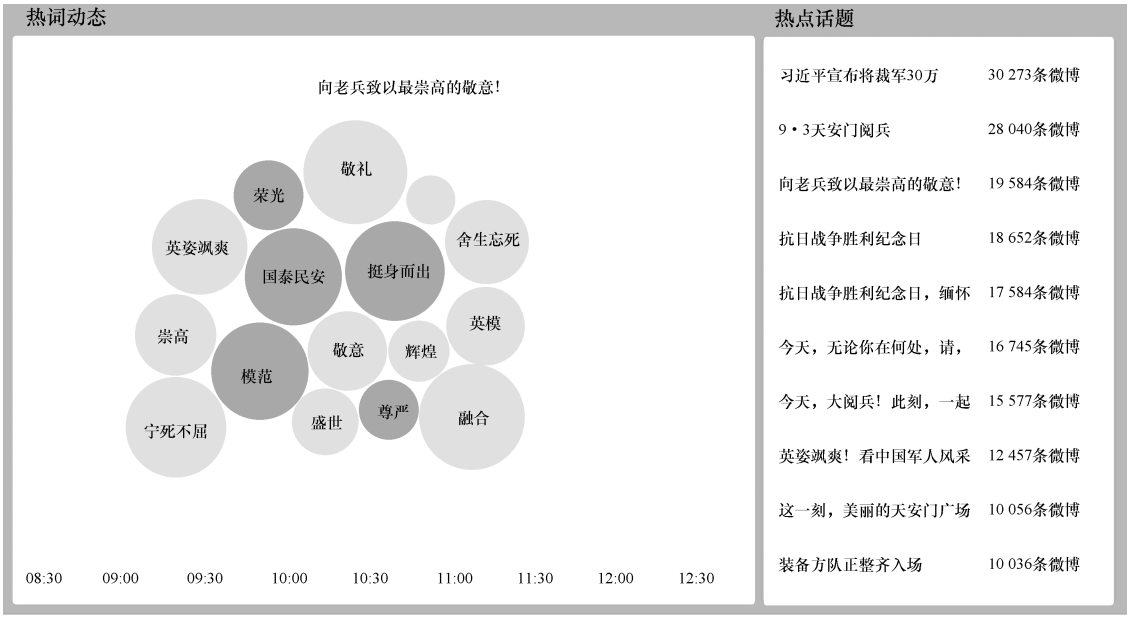


(a) “长江游轮倾覆”事件的各省用户微博数　(b) “长江游轮倾覆”事件的不同性别的用户微博数占比

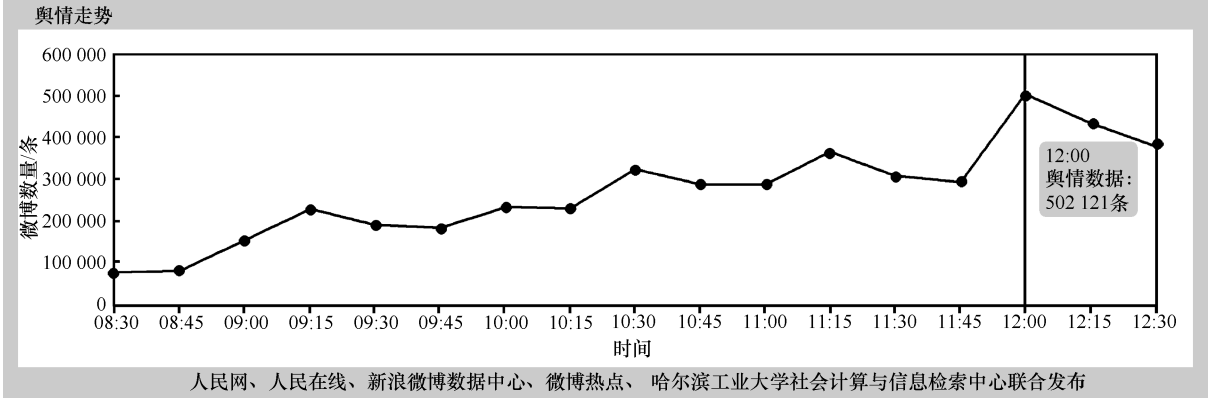
图4 “长江游轮倾覆”事件微博数量



（a）“9·3阅兵”全国各省的关注热度



（b）“9·3阅兵”中网民热议的话题



（c）“9·3阅兵”随时间变化的舆情走势

图 5 “网民情绪展示”平台的部分数据截图

4 结束语

本文介绍的“社会焦点透视镜系统”是微博大数据时代下的一种新型的舆情监测平台。该系统不仅可以像传统系统一样展示出社会热议的事情及民众的情绪分析，还可以深层透视焦点事件背后情绪分布的原因及其相应的用户群体，相信可以对当代社会治理方案的制定提供一定的技术支持。

参考文献:

[1] ZHAO J C, DONG L, WU J J, et al. MoodLens: an emoticon-based sentiment analysis system for Chinese Tweets in Weibo[C]//The 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 12-16, 2012, Beijing, China. New York: ACM Press, 2012: 1528-1531.

[2] WANG H, CAN D, KAZEMZADEH A, et al. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle[C]//The ACL 2012 System Demonstrations, July 8-14, Jeju Island, Korea. New York: ACM Press, 2012: 115-120.

[3] CIOT M, SONDEREGGER M, RUTHS D. Gender inference of Twitter users in non-English contexts[C]//The 2013 Conference on Empirical Methods in Natural Language Processing, October 18-21, 2013, Seattle, Washington, USA. Sofia: Association for Computational Linguistics, 2013: 1136-1145.

[4] DIAO Q M, JIANG J, ZHU F D, et al. Finding bursty topics from microblogs[C]//The 50th Annual Meeting of the Association for Computational Linguistics, July 8-14, Jeju Island, Korea. New York: ACM Press, 2012: 536-544.

[5] WILLIAMS J, KATZ G. Extracting and modeling durations for habits and events from Twitter[C]//The 50th Annual Meeting of the Association for Computational Linguistics, July 8-14, Jeju Island, Korea. New York: ACM Press, 2012: 223-227.

[6] ZHAO Y Y, QIN B, LIU T, et al. Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on Microblog[J]. Multimedia Tools and Applications, 2014: 1-18.

[7] MOHAMMAD S M, KIRITCHENKO S, ZHU X D. NRC-Canada: building the state-of-the-art in sentiment analysis of Tweets[C]//The International Workshop on Semantic Evaluation, June 2013, Atlanta, USA. New York: Association for Computational Linguistics, 2013: 321-327.

[8] SOCHER R, PERELYGIN A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//The Conference on Empirical Methods in Natural Language Processing(EMNLP 2013), October 18-21, 2013, Seattle, WA, USA. Sofia: Association for Computational Linguistics, 2013: 1631-1642.

[9] TANG D Y, WEI F R, YANG N, et al. Learning sentiment-specific word embedding for Twitter sentiment classification[C]// The 52th Annual Meeting of the Association for Computational Linguistics (ACL 2014), June 22-27, 2014, Baltimore, MD, USA. Baltimore: Association for Computational Linguistics, 2014: 1555-1565.

[10] ZHAO Y Y, QIN B, DONG Z J, et al. What causes different emotion distributions

of a hot event? A deep event-emotion analysis system on microblogs[C]//The 4th CCF Conference on Natural Language Processing & Chinese Computing(NLPCC 2015), October 9-13, 2015, Nanchang, China. Berlin: Springer, 2015: 453-464.

作者简介

赵妍妍(1983-),女,哈尔滨工业大学机电学院媒体技术与艺术系副教授、硕士生导师,中国中文信息学会社会媒体处理专委会委员,主要研究方向为社会计算、情感分析。

秦兵(1968-),女,哈尔滨工业大学计算机科学与技术学院教授、博士生导师,社会计算与信息检索中心副主任,中文信息学会信息检索专委会委员,中国计算机学会中文信息技术委员会委员,主要研究方向为社会计算、自然语言处理、文本挖掘。

刘挺(1972-),男,哈尔滨工业大学教授,社会计算与信息检索研究中心主任,中国计算机学会理事,中国中文信息学会常务理事、社会媒体处理专业委员会主任,国际会议ACL 2014、EMNLP 2015领域主席,主要研究方向为社会计算、信息检索和自然语言处理。

收稿日期:2016-01-20
基金项目:国家自然科学基金资助项目(No.61300113, No.61273321, No.61133012)
Foundation Items: The National Natural Science Foundation of China(No.61300113,No.61273321,No.61133012)