DeepOpinion:一种网络舆情深度分析与引导系统

张 文[1] 杜宇航[1] 李自强[1] 陈进东[2*]
(1. 北京化工大学 经济管理学院, 北京 100029; 2. 北京信息科技大学 经济管理学院, 北京 100192)

摘 要: 在广泛调研以往互联网舆情系统存在的一些问题和不足的基础之上,研发了一套互联网舆情深度分析与引导系统 DeepOpinion。该系统从用户、主题、情感三个方面对互联网舆情进行分析,并通过信息检索和内容定制技术对所关注的网络舆情进行适当引导。以天涯杂谈板块为网络舆情来源、以"辱母杀人案"事件为具体案例对系统功能进行了验证,结果表明 DeepOpinion 系统在互联网舆情深度分析与引导中发挥重要作用。

关键词: 网络舆情; 舆情分析; 舆情引导; DeepOpinion 系统
中图分类号: G350.7 **DOI**: 10.13543/j.bhxbzr.2018.04.017

引 言

网络舆情是指人们在受到某些经由互联网传播的事件刺激后而产生的对该事件的认知、态度、情感和行为倾向的集合体[1]。随着互联网的普及,网络日渐成为一种公众表达诉求的途径,网络舆情也成为社会舆情的重要组成部分,对社会和政治生活的影响日益突出,所以开展网络舆情监测成为越来越多机构用户的现实需求。

目前,国内外分别在舆情系统的设计与实现、舆情主题分析以及情感分析方面做了大量工作。在舆情系统研究方面,肖慧华[2]设计了一套政府网络舆情监测系统,能够实现对海量互联网信息的自动抓取、分类聚类。针对舆情主题,龚磊[3]在主题模型 LDA 挖掘的基础之上,利用可视化技术发现主题演化规律;Lo[4]利用 support vector machine (SVM)进行顾客抱怨信息自动分类,提高了客服的工作效率和客户的服务满意度。针对舆情情感分析,王英等[5]基于情感维度内容结合网络舆情情感特征,通过情感分析方法统计情感值,分析情感状况,得出评论高峰期内情感趋势和情感状态分布,实现网络舆情事件的预警研判。但是目前网络舆情系统存在两个问题:①分析维度单一,只能做词频统计、情感倾向分析或者聚类分析,对于舆情的真实价值分析与运用比较粗浅,并不能真正判断事件发展的程度、需要引导的方向和应对措施;②缺乏交互性,只能作单向的舆情分析而不能进行反向的舆情引导进而影响舆情的走向。

针对以上问题,本文开发了一种网络舆情深度分析与引导系统,该系统融合数据采集、用户、主题、情感等多维度分析方法,首次将舆情引导加入系统,通过多维度舆情分析,从大量的网络观点中挖掘到有价值的信息,对危险舆情进行有针对性的引导,以便整体掌握舆情事态的发展。

1 DeepOpinion 网络舆情深度分析与引导系统

1.1 总体架构

DeepOpinion 网络舆情深度分析与引导系统总体架构如图 1 所示。其基本思想是实时地采集互联网数据,并对采集的数据进行用户、主题、情感等多维度的分析,最后根据分析出的结果有针对性地进行舆情引导。

1.2 网络舆情数据采集

网络数据采集通过网络爬虫技术来实现。爬虫的主要目的是将互联网上的网页下载到本地形成一个互联网内容的镜像备份。采集过程为:从一个初始网页开始,获得初始网页上的 URL 后,不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一定停止条件,再将队列中的 URL 通过 HTMLParser 解析获取所需要的文本信息。采集流程如图 2 所示。

收稿日期: 2017-11-29
基金项目: 国家自然科学基金(6137046/61432001/71601023);
中央高校科研业务费(buctrc201504)
第一作者: 男,1981 年生,教授
*通信联系人
E-mail: j.chen@amss.ac.cn

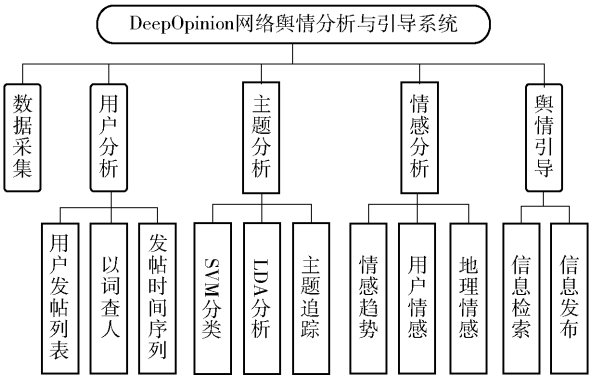

图 1　DeepOpinion 总体架构图

Fig. 1　Architecture diagram of DeepOpinion

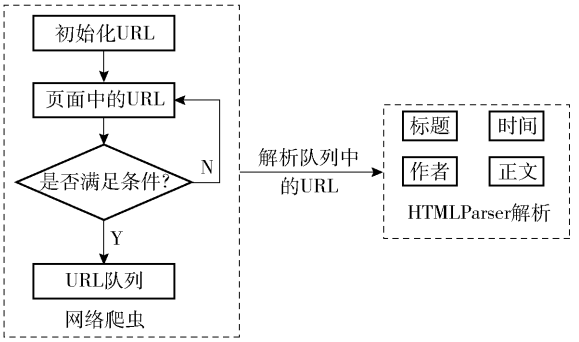

图 2　数据采集过程

Fig. 2　Data collection process

1.3　用户分析

用户分析是指通过分析活跃用户的舆情来把握当前舆情的走向，本文主要利用“以词查人”、“用户发帖列表”、“发帖时间序列”等方法来实现。首先通过“以词查人”利用话题中独有的关键字来查找该话题下的参与用户，并统计分析找出参与该话题最多的一些用户（也称意见领袖）；然后使用“用户发帖列表”分析意见领袖所发表的言论，他们往往不止参与一个话题，而是某一类话题或某几类话题，如社会突发类事件，当事件发生时，可以直接查看意见领袖对于该事件的言论，在事件发生初期就迅速掌握话题走向；最后通过“发帖时间序列”分析用户在单位时间内的发帖数量，以便了解话题热度及趋势、舆情开始时间、集中爆发时间以及整个事件发展过程。

1.4　主题分析

主题分析是指分析文本的主题类型、主题要素以及它们之间的相互关系[6]。浅层主题分析仅能确定主题边界，即区别文本所属的主题，比较复杂的分析能够在识别边界的基础上讨论主题内容。本文将从浅层 SVM 分类[7]、较为复杂的潜在狄利克雷模型（LDA）[8]以及主题追踪[9]3 个方面来进行主题分析。首先通过 SVM 训练出一个有效的分类器，将舆情分为不同类别；再利用 LDA 主题模型提取数据中的隐结构的统计模型，将主题模型方法应用于网络舆情文本集的主题挖掘，有助于在文本层面按照主题的思想分析理解海量舆论中的主要观点；最后通过主题追踪对主体感兴趣的主题发展趋势进行动态追踪，通过计算 SimHash 的海明距离[10]筛选出与所追踪主题相似的舆情，从主题层次的角度全面把握某个主题相关事件的时效性及其动态演化关系。

1.5　情感分析

情感分析是对给出的舆情文本的感情进行分析、归纳的过程[11]，即判断一条舆情中的观点持有者对某件事持有的态度（积极、消极或中立），以便了解舆情的情感分布，及时做出预警。本文使用台湾大学中文情感极性词典 NTUSD[12]来计算舆情情感，通过统计所获得文本中情感类别（积极和消极）的个数，利用情感公式计算每篇文档的情感值 *Sen*

$$Sen = Q_{正} - Q_{负} \tag{1}$$

式中 $Q_{正}$ 和 $Q_{负}$ 分别表示积极和消极词数量。

1.6　舆情引导

区别于传统的互联网舆情系统，DeepOpinion 加入了舆情引导的功能，在分析某一舆情事件后，根据其情感态势作出是否引导的决定，当消极情感超过该舆情的 70% 时，系统就会发出预警，提醒使用者对该事件进行引导。DeepOpinion 舆情引导主要通过两种方式进行：①通过内置 Chrome 浏览器进行舆情信息的检索，在百度新闻中查看积极信息，使用网络爬虫技术提取网页内容，并自动填入到转发区，可实现对内容的编辑；②自行填写转发区内容，再通过 HTMLClient 异步发表，实现舆情引导，使事件情感转向积极的方向。

2　实证分析

2.1　数据采集

为了验证本文系统的功能，通过系统的数据采集功能，利用 1.2 节中描述的爬虫算法将天涯杂谈网站的首页设为种子站点，将“http：//bbs. tianya. cn/post-free-”开头的 URL 放入队列，再解析队列中的 URL。共采集 2017 年 1 月 1 日至 6 月 30 日期间天涯杂谈总计 88 262 条数据，以此数据为基础，实现各项功能。

2.2 结果与讨论

2.2.1 用户分析

本文选取2017 年影响较大的事件“山东辱母杀人案”进行分析,利用文本关键词选择方法 TF-IDF[13] 选择出该舆情事件中的关键词,并使用 JUNG 开源工具绘制关键词网络如图 3 所示。可以看出,整个事件都在围绕几个关键词进行,其中“辱母”一词在关键词中较为突出,它贯彻整个事件的始终,与其他关键词联系也较为紧密。

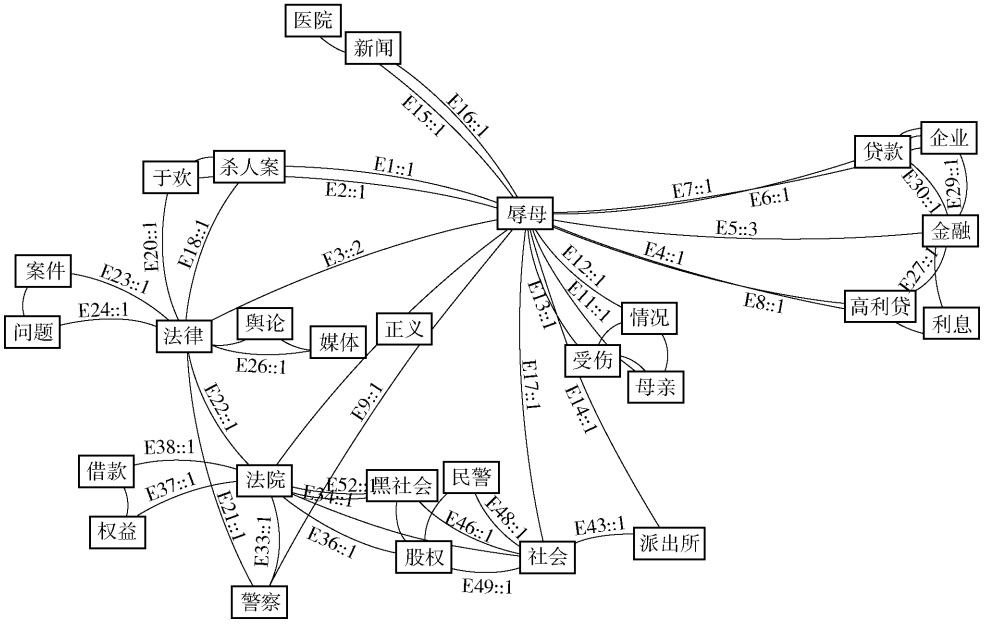

图 3 2017 年 3 月关键词网络图

Fig. 3 Keyword network diagram for March 2017

另外,通过该系统用户分析功能发现,这一事件在天涯杂谈总计 225 名用户参与讨论,总计发表舆情 290 条,并且随着事态的发展关注的用户越来越多,事件持续发酵。其中参与最多的用户发表言论达到 21 次,分析发现该用户所发表内容大多是关于社会问题的讨论,对社会问题有较为深刻见解,其可能作为意见领袖对舆情造成一定影响。

2.2.2 主题分析

已标注数据训练集构建的 SVM 分类器将舆情内容分为 7 大类别[14],见表 1。可以看出与该事件紧密相关(发帖数最多)的帖子主要有 3 类:讨论本事件经过的舆情归属于“社会稳定”;讲述本事件导火索“高利贷”的舆情归属于“金融经济”;探讨本事件中政府法院部门处理此事的过程归属于“政府执政”。

表 1 SVM 主题分类结果

Table 1 SVM classification

范围	日常生活	精神文明	政府执政	社会稳定	金融经济	资源环境	其他类型
全部	6 206	7 243	3 461	15 748	846	2 471	45 006
辱母杀人案	6	12	67	153	21	0	31

利用 LDA 将该事件分为 6 个主题(表 2),LDA 自动分类出的关键字可以帮助了解某一主题下的主要内容和相关内容。可以看出,6 个主题各自包含的信息分别为:主题 1 下的关键字表明“辱母杀人案”事件是由“高利贷”所引起的;主题 2 是事件发生后法院的判决情况;主题 3 描述政府在该事件中扮演的角色;主题 4 是“黑社会团伙”在该事件中的行为;主题 5 是事件发生后舆情的反应;主题 6 是案件发生后医院的处置方式。另外,在所有关于“辱母杀人案”的舆情中,有更多帖子是关于事件本身发展过程和用户对于该事件的个人观点。

2.2.3 舆情引导

通过情感分析得到的情感趋势图如图 4(a)所示,可以看出公众的态度普遍很消极。当消极帖子

表 2　LDA 主题分布

Tabel 2　LDA topic distribution

主题 1	主题 2	主题 3	主题 4	主题 5	主题 6
高利贷	警察	政府	公司	法律	医院
民间	母亲	村民	法院	社会	新闻
银行	讨债	派出所	股东	行为	事情
利率	儿子	土地	集团	案件	业主
借贷	人员	老人	黑社会	问题	护士
企业	法院	社会	民警	舆论	时候
金融	母子	老百姓	股权	媒体	电话
公司	行为	工作	农民	法官	事件
贷款	情况	问题	权益	正义	对方
利息	山东	部门	借款	杀人案	之后

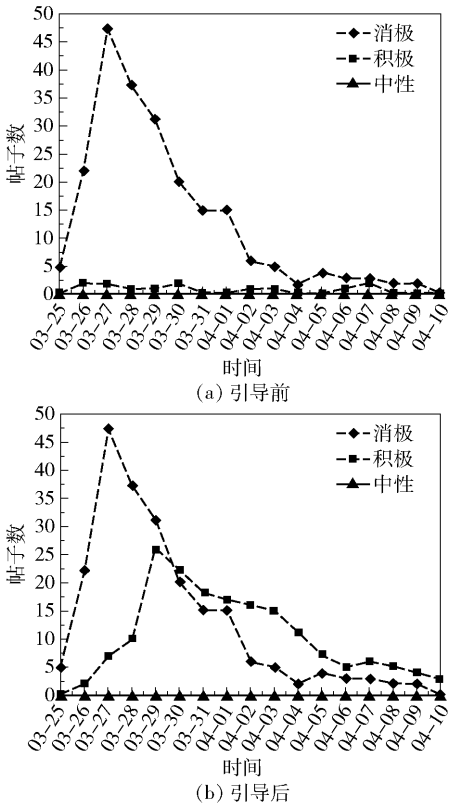

图 4　“辱母杀人案”舆情引导前后的情感走势图

Fig. 4　Public opinion before and after the guide to the emotional charts

超过 70% 的预警线时，系统开始预警并进行舆情引导。首先对该事件进行舆情检索，从百度新闻中找出能对事件发展态势起到缓和作用的文章，将其转载到舆情引导区，对文章内容作出情感判断并进行修饰编辑，然后转发到天涯社区。系统在 3 月 27 号开始对舆情进行引导，使一些偏激的公众舆论回归理性。舆情引导后的效果如图 4(b)所示，可以看出，从 3 月 27 日起网络舆情的积极情感明显上升，与消极的舆情情感平分秋色，表明舆情引导对于网络负面情绪起到了很好的正面引导作用。

3　结束语

在当前网络舆情爆炸式增长背景下，为了高效全面地掌握网络舆情动态并对其进行科学合理的引导，设计并实现了网络舆情分析与实时监测系统。该系统针对目前国内主流舆情系统分析维度单一的问题进行了改进，并首次将舆情引导加入系统。通过实例验证证明，本文系统通过用户、主题、情感的多维度分析，可以较全面地了解和掌控舆情内容及其发展态势，通过舆情引导可以明显改善网络舆情走向。本文系统对于政府部门掌握网络舆情及事态发展具有一定积极作用，可以为国内舆情系统的建设提供一种新的思路。

本文仅以天涯社区为例，利用现有技术开发了互联网舆情深度分析与引导系统，下一步将采集包括微博、微信在内的更多渠道的数据，植入更先进的技术工具对舆情进行分析。

参考文献：

[1] 刘波维，曾润喜. 网络舆情研究视角分析[J]. 情报杂志，2017，36(2)：91-96.
LIU B W，ZENG R X. The study on perspectives on internet public opinion in China[J]. Journal of Intelligence，2017，36(2)：91-96.（in Chinese）

[2] 肖慧华. 政府网络舆情监测系统的功能分析[J]. 科技广场，2017(3)：51-54.
XIAO H H. Function analysis of government network public opinion monitoring system[J]. Science Mosaic，2017(3)：51-54.（in Chinese）

[3] 龚磊. 基于 LDA 的主题发现及演化规律的可视化研究[J]. 现代计算机，2017(3)：42-44.
GONG L. Visualization of topic discovery and evolution based on LDA[J]. Modern Computer，2017(3)：42-44.（in Chinese）

[4] LO S H. Web service quality control based on text mining using support vector machine [J]. Expert Systems with Applications，2008，34(1)：603-610.

[5] 王英，龚花萍. 基于情感维度的大数据网络舆情情感倾向性分析研究——以“南昌大学自主保洁”微博舆情事件为例[J]. 情报科学，2017(4)：37-42.
WANG Y，GONG H P. Analysis of sentiment tendency of

big data online public opinion based on the sentiment dimension — taking "the independent cleaning of Nanchang university" weibo public opinion event as an example [J]. Information Science, 2017(4):37-42. (in Chinese)

[6] 杜阿宁. 互联网舆情信息挖掘方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2007.
DU A N. Publie opinion mining on the internet[D]. Harbin: Harbin Institute of Technology, 2007. (in Chinese)

[7] JOACHIMS T. Making large-scale SVM learning practical [R/OL]. (1998-06-15). http://hdl.handle.net/10419/77178.

[8] 陈晓美, 高铖, 关心惠. 网络舆情观点提取的 LDA 主题模型方法[J]. 图书情报工作, 2015, 59(21): 21-26.
CHEN X M, GAO C, GUAN X H. Extraction method of network public opinion based on LDA topic model[J]. Library and Information Service, 2015, 59(21):21-26. (in Chinese)

[9] YU Y, HU Z J, ZHANG Y H. Rearch on large scale documents deduplication technique based on simhash algorithm[C]// International Conference on Information Sciences, Machinery, Materials and Energy. Changsha, 2015.

[10] BUYRUKBILEN S, BAKIRAS S. Secure similar document detection with simhash[C]// 10th VLDB Workshop on Secure Data Management. Trento, 2013.

[11] 刘红玉. 网络舆情情感分析系统的设计与实现[D]. 成都:电子科技大学, 2013.
LIU H Y. Design and implementation of network public opinion sentiment analysis system[D]. Chengdu: University of Electronic Science and Technology of China, 2013. (in Chinese)

[12] 杨超, 冯时, 王大玲, 等. 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统, 2010, 31(4):691-695.
YANG C, FENG S, WANG D L, et al. Analysis on web public opinion orientation based on extending sentiment lexicon[J]. Journal of Chinese Mini-Micro Computer Systems, 2010, 31(4): 691-695. (in Chinese)

[13] JONES K S. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28(1): 11-21.

[14] TANG X J. Exploring on-line societal risk perception for harmonious society measurement[J]. Journal of Systems Science and Systems Engineering, 2013, 22(4): 469-486.

DeepOpinion: a system for deep analysis and guidance of internet public opinion

ZHANG Wen[1] DU YuHang[1] LI ZiQiang[1] CHEN JinDong[2*]

(1. College of Economics and Management, Beijing University of Chemical Technology, Beijing 100029;
2. College of Economics and Management, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: This paper proposes a system called DeepOpinion to monitor, track, analyze and guide internet public opinion by means of an extensive survey of existing studies on monitoring internet public opinion. Specifically, the DeepOpinion system analyzes internet public opinion from the viewpoints of users, topics and sentiments. The DeepOpinion system guides internet public opinion by using information retrieval and content customization. With the Tianya BBS forum as the source of internet public opinion, we demonstrate the functions of the DeepOpinion system in analyzing and guiding internet public opinion. Last but the most important, we use the example of the recent "murder for shaming mother" case in China as a case study to validate the practical usage of the proposed DeepOpinion system in analyzing and guiding internet public opinion.

Key words: internet public opinion; public opinion analysis; public opinion guidance; DeepOpinion system

(责任编辑:汪 琴)