基于搜索引擎关注度的网络舆情时空演化比较分析*

——以谷歌趋势和百度指数比较为例

陈 涛[1,2] 林 杰[1]

(1. 同济大学经济与管理学院 上海 200092;2. 宁波大学信息管理系 宁波 315211)

摘 要 突发事件网络舆情热度在时间和空间上会呈现一定的演变规律,文中利用搜索引擎的关注度指标对网络舆情热度时空演变的情况进行了研究,并以"小悦悦事件""郭美美事件"和"药家鑫事件"作为案例,比较了谷歌趋势和百度指数在关注度的时间和空间维度的变化特点。结论表明搜索引擎关注度指标比较有效地反映突发事件网络舆情的变化情况。

关键词 搜索引擎 关注度 网络舆情 谷歌 百度

中图分类号 TP319.3 G350 **文献标识码** A **文章编号** 1002-1965(2013)03-0007-04

Comparative Analysis of Temporal-Spatial Evolution of Online Public Opinion Based on Search Engine Attention

——Cases of Google Trends and Baidu Index

Chen Tao[1,2] Lin Jie[1]

(1. School of Economics and Management, Tongji University, Shanghai 200092;
2. Department of Information Management, Ningbo University, Ningbo 315211)

Abstract There are regular changes both in space and in time for the hotspots of online public opinion about emergencies. Through applying attention indicator on the cases of events of "little Yue-yue", "Guo Mei-mei" and "Yao Jia-xin", this paper analyzed the evolution of online public opinion and discussed the difference of the changes and characteristics between Google Trends and Baidu Index. The conclusion shows that the attention indicators of search engines can effectively reflect the changes of online public opinion.

Key words search engine attention online public opinion Google Baidu

0 引 言

近年来,突发事件发生的频率、产生的影响、造成的损失都越来越大,应急管理及其相关研究变得十分紧迫。互联网的普及和多种网络媒体的产生使网络媒体成为突发事件信息传播的重要渠道,形成了人们对于该事件的所有认知、态度、情感和行为倾向的集合,即网络舆情[1]。突发事件发生有其随机性的显著特点,并在网络中激发广大网民的集中关注。随着Web2.0的发展和普及,人们通过网络发表自己的观点见解的途径越来越多,突发事件被广大网民关注并形成广泛热议的深度和广度越来越便捷,网络舆情的热度演化在时间和空间两个维度上的监控尤为重要。

根据2012年7月艾瑞公司的统计数据,在目前中国的互联网搜索市场上,谷歌和百度共占据了95.6%的中文搜索流量。2012年第二季度中国搜索引擎市场规模68.7亿元,环比增长25.2%,同比增长55.0%[2]。搜索引擎已经成为人们日常生活必不可少的

收稿日期:2012-12-13 修回日期:2013-01-29

基金项目:国家社会科学基金项目"突发事件网络舆情演化的动态监测预警模式研究"(编号:12BTQ055);宁波大学预研项目"突发事件的网络舆情监控预警模式研究"(编号:XYY1001)资助。

作者简介:陈 涛(1973-),男,博士研究生,系主任,副教授,研究方向:网络舆情、数据挖掘;林 杰(1967-),男,教授,博士生导师,研究方向:管理信息系统、数据挖掘。

工具(像:国际上使用最广泛的搜索引擎有 Google、Yahoo 等,在国内有百度、搜狗等搜索引擎),搜索引擎在知识获取、科技查新、网络营销、博客搜索、地图检索等各方面有了比较深入的应用[3-7]。

搜索引擎用户关注度是以数千万网民在百度的搜索量为数据基础,以关键词为统计对象,科学分析并计算出各个关键词在搜索引擎中搜索频次的加权和,并以曲线图的形式展现。目前典型的搜索引擎如谷歌和百度均对用户搜索量进行了分析,并提供了相关关键词的热度比较分析。突发事件爆发以后,网民为了了解该突发事件的相关的新闻报道,往往会通过门户网站或者搜索引擎了解该事件的相关新闻报道,这种主动的搜索行为体现出了突发事件的被关注情况。网络舆情中的舆情事件名称往往成为了搜索引擎的热点关键词,从舆情的热度分析角度考虑,如果用户通过搜索引擎进行查询的次数越多,说明该舆情事件被网民的关注度越高,从而该事件的热度就越高。

1 搜索引擎关注度

1.1 **谷歌趋势** 谷歌趋势(Google Trends)是谷歌开发的一款分析用户在谷歌中搜索过的关键词并展示该关键词的关注度的服务。分析的结果会在地图上显示出对于关键词的地区关注度差异[8]。

谷歌趋势中的搜索量指数(Search Volume Index,SVI)体现了在一定区域内和一定时间段中针对某关键词 T_i 实际搜索数与平均搜索量之间的比例关系[9],即

$$SVI^{T_i}_{\text{timepreriod}} = V^{T_i}_{\text{now}} / EV^{\text{zone}_j}_{\text{timepreriod}} \quad (1)$$

其中 $V^{T_i}_{\text{now}}$ 表示当前针对关键词 T_i 的搜索引擎搜索数量, $EV^{\text{zone}_j}_{\text{timeprepriod}}$ 表示在考察搜索量的一定区域和一定时间段中所有关键词的平均搜索数量,该值越大,说明该关键词 T_i 的被关注度越高。搜索量指数(SVI)采用相应区域的总流量对数据进行了标准化处理,搜索量指数表明了针对某关键词对比当期平均搜索数量的相对上升和下降比例。同时,当使用者改变了搜索数据的统计时间跨度将看到 SVI 序列的变化情况。例如,如果某一关键词 T_i 当前的真实搜索数为 1000,而所有关键词 2012 年平均搜索数是 800,则 $SVI^{T_i}_{2012} = 1000/800 = 1.25$;同理,若在 2004-2012 年所有关键词的平均搜索数降低到 500,则 $T^{T_i}_{2004-2012} = 1000/500 = 2$。

利用搜索量指数谷歌趋势可以对多个不同的关键词的搜索行为进行比较,也可以针对一个关键词在不同的地区和时间上的搜索行为进行比较;还提供一些关键词未来的搜索趋势预测。同时还提供一项新功能,向用户提供关键词搜索分析的 HTML 代码,这样用户就可以在自己的 Web 页面中嵌入关键词的搜索分析结果。谷歌趋势可以利用关键词搜索量指数应用于电子商务领域的热门商品统计与预测。搜索量指数相对地反应了网民对某一关键词的关注程度,本文将利用搜索量指数变化规律探讨热点突发事件的时空演变规律及特点。

1.2 **百度指数** 百度指数与谷歌趋势相类似,2006 年正式推出百度指数的数据分析功能模块,百度指数是用以反映关键词在过去 30 天内的网络曝光率及用户关注度。它能形象地反映该关键词每天的变化趋势。百度指数是以百度网页搜索和百度新闻搜索为基础的免费海量数据分析服务,用以反映不同关键词在过去一段时间里的"用户关注度"和"媒体关注度"。通过百度指数可以发现、共享和挖掘互联网上最有价值的信息和资讯,直接、客观地反映社会热点、网民的兴趣和需求。百度指数每天更新一次,并且提供自 2006 年 6 月至今任意时间段的用户关注度数据[10]。同时,根据不同的关键词,机器自动从百度新闻搜索中获取与该关键词最相关的 10 条热门新闻,并将新闻按时间顺序均匀分布在"用户关注度"的曲线图上,以字母标识,每个字母对应一条新闻。百度指数是综合反映该关键词在过去 1 天用户对它的关注和媒体对他的关注的一个参考值。任意关键词的百度指数都是该关键词在比较期的数值/该关键词在基期的数值。比较期的数值和基期的数值是通过当天的用户搜索量和百度新闻中过去 30 天相关的新闻数量相比得来的。百度指数信息服务出现滞后于谷歌趋势。但同样体现了某一关键词在特定时间段中被用户和媒体关注的强度。

2 突发事件网络舆情的搜索引擎时空热度案例比较分析

网络舆情热度的涵义是,当突发事件发生后,网络媒体和网民对事件的报道、讨论以及政府或者网络监管部门提供的引导机制在网络上所形成的突发事件舆情高涨程度[11]。网络舆情的热度不仅仅体现在时间维度,同时还体现在不同地域网民对突发事件的关注强度。

2.1 **案例比较分析** 根据谷歌趋势、百度指数的定义和计算方法。突发事件的发生发展过程中网民对该事件的关注程度可以通过搜索引擎查询数体现出来。尤其是涉及到社会热点和话题的突发事件的谷歌搜索量指数以及百度指数都应能清晰地体现出来。网络舆

情往往与网民社会生活中相关的法律、道德、自然灾害、战争等方面事件较为紧密,尤其是涉及到法律、道德等方面的信息在网络上传播和发酵的广度和深度更为明显。本文搜集了几个最典型的突发事件包括小悦悦事件、药家鑫事件以及郭美美事件。之所以以上述三个典型的突发事件作为分析案例,这三个事件比较突出地体现了网民对相关事件的道德良知和法律规范认识,其时空热度变化比较显著。为了有效地对三事件类别进行分类,对实验数据进行了预处理,首先利用网络爬虫从新浪网三事件网络新闻网民评论中抓取相关评论,然后利用停用词表过滤掉停用词,进而获取了网民评论中的类别高频实词(词频大于 50),从获取的词语(包裹名词、形容词和动词等实词)语义特征可以观察到,三事件网民评论中的类别特征词如表 1 所示。

表 1　三突发事件网民评论中的高频类别特征词

	法律类别	道德类别
小悦悦事件(4076 条网友评论)	罪恶、判刑、赔偿、法律、交通事故、法院、公正、判决	道德、良知、美德、冷淡、报应、素质、无视、义务、良心、反思、人性、向善、善良、灵魂、谴责、宽容、文明、舆论、相信、泯灭、报应、责任、忏悔、指责、见死不救、冷漠、冷淡、教育、缺德
药家鑫事件(2985 条网友评论)	法律、法治、法制、公平、公正、正义、自首、赔偿、死刑、剥夺、威严、违法、判决、犯法、案件、诉讼、处死、律师、法院、罪孽、剥夺、枪毙、司法、制度、罪恶、违法、执行、赔偿、十恶不赦、罪有应得、宽大、杀人偿命	良知、教训、宽容、道德、人性、败类、冷血、沦丧、价值观
郭美美事件(6856 条网友评论)	严惩、法制、法治、检察院、侦查、替罪羊、真相、贪官、罪名、公安机关、立案、贪污、腐败、法律、罪行、立法、合法、司法、贪腐、判刑、判决、严惩、法规、公正、受害人、追究、追查、黑幕、内幕	慈善、捐款、谎言、信任、相信、慎重、质疑、炒作、爱心、常理、无耻、鄙视、爱心、缺德、炫富、炫耀、恶心、脑残、黑暗、恶毒、解释、狡辩、善款、廉耻、质疑、怀疑、荒唐、龌龊、约束

从表 1 的数据观察到,根据已经构建的语义词典[12],小悦悦事件中 18 名路人路过但都视而不见,漠然而去引起了广泛的社会道德热评,抽取的高频实词以道德类别为主;药家鑫事件则体现了广大网民对法律法规的认识,抽取的高频实词以法律类别为主;而郭美美事件则介于道德和法律之间。从图 1 展示了谷歌趋势对三个热点事件网民的关注度事件变化趋势情况。

从图 1 中谷歌搜索趋势中可以比较清楚地观察到,介于道德和法律之间的热点事件网民的关注度最高,倾向于道德的小悦悦事件则相对较高,而涉及到法律规范的药家鑫事件则最低。

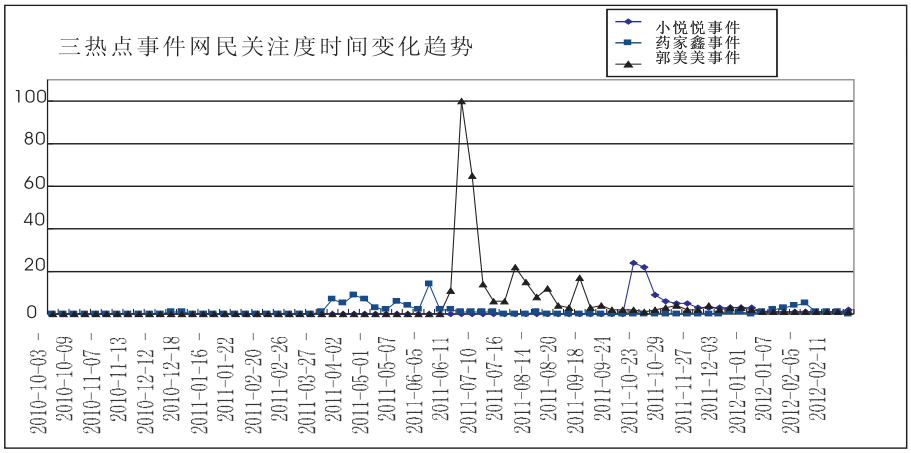

图 1　三热点事件网民关注度时间变化趋势
(利用谷歌趋势提取数据整理结果)

图 2 显示了三事件热点搜索的次级区域的关注情况,从图 2 中可以观察到,网民对相关事件的发生地的关注度相对高,其中药家鑫事件发生在陕西;小悦悦事件发生在广东。而郭美美事件的热点搜索次级区域则有别于前两个事件,受到了国内大部分省份网民的广泛关注,尤其是对内蒙古关注度最高,较高关注的次级区域还包括云南、甘肃、新疆、广西等地。究其原因,郭美美事件涉及到的公益慈善组织与经济欠发达地区关系更为密切,受到该地区的网民的关注度则更高。

利用百度指数分析功能,在相关的网页界面中输入三事件关键词,可以得到用户关注度的变化情况如图 3 所示,并展示了相关事件的最新新闻链接,在页面中部对相关检索词的变化情况作了统计,并在页面底部对地区分布用地图的方式进行展示。

由图 3 可以观察到,百度指数的用户关注度随时间变化的趋势情况与谷歌趋势结果基本类似,但是在地区分布中百度指数的展示结果中与谷歌搜索结果的区域关注情况不完全一致。百度指数进一步统计了关注人群的年龄结构、职业分布和学历分布情况并进行了可视化展示,如图 4 所示,该项数据的统计体现了百度在关注人群细分方面领先于谷歌趋势。

2.2　**结　论**　从上述三热点事件的谷歌趋势和百度指数关注度比较分析中,可以发现针对突发事件的网络舆情存在下述几点规律:a. 突发事件的发生会导致网民关注度出现脉冲式增长,并在一段时间内逐渐衰减,这反应了突发事件网络舆情的随机性和爆发性,随着信息传播和披露范围的扩展,舆情热度会逐渐降低;b. 网络舆情往往与网民社会生活中相关的法律、道德、自然灾害、战争等方面事件较为紧密,尤其是涉及到法律、道德等突发事件,即使事件的直接关联人群范

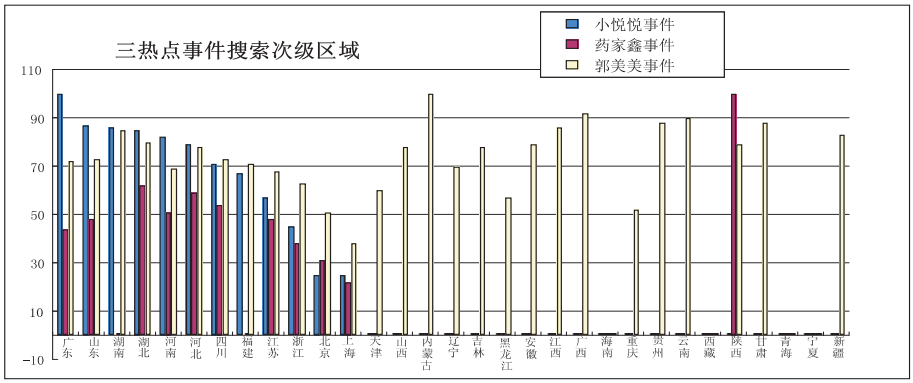

图2　三事件热点搜索次级区域分布图以及利用谷歌趋势提取数据整理统计结果

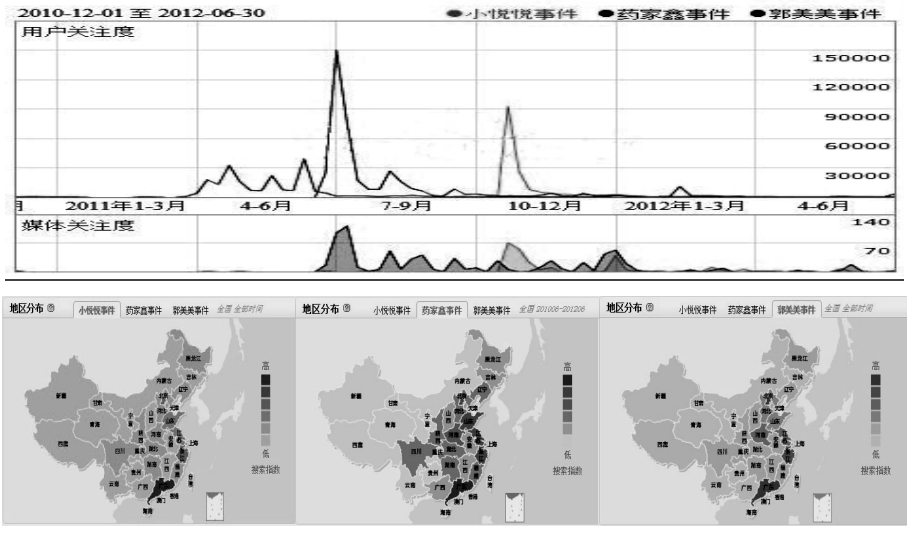

图3　三事件百度指数随时间变化用户关注度统计图以及关注地区分布图

图4　百度指数关注人群属性统计图

围并不大,仍然会引起广大网民的广泛关注,同时突发事件被网民或媒体的关注度幅值往往与突发事件类别有关,从上述三个案例分析观察到,介于法律和道德范畴之间的热点事件,更容易引起关注并进一步发酵;c. 突发热点事件被关注的地区属性,往往与该热点事件所在地理位置以及事件涉及的关联地区有关,文中案例所涉及的三事件的发生地网民对事件的关注热度都较高,"郭美美事件"涉及的"慈善捐款"的关联地区产生的反响也较明显。

同时,利用谷歌趋势和百度指数对三个热点事件关键词进行分析的结果表明,谷歌趋势和百度指数存在以下几方面的共同点和区别:a. 谷歌趋势和百度指数衡量关注度指标的具体算法和数据展示虽不尽相同,但二者的思路是一致的,区别在于谷歌趋势的搜索量指数采用了相关关键词对比当期平均搜索数量的相对升降比例,谷歌趋势提供了相关数据下载链接,而百度指数则采用了"用户关注度"和"媒体关注度"的绝对数值衡量关注热度(仅以图形化方式展示),两者均为研究突发事件网络舆情监控提供了时间和空间两个维度的数据支持,并可以利用事件关键词方式对多个事件进行比较,但其中百度指数最多只能输入三个关键词进行比较;b. 从针对三事件的谷歌趋势和百度指数统计结果观察,两者在关注度时间变化趋势基本上一致,在关注地区的数据统计图存在一些差异(尤其是郭美美事件关注地区图差异比较显著,这与网络用户使用不同搜索引擎习惯有关),但二者在相关突发事件发生地的关注热度均比较高,为突发事件网络舆情监控提供了时间和空间维度上数据支持;c. 在百度指数中增加了关注人群

(下转第16页)

(上接第 10 页)

属性的统计图,较好地展示了关注群体的性别比例、年龄分布、职业分布和学历分布情况,为网络舆情疏导过程中对象目标群体选择提供了统计数据支持,而谷歌趋势在关注人群的细分方面未给出具体的统计数据,网络热点事件的关注人群具体情况是网络舆情监控的重要指标,在这一指标上,百度有比较明显的优势。

3　结束语

突发事件网络舆情搜索引擎用户关注度能比较直观地刻画网络舆情变化的时空特点;同时,突发事件由于其内生的随机性以及在网络媒体扩散的内在规律,会导致用户关注情况出现急速爆发并在较短的时间内迅速衰减的情况,利用搜索引擎用户关注度度量方法可以监控不同突发事件的热度差异,并利用公共媒体进行疏导;另外,搜索引擎关注度的次级区域分布结果体现了突发事件所在地往往是舆情监控的重要节点,对网络舆情传播和扩散中正向引导起着至关重要的作用。进一步的研究工作可以针对相关突发事件的网民关注度的联动变化以及相互影响方面深入和展开。

参 考 文 献

[1]　曾润喜. 网络舆情信息资源共享研究[J]. 情报杂志,2009(8):187-191

[2]　艾瑞咨询:"2012 年 Q2 搜索引擎市场规模达 68.7 亿,同比增速放缓"[EB/OL]. [2012-07-27]. http://www.iresearch.com.cn/Report/view.aspx? Newsid=177708

[3]　张小娣,宋余庆. 基于科学知识图谱的搜索引擎前沿分析[J]. 科技管理研究,2011(18):226-230

[4]　李建婷. 网络搜索引擎在科技查新中的应用[J]. 情报杂志,2011,30(增):170-171

[5]　姜旭平,王　鑫. 影响搜索引擎营销效果的关键因素分析[J]. 管理科学学报,2011(14):37-44

[6]　王婵娟. 专业搜索引擎之博客搜索[J]. 图书馆学研究,2009(6):54-56

[7]　唐　曦, 黄　燕, 邱菲菲等. 联网地图搜索引擎视觉质量的模糊评价与可视化分析[J]. 测绘科学,2011(3):40-43

[8]　百度百科:谷歌趋势[EB/OL]. [2012-12-13]. http://baike.baidu.com/view/4583508.htm

[9]　Zhi Da, Joseph Engelbergz, Pengjie Gao et al, In Search of Earnings Predictability[C]. China International Conference in Finance,2011:1-32

[10]　百度百科:百度指数[EB/OL]. [2012-12-13]. http://baike.baidu.com/view/1235.htm

[11]　张一文,齐佳音,方滨兴,等. 非常规突发事件网络舆情热度评价体系研究[J]. 情报科学,2011(9):1419-1424

[12]　陈　涛,孙茂松. 基于 SOM 的语义词典自动构建实验研究[J]. 情报学报,2007,26(1):77-83.　　(责编:刘影梅)