面向大数据的网络舆情热度动态预测模型研究*

兰月新[1] 刘冰月[2] 张 鹏[1] 夏一雪[1] 李昊青[1]

(1.中国人民武装警察部队学院 廊坊 065000;
2.天津交通职业学院 天津 300132)

摘 要 [目的/意义] 面向大数据研究网络舆情热度模型以及热度预测模型,能够准确把握大数据环境下网络舆情热度,并可以决定网络舆情应对和舆论引导措施的轻重缓急,具有重要的理论意义。[方法/过程] 定性分析大数据环境下网络舆情热度影响因素,通过定义最大关联度向量,基于灰色关联度方法构建网络舆情热度模型,并在此基础上构建多维度logistic模型对各个媒体平台舆情信息开展预测,通过灰色关联度得出动态预测方法。[结论/结果] 经过理论建模和实证分析得出构建的热度模型和热度动态预测模型是可行的,以上理论研究可为政府准确把握大数据环境下网络舆情热度,制定网络舆情引导策略提供参考依据。

关键词 大数据 网络舆情 灰色关联度 热度预测 logistic
中图分类号 C912.6 **文献标识码** A **文章编号** 1002-1965(2017)06-0105-06
引用格式 兰月新,刘冰月,张 鹏,等.面向大数据的网络舆情热度动态预测模型研究[J].情报杂志,2017,36(6):105-110,147.
DOI 10.3969/j.issn.1002-1965.2017.06.019

The Internet Public Opinion Hot-degree Dynamic Prediction Model Oriented to Big Data

Lan Yuexin[1] Liu Bingyue[2] Zhang Peng[1] Xia Yixue[1] Li Haoqing[1]
(1.The Chinese People's Armed Police Force Academy, Langfang 065000;
2.Transportation Vocational College, Tianjin 300132)

Abstract [Purpose/Significance] It is of important theoretical significance to conduct researches on the Internet public opinion hot-degree model and the hot-degree dynamic prediction model oriented to big data, which help grasp the network public opinion hot-degree accurately and determine the activities' priorities on guiding and controlling the development of the public opinions. [Method/Process] This paper conducted a qualitative analysis of the factors of Internet public opinion hot-degree oriented to big data, and through defining the maximum relevance vector, built the Internet public opinion hot-degree model based on grey correlation method, then carried out further research on multidimensional logistic model of predicting Internet public opinion on various media platforms, which comes to a dynamic prediction model combined with gray correlation degree. [Result/Conclusion] The feasibility of the two models was verified by an empirical analysis and a theoretical modeling. The findings can provide significant reference for the government to grasp the Internet public opinion hot-degree oriented to big data and develop the appropriate Internet public opinion guiding strategies.

Key words big data the internet public opinion grey correlation degree hot-degree prediction logisitic

收稿日期: 2017-03-20 **修回日期:** 2017-04-19
基金项目: 河北省科技计划项目"面向大数据的网络舆情对抗关键技术研究"(编号:16215604);国家社会科学基金青年项目"公共安全视角下网络舆情风险建模与对策研究"(编号:15CXW015);河北省社会科学发展研究课题"基于涉恐舆情综合研判的反恐情报预警研究"(编号:201604120504);河北省统计科学研究计划重点项目"大数据环境下网络舆情数据分析与决策支持研究"(编号:2016HZ09)。
作者简介: 兰月新(ORCID:0000-0002-4791-5094),男,1981年生,副教授,硕士生导师,研究方向:网络舆情;刘冰月(ORCID:0000-0002-3874-7754),女,1989年生,硕士,助教,研究方向:数据分析;张 鹏(ORCID:0000-0002-8664-5058),男,1981年生,博士,讲师,研究方向:网络舆情;夏一雪(ORCID:0000-0002-8044-0553),女,1983年生,讲师,博士,研究方向:公共管理;李昊青(ORCID:0000-0002-1780-1184),男,1983年生,馆员,硕士,研究方向:情报学。

1 现状分析

截至2016年12月,中国网民规模达到7.31亿,互联网普及率为53.2%,其中手机网民比例为95.1%[1]。随着移动宽带互联网的普及,网络舆情在数据体量、复杂性和产生速度等方面发生巨大变化,已经呈现大数据环境。任何网络话题在互联网上都会形成规模或大或小的网络舆情,其受关注程度往往利用网络舆情热度来衡量。如何准确把握大数据环境下网络舆情热度,决定着网络舆情应对和舆论引导措施的轻重缓急,具有重要的理论意义和实践价值。

1.1 国内外文献综述 网络舆情热度研究涉及数学、统计学、应用社会学、政治学、传播学、管理学、情报学等多个学科的理论和方法,属于交叉学科研究。目前国内外学术研究情况如下:

网络舆情热度研究主要包括三个方面。第一,建立网络舆情热度评价指标体系,其中绝大部分研究是专门针对一个传播平台(微博、论坛等)开展热度研究[2-8],主要包括信息发布数、转发数、评论数等定量指标,也包括事件敏感度、网民参与度、网民情绪等定性指标,评价模型则以层次分析法和BP神经网络分析为主。第二,建立定量模型研究网络舆情热度,并给出网络舆情引导策略,主要包括系统动力学模型、微分方程模型、最优化理论模型、定性数据挖掘模型、文本聚类分析等[9-13]。第三,全面考虑各个传播平台发布的信息对网络舆情热度的影响,给出热度计算公式,例如由清华大学新闻与传播学院新媒体中心与新浪微舆情联合推出网络传播热度指数算法1.0版[14],热度指数算法如下:

$$R = 0.4Y_1 + 0.45Y_2 + 0.05Y_3 + 0.1Y_4$$

$$Y_i = \left(\frac{2}{1 + a^{-X_i}} - 1\right) \times 100$$

其中 X_1 = 新闻* 0.189+报刊* 0.175+APP* 0.187+微信* 0.182+政务* 0.167+外媒* 0.1; X_2 = 微博* 0.319+论坛* 0.355+博客+0.326; X_3 = 视频; X_4 = 其他网站。

网络舆情热度预测方面主要通过构建定量预测模型来开展研究,其中马尔科夫预测是常用的方法[15-19]。除此之外,网络舆情热度预测研究还包括:根据信息传播的二八定律,通过舆情传播关键节点预测热点舆情演化趋势[20];通过社会惯性理论和时间序列分析研究网络舆情热度预测模型[21];研究K-近邻的网络话题热度预测算法[22];使用有监督的机器学习算法提取训练样本静态和动态特征训练热度预测模型[23];利用Gamma分布模型提出融合观点倾向的话题热度趋势建模方法[24];构建离散话题热度预测模型DTPM模型[25]等。

1.2 国内应用情况综述 人民网、新华网、清博舆情研究院等机构均定期发布互联网舆情分析报告,但是报告中同一事件的热度排名却各有不同。整理人民网《2016年中国互联网舆情分析报告》[26]、新华网《2016年度社会热点事件网络舆情报告》[27]、清博研究院《2016上半年舆情发展态势及规律全景报告》[28]数据得到热度排名前20的舆情事件(见表1)。由热度排名对比表发现,三个报告的舆情热度排名,差距较大,排名重合度低,例如"和颐酒店女子遇袭事件"排名分别为16、14和1。除了排名不全、监测数据不全、传播平台不全等差异外,三类排名热度计算方法均是对不同平台的数据赋予权重,然后综合考虑各个传播平台数据得出最终热度值,这种方法受主观影响程度较大,而且在不考虑各个平台之间数据的相互影响,不考虑舆情事件的相互影响的情况下,最终导致热度排名存在差异。所以,在实践应用层面需要统一热度计算方法,使舆情分析报告能够更好地辅助决策。

表1 三类报告热度排名对比

热度排名	人民网舆情全年报告	新华网舆情全年报告	清博舆情上半年报告
1	杭州G20峰会	全面放开落户限制	和颐酒店女子遇袭事件
2	南海仲裁事件	哈尔滨天价鱼	魏则西事件
3	雷洋事件	山东问题疫苗事件	2016全国两会
4	2016美国大选	常州外国语学校毒地事件	山东问题疫苗事件
5	王宝强离婚事件	魏则西事件	雷洋事件
6	魏则西事件	雷洋事件	帝吧出征
7	女排奥运夺冠	河北大贤村洪灾事件	广东弑医案
8	网络直播带动网红	网约车新规出台	全面开放二孩政策
9	A股熔断机制	王宝强离婚事件	高考减招风波
10	南方洪灾	徐玉玉电信诈骗案	常州外国语学校毒地事件
11	山东问题疫苗事件	罗尔涉嫌诈捐舆情风波	哈尔滨天价鱼
12	网约车新规出台	东北女孩怒斥广安门医院号贩子	快播案
13	校园毒跑道	河南女孩王娜娜被顶替上大学事件	A股熔断机制

续表 1　三类报告热度排名对比

热度排名	人民网舆情全年报告	新华网舆情全年报告	清博舆情上半年报告
14	杨改兰案与《盛世中的蝼蚁》	和颐酒店女子遇袭事件	血友病百度贴吧经营权卖出
15	赵薇新片引风波	安徽男子“右肾失踪”风波	任志强微博被关闭
16	和颐酒店女子遇袭事件	湖北仙桃民众抗议垃圾焚烧项目	——
17	朴槿惠闺蜜门	不影响执法前提下警察不得干涉拍摄	——
18	高考减招风波	南方洪灾	——
19	帝吧表情包大战	中关村二小疑似“校园欺凌”事件	——
20	连云港反核群体性事件	取消 61 项职业资格认证	——

虽然学术层面和实践应用层面取得了很多研究成果，但是网络舆情热度及热度预测研究仍有很多问题需要解决，例如①网络舆情热度指标及其量化受主观影响较大，导致同一事件热度差异过大; ②网络舆情热度中涉及的权重问题主要依靠专家打分，专家数量和问卷设计影响评估结果; ③大部分热度预测模型不考虑网络舆情传播规律，大多从统计学和机器学习等视角的开展案例研究，导致预测模型符合历史和当下数据，而对未来的预测效果较差。综合以上问题，本文定性分析大数据环境下网络舆情热度影响因素，通过定义最大关联度向量，并基于灰色关联度方法来构建网络舆情热度模型，同时在此基础上构建多维度 logistic 模型对各个媒体平台舆情信息开展预测，基于灰色关联度得出动态预测方法，以上理论研究可以为政府准确把握大数据环境下网络舆情热度，制定网络舆情引导策略提供参考依据。

2　面向大数据的网络舆情热度

2.1　**大数据环境下的网络舆情**　大数据环境下，网络舆情特征可以总结为“事件类型多、舆情主体多、信息格式多、上网速度快”。政府和网民都成为信息发布主体，并且后者发布的信息数量远远超出前者，改变了由政府作为信息发布主体的格局。我国存在网民结构复杂且发表网络言论视角众多、网民水平参差不齐，网民心理多元化等问题。除此之外，庞大的媒体数量使得信息传播平台越来越多，网络新闻、网络报刊、微博、微信、新闻客户端、论坛、视频网站等互联网平台已经成为网络舆情信息传播的主要载体，并且绝大部分微博、论坛、社交网站、视频网站等网络媒体的信息均可相互共享、转发，网络舆情信息交互便捷、信息异化程度不断加大，导致舆情信息量激增而形成大数据环境(见图 1) 。

2.2　**网络舆情热度**　网络舆情热度，即网络舆情事件受关注的程度。网络舆情热度值决定着网络舆情应对和舆论引导措施的轻重缓急。首先，大数据环境下，某个网络话题被网民关注后，舆情信息会由一个传播平台向多个平台扩散，在这个过程中信息量在不断变化(见图 2) ，网络舆情热度也在不断升温，所以网络新闻、网络报刊、微博、微信、新闻客户端、论坛、视频网站等互联网平台上的舆情信息量决定着网络舆情热度; 其次，定量确定网络舆情热度的目的是为了使网络舆情主体能够宏观把握当前的舆情态势，进而制定舆情引导策略，所以某个网络话题的网络舆情热度属于相对量，在计算时需要系统考虑某短时间内的所有网络话题; 最后，大数据的核心和目标就是预测，所以需要根据监测数据开展热度预测，并且随着新数据的加入而动态修正预测结果，实现对网络舆情热度的动态预测。综合以上三点，本文通过灰色系统理论确定热度，并根据网络舆情传播规律开展热度预测。

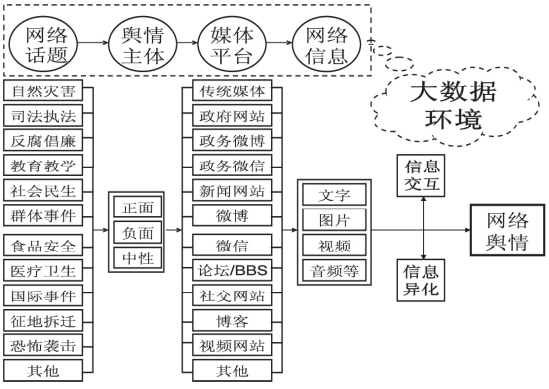

图 1　大数据环境下的网络舆情

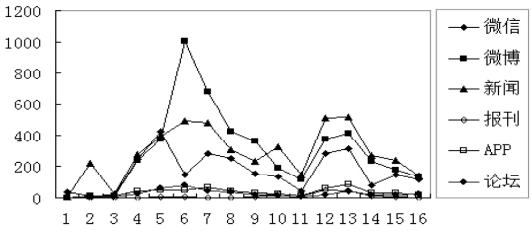

图 2　美国退出 TPP48 小时舆情数据

(2017.1.24 00: 00-2017.1.25 21: 00)

3　面向大数据的网络舆情热度预测模型

3.1　网络话题热度模型

a.基本假设。假设某个时间段内，网民关注的网络话题数量为个，每个网络话题在 n 个传播平台传播形成网络舆情。网民针对第 i 个网络话题在网络新闻、论坛、微博、微信等传播平台发布的信息总量为

$x_i(j)$，构建第 i 个网络话题的热度向量

$$X_i = (x_i(1), x_i(2), \cdots, x_i(n))$$

其中 $1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n$。由于网络舆情热度与舆情信息总量成正比，在某个时间段内分别取 n 个传播平台信息总量的最大值构建最大热度向量

$$X_{\max} = (\max_{1\leqslant i\leqslant n} x_i(1), \max_{1\leqslant i\leqslant n} x_i(2), \cdots, \max_{1\leqslant i\leqslant n} x_i(n))。$$

$$\eta_i(k) = \frac{\min_i \min_k |\max_{1\leqslant i\leqslant n} x_i(k) - X_i(k)| + \rho \max_i \max_k |\max_{1\leqslant i\leqslant n} x_i(k) - X_i(k)|}{|\max_{1\leqslant i\leqslant n} x_i(k) - X_i(k)| + \rho \max_i \max_k |\max_{1\leqslant i\leqslant n} x_i(k) - X_i(k)|}$$

其中，$\rho \in (0, +\infty)$ 称为分辨率，ρ 越小，分辨力越大。关联系数得到的是每个网络话题的热度向量与最大热度向量在各个传播平台的关联系数值，结果较多且信息过于分散，所以计算关联系数的平均数得到第 i 个网络话题的灰色关联度

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \eta_i(k)。$$

不难发现，$0 < r_i < 1$，所以本文将某个网络话题的热度向量与最大热度向量的关联度定义为某个网络话题的网络舆情热度。

3.2 网络舆情热度动态预测模型

a.基本预测模型。网络舆情热度与信息总量呈正比，所以需要对网络舆情统计数据进行累加，进而得到随时间变化的网络舆情信息总量数据。通过雾霾微博舆情实例分析以及过往研究，容易得出统计数据累加后呈现出“S”形曲线特征（见图 3），基于此，本文选取 logistic 作为基本模型来开展动态预测。

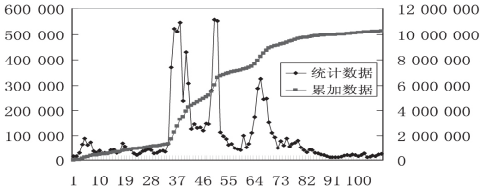

图 3 雾霾微博舆情统计数据与累加数据对比图
（2016.11.1-2017.2.16）

Logistic 模型

$$\frac{dx}{dt} = rx\left(1 - \frac{x}{K}\right)$$

对应的差分方程为

$$\Delta x_k = rx_k\left(1 - \frac{x_k}{K}\right) = rx_k - \frac{r}{K}x_k^2$$

其中 x 代表累计信息量，r 代表内禀增长率，K 为累计信息量上限，$\Delta x_k = x_k - x_{k-1}$，$k = 1, 2, \cdots, n$。不难看出，$\Delta x_k$ 是关于 x_k 和 x_k^2 的二元线性关系，应用 EXCEL 或者 MATLAB 进行回归分析，即可得出回归系数 r 和 $\frac{r}{K}$，从而得到参数 K 和 r，简称此种方法为差分回归法。

所以，通过计算每个网络话题的热度向量与最大热度向量的“距离”，可以得出每个网络话题的热度。

b.网络话题的热度。根据灰色关联分析[29]，定义每个网络话题的热度向量与最大热度向量的关联系数为

b.网络舆情热度动态预测模型。假设第 i 个网络话题在 t 时刻的热度向量为

$$X_i^t = (x_i^t(1), x_i^t(2), \cdots, x_i^t(n))，$$

构建第 i 个网络话题在 n 个传播平台的动态预测模型为

$$\begin{cases} \dfrac{dx_i^t(1)}{dt} = rx_i^t(1)\left(1 - \dfrac{x_i^t(1)}{K}\right) \\ \dfrac{dx_i^t(2)}{dt} = rx_i^t(2)\left(1 - \dfrac{x_i^t(2)}{K}\right) \\ \cdots\cdots \\ \dfrac{dx_i^t(n)}{dt} = rx_i^t(n)\left(1 - \dfrac{x_i^t(n)}{K}\right) \end{cases}.$$

网民针对网络话题开展交流从而形成网络舆情，经过一段时间的监测，获取用于预测的历史数据，通过差分回归法分别预测网络舆情在 n 个传播平台的传播趋势，然后应用本文构建的热度模型，进而计算网络舆情未来的热度。

4 实证研究

4.1 热度模型验证

a.数据来源。在 2017 年社会蓝皮书《中国社会形势分析与预测》中，中国互联网舆情分析报告针对 2016 年每月排名前 50 的 600 件舆情热点事件进行分析，得出 2016 年排名前 20 的舆情事件。报告针对网络报刊、网络新闻、论坛、博客、微博、微信、APP 七类传播平台，运用德尔菲法和层次分析法计算网络舆情热度（见表 2）。本文将在此数据上应用灰色关联度计算舆情热度，并与中国互联网舆情分析报告开展对比研究。

表 2 2016 年热点舆情事件统计数据（单位：千篇）

编号：事件	报刊	新闻	论坛	博客	微博	微信	APP	热度排名
1：杭州 G20 峰会	36.1	602.6	59.5	47.2	80	327.3	28.9	1
2：南海仲裁事件	18.4	411.6	170	65.3	307.6	240	37.7	2
3：雷洋事件	16.2	237.5	66.6	43.4	67.5	292.1	19.3	3
4：2016 美国大选	9.5	443.5	20.3	57.4	54.6	158.3	18.1	4

续表 2　2016 年热点舆情事件统计数据（单位：千篇）

编号：事件	报刊	新闻	论坛	博客	微博	微信	APP	热度排名
5：王宝强离婚事件	4.2	220.7	55.7	24.5	328.5	175.3	18.6	5
6：魏则西事件	10.2	169.1	40.3	40	104.1	195.8	10.6	6
7：女排奥运夺冠	9.4	127.6	20.2	11.2	67.3	118.9	17.9	7
8：网络直播带动网红	5.7	240.8	22	21	5.6	122.9	13.6	8
9：A 股熔断机制	7.3	203.3	44.8	45.3	18.2	52.9	5.1	9
10：多省份暴雨洪灾	8.7	126.2	13.6	10.4	13.2	67.4	4.3	10
11：山东问题疫苗事件	5.3	117	15.6	9.3	16.8	94.6	2.5	11
12：网约车新规出台	2.9	100.4	7.9	6.5	15.7	56.4	4.6	12
13：校园毒跑道	2.4	42.4	9.3	6.4	26.4	20.7	6.1	13
14：杨改兰案与《盛世中的蝼蚁》	2.3	49.6	11	8.9	7.9	59.5	1.7	14
15：赵薇新片引风波	0.5	33	10.5	4.7	121.4	45.2	5.3	15
16：和颐酒店女子遇袭事件	1.2	29.4	4	2.5	57.1	14.7	1.2	16
17：朴槿惠闺蜜门	1.3	31.6	17	1.2	16	4.6	1	17
18：高考减招风波	0.8	17.1	5.3	13.1	11	16.1	0.7	18
19：帝吧表情包大战	0.4	21.3	4.9	2	9.9	29.1	1.5	19
20：连云港反核群体性事件	0.5	12.9	4.1	2.6	13.7	11.9	0.4	20

b.基于灰色关联度的网络舆情热度模型。根据表 2 数据，构建最大热度向量

$(36.1, 602.6, 170, 65.3, 328.5, 327.3, 37.7)$。

选取分别率 $\rho = 0.5$，计算 20 个热点舆情与最大热度舆情的关联系数（见表 3）。

表 3　关联系数表

编号	报刊	新闻	论坛	博客	微博	微信	APP
1	1.0000	1.0000	0.7274	0.9422	0.5427	1.0000	0.9710
2	0.9434	0.6069	1.0000	1.0000	0.9338	0.7716	1.0000
3	0.9368	0.4468	0.7404	0.9309	0.5304	0.8933	0.9413
4	0.9173	0.6495	0.6633	0.9739	0.5184	0.6357	0.9377
5	0.9024	0.4357	0.7206	0.8784	1.0000	0.6598	0.9392
6	0.9193	0.4048	0.6945	0.9210	0.5678	0.6916	0.9158
7	0.9170	0.3830	0.6631	0.8450	0.5303	0.5859	0.9371
8	0.9065	0.4490	0.6658	0.8694	0.4773	0.5906	0.9244
9	0.9110	0.4248	0.7019	0.9365	0.4872	0.5180	0.9004
10	0.9150	0.3823	0.6534	0.8430	0.4832	0.5315	0.8982
11	0.9054	0.3778	0.6563	0.8404	0.4861	0.5589	0.8933
12	0.8988	0.3699	0.6453	0.8337	0.4852	0.5212	0.8991
13	0.8974	0.3448	0.6472	0.8335	0.4939	0.4902	0.9032
14	0.8972	0.3478	0.6497	0.8394	0.4791	0.5240	0.8912
15	0.8923	0.3411	0.6489	0.8295	0.5874	0.5110	0.9010
16	0.8942	0.3397	0.6398	0.8244	0.5207	0.4854	0.8898
17	0.8944	0.3405	0.6584	0.8214	0.4855	0.4775	0.8893
18	0.8931	0.3349	0.6416	0.8496	0.4815	0.4865	0.8885
19	0.8920	0.3365	0.6410	0.8233	0.4806	0.4972	0.8907
20	0.8923	0.3333	0.6399	0.8246	0.4836	0.4832	0.8877

通过关联系数计算关联度，并与中国互联网舆情分析报告中热度排名进行对比（表 4），容易看出排名差值大于或等于 2 的有 5 个事件，排名差值为 1 的有 10 个事件，剩余 5 个排名相同，说明根据不同热度方法计算的结果相近。但是，相对于依靠专家打分、赋值权重的德尔菲法和层次分析法，本文基于关联度计算热度的方法全面、系统考虑每个舆情事件在总体中的相对热度，避免了专家的主观影响，相对客观的反映网络舆情热度，效率更高，便于操作，更加贴近实际。

表 4　关联度与热度排名

编号：事件	本文热度	热度排名	报告排名	排名差值
1：杭州 G20 峰会	0.8833	2	1	1
2：南海仲裁事件	0.8937	1	2	−1
3：雷洋事件	0.7743	4	3	1
4：2016 美国大选	0.7565	5	4	1
5：王宝强离婚事件	0.7909	3	5	−2
6：魏则西事件	0.7307	6	6	0
7：女排奥运夺冠	0.6945	9	7	2
8：网络直播带动网红	0.6976	7	8	−1
9：A 股熔断机制	0.6971	8	9	−1
10：多省份暴雨洪灾	0.6724	12	10	2
11：山东问题疫苗事件	0.6740	10	11	−1
12：网约车新规出台	0.6647	13	12	1
13：校园毒跑道	0.6586	15	13	2
14：杨改兰案与《盛世中的蝼蚁》	0.6612	14	14	0
15：赵薇新片引风波	0.6730	11	15	−4
16：和颐酒店女子遇袭事件	0.6563	16	16	0
17：朴槿惠闺蜜门	0.6524	18	17	1
18：高考减招风波	0.6537	17	18	−1
19：帝吧表情包大战	0.6516	19	19	0
20：连云港反核群体性事件	0.6492	20	20	0

4.2　网络舆情热度预测应用研究　a.数据来源　2016 年 11 月 21 日 9 时许，京昆高速山西平阳段发生多车相撞交通事故，通过清博舆情监测软件（http://yuqing.gsdata.cn）及时获取网络报刊、网络新闻、论坛、微博、微信、APP 六类传播平台的网络舆情数据（图 4），数据采集时段为 2016.11.21 9:00—2016.11.23 9:00。本文将以中国互联网舆情分析报告中 20 个热点舆情事件数据（见表 1）为基本数据库，再分别融入京昆高速交通事故舆情数据，计算网络舆情热度。

b.网络舆情热度动态预测　分别选取六类传播平台的 24 组数据、30 组数据、36 组数据、42 组数据、48

组数据，通过本文构建的网络舆情热度动态预测模型预测六类传播平台数据的变化趋势，然后应用灰色关联度计算热度(图5)。通过对比实际热度和5种情况的预测热度容易发现，随着数据量的增多，热度预测结果与实际热度越接近，从而说明本文构建的模型开展热度动态预测可行的。

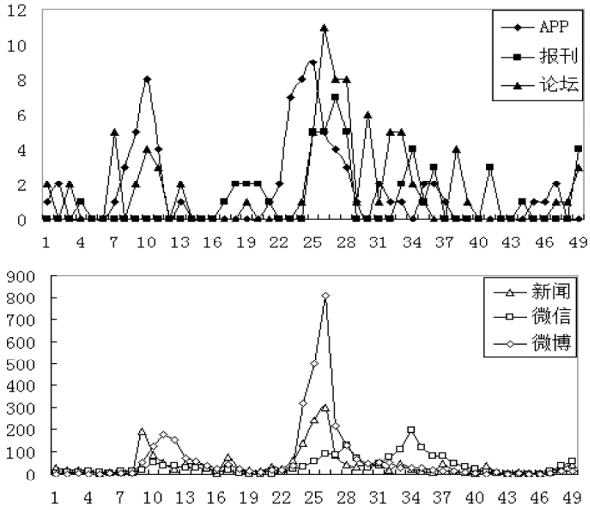

图4 京昆高速交通事故六个平台网络舆情统计数据(单位:条数)

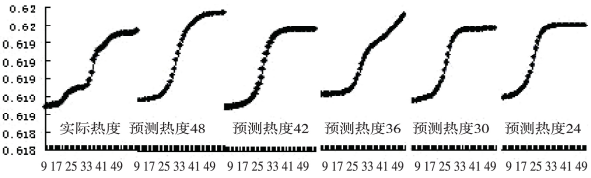

图5 京昆高速交通事故网络舆情热度动态预测

c.提升预测精度的方法 传统预测方法在预测舆情趋势时，需要大量统计数据支撑，而应用差分回归法预测趋势时，则只需要部分数据即可。网络舆情形成一段时间后，便可预测舆情在 n 个传播平台的传播趋势。然而，除了预测方法外，在实际工作中还可从完善预测基础数据角度来提升预测精度。

第一，增加数据库的案例。本文构建灰色关联度模型中的案例库只有20个，在实际工作中需要增加案例数量提升热度计算精度。.

第二，增加媒体平台个数，获取更多监测数据。本文建模和实证分析仅仅涉及六个平台，在实际工作中，还应考虑视频网站等媒体平台，还需考虑评论数、点赞数网络舆情热度的影响。

5 研究结论与展望

本文在定性研究大数据环境下网络舆情热度影响因素的基础上，基于灰色关联度模型研究了网络舆情热度问题，通过定义最大关联度向量，定量确定了网络舆情热度，避免了主观影响，并在此基础上构建多维度logistic模型开展各个媒体平台舆情信息预测模型，通过灰色关联度得出动态预测方法，最后通过实证分析验证了本文热度模型和热度动态预测的可行性，然而在本文研究基础上，还有很多问题需要深入研究，例如各个媒体平台在网络舆情传播过程中发挥的作用各有不同，在计算关联度时可以由定权关联度变权关联度，即将灰色关联度

$$r_i = \frac{1}{n}\sum_{k=1}^{n}\eta_i(k)$$

变为

$$r_i = \sum_{k=1}^{n}\lambda_i(k)\,\eta_i(k)\ ,$$

其中 $\lambda_i(k)$ 为第 i 个媒体平台的权重。

参 考 文 献

[1] 中国互联网络信息中心.第39次中国互联网络发展状况统计报告[R].http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201701/t20170122_66437.htm,2017-3-20.

[2] Lax JR,Phillips J H.How should we estimate public opinion in the states[J].American Journal of Political Science,2009,52(1):107-121.

[3] 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指标体系构建[J].情报杂志,2010,29(11):71-76.

[4] WuSiyuan.A discussion on theoretical base of internet opinion[C].Proceeding of 2nd International Conference on Artificial Intelligence,Management Science and Electonic Commerce,Deng Long,August 8-10,2011.

[5] 王长宁,陈维勤,许 浩.对微博舆情热度监测及预警的指标体系的研究[J].计算机与现代化,2013,209(1):126-129.

[6] 曹学艳,张 仙,刘 樑,等.基于应对等级的突发事件网络舆情热度分析[J].中国管理科学,2014,22(3):82-89.

[7] 孙飞显,程世辉,靳晓婷,等.政府负面网络舆情热度定量评价方法[J].情报杂志,2015,34(8):137-141.

[8] 林文声,姚一源,王志刚.食品安全事件网络舆情热度评价研究[J].现代管理科学,2016(9):30-32.

[9] Gil-Garcia R,Pons-Porrata A.Dynamic hierarchical algorithms for document clustering[J].Pattern Recognition Letters,2010(31):469-477.

[10] 王慧军,石 岩,胡明礼,等.舆情热度的最优监控问题研究[J].情报杂志,2012,31(1):71-75.

[11] 卢珺珈,张宏莉,张 玥.基于BBS的热点话题发现与态势预测技术的研究[J].智能计算机与应用,2012,2(2):1-5.

[12] 杨 雄.基于因果回路图的网络舆情热度演化模型研究[J].常州工学院学报,2013,26(12):21-26.

[13] 兰月新,王 芳,董希琳,等.公共危机事件网络舆情热度模型研究[J].情报科学,2016,34(2):32-36.

[14] 红网舆情.新浪微舆情与清华大学共推网络传播热度指数算法[EB/OL].http://yuqing.rednet.cn/Article.asp?id=317385,2017-3-20.

[15] LIU Kan,LI Jing,LIU Ping.Trend analysis of public opinion based on Markov chain.Computer Engineering and Applications,2011,47(36):170-173.

(下转第147页)

nyc.gov/assets/doitt/downloads/pdf/nyc_open_data_plan.pdf.
[13] Agency & Office Officials [EB/OL]. [2016-08-30]. http://www1.nyc.gov/office-of-the-mayor/admin-officials.page.
[14] 广东省经济与信息化委员会主要职责内设机构和人员编制规定 [EB/OL]. [2016-07-16]. http://zwgk.gd.gov.cn/006939748/201402/t20140226_480387.html.
[15] 南海率先成立数据统筹局 [EB/OL]. [2016-07-16]. http://www.gdei.gov.cn/zwgk/mtbd/2014/201406/t20140603_112671.htm.
[16] 贵阳市大数据发展管理委员会主要职责内设机构和人员编制规定 [EB/OL]. [2016-07-16]. http://www.gygov.gov.cn/art/2015/12/30/art_18325_881464.html.
[17] Malone T W, Laubacher R, Dellarocas C. The collective intelligence genome [J]. IEEE Engineering Management Review, 2010, 38(3): 38-52.
[18] Research: Open data means business [EB/OL]. [2016-12-10]. http://theodi.org/open-data-means-business.
[19] What is open data? [EB/OL]. [2016-12-17]. http://opendatahandbook.org/guide/en/what-is-open-data/.
[20] NYC big apps [EB/OL]. [2016-06-27]. http://www.nycedc.com/program/nyc-bigapps.
[21] My city way [EB/OL]. [2016-06-28]. https://www.crunchbase.com/organization/my-city-way.
[22] Embark NYC [EB/OL]. [2016-06-28]. http://letsembark.com.
[23] NYC BIGAPPS 2013 [EB/OL]. [2016-06-27]. http://2013.nycbigapps.com.
[24] 上海开放数据创新应用大赛 [EB/OL]. [2016-06-27]. http://soda.datashanghai.gov.cn.
[25] Open data: Unlocking innovation and performance with Liquid Information [EB/OL]. [2016-06-30]. http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information.
[26] Supporting and encouraging startups [EB/OL]. [2016-07-08]. http://theodi.org/about.
[27] 王　延. 多个项目实现落地孵化, 沪开放数据大赛成效显著 [N]. 浦东时报, 2014-04-20(03).
[28] 刘　祎. 贵阳市交通大数据孵化器开通 [N]. 贵阳日报, 2015-09-29(02).
[29] 5 Star Open Data [EB/OL]. [2016-06-30]. http://5stardata.info/en/.
[30] 司　莉, 李　鑫. 英美政府数据门户网站科学数据组织与查询研究 [J]. 图书馆论坛, 2014(10): 110-114.
[31] LC linked data service(library of congress) [EB/OL]. [2016-07-03]. http://id.loc.gov.
[32] 钱国富. 基于关联数据的政府数据发布 [J]. 图书情报工作, 2012, 56(5): 123-127.
[33] Why linked data for data.gov.uk? [EB/OL]. [2016-07-10]. http://www.jenitennison.com/2010/01/26/why-linked-data-for-data-gov-uk.html.
[34] 北京地图 API [EB/OL]. [2016-07-10]. http://www.beijing-map.gov.cn/API.
[35] API 服务 [EB/OL]. [2016-07-10]. http://data.qingdao.gov.cn/data/service/index.htm.
[36] 张　涵, 王　忠. 国外政府开放数据的比较研究 [J]. 情报杂志, 2015(8): 142-146, 151.
[37] 周志峰, 黄如花. 国外政府开放数据门户服务功能探析 [J]. 情报杂志, 2013(3): 144-147.

(责编: 贺小利; 王平军)

(上接第 110 页)

[16] 屈启兴, 齐佳音. 基于微博的企业网络舆情热度趋势分析 [J]. 情报杂志, 2014, 33(6): 133-137.
[17] 王新猛. 基于马尔可夫链的政府负面网络舆情热度趋势分析 [J]. 情报杂志, 2015, 34(7): 161-164.
[18] 彭　艺, 李　磊. 一种基于马尔可夫链的微信舆情热度预测模型 [J]. 信息技术, 2016(8): 81-84.
[19] 汪　鹭, 陆朝阳. 基于马尔可夫模型的网络舆情热度趋势分析 [J]. 河南工程学院学报(自然科学版), 2016, 28(3): 59-63.
[20] Watts D J, Dodds P S. Influential, networks, and public opinion information [J]. Journal of Consumer Research, 2007(34): 441-458.
[21] Cheng Hui, Liu Yun. An online public opinion forecast model based on time series [J]. Journal of Internet Technology, 2008, 9(5): 429-432.
[22] 聂恩伦, 陈　黎, 王亚强. 基于 K 近邻的新话题热度预测算法 [J]. 计算机科学, 2012, 39(6A): 247-260.
[23] Zhai Xiaofang, Liu Quanming, Cheng Yaodong, et al. Research on hotness prediction in sina microblog based on forward level analysis [J]. Computer Engineering, 2015, 41(7): 31-35
[24] Wang Pengcheng, Xiao Zheng, Liu Hui. Research on topic hotness trend modeling fused with opinion tendency [J]. Computer Engineering, 2015, 41(7): 66-70.
[25] 裴可锋, 陈永洲, 马　静. 基于 DTPM 模型的话题热度预测方法 [J]. 情报杂志, 2016, 35(12): 52-58.
[26] 李培林, 陈光金, 张　翼. 社会蓝皮书: 2017 年中国社会形势分析与预测 [M]. 北京: 社会科学文献出版社, 2016.
[27] 新华网舆情在线. 2016 年度社会热点事件网络舆情报告 [R]. [2017-03-20]. http://www.xinhuanet.com/yuqing.
[28] 清博舆情. 2016 上半舆情发展态势及规律全景报告 [R]. [2017-03-20]. http://mp.weixin.qq.com/s/qj4McrVftrua KjqPxsJsWQ.
[29] 宁宣熙, 刘思峰. 管理预测与决策方法 [M]. 北京: 科学出版社, 2008.

(责编: 王　菊; 校对: 贺小利)