应用研究。

# 主成分分析与 BP 神经网络在微博舆情预判中的应用

# 饶 浩 陈海媚2

(1. 韶关学院信息管理系,广东 韶关 512005; 2. 韶关学院应用数学系,广东 韶关 512005)

〔摘 要〕首先通过主成分分析消除原始指标之间的相关性,使指标数量变少且相互之间不相关,从而构建综合预判指标,再利用 BP 神经网络建立微博舆情预判模型。实验选取 2013 年微博热门话题作为训练样本,选取 2014 年的话题作为预测。实验结果表明,主成分分析有助于去除原始样本数据的冗余,简化了网络的复杂度,所得到的结果更加准确。因此,该模型较仅使用 BP 神经网络的准确性更高。

(关键词) 主成分分析; BP 神经网络; 微博舆情; 预判模型DOI: 10.3969/j. issn. 1008 - 0821. 2016. 07. 011

(中图分类号) G206 (文献标识码) A (文章编号) 1008-0821 (2016) 07-0058-05

# Application of Micro – blog Public Opinion Prediction Based on Combination of Principal Component Analysis and BP Neural Network

Rao Hao<sup>1</sup> Chen Haimei<sup>2</sup>

- (1. Department of Information Management, Shaoguan University, Shaoguan 512005, China;
  - 2. Department of Applied Mathematics , Shaoguan University , Shaoguan 512005 , China)

(Abstract) Firstly, the paper reduced the correlation of original indicators by principal component analysis and the number of indicators and made them separate from each other. Secondly, it established comprehensive indicators for prediction. BP neural network is used to establish the model of micro – blog public opinion prediction. The experiment using 2013 micro – blog hot topics to predict 2014 hot topics. Experimental results showed that principal component analysis helped reduce redundancy of the statistical data and made predictions more accurate. This model was more accurate than the model using only BP neural network.

(Key words) principal component analysis; BP neural network; micro - blog public opinion; prediction model

针对舆情的研究,国外起步较早,从 19 世纪中期初级 发展,到 20 世纪中期已经走向成熟<sup>[1]</sup>。在我国,专门针对 舆情的研究,大致始于 2004 年,党的十六届四中全会提 出 "建立舆情汇集和分析机制,畅通社情民意反映渠道"后,"舆情"这一词才逐渐被社会各界所熟悉并使用,此后在我国成立了有关舆情的研究中心。

目前我国针对舆情分析指标的研究有很多,如曾润喜等在建立网络舆情分级预警机制的基础上,构建了警源、警兆、警情3类指标体系<sup>[2]</sup>。高承实等结合信息空间模型构建了微博舆情的三维空间,建立了社会类指标与技术类指标、舆情主体与舆情受众之间的关系<sup>[3]</sup>。方洁等从利益相关者的视角出发,结合信息空间模型分析微博舆情传播

的过程,构建微博舆情监测指标<sup>[4]</sup>。王长宁等分析微博舆情的传播特点,在此基础上建立微博舆情预警体系<sup>[5]</sup>。也有众多学者针对预警模型进行研究,如张华采用离散的时间序列描述微博舆情的趋势,对微博文本中的热点话题进行提取、分析并对微博舆情进行预测<sup>[6]</sup>。曹帅等建立了微博舆情发展趋势的直线修正模型,阐释了移动平均法能够初步分析微博舆情的发展趋势并对其进行了预测<sup>[7]</sup>。张金伟等建立了一种基于性格、心情和情感空间的多层次心理预警模型,有效地实现了对微博情感的分析和描述<sup>[8]</sup>。邱智伟等对各维度包含的指标与微博舆情进行相关性检验,经回归分析构建舆情微博数预测方程式,通过舆情微博数的增减确定微博舆情的涨落,从而构建微博舆情预警模

收稿日期: 2016-02-04

基金项目: 2013 年教育部人文社会科学研究青年基金项目 "社交媒体潜在舆情发现及导控机制研究"(项目编号: 13YJCZH144) 研究成果之一。

<sup>~~~~</sup> 作者简介: 饶 浩(1976 –),男,副教授,硕士,研究方向: 网络技术,信息系统分析。

型<sup>[9]</sup>。林琛设计了包含网络舆情监测指标、网络舆情评估指标与网络舆情预警指标的三层网络舆情指标体系<sup>[10]</sup>。

此外还有许多针对微博舆情的研究,例如李天龙等引入系统研究方法,界定了微博舆情生成机制<sup>[11]</sup>。张亚明等结合基于 Vague 集的 AHP 评估方法对舆情进行实证分析<sup>[12]</sup>。唐晓波提出用依存句法分析来改进传统文本相似矩阵,以提高微博数据聚类准确性<sup>[13]</sup>。王洪亮等同时从情报学和传染病学的角度来研究突发自然灾害事件微博舆情蔓延特征<sup>[14]</sup>。高承实等结合信息空间模型构建了微博舆情的三维空间,运用层次分析法,建立了微博舆情监测指标体系<sup>[15]</sup>。

建立科学的微博监测指标是进行舆情预判的前提。虽然针对微博监测指标的研究已有很多,但是许多学者的研究只涉及到了定性的层面,而没有更深入的进行定量分析。也有一些学者涉及到了定量的研究,但是研究所得到的指标难以获取,导致很难进行预判监控;并且得到的定量的指标之间可能还会存在着一定的相关性,导致预判监控的结果准确率不够高。

#### 1 主成分分析与 BP 神经网络模型的建立

本研究尝试建立基于主成分分析与 BP 神经网络的微博网络社群突发舆情预判模型,通过主成分分析消除指标之间的相关性,再利用 BP 神经网络人工智能的特点,通过最速下降法的学习,输出话题等级。主成分分析使指标数量变少且相互之间不相关,再利用 BP 神经网络强大的非线性映射,有效处理这些内部机制复杂的舆情管控问题。研究以期帮助相关部门及时了解微博的舆情,有助于其对正面信息加大力度传播,对负面信息加以预警,从而保障群众对相关部门的信任与支持,利于国家和谐稳定可持续的发展[16]。

BP(Back Propagation)神经网络,即误差反传误差反向传播算法的学习过程,由信息的正向传播和误差的反向传播两个过程组成。输入层各节点接收输入信息,并传递给隐含层各节点;隐含层负责信息变换,考虑到增加隐含层的节点数比增加隐含层的层数更便于计算,不会使网络太过于复杂化,因此本文的隐含层为单隐层结构;隐含层传递信息到输出层各节点,完成一次信息的正向传播。当实际输出与期望输出不符时,误差通过输出层,按误差梯度下降的方式修正各层权值,向隐含层、输入层逐层反传。信息正向传播和误差反向传播的交替进行,使各层权值不断调整,一直到网络输出的误差减少到可以接受的范围或者预先设定的学习次数为止[17]。

本研究采用主成分分析构建微博网络舆情预判指标体系,尽可能多的考虑对预判结果有影响的指标。可是涉及的多个指标之间可能会存在一定的相关性,比如说:活跃粉丝多的微博用户可能发出一条微博会有很多的评论量。因此找出几个为原来指标的线性组合的综合指标就显得很重要了。综合指标有以下几个特点:(1)综合指标是原始指标的线性组合,因此保留了原始指标的主要信息;(2)彼此之间不相关;(3)比原始指标具有某些优越性质。这使得在研究指标体系时更加容易[18]。

通过主成分分析可以消除指标之间的相关性,用原始指标的线性组合形成的相互之间不相关的综合指标来代替原始指标,选取特征值大于1并且原始指标的大部分信息可被解释的若干个综合指标,此时得到的综合指标的数量会少于原始指标,从而减少指标,去除原始样本数据的冗余,利于数据的整理与计算;再利用 BP 神经网络,分别用原始指标和综合指标作为输入,通过数据流的正向传输和反向传输两个过程的交替进行,分别输出话题等级,比较两种输入所得到的输出结果[18]。所建立的综合模型见图 1。

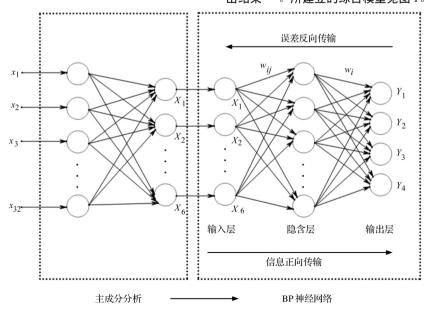


图 1 主成分分析与 BP 神经网络模型

## Journal of Modern Information

# 2 主成分分析法与 BP 神经网络的结合应用

## 2.1 话题表

由于该类问题没有 Benchmark 算例,此处选取的训练 样本与预测样本仅为说明模型使用过程,具体如下:根据 新浪微博数据中心提供的 2013 年 1 月 - 2013 年 12 月微博 热门话题盘点及新浪微博 2013 年热门事件话题大盘点,选 取 2013 年十大热点话题为预警度最高的 I 级(特别严重),在各月中排在前面但没在十大话题出现的作为 II 级(严重),在各月中排在中间位置的作为 II 级(较重),在各月中排名相对较后的作为 IV 级(一般),训练话题表如表 I 所示。以同样的方法选取 I 2014 年的话题作为预测,预测话题表如表 I 所示。

表1 各级训练样本话题表

 话 题	预警等级	话题	预警等级	话 题	预警等级	话 题	预警等级
土豪	I级	致青春	Ⅱ级	201314	Ⅲ级	南方供暖	IV级
双十一	Ι级	快乐男声	Ⅱ级	找力宏	Ⅲ级	蓝可儿	IV级
雅安地震	Ι级	薛蛮子嫖娼	Ⅱ级	315	Ⅲ级	撒切夫人	IV级
中国好声音	Ι级	李天一强奸	Ⅱ级	马上有	Ⅲ级	凯特王妃	IV级
爸爸去哪儿	Ι级	禽流感	Ⅱ级	520 表白	Ⅲ级	定西地震	IV 级
小时代	Ι级	兰陵王	Ⅱ级	高 考	Ⅲ级	公务员考试	IV级
辣妈正传	Ι级	国足	Ⅱ级	全国高温	Ⅲ级	腾格尔新歌《桃花源》	IV级
春晚	Ι级	小爸爸	Ⅱ级	七夕	Ⅲ级	小龙女陈妍希	IV级
屌 <b>丝</b>	Ι级	中国梦之声	Ⅱ级	iPhone5s	Ⅲ级	龙门镖局	IV级
我是歌手	Ι级	雾霾	Ⅱ级	余姚水灾	Ⅲ级	波音 777 坠机	IV级

表 2 各级预测样本话题表

 话 题	预警等级	话题	预警等级	话 题	预警等级	话 题	预警等级
世界杯	Ι级	杨幂结婚生子	II 级	APEC	Ⅲ级	杨幂刘恺威大婚	IV级
春晚	Ι级	古剑奇谭	Ⅱ级	欧 冠	Ⅲ级	北京爱情故事	IV级
雾霾	Ι级	爱情公寓	Ⅱ级	花儿与少年	Ⅲ级	妻子的秘密	IV级
爸爸去哪儿	Ι级	微博上市	Ⅱ级	柯震东吸毒	Ⅲ级	黄奕离婚	IV级
来自星星的你	Ι级	舌尖上的中国	Ⅱ级	步步惊情	Ⅲ级	秦时明月	IV级
iPhone6	Ι级	周一见	Ⅱ级	鲁甸地震	Ⅲ级	金玉良缘	IV级
中国好声音	Ι级	冰桶挑战	Ⅱ级	奔跑吧兄弟	Ⅲ级	舌尖上的宿舍	IV级
我是歌手	Ι级	NBA	Ⅱ级	青奥会	Ⅲ级	花千骨	IV级
马航	Ι级	冬奥会	Ⅱ级	东莞扫毒	Ⅲ级	离婚律师	IV级
春 运	Ι级	亚运会	Ⅱ级	小 冰	Ⅲ级	宁财神吸毒	IV级

# 2.2 对数据进行主成分分析

微博热度由广播数  $x_1$ 、收听数  $x_2$ 、听众数  $x_3$ 、平均转播数 and 平均评论数  $x_4$ 、听众的平均收听人数  $x_5$ 、听众的平均听众人数  $x_6$ 、听众的平均等级  $x_7$ 、收听人的平均收听人数  $x_8$ 、收听人的平均听众人数  $x_9$ 、收听人的平均等级  $x_{10}$ 、手机达人  $x_{11}$ 、QQ 会员  $x_{12}$ 、转播达人  $x_{13}$ 、天下无双  $x_{14}$ 、微博劳模  $x_{15}$ 、青春正能量  $x_{16}$ 、秒微创意  $x_{17}$ 、事实派  $x_{18}$ 、微生活  $x_{19}$ 、微爱校园行  $x_{20}$ 、投票达人  $x_{21}$ 、礼物达人

 $x_{22}$ 、新鲜达人  $x_{23}$ 、热心达人  $x_{24}$ 、奇吃妙享  $x_{25}$ 、评论达人  $x_{26}$ 、点赞狂魔  $x_{27}$ 、心情签到达人  $x_{28}$ 、动感勋章  $x_{29}$ 、新星 主播  $x_{30}$ 、沙发王  $x_{31}$ 、蒲公英印记  $x_{32}$  这 32 个原始指标组成。随机选取 324 个名人,获取上述数据。经过主成分分析处理之后得到综合指标  $X_1$ 、 $X_2$ 、 $X_3$ 、 $X_4$ 、 $X_5$ 、 $X_6$ ,这 6个综合指标能概括 32 个原始指标 77. 81% 的信息,达到了降低复杂度的同时尽量保留大部分信息的目的。

# 2.3 建立 BP 神经网络模型

以话题为关键词,获取该话题转发量最多的 10 位名人的 32 个原始指标(其中有些指标是该话题所特有的,比如: 平均转播数 and 平均评论数; 还有一些指标是该名人所特有的,比如: 广播数、收听数),分别计算 32 个原始指标的平均值作为该关键词的相应原始指标。

- (1) 直接用 32 个原始指标  $x_1$ 、 $x_2$   $\cdots x_{32}$  作为 BP 神经网络的输入数据。以训练样本话题的数据为样本对 BP 神经网络进行训练,然后用预测样本的数据进行输入,比较输出结果与真实结果。
- (2) 32 个原始指标先根据主成分分析得到的公式算出 6 个综合指标,再把得到的 6 个综合指标  $X_1$ 、 $X_2$ 、 $X_3$ 、 $X_4$ 、  $X_5$ 、 $X_6$  作为 BP 神经网络的输入数据,经过训练学习后把输出结果与真实结果进行比较。

两种方法的运行结果对比如图 2 所示。

# 2.4 结果分析

为了使结果看起来更直观。将准确率分为一级准确率、

二级准确率、三级准确率、四级准确率。其中,一级准确 率: 输出结果与真实结果相同的话题数/总话题数; 二级准 确率:输出结果与真实结果相差一个等级(例:真实结果 为Ⅰ级预警,而预测结果为Ⅱ级预警,反过来亦然)的话 题数/总话题数; 三级准确率: 输出结果与真实结果相差两 个等级(例: 真实结果为Ⅰ级预警,而预测结果为Ⅲ级预 警)的话题数/总话题数;四级准确率:输出结果与真实 结果相差3个等级(例: 真实结果为 [级预警, 而预测结 果为Ⅳ级预警)的话题数/总话题数。根据定义,计算出 两种方法所得结果的准确率: 法一得到的一级准确率为 22/40=55%, 二级准确率为11/40=27.5%, 三级准确率 为4/40 = 10%,四级准确率为3/40 = 7.5%。而法二得到的 输出结果一级准确率为35/40 = 87.5% , 二级准确率为2/40 = 5% , 三级准确率为1/40 = 2.5% , 四级准确率为2/40 = 5% 。 由此可以得到,经过主成分分析后预警结果的一级准确率比没 有进行主成分分析提高了32.5%;而误差比较明显的三级准 确率和四级准确率比没有进行主成分分析降低了10%。发

Neural Network

Algorithms

Training:

Derivative: Progress

Epoch:

Time:

Performance:

Validation Checks:

Performance

Training State

Regression

Minimum gradient reached.

Plot Interval:

Gradient:

♠ Neural Network Training (nntraintool)

Mean Squared Error (mse)

Default (defaultderiv)

0

0

(plotperform)

(plottrainstate)

(plotregression)

0.356

0.0662

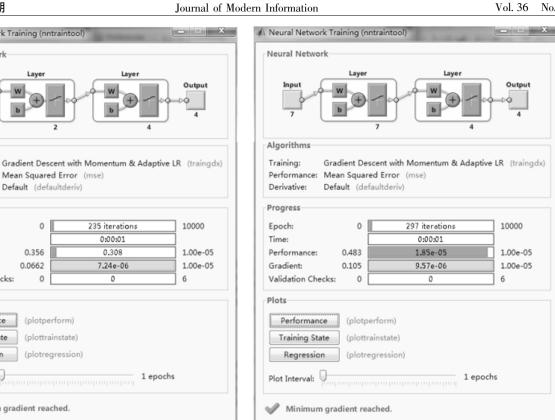
235 iterations

0:00:01

0.308

7.24e-06

Stop Training



一、没有进行主成分分析

二、主成分分析

Stop Training Cancel

图 2 运行结果对比图

Cancel

现因为主成分分析有助于去除原始样本数据的冗余,简化 了网络的复杂度,所以所得到的结果更加准确。

另一方面,方法一训练时间为1秒,进行了235次迭 代,错误率为0.308。方法二训练时间为1秒,进行了297 次迭代,错误率为1.85\*10^(5)。虽然训练时间与迭代次 数都差不多,可是误差却是几个数量级的区别。可见简化 网络的复杂度可以在同样的时间内进行更加有效的学习。

由此可以得出,为了实验的准确性,需要获取尽可能多 的原始数据,可是过多的原始数据之间不可避免的可能会有 一定的相关性,这就导致了矛盾的存在。而采用主成分分析 的方法,可以有效的减少这种矛盾,使结果更加准确。

#### 3 结 语

本研究结合主成分分析与 BP 神经网络来构建微博网 络社群突发舆情危机预判模型,通过实例来验证模型的准 确性。此模型较仅使用 BP 神经网络的准确性更高。此外, 该模型也会存在着一些缺点。首先是进行主成分分析的 324 位名人是随机选取的,在此进行改进(如把名人细分: 体育类、新闻类、综艺类等) 并增加名人的数量,预判结 果的准确性可能会有所增长。其次是以话题为关键词获取 原始指标时为了获取的方便,是以转发量为考虑对象的, 而没有充分结合所有方面进行考虑,且获取的名人数量不 够多。再次此刻获取 2013 年某个关键词的数据,获取到的 是当前的数据,用当前数据的值来分析2013年的情况,不 可避免的存在一定的误差。最后也存在 BP 神经网络固有

的缺点,就是容易陷入局部极小值点,结果也会根据隐含 层的元素个数而改变。

该模型的预判结果的准确性将随着原始指标的合理性、 训练集数量的增加而增长。研究所得到的结果可以为舆情 的管理提供有效的指导。

#### 考文献

- [1] 艾新革. 国内外舆情研究述略 [J]. 图书馆学刊, 2011, (9): 140 - 142.
- [2] 曾润喜. 网络舆情突发事件预警指标体系构建 [J]. 情报理 论与实践, 2010, (1): 77-80.
- [3] 高承实,荣星,陈越. 微博舆情监测指标体系研究 [J]. 情 报杂志,2011,(9):66-70.
- [4] 方洁,龚立群,魏疆. 基于利益相关者理论的微博舆情中的 用户分类研究 [J]. 情报科学,2014,(1): 18-22.
- [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, (1): 126-129.

(下转第70页)

- [13] 顾忠伟,徐福缘,卫军.可穿戴商务消费者初始信任影响因素的实证研究[J].管理评论,2015,27(7):168-176.
- [14] Min Q, Shaobo J I, Gang Q U. Mobile Commerce User Acceptance Study in China: A Revised UTAUT Model [J]. Tsinghua Science & Technology, 2008, (3): 257-264.
- [15] 陈洁,朱小栋. 基于 UTAUT 和 ITM 整合视角的微信支付用户使用行为影响因素研究 [J]. 现代情报,2015,35(4):
- [16] 万君,郭婷婷,吴正祥. 基于 UTAUT 模型的消费者移动互 联网业务使用意愿影响研究 [J]. 资源开发与市场,2015, 31 (10): 1224-1227.
- [17] Herath T, Rao HR. Protection motivation and deterrence: A framework forsecurity policy compliance in organisations [J]. Eur J Inform Syst, 2009, 18: 106-125.
- [18] Prentice Dunn, Steven, Rogers RW. Protection Motivation Theory hand preventive health: beyond the Health Blief Meodel [J]. HealthEduc Rs, e 1986, 1 (3): 153-161.
- [19] Rogers R W. A protection motivation theory of fear appeals and attitude changel [J]. The journal of psychology , 1975 , 91 (1): 93-114.
- [20] Johnston AC, Warkentin M. Fear appeals and information security behaviors: Anempirical study [J]. MIS Q 2010, 34: 549 566.
- [21] 王芸,肖霞,郑频频,等. 保护动机理论在个体行为改变中的应用和发展 [J]. 中国健康教育,2009,25 (11):853-855.
- [22] 刘彩. 保护动机理论在健康行为解释、干预与预测中的应用研究 [J]. 医学与社会,2015,28 (7):77-79.
- [23] Pechmann C, Zhao G, Goldberg ME, Reibling ET. What to convey in antismokingadvertisements for adolescents: The use of protection motivation theory toidentify effective message themes [J]. J Market, 2003, 67: 1-18.
- [24] Bulgurcu B , Cavusoglu H , Benbasat I. Information security pol-

- icy compliance: An empirical study of rationality based beliefs and information security awareness [J]. MIS Q , 2010 , 34: 523 – 548.
- [25] Anderson CL, Agarwal R. Practicing safe computing: A multi-media empirical examination of home computer user security behavioral intentions [J]. MIS Q, 2010, 34: 613-643.
- [26] Yoon C, Kim H. Understanding computer security behavioral intention in theworkplace: An empirical study of Korean firms [J]. Inform Technol People, 2013, 26: 401-419.
- [27] Lv X , Guo X , Xu Y , et al. Explaining the Mobile Health Services Acceptance from Different Age Groups: A Protection Motivation Theory Perspective [J]. International Journal of Advancements in Computing Technology , 2012 , 4 (3): 1-9.
- [28] Guo X , Han X , Zhang X , et al. Investigating m Health Acceptance from a Protection Motivation Theory Perspective: Gender and Age Differences [J]. Telemedicine and e Health , 2015 , 21 (8).
- [29] 刘炜. 基于扩展 TTF 和 UTAUT 模型的老年用户社会化网络服务采纳行为研究 [J]. 软科学, 2015, 29(3): 120-124.
- [30] Li D, Browne G J, Wetherbe J C. Why Do Internet Users Stick with a Specific Web Site? A Relationship Perspective [J]. International Journal of Electronic Commerce, 2006, 10 (4): 105 – 141.
- [31] Maddux J E , Rogers R W. Protection motivation and self efficacy: A revised theory of fear appeals and attitude change [J]. Journal of experimental social psychology , 1983 , 19 (5): 469 – 479.
- [32] Koufaris M , Hampton Sosa W. The development of initial trust in an online company by new customers [J]. Information & Management , 2004 , 41 (3): 377 397.

(本文责任编辑: 郭沫含)

# (上接第62页)

- [6] 张华. 基于优化 BP 神经网络的微博舆情预测模型研究 [D]. 武汉: 华中师范大学, 2014.
- [7] 曹帅, 兰月新, 苏国强, 等. 基于移动平均法的微博舆情预测模型研究[J]. 湖北警官学院学报, 2014, (3): 40-42.
- [8] 张金伟,刘晓平.基于心理预警模型的微博情感识别研究 [J].合肥工业大学学报:自然科学版,2013,(11):1318-1322
- [9] 邱智伟. 基于混沌理论的微博舆情预警模型研究 [D]. 广州: 华南理工大学, 2015.
- [10] 林琛. 基于网络舆论形成过程的舆情指标体系构建研究 [J]. 情报科学, 2015, (1): 146-149.
- [11] 李天龙,李明德,张宏邦. 微博舆情生成机制研究 [J]. 情报杂志,2014,(9): 117-122.
- [12] 张亚明,刘婉莹,刘海鸥. 基于 Vague 集的微博舆情评估体

系研究 [J]. 情报杂志, 2014, (4): 84-89.

- [13] 唐晓波,肖璐. 基于依存句法分析的微博主题挖掘模型研究 [J]. 情报科学,2015,(9):61-65.
- [14] 王洪亮,周海炜. 突发自然灾害事件微博舆情蔓延规律与控制研究 [J]. 情报杂志,2013,(9):23-28.
- [15] 高承实,荣星,陈越. 微博舆情监测指标体系研究 [J]. 情报杂志,2011,(9):66-70.
- [16] 潘芳,张霞,仲伟俊. 基于 BP 神经网络的微博网络社群突发舆情的预警监控 [J]. 情报杂志,2014,(5): 125-128.
- [17] 杨淑娥,黄礼. 基于 BP 神经网络的上市公司财务预警模型 [J]. 系统工程理论与实践,2005,(1): 12-18.
- [18] 陈建宏,刘浪,周智勇,等.基于主成分分析与神经网络的采矿方法优选[J].中南大学学报:自然科学版,2010,(5):1967-1972.

(本文责任编辑:郭沫含)