Doi:10.3969/j.issn.1672-0105.2019.04.013

基于文本挖掘的大学生网络舆情检测和预警模型*

金慧峰，程振设
（浙江工贸职业技术学院，浙江 温州 325003）

摘 要：为了实现大数据时代大学生网络舆情的动态监测和预警，从百度贴吧里抓取近几年学院大学生网络聊天记录，建立了网络舆情指数、预警级别划分和特征词指数，成功实现了网络舆情的实时动态检测和异常舆情的预警，成功捕捉到热门主题及其关键词，为高校舆情防控和利用提供了实证支持。

关键词：大数据；大学生；网络舆情；检测；预警
中图分类号：G641 **文献标识码：**A **文章编号：**1672-0105(2019)04-0057-06

Detection and Early Warning Model of College Students' Internet Public Opinion Based on Text Mining

JIN Hui-feng, CHENG Zhen-she
（Zhejiang Industry & Trade Vocational College, Wenzhou, 325003, China）

Abstract: In order to realize the goal of dynamic monitoring and early warning of college students' internet public opinion in the era of big data, the college students' internet chat posts in recent years have been crawled from the Baidu Post Bar, to establish some indexes, including the internet public opinion index, warning level classification, and feature words index. It has successfully achieved the real-time dynamic detection of network public opinion and early warning of abnormal opinions, as well as hot topics and their relevant key words, which provides empirical support for the prevention and control of public opinion in colleges and universities.

Key Words: big data; college student; network public opinion; detection; early warning

0 引言

互联网运营模式的不断创新、线上线下服务融合的加速、公共服务线上化步伐的加快，使得几乎所有大学生成为网民。微博、微信、论坛、贴吧等社交网络的繁荣发展，使得大学生在这些社交网络上发帖、转发、评论等行为已经成为常态。在传统数据时代，研究者主要通过抽样调查、内容分析等方法获取有限的、有代表性的舆情样本信息，并运用统计学方法进行分析。在大数据时代，随着海量舆情信息的涌现和数据采集技术的进步，样本分析被总体分析所取代，传统的抽样分析和检测预警手段已无法适应大数据时代的发展趋势，网络舆情大数据的分析、检测和预警成为社会管理的客观需求。

目前关于大数据时代高校学生网络舆情监测和预警机制的研究成果较少，主要分为两个层面。其一是理论层面，根据大学生网络舆情传播的特点和现状，提出了高校网络舆情管理的思路、策略和路径[1-4]；其二是技术层面，主要集中于网络检测系统的设计[5-8]、网络舆情挖掘技术[9-10]等。不论理论层面还是技术层面，均没有针对大学生网络舆情的定量化监测的成果，主要原因可能在于海量文本信息不但对当前计算机性能提出了较大挑战，而且对文本挖掘技术也提出了较高的要求。

1 相关理论简介

1.1 文本表示方法

目前，基于统计的文本挖掘方法[11-12]中，文本

收稿日期：2019-10-12
基金项目：2018年教学创新项目“软件技术专业现代学徒制探索与实践”（浙工贸〔2018〕61号）；2018年度党建研究课题重点项目“大数据时代高校学生网络舆情检测与预警机制研究”（浙工贸党办〔2018〕10号）；浙江省高校“十三五”优势专业建设项目（浙教高教〔2016〕164号）
作者简介：金慧峰（1976—），男，浙江永嘉人，浙江工贸职业技术学院高级工程师，硕士，主要研究方向：计算机软件及计算机应用、职业教育；程振设（1969—），男，浙江永康人，浙江工贸职业技术学院讲师，硕士，主要研究方向：思政教育。

是以向量形式表示的，向量的分量是特征词的频数，特征词是根据文本挖掘的任务或目标来确定的，可以是名词、动名词或形容词，等等。因此，要将文本表示为向量，首先就要将文本分词。

1.2 文本分词

目前国内常用的分词方法[11-12]有：机械分词法、词库匹配法、词频统计法、语义分析法、神经网络分词法、联想-回朔法、联想词群法、知识与规则法等。这些分词算法可以归为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。本文采用中科院计算所的汉语词法分析系统ICTCLAS进行分词，该系统的功能有中文分词、词性标注、未登录词识别，分词正确率高达97%以上，未登录词识别召回率均高于90%，其中中国人名的识别召回率接近98%，处理速度为31.5Kb/s。

1.3 特征词选择

所有文本分词之后，形成词语集，词的总数通常都很大，这会使得表示文本的向量空间的维数也相当大，因此需要降维。降维技术有两类：特征选择和特征重构。

特征选择是指去除冗余的和不能表达文本挖掘任务信息的词，或者选择那些能够表达文本挖掘任务信息的词（称之为特征词），从而减少词语总量，达到降维目的。特征选择的结果为原始词语集的子集。特征选择方法：根据词频来判断，当词频小于或大于给定的阈值时就去掉。

特征重构是指将原始词语集经过数学变换构造出新的词语集，以此达到降维的目的。新的词语集不是原始词语集的子集。比较常用的特征重构方法是潜在语义分析。

2 实证研究

2.1 研究设计

本文以百度贴吧里近几年浙江工贸职业技术学院（下称学院）大学生网络聊天的文本信息为研究对象，开展大学生网络舆情的监测和预警。相关工作主要有五步：第一，使用python爬虫软件从百度贴吧抓取近几年的聊天帖子，数量将超过万条。每个帖子的信息包括帖子ID、主题、作者、跟帖数量、跟帖内容、跟帖作者、跟帖日期和时间。第二，对抓取到的文本信息作总体特征分析、热门主题及其作者搜寻、热门主题的内容分析等。第三，建立舆情指数，度量网络舆情的大小，形成动态直观的网络舆情走势图。第四，设置“黄色、橙色和红色”三个预警级别，对网络舆情进行预警。第五，建立特征词指数，实现对热门主题的热点关键词的捕捉。

2.2 数据采集与初步分析

编写python爬虫软件，从浙江工贸百度贴吧（http://tieba.baidu.com/f?kw=浙江工贸）抓取到2007年5月4日到2018年2月28日大学生的“精品”帖子，一共6551条文本评论。每个帖子的信息包括帖子ID、主题、作者、跟帖数量、跟帖内容、跟帖作者、跟帖日期和时间。

2.3 大学生网络舆情检测模型

以天为计时单位。设 a_i 表示第 i 天的衍生贴数量（个），$\overline{a_i}$ 表示第 i 天的历史平均衍生贴数量（个），则第 i 天的舆情指数为

$$u_i=\begin{cases}\dfrac{a_i}{\overline{a_i}}, & \overline{a_i}\neq 0\\ 1, & \overline{a_i}=0\end{cases} \tag{1}$$

$u_i\in[0,+\infty)$。

统计出每天的舆情指数 $u_1,u_2,\ldots$，就形成了动态指数，如表1所示。

如果以时刻 i 为横轴，以舆情指数为纵轴，可以画出动态指数图。

从2007年5月4日至2018年3月17日的动态指数，如图1所示（剔除了指数为0）。

另外，从2007年5月4日至2018年3月17日的最大指数是53.9，具体日期是2013年8月19日，意味着这一天的帖子数量是历史平均值的53.9倍，其主题是“亲，你遇到了么?”，进一步查看帖子内容（略），大部分是关于寻找在温州的老乡的帖子。中国人普遍具有浓重的老乡情节，当大学生收到录取通知书之后，即将从全国各地来到陌生的温州，此时如果能够遇到老乡，那么就有了类似于亲人一样的、可以互相依赖和帮助的朋友，于是通过网络查找老乡就成为一条便捷的途径。

2.4 大学生网络舆情预警

为了预警，需要确定舆情指数的合理界限。如果舆情指数超过了这个界限，就发出预警信号。从表1和图1可知，一方面，舆情指数为0的指数占比

表1 部分舆情指数

时间	舆情指数	时间	舆情指数	时间	舆情指数	时间	舆情指数
2017/7/2	3.030 8	2017/8/4	0.606 2	2017/8/29	3.030 8	2018/1/23	0.606 2
2017/7/3	1.212 3	2017/8/6	3.637	2017/8/30	1.818 5	2018/2/2	0.606 2
2017/7/4	0.606 2	2017/8/9	1.818 5	2017/8/31	3.637	2018/2/3	0.606 2
2017/7/6	1.818 5	2017/8/10	1.818 5	2017/9/1	1.818 5	2018/2/4	1.212 3
2017/7/7	0.606 2	2017/8/11	1.818 5	2017/9/2	1.818 5	2018/2/7	0.606 2
2017/7/10	0.606 2	2017/8/12	3.030 8	2017/9/3	1.212 3	2018/2/9	0.606 2
2017/7/11	1.212 3	2017/8/13	1.212 3	2017/9/4	0.606 2	2018/2/23	0.606 2
2017/7/12	0.606 2	2017/8/14	0.606 2	2017/9/5	1.818 5	2018/2/24	3.030 8
2017/7/18	0.606 2	2017/8/15	10.911	2017/9/6	0.606 2	2018/2/25	1.212 3
2017/7/20	1.818 5	2017/8/16	3.637	2017/9/9	2.424 7	2018/2/27	0.606 2
2017/7/21	0.606 2	2017/8/17	0.606 2	2017/9/10	0.606 2	2018/2/28	0.606 2
2017/7/23	0.606 2	2017/8/18	1.212 3	2017/9/13	0.606 2	2018/3/1	1.818 5
2017/7/24	0.606 2	2017/8/19	3.030 8	2017/9/14	0.606 2	2018/3/2	2.424 7
2017/7/25	0.606 2	2017/8/22	1.818 5	2017/10/18	0.606 2	2018/3/3	1.212 3
2017/7/27	0.606 2	2017/8/24	4.243 2	2018/1/11	0.606 2	2018/3/5	0.606 2
2017/7/28	1.212 3	2017/8/25	12.123 3	2018/1/17	0.606 2	2018/3/15	0.606 2
2017/7/31	0.606 2	2017/8/26	1.818 5	2018/1/19	0.606 2	2018/3/16	1.212 3
2017/8/1	1.818 5	2017/8/27	0.606 2	2018/1/20	1.818 5	2018/3/17	1.212 3
2017/8/2	1.818 5	2017/8/28	0.606 2	2018/1/21	3.030 8		

注：(1)时间区间为2017年7月1日至2018年3月17日；(2)不含舆情指数0。

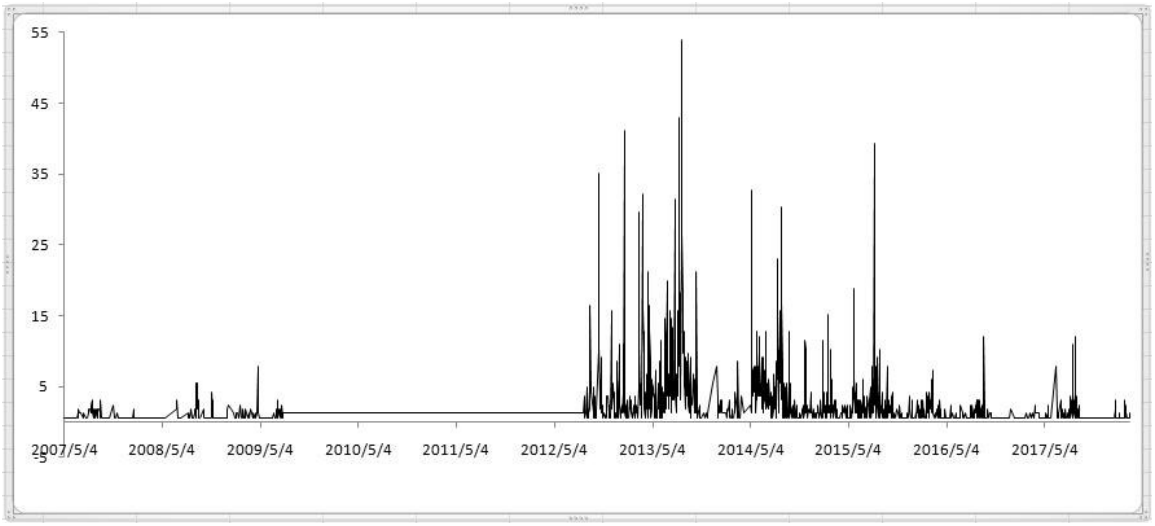

图1 2007/5/4—2018/3/17的舆情指数

很大，是32.8%；另一方面，舆情指数的极差也很大，是53.9。于是将原指数中的0指数剔除，并针对非0指数实施以7天为窗口的移动平均，再画出舆情指数的直方图，如图2所示。

从图2可知，非0指数呈现负指数分布。给定显著性水平 $\alpha=0.01$，估计其均值得 $\mu=3.0475$，指数分布的参数 $\lambda=1/\mu\approx0.328$，指数分布的概率密度函数为

$$f(x)=\begin{cases}0.328e^{-0.328x}, & x\geqslant 0;\\ 0, & x<0.\end{cases} \tag{2}$$

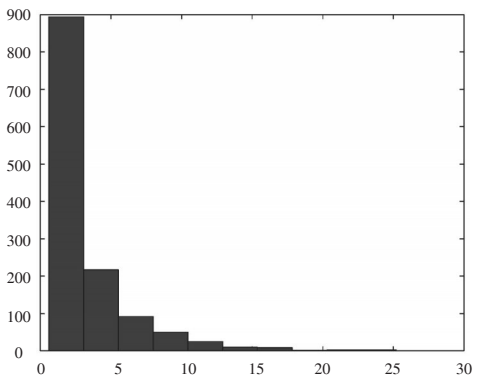

图2 舆情指数直方图

2.4.1 大学生网络舆情预警级别的设置

本文将预警级别设定为“黄色、橙色和红色”三个级别。

给定显著性水平 α，置信度 $1-\alpha$ 对应的分位数记作 μ_α。如果舆情指数超过分位数 μ_α，则发出预警信号。于是给定三个不同的显著性水平 $\alpha=0.1$、0.05、0.01，预警级别的临界值即可确定，如表2所示。

表2 预警级别临界值

预警状态	正常	黄色	橙色	红色
判定标准	$u_i \leqslant \mu_{0.1}$	$u_i > \mu_{0.1}$	$u_i > \mu_{0.05}$	$u_i > \mu_{0.01}$

2.4.2 大学生网络舆情预警级别的设置结果

不同显著性水平下的指数分布检验、分位数和均值估计结果如表3所示。

表3 指数分布检验、分位数和均值估计

显著性水平	相伴概率	统计量	分位数	均值
0.01	0.000 0	0.193 4	14.034 1	3.047 5
0.05	0.000 0	0.193 4	9.129 4	3.047 5
0.1	0.000 0	0.193 4	7.017 1	3.047 5

从表3可知，在0.01的显著性水平下，非0指数服从指数分布。于是，舆情指数预警的临界值如表4所示。

表4 预警临界值

预警状态	正常	黄色	橙色	红色
判定标准	$u_i \leqslant 7$	$u_i > 7$	$u_i > 9$	$u_i > 14$

从2017年7月1日至2018年3月17日的非0舆情指数预警图，如图3所示。

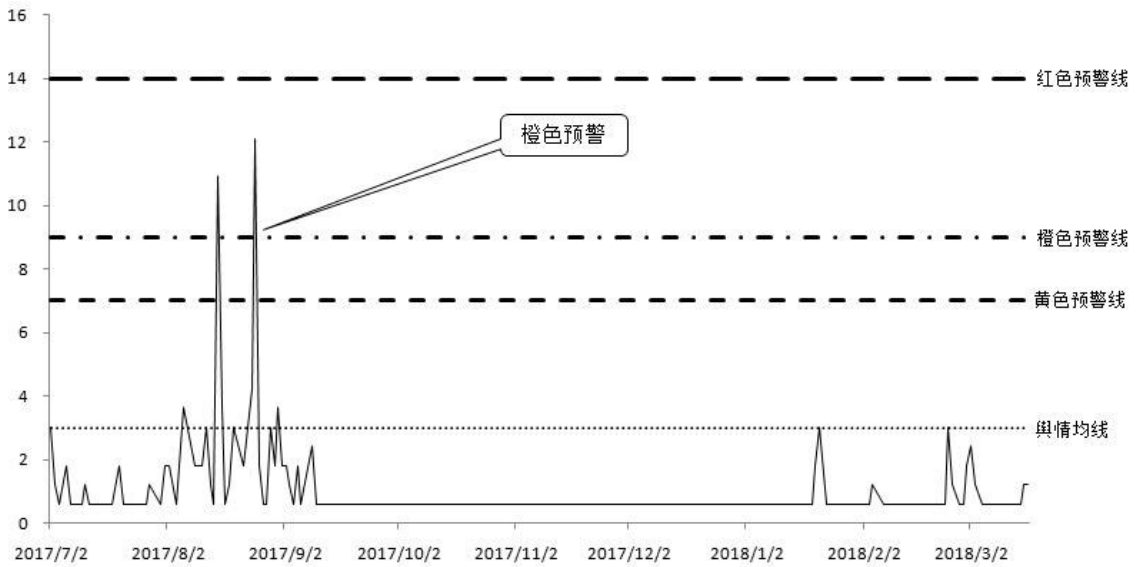

图3 2017/7/1—2018/3/17的舆情指数及预警线

从图3和表1可知，在2017年8月15日和25日分别发出了橙色预警信号，需要引起关注。

查看2017年8月15日的发帖主题，分别是“毕业老学姐解答专升本疑惑”和“在浙工贸的70件事”。查看2017年8月25日的发帖主题，分别是“毕业老学姐解答专升本疑惑”、“开学骗术多——揭秘那些常见骗术”、“掉进染色桶里的工贸”和“在浙工贸的70件事”。

可见，跟帖增多的原因是学生对“专升本”话题很感兴趣，对“开学骗术”和工贸学院话题很关注。究其原因，首先，大二学生即将升入大三，一部分学生开始考虑专升本的诸多问题了；其次，新生即将报到，为了防止被骗对开学骗术自然就很关注；第三，毕业生可以回顾在大学的三年期间发生的历历往事，记录美好瞬间、回味幸福时刻、抒发离愁别绪，每一件事都成为工贸学院的特写，也成为即将来到工贸学院的准大学生们感兴趣的事件，引起他们的关注就不足为奇。

2.5 热门主题的关键词搜寻

将触发预警的主题称为热门主题。对于热门主题，我们需要进一步确定吧友们讨论的关键词是什

么，例如对于专升本这个热门主题，关键词是“辅导、考试、报志愿、高等数学”里的哪一个？因此需要建立关键词搜寻模型。

2.5.1 文本预处理

采用中科院计算所的汉语词法分析系统ICTCLAS对文本进行分词，形成词语集，然后选择名词、动词和形容词作为特征词，一共340 6个。以向量表示文本，设 X 表示一条文本，则

$$X=(w_1,w_2,...,w_m)^T \tag{3}$$

其中，w_i 表示第 i 个特征词的频数，m 是特征词的个数。

2.5.2 特征词指数

由于舆情指数反映了衍生贴的相对数量，而每一个衍生贴是由特征词表示的，在热点帖子已经确定的情况下，如何测量特征词的热度呢？本文使用特征词指数来度量特征词的热度。

设有 m 个特征词，有 n 个文本，第 i 个特征词在第 j 个文本中的频数记作 a_{ij}，$a_{ij}\geqslant 0$，$i=1,2,...,m$，$j=1,2,...,n$。

对于第 i 个特征词，在第 j 个文本中出现的次数越多，说明其反映大学生的心理愿望越强烈，则热度越大，于是第 i 个特征词在第 j 个文本中的热度使用频率来度量，即

$$b_{ij}=\frac{a_{ij}}{\sum_{i=1}^{m}a_{ij}},\ i=1,2,...,m,\ j=1,2,...,n \tag{4}$$

第 i 个特征词的平均热度为

$$c_i=\frac{1}{n}\sum_{j=1}^{n}b_{ij},\ i=1,2,...,m \tag{5}$$

对于第 i 个特征词，在各个文本中出现的次数越多，说明讨论它的大学生越多，则热度越大，于是第 i 个特征词的权系数为

$$d_i=\frac{1}{n}\sum_{j=1}^{n}e_{ij},\ i=1,2,...,m \tag{6}$$

其中，

$$e_{ij}=\begin{cases}1, & a_{ij}>0;\\ 0, & a_{ij}=0.\end{cases},\ i=1,2,...,m,\ j=1,2,...,n \tag{7}$$

第 i 个特征词的加权热度为

$$f_i=c_i d_i,\ i=1,2,...,m \tag{8}$$

第 i 个特征词的归一化加权热度为

$$g_i=\frac{f_i}{\sum_{i=1}^{m}f_i},\ i=1,2,...,m \tag{9}$$

归一化加权热度 $g_i\in[0,1]$。

称归一化加权热度超过某阈值的特征词为关键词。于是，通过设置一个合适的阈值 ε，可将关键词筛选出来。

2.5.3 关键词搜寻结果

以2017年8月15日引起橙色预警为例，针对主题“毕业老学姐解答专升本疑惑”，设置阈值 $\varepsilon=0$，并删除无意义的词，关键词搜寻结果如表5所示。

将表5中这些关键词联系起来分析，可以推测吧友们讨论的主要话题，比如：“专升本报考的学校和专业”“考试要求”“会计”“数学”“找到女朋友”“难易”，等等。作为即将专升本的学生，他们关心的话题自然是考试要求、考试内容、难易程度、报考学校以及专业；由于工贸学院的会计专业学生的入门录取分数高，学生基础扎实，所以专升本的学生自然就多；在专升本的考试科目中，数学

表5 关键词搜寻结果

名词	词语	专升	报	专业	会计	问题	学校	女朋友	朋友	学期	数	办法
	指数	0.107 6	0.023 3	0.023 1	0.006 5	0.003 8	0.003 6	0.003 5	0.003 5	0.001	0.000 7	0.000 5
	排序	1	2	3	5	6	7	8	9	11	12	13
动词	词语	升	学	考	找到	要求	考试	参考	解答	喜欢		
	指数	0.150 3	0.054 1	0.050 5	0.003 5	0.003 1	0.001	0.000 8	0.000 7	0.000 5		
	排序	1	2	3	4	5	6	7	9	11		
形容词	词语	难	大	全	易	小	新					
	指数	0.002 6	0.001	0.001	0.000 8	0.000 7	0.000 7					
	排序	1	2	3	4	6	7					

注：主题“毕业老学姐解答专升本疑惑”。

是关键，既容易得分又容易失分，区分度大，数学自然成为学生讨论的话题；至于“找到女朋友”，可能是某些男生希望专升本之后快速的找到女朋友吧。

3 研究结论

本文以百度贴吧里从2007年5月4日至2018年3月17日的学院大学生网络聊天文本信息为研究对象，建立了网络舆情检测模型，实现了大学生网络舆情的定量检测。然后设置了三级预警反应机制，实现了网络舆情异常状况的预警和热门主题的捕捉。最后，建立了特征词指数，实现了对热门主题里的关键词捕捉。获得的结论如下：

（1）最大指数发生的日期是2013年8月19日，其主题是“亲，你遇到了么?”，帖子内容是寻找在温州的老乡。

（2）在2017年8月15日和25日分别发出了橙色预警信号，其中，2017年8月15日的发帖主题分别是“毕业老学姐解答专升本疑惑”和“在浙工贸的70件事”；2017年8月25日的发帖主题分别是“毕业老学姐解答专升本疑惑”、“开学骗术多——揭秘那些常见骗术”、“掉进染色桶里的工贸”和“在浙工贸的70件事”。

（3）搜寻热门主题“毕业老学姐解答专升本疑惑”的关键词，分别是“专升本报考的学校和专业”“考试要求”“会计”“数学”“找到女朋友”“难易”，等等。

综上所述，通过研究高校网络舆情，建立和健全舆情监测和预警机制，可以实时掌握大学生的思想动态，及早发现突发事件的苗头，主动解决学生的思想问题，优化高校思想政治教育方法，对于维护校园和谐发展，促进社会稳定具有重要意义。

参考文献：

[1] 周丽梅.大数据背景下的高校网络舆情管理策略研究[J].东南传播,2017(2):46-48.
[2] 魏伟华.基于大数据背景下的高校学生网络舆情管理机制研究[J].中国成人教育,2017(17):69-73.
[3] 徐江虹.基于大数据的高校网络舆情应对研究[J].学校党建与思想教育,2017(23):44-46.
[4] 张雯鑫.大数据背景下大学生网络舆情管理研究[D].中国矿业大学,2017.
[5] 王树辰.基于海量舆情信息的话题检测系统的设计与实现[D].中山大学,2013.
[6] 郑岩.高校网络舆情监测系统的目标定位、评判依据与运行保障研究[J].情报科学,2015(6):81-85.
[7] 杨秋平.内容分析技术在网络舆情智能检测中的应用[J].制造业自动化,2011,33(6):53-55.
[8] 杨秋平.网络舆情智能检测与分析系统的设计[J].电脑知识与技术,2011,07(4):759-761.
[9] 赵旭东.互联网舆情指数挖掘方法研究[D].哈尔滨工业大学,2007.
[10] 宋保江.网络舆情检测与控制关键技术研究[D].哈尔滨工业大学,2010.
[11] 万源.基于语义统计分析的网络舆情挖掘技术研究[D].武汉理工大学,2012.
[12] 韩开旭.基于支持向量机的文本情感分析研究[D].东北石油大学,2014.

（责任编辑：黎浩宏）