文章编号：2096-1472(2022)-11-44-04
DOI:10.19644/j.cnki.issn2096-1472.2022.011.010

基于弹幕的网络舆情文本挖掘与情感分析

白 健，洪小娟

(南京邮电大学管理学院，江苏 南京 210003)
✉1535179246@qq.com; 1291823970@qq.com

摘 要：针对传统评论方式依赖整体感知且相对滞后的问题，以弹幕这一新兴短信息表达方式为研究对象，采用文本挖掘与情感分析的方式研究弹幕与网络舆情之间的潜在联系。采用网络爬虫技术采集网络舆情弹幕数据，使用Jieba库实现分词、去停用词及高频词统计，基于WordCloud库绘制词云图，实现可视化，并使用SnowNLP库计算网络舆情弹幕的情感得分，运用隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)模型进行主题词提取，实现对网络舆情弹幕的情感分类和主题分析。实验结果表明，该方法可多维展现网民的情感倾向与关注焦点，是对传统评论文本研究的有效补充。

关键词：网络舆情；情感分析；文本挖掘；SnowNLP；LDA
中图分类号：TP391.1 **文献标识码**：A

Text Mining and Sentiment Analysis of Network Public Opinion based on Bullet Screen

BAI Jian, HONG Xiaojuan

(*School of Management*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)
✉1535179246@qq.com; 1291823970@qq.com

Abstract: In view of the problem that the traditional review methods rely on the overall perception and are relatively lagging behind, this paper takes the bullet screen, a new short message expression, as the research object, and proposes to use text mining and sentiment analysis to further study the potential relationship between bullet screen and network public opinion. Firstly, network crawler technology is used to collect the bulletin data of network public opinion. Secondly, Jieba library is used to realize word segmentation, stop words and high frequency word statistics, and WordCloud library is used to draw word cloud map to realize visualization. Finally, SnowNLP library is used to calculate the sentiment score of the network public opinion bullet screen, and LDA (Latent Dirichlet Allocation) model is used to extract the keywords to realize the sentiment classification and theme analysis of the network public opinion bullet screen. The experimental results show that this method can show the sentiment tendency and focus of netizens in multiple dimensions, and it is an effective supplement to traditional review text research.

Keywords: network public opinion; sentiment analysis; text mining; SnowNLP; LDA

1 引言(Introduction)

随着新媒体技术的不断蓬勃发展，人们获取信息和表达情绪的方式更加多元化。以Bilibili为代表的新媒体传播平台在传统评论的基础上引入弹幕评论，为网民提供全新表达途径的同时，也构建了全新的网络舆情空间，逐渐成为新的“网络舆情传播载体”[1]。

传统评论是网民基于整体感知做出的“滞后”评论，因而更加偏于“理性”表达[2]。而弹幕作为一种新媒体时代下的短信息表达方式，以实时评论的方式表达了用户对于当前视频的即刻认知与行为倾向，相比于传统评论方式具有更强的

情感色彩和时效性[3-4]，这对于网络舆情情感分析研究具有独特的研究价值。通过对弹幕内容进行数据可视化、情感分析以及主题分类，有助于动态把握网络舆情态势走向，追踪网民关注热点，寻找弹幕背后所蕴含的情感倾向和舆情热点，为防范化解网络舆情风险，完善舆情分析机制，构建和谐稳定网络空间做出贡献。

2　研究设计(Research design)

本文研究设计思路：首先，使用Python编写网络爬虫技术代码进行网络舆情弹幕文本数据采集和数据清洗；其次，使用中文分词组件Jieba进行弹幕数据的分词、去停用词以及高频词统计，得到网络舆情的高频关键词及权重；再次，调用WordCloud库设置词云图样式并将经过Jieba分词器处理的弹幕数据进行词云图呈现；最后，基于SnowNLP进行情感分析，判断弹幕数据中积极、消极、中性的情感比例并进行分析，得出情感分析占比图、直方图和波动图，并基于LDA主题模型得到焦点主题。具体研究流程如图1所示。

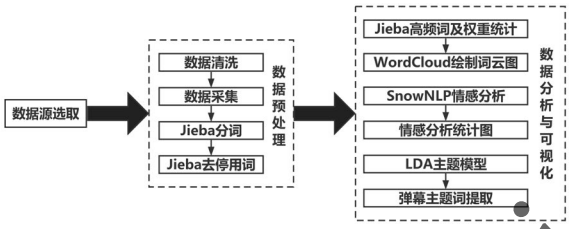

图1 基于弹幕的网络舆情文本挖掘与情感分析流程

Fig.1 Text mining and sentiment analysis process of network public opinion based on bullet screen

3　数据采集与数据清洗(Data acquisition and data cleaning)

3.1　数据源选取

Bilibili作为一个快速崛起的新媒体平台，具有超过3亿的用户数量，其活跃用户群体大，弹幕数量丰富且具有较好的包容性[5]，因而本文选择Bilibili作为数据源，进行数据采集操作。

3.2　数据采集

首先使用开发者工具获取视频弹幕的异步请求包，观察和分析网页变化规律，找到网络数据来源[6]。通过对目标网页数据来源地解析，从Headers中获取爬虫所需的URL、Cookie及User-agent。其次，使用Python的Requests第三方库，使用解析获得的Cookie以及User-agent构建headers{}请求头，结合URL地址调用request.get()方法获取原始弹幕数据；最后，使用Python内置Re库的正则表达式re.findall()函数精确匹配要爬取的内容，剔除无关数据，并将弹幕数据进行存储。

3.3　数据清洗

数据清洗是网络爬虫的重要一环，通过剔除原始弹幕文本中的表情符号、数字、空白值等无效信息，可以有效提升数据质量[7]。

4　高频词统计与数据可视化(Statistics and data visualization of high frequency words)

在完成数据采集以及数据清洗后，调用第三方Jieba、WordCloud库实现高频词统计与数据可视化。

4.1　Jieba分词、去停用词及高频词统计

Jieba分词器是目前Python中最好的中文分词组件，主要利用中文词库确定汉字间的相关概率，进而产生正确分词结果，此分词方式的准确率超过了97%，能够很好地协助使用者完成主题词抽取、潜在主题发现等工作，尤其适用于中文文本分类。Jieba支持用户词典和停用词字典功能，这能够在较大程度上提升分词结果的准确度，对分词结果不太理想的词组，也能够采取引入用户自定义字典的方法加以处理[8]。因而本文选择使用Jieba分词器进行弹幕文本数据的分词、去停用词及高频词统计。

首先，使用Pandas库的read_csv()方法导入经过简单数据清洗的弹幕文本数据，并通过Jieba库的jieba.lcut()方法实现对弹幕文本的分词操作；其次，使用stopwords=[line.strip() for line in open().readlines()]导入停用词词典，并通过遍历循环将“增加热度、增热专用、1、2”之类无效弹幕进行剔除；最后，使用jieba.analyse.extract_tags()方法提取弹幕文本“Top10关键词及权重”并通过遍历操作实现存储。

4.2　WordCloud词云图绘制

WordCloud库以WordCloud对象为基础，以词语为基本单位进行词云图绘制。首先，通过wordcloud.WordCloud()函数进行词云图参数设置，本文设置width=1200，height=900，font_path='msyh.ttc'，background_color="white"，max_words=1500，stopwords=stopwords，确定词云图的形状、尺寸、背景色、字体等；其次，使用wordcloud.generate_from_text()方法将Jieba分词处理后的弹幕文本数据传入词云图中；最后通过wordcloud.to_file()方法输出词云图。

5　弹幕情感倾向分析(Sentiment tendency analysis of bullet screen)

5.1　SnowNLP情感分析原理

传统的Python自然语言处理库大多都面向英文，对于中文文本处理兼容性较差，而SnowNLP库的出现很好地弥

补了这一点[9]。SnowNLP库自带中文正负情感训练集，可以通过朴素贝叶斯原理实现情感分析、词性标注、文本分类等操作，很好地适用于中文文本数据的处理，故本文选取SnowNLP进行网络舆情的情感分析。通过SnowNLP情感分析可以获得情感分析占比图、直方图、波动图以及情感得分表等可视化结果。SnowNLP情感预测基本原理如下。

假设情感分类有积极态度A_1和消极态度A_2两类，且单一弹幕中有$B_1, B_2, \cdots, B_n$共n个相互独立的单词。

$$P(A_1|B_1,\cdots,B_n)=\frac{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}{P(B_1,\cdots,B_n)} \tag{1}$$

由朴素贝叶斯公式(式1)可知：具有$B_1, B_2, \cdots, B_n$单词的弹幕属于积极态度A_1的概率=如果弹幕是积极态度A_1，那么该句子具有$B_1, B_2, \cdots, B_n$单词的概率×弹幕是积极态度A_1的概率/弹幕里具有$B_1, B_2, \cdots, B_n$单词的概率。

同时，根据全概率公式$P(B)=P(B|A)\cdot P(A)+P(B|A')\cdot P(A')$，可以得到$P(B_1,\cdots,B_n)=P(B_1,\cdots,B_n|A_1)\cdot P(A_1)+P(B_1,\cdots,B_n|A_2)\cdot P(A_2)$，即弹幕里具有单词$B_1, B_2, \cdots, B_n$的概率=如果句子是积极态度$A_1$，那么该句子具有$B_1, B_2, \cdots, B_n$单词$A_1$的概率×句子是积极态度的概率+如果句子是消极态度$A_2$，那么该句子具有$B_1, B_2, \cdots, B_n$单词的概率×句子是消极态度$A_2$的概率，进而可以将式(1)转化为式(2)。

$$P(A_1|B_1,\cdots,B_n)=\frac{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)+P(B_1,\cdots,B_n|A_2)\cdot P(A_2)} \tag{2}$$

式(2)即为SnowNLP情感预测过程使用的基本式，该式还可以进一步简化为式(3)。

$$\begin{aligned}P(A_1|B_1,\cdots,B_n)&=\frac{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)+P(B_1,\cdots,B_n|A_2)\cdot P(A_2)}\\&=\frac{1}{1+\frac{P(B_1,\cdots,B_n|A_2)\cdot P(A_2)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}}\\&=\frac{1}{1+\exp\left[\lg\left(\frac{P(B_1,\cdots,B_n|A_2)\cdot P(A_2)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}\right)\right]}\\&=\frac{1}{1+\exp[\lg(P(B_1,\cdots,B_n|A_2)\cdot P(A_2))-\lg(P(B_1,\cdots,B_n|A_1)\cdot P(A_1))]}\end{aligned} \tag{3}$$

其中分母中的1可以改写为

$$1=\exp[\lg(P(B_1,\cdots,B_n|A_1)\cdot P(A_1))-\lg(P(B_1,\cdots,B_n|A_1)\cdot P(A_1))] \tag{4}$$

5.2 LDA主题模型原理

LDA是潜在语义分析和概率语义分析的扩展，在文本数据挖掘等领域广泛使用。LDA模型可以自动将文本自动编码为一定数量具有实质性意义的主题，可极大减少人为干预负担。运行LDA模型，可以获得每个主题下的词语分布概率，以及文档对应的主题概率，其模型结构如图2所示。

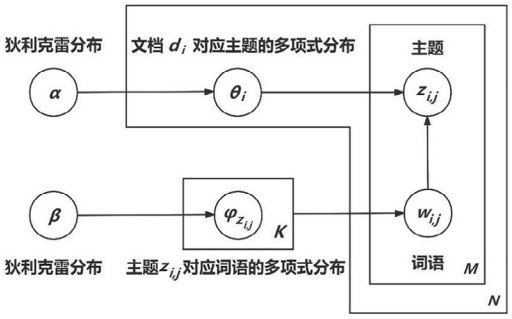

图2 LDA模型结构示意图

Fig.2 Structure diagram of LDA model

LDA模型分为文档、主题和词语三层，是典型的生成式主题模型，具体文档生成过程如下。

(1)按照先验概率$P(d_i)$选择一篇文档d_i。

(2)从以参数为α的狄利克雷分布中随机生成文档d_i对应主题的多项式分布θ_i。

(3)从文档d_i对应主题的多项式分布θ_i中随机生成第j个词的主题$z_{i,j}$。

(4)从以参数为β的狄利克雷分布中随机生成主题$z_{i,j}$对应词语的多项式分布$\varphi_{z_{i,j}}$。

(5)综合主题$z_{i,j}$对应词语分布情况$\varphi_{z_{i,j}}$生成词语$w_{i,j}$。

其中，参数α、β及主题对应的主题数K一般事先给定，图中向量边表示依存关系，矩形表示重复，矩形中的字母M、N、K代表重复次数。

5.3 实验与分析

为了验证基于弹幕的网络舆情文本挖掘与情感分析的可行性以及可靠性，以“鸿星尔克捐款”为主题构建实验数据，进行效果检验。

首先，爬取相关弹幕并对数据进行清洗，获得视频地址、弹幕地址、弹幕时间以及弹幕内容等数据，如图3所示。

视频地址	弹幕地址	弹幕时间	弹幕内容
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2022/7/2 19:20	满分作文
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2022/3/26 19:30	我就要买国货，国货之光，再上一百万个缝纫机也不够，支
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2022/3/3 19:31	我就要多买国货，支持国货
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2021/10/12 16:23	我就要支持国货！
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2021/9/9 2:49	算量！！！
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2021/9/5 14:46	较大指华人圈震动
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2021/9/4 12:03	好
https://www.bilibili.com/video/BV1yE4y1c7sy	http://comment.bilibili.com/376358642.xml	2021/8/23 8:50	你的建议我虚心接受，我要支持国货！

图3 爬虫结果展示(部分)

Fig.3 Crawler results show (partial)

其次，经过Jieba分词、去停用词、高频词统计，获得“Top10关键词及权重”表，详见表1。其中，“国货、格局、鸿星尔克、支持”等网络舆情关键词赫然在列，其权重分别为1.426044、1.144364、0.934489、0.518985。同时，通过WordCloud绘制词云图，可以得到以“鸿星尔克捐款”为主题的弹幕词云图，如图4所示。图中“支持国货、格局、鸿星尔克”等关键词词频较高。

表1 Top10关键词及权重

Tab.1 Top10 keywords and their weights

关键词	权重	关键词	权重
国货	1.426044	野性消费	0.158287
格局	1.144364	理性消费	0.149901
鸿星尔克	0.934489	老板	0.139580
支持	0.518985	就要	0.124475
之光	0.226227	衣服	0.081145

图4 词云图

Fig.4 Word cloud

最后，调用SnowNLP和LDA进行最为重要的弹幕情感倾向分析和主题提取。通过SnowNLP情感分析，可以得到与“鸿星尔克捐款”相关的网络舆情弹幕情感分析占比图、直方图和波动图，如图5—图7所示。图5从情感得分占比的角度给出了情感分析数据，可以直观看出积极、消极及中性情感分别占比为87.93%、10.66%和1.41%。图6以直方图的形式呈现了情感得分的区间分布，从图中可以看出整体情感分布靠右，说明网民对于该网络舆情事件呈现较为积极的态度。图7以波动图的形式呈现了弹幕时间与情感得分的关系。图中，横轴为弹幕时间，纵轴为弹幕情感得分，波动曲线整体分布靠上，且随着时间推移越发稳定于上侧区间，一方面说明情感得分均值高于0.5，网民对该事件大多持积极观点，另一方面说明随着时间推移持有积极观点的网民逐渐占据多数。

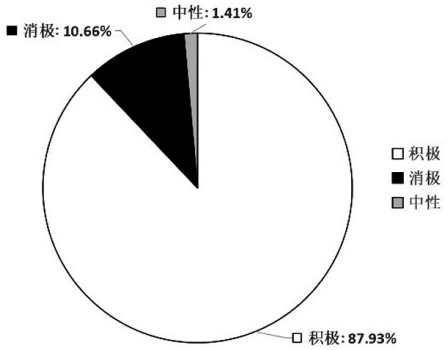

图5 鸿星尔克情感分析占比图

Fig.5 Proportion chart of sentiment analysis for Hongxing Erke

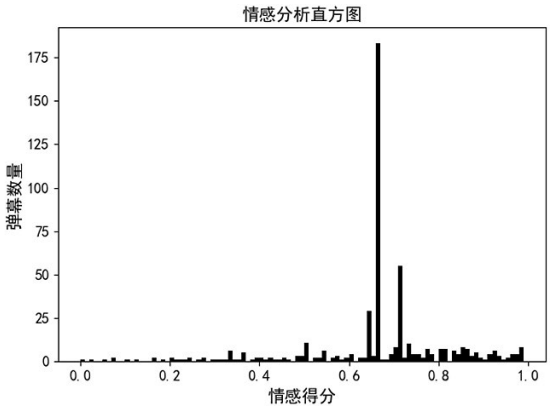

图6 情感分析直方图

Fig.6 Histogram sentiment analysis

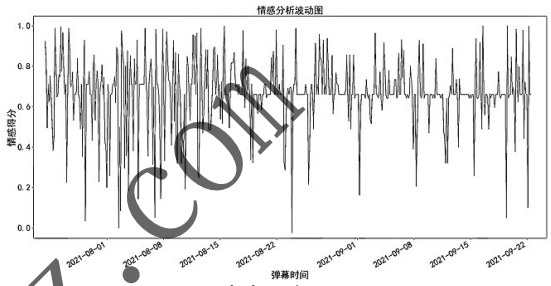

图7 情感分析波动图

Fig.7 Fluctuation graph of sentiment analysis

通过实验分析不难发现，情感分析占比图、直方图和波动图三者分别从情感得分占比、分布和时间三个角度呈现了弹幕背后蕴含的舆情信息，说明以“鸿星尔克捐赠”为主题的相关网络舆情，整体情感倾向较为积极，网络舆情态势稳步向好。

通过LDA进行主题词提取可以获得弹幕数据主题聚类表，详见表2。从表2可以看出“格局、国货、鸿星尔克、吴荣照老板”四个主题是弹幕背后隐藏的核心主题词，是网民真正的关注焦点。

表2 主题聚类表

Tab.2 Theme clustering table

主题	特征词
格局	格局、理性、衣服、消费、老板……
国货	国货、理性、衣服、消费、老板……
鸿星尔克	鸿星、消费、理性、国货、衣服……
吴荣照老板	老板、衣服、理性、消费、红星……

不难发现，利用LDA进行主题词提取获得的主题聚类表与Jieba分词获得的“Top10关键词及权重”表及WordCloud绘制的词云图所示结果一致，三者相互印证，说明无论是词频角度还是聚类角度，“格局、国货、鸿星尔克、吴荣照老板”均为该网络舆情的核心焦点，进一步呈现了弹幕与网络舆情之间的潜在联系，这对于切实把握网民关注焦点，防范化解衍生舆情具有重要意义。

6 结论(Conclusion)

弹幕相较于传统评论具有更强烈的情感色彩以及更强的时效性，本文通过对弹幕数据的文本挖掘和情感分析探索隐藏在弹幕背后的网络舆情信息。实验结果显示，历经网络爬虫、数据清洗、数据可视化、SnowNLP情感分析和LDA主题词分类等步骤后，获得的网络舆情弹幕词云图、情感分析占比图、直方图、波动图及LDA主题聚类表等结果较好地呈现了网民的情感倾向与关注焦点，这对于把握网络舆情动态走向、防范化解网络舆情风险具有一定的现实意义。新媒体时代下，网民群体意见表达渠道更加多元化，弹幕这一新兴情感表达方式的出现，是对现有舆情研究的良好补充，通过深入对网络舆情弹幕的研究可以更好地响应网民合理关切，完善舆情分析机制，进而为构建更加和谐清明的网络空间做出贡献。

参考文献(References)

[1] 包雅玮.新媒体环境下青年爱国表达的新特征——以“B站”弹幕文化为例[J].中国青年研究,2021,7(02):96-101,109.

[2] 李知谕,杨柳,邓春林.基于弹幕与评论情感倾向的食品安全舆情预警研究[J].科技情报研究,2022,4(03):33-45.

[3] 孙晓宁,姚青.多元主题场景下的用户弹幕与评论特征比较研究:基于Bilibili网站[J].情报理论与实践,2021,44(09):135-141,121.

[4] YU L, WU Y, YANG J, et al. Bullet subtitle sentiment classification based on affective computing and ensemble learning[J]. Wireless Communications and Mobile Computing, 2021, 2(01):1-9.

[5] 刘臻睿.B站弹幕文化与“Z世代”集体记忆的建构[J].新媒体研究,2022,8(05):75-77,88.

[6] 陈琳,任芳.基于Python的新浪微博数据爬虫程序设计[J].信息系统工程,2016,7(09):97-99.

[7] 邹墨馨,辛雨璇.基于文本挖掘的影视弹幕情感分析研究[J].科技创新与应用,2021,11(24):51-53.

[8] 江涛,黄昌昊,孙斌.基于文本挖掘的弹幕情绪分析研究[J].智能计算机与应用,2022,12(08):60-64,69.

[9] 刘策,李贞,颜明会.面向大众点评网评论的文本情感分析研究[J].现代信息科技,2021,5(19):37-39.

作者简介：

白 健(1999-)，男，硕士生.研究领域：网络舆情.

洪小娟(1975-)，女，硕士，教授.研究领域：网络舆情，信息化建设.

(上接第62页)

台计算机之间的防火墙。防火墙在保护计算机方面可以起到更实质性的作用，因为毕竟所有的数据流都需要通过防火墙进行过滤[11]。一般来说，防火墙有以下功能：首先，防火墙可以防止其他无关的用户进入私人电脑；其次，即使有人从外部进入我们的系统，防火墙也可以阻止他接近你的防御设施；最后，防火墙可以阻止我们访问特殊站点。

7 结论(Conclusion)

计算机网络安全是每个计算机用户都需要关注的问题。我们生活中一定要注意对钓鱼网站、非法链接、垃圾邮件等的清理，不要因为疏忽而给不法分子机会。另外，计算机网络安全的技术发展要尽快跟上，从技术上对不法分子进行压制。未来计算机网络安全技术的发展还有很长的路要走。要尽快实现技术突破，完善安全防护措施。

参考文献(References)

[1] 什么是网络安全和信息安全[EB/OL].(2022-03-28)[2022-05-22].https://baijiahao.baidu.com/s?id=1728511997141647939&wfr=spider&for=pc.

[2] 杨海红.大数据时代计算机网络安全及有效防范措施探究[J].数字通信世界,2021(12):140-141,150.

[3] 刘娟.大数据时代的计算机网络安全及防范措施[J].无线互联科技,2021,18(14):31-32.

[4] KIM H Y, LEE K G. IIoT malware detection using edge computing and deep learning for cybersecurity in smart factories[J]. Applied Sciences, 2022, 12(15):7679.

[5] 宋吉书.大数据时代的计算机网络安全及防范措施[J].新型工业化,2021,11(05):84-85,88.

[6] 梅彬.基于人工智能理论的网络安全管理关键技术研究[J].信息网络安全,2021(S1):66-69.

[7]朱军红,周海军,唐明根.大数据时代下计算机网络安全及防范措施探究[J].无线互联科技,2021,18(07):21-22.

[8] LI L F. Application of data encryption technology in computer network information security[J]. Security and Communication Networks, 2022(10):15-20.

[9] 宋涛,李秀华,李辉,等.大数据时代下车联网安全加密认证技术研究综述[J].计算机科学,2022,49(04):340-353.

[10] 李宽荣,牛志杰,高宇,等.车联网大数据的安全访问控制模型设计[J].单片机与嵌入式系统应用,2022,22(04):29-33.

[11] 迟克群.大数据背景下计算机网络安全策略分析[J].网络安全技术与应用,2022(05):168-169.

作者简介：

贺军忠(1982-)，男，硕士，副教授.研究领域：网络组建与信息安全，网络营销.