doi:10.3772/j.issn.2095-915x.2020.02.003

突发事件评论集中的情报甄别方法初探

开放科学（资源服务）标识码（OSID）

张运良[1,2] 丁思媛[1,2] 高雄[1,2]

1. 中国科学技术信息研究所 北京 100038；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘要：本文研究在突发事件评论集中甄别情报的方法，有助于补充和完善应急情报体系，也能够促进情报甄别方法的发展。以科技类突发事件评论集为研究对象，研究了基于内部评论内容、外部数据特征、内外部数据一致性的情报甄别方法；探讨了复杂网络分析、情感色彩分析、主题标引、命名实体抽取等技术在情报甄别中的作用。在 3 个科技类突发事件评论集中，甄别出了有价值的情报信息，说明相关方法在有关联非规范短文本情报甄别中具有一定的作用。

关键词：情报甄别；评论集；评论内容；外部数据；内外部数据一致性

中图分类号：G353.1

A Preliminary Study on the Method of Information Screening in Emergency Comment Set

ZHANG Yunliang[1,2] DING Siyuan[1,2] GAO Xiong[1,2]

1. Institute of Scientific & Technical Information of China, Beijing 100038, China;
2. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, China

Abstract: To study the method of information screening in emergency comments is helpful to supplement and improve the emergency intelligence system, and also to promote the development of information screening methods. Taking the scientific and technological emergency comment sets as the research base, this paper studies the information screening method based on the internal comment content, external data characteristics, internal and external data consistency, and discusses the role of complex

基金项目：中国科学技术信息研究所创新基金面上项目“知识驱动的评论类情报甄别方法研究”（MS2020-05）；中国工程科技知识中心项目“知识组织体系建设”（CKCEST-2020-1-19）；融智库 2019 年重点课题“新闻出版业知识服务发展研究”。

作者简介：张运良（1979-），博士，研究员，研究方向：知识组织、自然语言处理、情报工程，E-mail：zhangyl@istic.ac.cn；丁思媛（1996-），硕士研究生，研究方向：情报工程、知识组织；高雄（1990-），硕士，助理研究员，研究方向：自然语言处理、知识工程、文本挖掘。

doi:10.3772/j.issn.2095-915x.2020.02.003

network analysis, sentiment analysis, indexing, named entity extraction and other technologies in information screening. The valuable information has been identified in three scientific and technological emergency comment sets. The results show that the relevant methods play a certain role in the information screening on interrelated non-standard short texts.

Keywords: Information screening; comment set; comment content; external data characteristics, internal and external data consistency

引言

突发事件一般是指那些突然发生，造成或者可能造成严重社会危害，需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件[1]。突发事件具有发生不确定、发展途径和演变规律不易把握、危害程度难以预计、常规防治手段失效等特征[2]。在这样的条件下， 应急决策情报体系的建设迫在眉睫，解决核心情报的获取问题是提升突发事件应急决策能力的关键之一，在专业的情报服务机构、信息服务机构和智库之外，其他可能的情报源也值得重视[3]。论坛中针对突发事件相关的评论通常实时性比较强，版面划分相对保证了讨论问题的专业性，而用户等级、积分等管理制度从一定程度上使得其评论好于一般匿名评论，而且由于具有汇聚众智的特点，能够形成对传统情报的补充。但是论坛中的评论数量较大，质量参差不齐，且并非所有的评论都具有情报价值，情报甄别的目的即是将素材中潜在价值较高的评论内容遴选出来，并通过后续处理，补充和丰富应急情报体系中的情报资源。由于这些评论以初始评论为核心组织在一起形成一个评论集，评论集尺度适中，评论之间相互关联依存，为研究针对性的情报甄别方法提供了可能。

尽管科技类突发事件没有列入现有突发事件的四个大类之中，但由于科技事件能够影响小到把握趋势、研究选题[4]，大至资源配置[5]、社会发展[6]等活动，突发的尤其是涉及国际竞争的科技事件，则更可能打破国家之间的科技以至政治力量的平衡，影响国家安全，重要性不言而喻。本文选择科技类突发事件评论集为基础，主要从个案评论集的内部评论内容、外部评论信息以及内外部数据一致性等角度探索情报甄别方法。情报甄别以效用评价为核心，内容评价用以辅助效用评价，研究既是对现有科技情报服务的拓展，同时也能为其他类型突发事件情报甄别及服务提供借鉴，通过持续的研究和工具化将会形成对大规模的突发事件评论集合情报甄别的能力。

1 相关研究进展

目前尚未检索到与本文研究对象完全一致的研究，但是一些相关研究值得关注。在情报甄别方面，尤其是竞争情报应用中有一些基础研究[7-10]，中国科学技术信息研究所研制了包含权威性、影响力、关注度、领域相关度、完整性、时效性、新颖性、领域交叉性等面向情报工程的数据价值二级指标的计算方法，并在论文、专利、新闻、立项数据等情报资源上进行

了数据价值的综合评价 [11]。但是该方法基于深度学习和词嵌入技术，而且数据集合之间的关系主要基于相同主题，与评论集中的关系不同，对于评论类情报甄别目前还难于适用。

在一般评论类情报甄别方面，研究者往往侧重真实性甄别 [12-15]，评论之间的关联主要是基于评论对象，如商品或饭店、酒店等服务 [16]，从而对评论事件整体刻画不足。在已知论坛评论相关研究中，缺乏基于评论集做甄别的研究。较多的研究基于金融类论坛，如东方财富股吧等，但是研究多是基于社会互动等相关理论 [17]，从宏观层面基于论坛发帖的模式 [18-19]，以及发帖内容的情感和数量来了解用户的情绪，进而通过模型分析证券的走势 [20-22]。也有一些研究论坛内部的互动行为 [23] 或者根据论坛内容为新用户做推荐 [24]。

尽管以上研究的内容规范性、评估尺度或甄别方法与本文有明显的不同，但是相关研究思路和成果可以借鉴。

2 实验数据准备

本文采用的数据从水木社区 (www.newsmth.net) 的航空航天 (Aero) 和微电子技术 (METech) 两个板块采集，两个板块都是属于讨论科学技术相关话题较为集中的板块。本文以航空航天为主，微电子技术为辅，分别从中选择 2 篇和 1 篇回复用户较多的评论及其全部回复，构成 3 个话题评论集。3 个话题分别关于美国停供国产大飞机发动机（以下称为话题 A）、SpaceX 公司再次发射星链项目卫星（以下称为话题 B）以及对极紫外（EUV）光刻机逆向的想法（以下称为话题 C）。

2.1 数据采集

由于采集的评论本身数量不多，采集过程和研究思路形成过程重叠，因此本研究采用自动和手工采集相结合的方式。采集工具主要采用在基于 Chromium 内核的 360 急速浏览器中安装 Web Scraper 扩展①的方式来实现，基于爬取数据量及费用的考虑，采用本地爬取方式。结合网站的结构和特点，主要采集评论者、评论内容和评论楼层信息。其中评论者又根据网站提供的信息，采集了其中的身份、发文数、作者等级、作者积分等信息；评论内容则囊括了评论内容以及包括发信人、信区、发信站、时间、内容（可能包含引用）、发文工具、签名档、修改信息、来源及 IP 地址等；板块评论楼层信息主要用于后续的数据清洗和评论标识。导出数据格式为 CSV，可以方便后续加工。3 个话题评论统计如表 1 所示，大部分评论都是对原始评论的单开贴回复评论，少量是附在原始评论贴上的直接评论，两种评论主要是评论形式的差异，因此在分析时不做特别的区分处理。

表 1 评论集的评论数量统计

评论集	原始评论	单开贴回复评论数量	直接评论数量	评论总数量
话题A	1	99	8	108
话题B	1	65	0	66
话题C	1	54	2	57

① https://www.webscraper.io/

doi:10.3772/j.issn.2095-915x.2020.02.003

2.2 数据清洗

对于每个话题，需要为其下的评论赋予编号，本文采取数字编号的方式。对于标记为楼主的评论编号设为 0，其余各评论按照显示在发帖中的“第 X 楼”字样提取 X 代表的数字作为其编号，依据编号对所有爬取的次序混乱的数据进行重新排序，去除无用的页面信息，对于在原帖上直接评论的情况，在其它评论编号完成后，排在最后，依据发表顺序，编号依次递增。

在采集的过程中，由于评论内容中包含信息较多，对于其中正文内容、引用评论者和 IP 等基于其模式特征，在 EmEditor 文本工具中利用正则表达式进行处理和提取，对于一些规则性不强的内容，则人工对结果进行订正和补充。

由于网页中没有附在原始评论贴上的直接评论者的信息，因此通过在论坛查询的方式，人工补充其发文数量、等级和积分等信息。对所有的作者按照其发表评论的编号顺序依次从 0 开始编号，3 个话题的作者数分别为 58，19 和 39。

3 基于评论内容的情报甄别

3.1 评论的原创性

一般来讲，原创评论可能更具情报意义，引用和转载的评论意义就要打折扣。区分原创评论至少有两条路径。其一是引用的标志，如包括特定 URL、“据……报道”和一些典型的媒体，如 WSJ，Weibo 等。其二是附和型评论，这些评论本身没有特别重要的意义，除非引入了新的观点，或者对被引用评论做了补充完善。典型的附和型评论可能包括一些提示词，从分析的这 3 个话题中找到一些典型提示词，如“对了”、“有道理”、“可以（了）”、“是的”、“必须（的）”等。对以上类型的评论进行剔除，能够找到更有价值的评论候选。

3.2 评论的长度

对 3 个话题中评论的字数统计见表 2，不同话题之间具有较大的差异。从直观上字数较少的评论通常歧义较大，不易解读，而字数较多的评论相对可能更容易解读，信息量也更大。当然存在不同的情况，有一些长的评论也可能是表达冗余，比如反复诉说，用三个连续的句号表示省略号等。但是无论如何，相对较长的评论为分析和甄别其价值带来了更多的便利和可能。

表 2 对不同话题中评论字数统计

所属话题	最长评论字数	最短评论字数	平均字数
话题A	268	2	26.17
话题B	134	5	34.44
话题C	169	3	46.57

3.3 评论主题揭示程度和实体论及程度

如果运用更多的主题术语，则理论上该评论的主题揭示程度越高。尽管 3 个话题都是科技创新相关的，但是分别来自航空航天和微电子技术两个领域，利用标引工具能够一定程度上揭示每条评论主题的揭示情况。理论上对不同领域进行主题标引需要采用不同的词表，并进行训练，工作量较大。本文首先计划利用通用的《汉语主题词表》服务系统自动标引工具[②]进行标引。该工

② https://ct.istic.ac.cn/site/term/automaticIndexing

具能够覆盖不同的工程科技领域，也能够给出若干可能的学科分类，两个典型的标引结果如表 3 所示。但是在受限于评论内容为短文本或书写不规范情况下无法给出结果，如评论内容中“rr”代表罗尔斯·罗伊斯公司，通常为 RR，且其为实体，如果无知识图谱关联，很难得出其航空相关的标注。也有一些标引结果解释性不强，比如在讨论飞机的评论中，标引“高铁”是否合适值得商榷。此外评论中经常喜欢举例子，比如在话题 C 讨论逆向工程的评论中也谈及发动机，中国结，打乒乓球等，这种情况下可能这些非微电子技术的主题词就不需要标注。

表 3 《汉语主题词表》服务系统自动标引工具部分结果示例

评论文本内容	主题标引结果				学科分类标引结果
	序号	主题词	文中词汇	相关度	
当然可以了//*但是那样必须强制国内支线航线//使用自己知识产权的发动机的飞机//对此，我是全心全意支持的//反正我出行就走高铁，不坐飞机	1	发动机	发动机	0.75	U491.11交通调查 V271飞机 V32航空飞行 G30科学研究理论 U238高速铁路
	2	航线	航线	0.75	
	3	高速铁路	高铁	0.75	
	4	飞机	飞机	0.75	
	5	知识产权	知识产权	0.75	
应该本来就有rr的设计吧	无标引结果				无标引结果

* 原评论中的换行用 // 代替，下同

考虑到上述问题，在研究中采用人工方式标注。除了主题词之外，因为实体论及程度对于评论同样具有重要意义，如果不论及任何实体，则评论不具有针对性，其真实性也要打折扣，于是在评论中用到更多的实体，包括使用精确的数量，则严谨性会提高，因此将数量纳入到实体之中一并标注。

实体标注中需要克服一定的困难，除了上文论及的“rr”“ge”这类英文简写，中文实体名称同样复杂，关于美国总统“特朗普”有多种说法，“川普”是其中一种，还有很多网友使用调侃式的说法，此外由于评论很不严谨，也有评论者将连词“不但”写成了“不丹”，容易误识别为国家，在后续定制实体识别工具中这些问题都需要注意。

直接统计主题词和术语的数量有助于识别有价值的评论，但是由于评论本身的长度不同，仅仅识别其数量不够科学，需要识别其相对比值。抽取出一些比值比较高的例子如表 4，可以看到尽管整体还是比较短，但是同样是提供了一些高质量评论候选，如果希望将相对较长的评论提取出来，可以考虑结合数量与比例的指标。

表 4 高主题和实体占比的评论示例

评论内容	主题词和实体词累计占全部内容的比例
rr好像是跟航发有合作929的发动机。	0.50
一颗卫星管9万平方公里	0.91
不只是机器，芯片的很多化工原材料也得进口	0.58

3.4 评论的情感色彩

由于论坛中的评论具有不同的功能，本研究假设具有强烈感情色彩的评论主要用于情感的宣泄，其讨论问题的客观程度则有所降低，因此可以筛选出具有相对中性的情感色彩的评

doi:10.3772/j.issn.2095-915x.2020.02.003

论作为高价值评论的候选。本研究中没有开发新的工具，而是采用了基于开放的工具包调用的方式，主要包括 Jiagu[③]，SnowNLP[④]，Baidu Senta[⑤]，Baidu API[⑥]等。由于文本较短，一定程度上影响了情感分析的效果，经过人工比对，发现 Baidu Senta 的结果相对更接近人工感觉，但是仍然有分析不准确的地方，比如对“美国人哪有那么傻”的得分为0.0194，为极端消极（正负分界线为 0.5），“好好去研究一下 17 年的事情”得分为 0.9612，情感为极端积极。

本研究从中选择情感色彩较弱的，即结果取值在 0.5 ± 0.1 范围内的评论，如表 5 所示，发现具有一定参考价值，如果需要甄别还需要用到其他的手段。

表 5 情感色彩较弱的评论样例

所属话题	ID	情感值（Positive）	文本内容
话题A	0	0.4949	华尔街日报
	15	0.5049	我们不用加入，逐步脱钩，只维护不新购
	22	0.5706	老谣言要自己搜答案
	24	0.5244	互认了
	36	0.435	那中国的标准是按照FAA指定的，是不是就是互认了。
	41	0.5462	咱们的发动机可以用了？
	44	0.5183	哦，我以为你说中国军队呢……
	61	0.4469	主要是为了737max复飞
	67	0.4698	有道理
	84	0.553	操作系统芯片已经成为大家笑柄了
话题B	2	0.5497	第一批星链打的也是较高轨道
	3	0.404	260到280公里，很低了
	5	0.5573	会降低吧，目前的稳定高度是在280
	21	0.5334	差不多吧　应该能覆盖这么大的区域
	30	0.4778	海事卫星和铱星已经能覆盖全球，最大问题是费用
	37	0.4881	每辆tesla都是一个基站
	55	0.4846	现在的终端很小了，看起来就是个老式手机
话题C	15	0.4591	不说别的，就光路的那些镜片都搞不定，全世界就蔡司能搞，asml也得等着蔡司慢慢搞

3.5 基于评论引用网络的分析

由于本文采用的评论集中可以体现一些引用关系，因此构造引用网络，希望通过引用网络甄别出有价值的评论。

在一条具体评论中往往会体现本条评论针对的对象，也即另一条评论。基于此可以构建评论引用网络，在本文的构建过程中，由于可能涉及在一个评论的引用评论中又引用了其他

③ https://github.com/ownthink/Jiagu
④ https://github.com/isnowfy/snownlp
⑤ https://github.com/baidu/senta
⑥ https://github.com/baidu/senta

评论，即评论引用的嵌套，引用关系自动确定的规则可能比较复杂，因此暂时用人工方式确认。除了初始评论外，每条评论均会形成一条数据，数据模式如下：“SID,TID”，其中 SID 和 TID 都是评论的编号，这样的一条数据代表编号 SID 的评论引用了编号为 TID 的评论。但由于部分用户发表评论的时候会删掉引用信息，因此无法判断其 TID，针对这种情况，统一设置其 TID 为 0，表示其引用了初始评论。为了更好的展示引用网络，利用 graphviz2.3.8⑦软件和 pygraphviz1.5 包，构造引用网络图见图 1 ~ 3 所示，图中节点的大小代表在网络中的重要程度，节点越大，和其它节点的关联越多。话题 A 和话题 C 的模式类似，初始评论话题获得大量的关注，话题 B 则明显不同，存在数个节点与初始评论节点相差不大，主要在于针对初始评论的评论相对较少。一般而言初始评论的信息量都较大，本身需要关注，除此之外，还应该关注其他在网络中处于较重要位置的节点，如话题 A 中的 2 号、3 号和 10 号评论，话题 B 中的 37 号评论和 41 号评论，话题 C 中的 27 号，29 号和 32 号评论等。

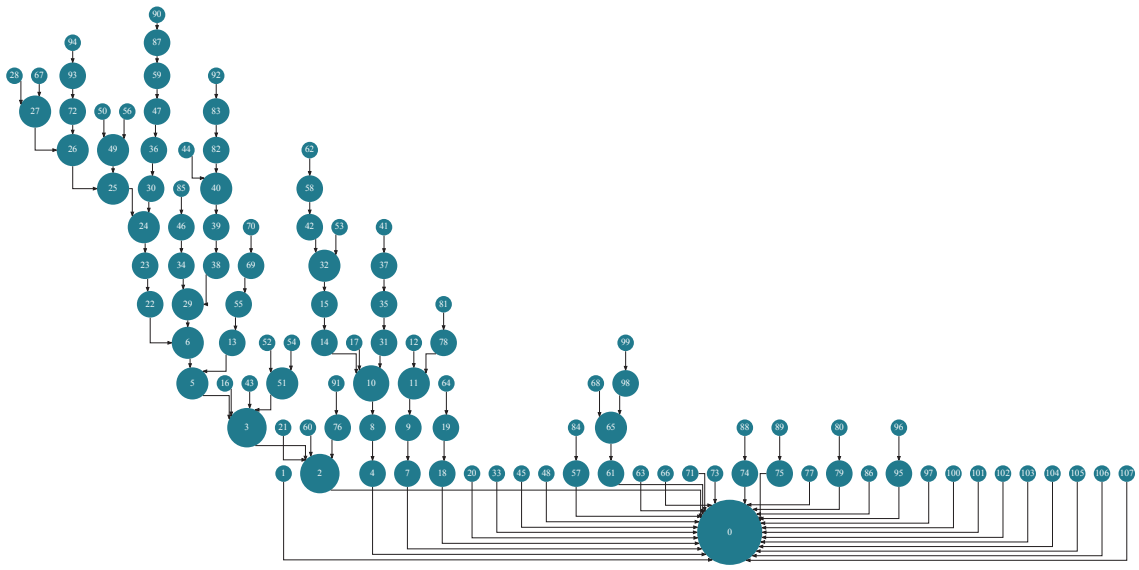

图 1 话题 A 中的评论引用关系图

但是仅仅依靠这种关系还是不够的，还需要看评论本身的质量以及引用评论对被引用评论的意见，因此需要对这两个指标进行度量。

首先对一条评论 c_i 本身的质量进行度量，主要度量其是否具有观点，设计度量指标 V_V（Value of Viewpoint），出于简单，将其设为 3 级，具体定义如式（1）所示。一些典型示例如表 6 所示。

$$V_V(c_i)\begin{cases}0:\text{无观点或者观点与主题无关}\\1:\text{有与主题相关观点但语气不肯定}\\2:\text{有与主题相关观点并语气肯定}\end{cases} \quad (1)$$

⑦ http://www.graphviz.org/

doi:10.3772/j.issn.2095-915x.2020.02.003

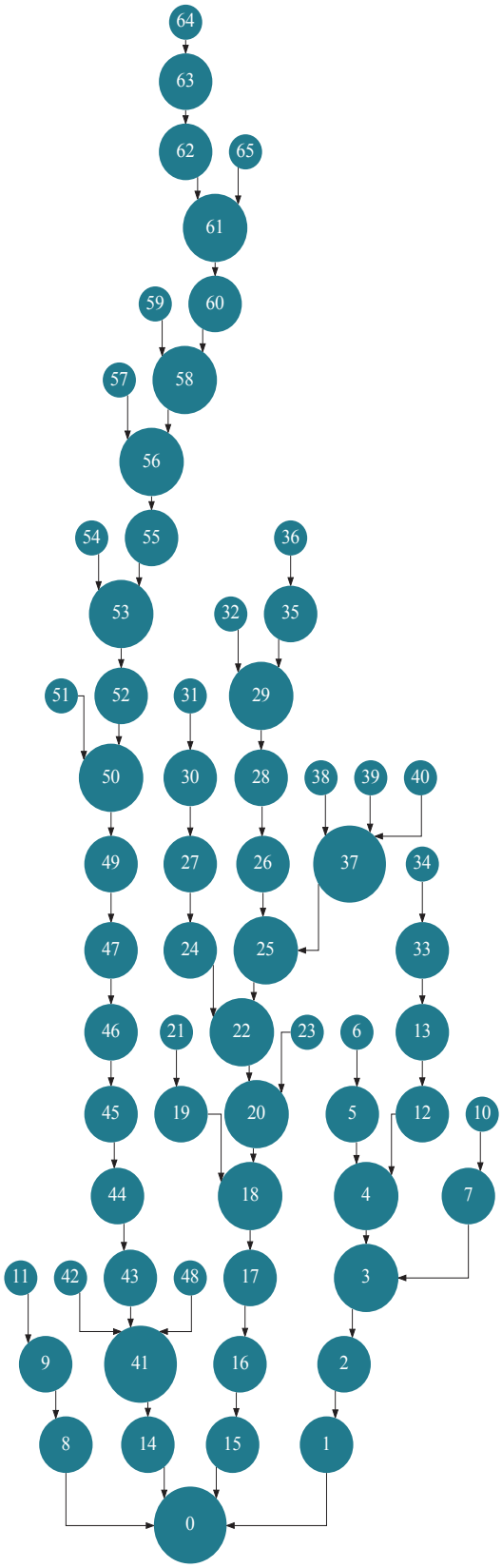

图 2 话题 B 中的评论引用关系图

关于对引用网络的评价，设计了引用度指标 D_R（Degree of Refer），可能包含三种情况，支持，反对，中立（或者无关），对于支持和反对分别给出两个等级，则共有从 -2 到 2 共 5 个整数取值，定义如式（2）所示，一些典型示例如表 7 所示。

$$D_R(c_{ij})\begin{cases}2:\text{评论}c_j\text{与评论}c_i\text{相关且对}c_i\text{较强支持}\\1:\text{评论}c_j\text{与评论}c_i\text{相关且对}c_i\text{一般支持}\\0:\text{评论}c_j\text{与评论}c_i\text{无关或对}c_i\text{中立}\\-1:\text{评论}c_j\text{与评论}c_i\text{相关且对}c_i\text{一般反对}\\-2:\text{评论}c_j\text{与评论}c_i\text{相关且对}c_i\text{较强反对}\end{cases} \quad (2)$$

这里需要说明的是，对于支持度判断有一些不能直接从本评论中直接判别，还需要联合被引用评论来分析。

因此首先选择较高引用的评论 c_i，判断本身是否有观点，对于 V_V 为 1 和 2 的情况，则进一步分析其引用评论 c_{ij} 对该评论支持程度 D_R，综合计算评论的价值 V_C（Value of Comment），如式（3）所示。一些具有较高 V_C 绝对值的评论示例如表 8 所示。

$$V_C(c_i)=V_V(c_i)\times\sum_{j=0}^{n}D_R(c_{ij}) \quad (3)$$

表 8 中具有较高 V_C 绝对值的评论发帖可能具有相对较高的参考价值，取值为正的具有正向参考价值，取值为负的具有反向的参考价值。这些值的高低是和具体的评论集合有关，集合间差异较大，另外引用该评论的评论更多，则结果参考价值更大。反之可能由于引用者较少，少量引用者的态度就能够改变 V_C 值，从而影响其参考价值。实际上对评论本身的 V_V 的判别以及 D_R 的自动计算都是需要进一步研究。在自动实现中，短文本以及话题转换，前后评论之间话题的衔接都是需要重点关注。

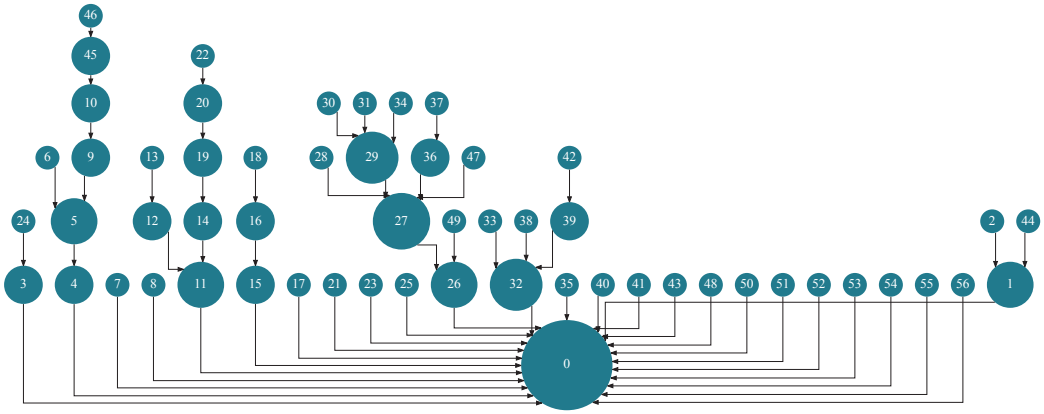

图 **3** 话题 **C** 中的评论引用关系图

表 **6** 具有不同观点度量值的评论示例

观点度量值（V_V）	评论示例（c_i）内容
0	互认了
0	有道理
0	lz傻……
0	【美国怕中国逆向研发 或禁止LEAP-1C发动机出口】 据WSJ报道，美国政府正在考虑取消CFM向中国出口LEAP-1C发动机的许可，LEAP-1C是CFM专门为中国国产大型客机C919生产的发动机型号。白宫方面……
1	也就是FAA的认证是没戏了?
1	应该本来就有rr的设计吧
1	260到280公里，很低了
1	要过要扼杀一开始不许可好了，即使现在断供，也不是只有这一家可以选，最多又是自己制造谈判筹码而已
2	做不到工艺
2	逆向是借口，真实意图是扼杀中国客机产业
2	原理上眼一看就能明白吧。材料参数，软件机理分析了才能知道。
2	通话没问题，现在的高轨卫星电话就普通手机大小，加个比较粗的天线，海事宽带的加个ipad大小的面阵天线，速率1M……

* 考虑到篇幅和对一些不文明用语的过滤，部分内容用省略号代替，下同。

表 **7** 具有不同引用度的评论示例

引用度（D_R）	引用评论	被引用评论
-2	拍脑袋	【对极紫外光刻机，不可能逆向吗】大不了我们多买十台，把它彻底扒一扒
-2	所谓的弯道超车行不通	
-1	不是说买不到吗?	
0	你还别说，很多领导就是你这种想法	
-1	要过要扼杀一开始不许可好了，即使现在断供，也不是只有这一家可以选，最多又是自己制造谈判筹码而已	逆向是借口，真实意图是扼杀中国客机产业
1	确实，借口而已	
2	必须的	
0	为了保护波音公司而已	

doi:10.3772/j.issn.2095-915x.2020.02.003

表 8 具有相对较高 V_C 绝对值的评论示例

所属话题	评论 ID	评论内容	V_C
话题 A	3	逆向是借口，真实意图是扼杀中国客机产业	4
话题 A	27	最麻烦的是造发动机的美国企业，不卖给中国，可以多卖给波音 737MAX 啊 // 反而是中国背景的企业有福了，美国企业卖不掉产品，必然会想办法把技术折价卖给有路子卖给中国的第三方	4
话题 A	25	对 // 从一开始中国的目标就不是认证某款机型，而是获得和 FAA 等价的认证能力。这个能力是科学，和政治权力是截然不同的概念。	-2
话题 B	37	每辆 tesla 都是一个基站	10
话题 B	0	【星链目前来看就是个铱星系统】但是终端还没发售 // 铱星每分钟网内通话费 1 刀，卫星重点数量少点	-10
话题 B	41	只要产业化能成规模，这些都不是问题 //　现在马斯克这个，搞好是超级创新。。。搞不好就是铱星第二。// 不过靠卫星这个思路，倒是非常适合战争状态下。。。有点像当初互联网的产生了。// 天朝要高度戒备，小心。。。先模仿跟随之。。。	-6
话题 B	25	问题是卫星电话密度才多大使用频率才多高？　手机几乎人手一部，随便一个基站的电费都不是一个小数目，一个哪怕是 1 吨的卫星带那点小破太阳能电池哪来那么高功率电能支撑这样用？	-6
话题 B	56	电池够用吗？还是经常放到底座上充电备用？总觉得手机大小的天线发射信号给卫星不靠谱，信噪比太低了。	-6
话题 B	61	就通信、上网来说，目前的高轨海事之类的已经满足偏远地区、海上需求，真不明白建个低轨通信星座的意义何在，几百上万颗卫星，技术也不成熟，有点吃饱了撑的意思	-6
话题 C	0	【对极紫外光刻机，不可能逆向吗】大不了我们多买十台，把它彻底扒一扒	-68
话题 C	5	原理上眼一看就能明白吧。材料参数，软件机理分析了才能知道。	-6
话题 C	11	有可能可以加速研究的过程，但是不可能马上实现。……而光刻机难度个比这个难得太多太多。　而这个缩短的时间有可能是 40 年减少到 30 年？或者其他不短的尺度。……	6

4 基于外部数据的情报甄别

4.1 评论者基本属性

通过对 3 个话题的用户身份进行分析，得到统计如表 9 所示，其中大部分为（普通）用户，少部分为版主和核心驻版。通过人工阅读分析，未发现版主和核心驻版的评论更具意义，当然这可能与版主与部分核心驻版并非服务本版，因而在本版的言论相对随意有关，即便是本版的核心驻版发文质量也不一样，未见与普通用户显著区别。

尽管在系统中可以查询性别信息或者依据级别命名的不同区分男女，但是由于系统本身不具有审核功能，因此无法确定性别信息是否准确，也无法据此判断。依据评论者等级名称的性别区分，得到评论者性别信息如表 10 所示，从这些信息判断，女性比例较低，并且与该论坛前期采集另一板块的话题相比更低，可能与女性参与工程科技有关的话题的积极性较低有关。但是评论质量与性别并无显著关联。

表 9 评论者身份属性信息统计

所属话题	版主	核心驻版	用户	全部
话题A	2（非本版）	2（本版）	53	57
话题B	0	3（其中2为本版）	16	19
话题C	1（非本版）	0	38	39

表 10 评论者性别属性信息统计

所属话题	男	女	全部
话题A	55	2	57
话题B	18	1	19
话题C	38	1	39

此外，本文也对发文数、积分以及等级进行了统计，经过对 3 个话题个案考察和分析，对评论价值的影响几乎可以忽略。

表 11 评论者发文数、积分、等级等最高和最低信息统计

所属话题	最多发文数	最少发文数	积分最高	积分最低	等级最高	等级最低
话题A	179657	16	68087	1777	15	4
话题B	109702	529	64826	489	15	4
话题C	121559	72	62447	3342	15	3

4.2 评论者引用网络

参考 3.5 节，依据对全部评论者设置对应的 ID，根据帖子的情况，构造评论者引用网络，3 个话题的作者引用网络，如图 4 ~ 6 所示，其中节点的大小代表用户的互动情况，边上的方向代表用户之间相互引用评论的方向，有箭

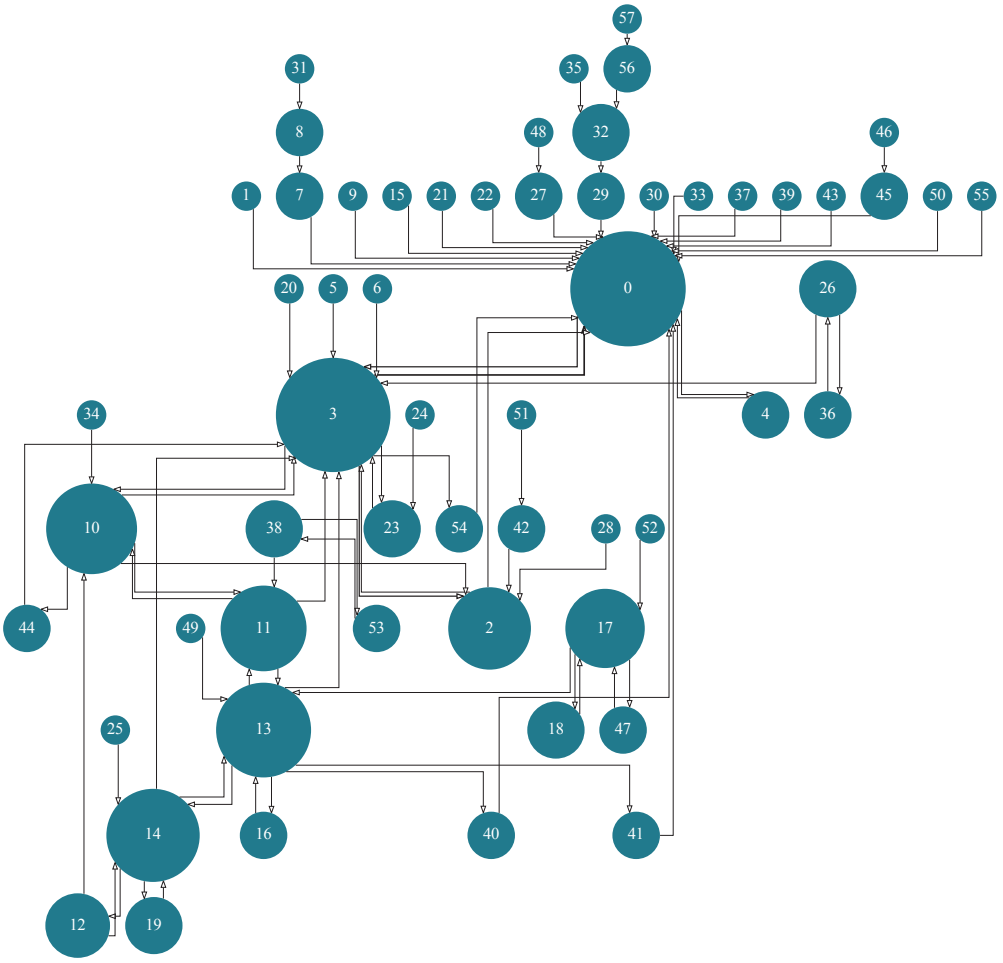

图 4 话题 A 中的评论者引用关系图

doi:10.3772/j.issn.2095-915x.2020.02.003

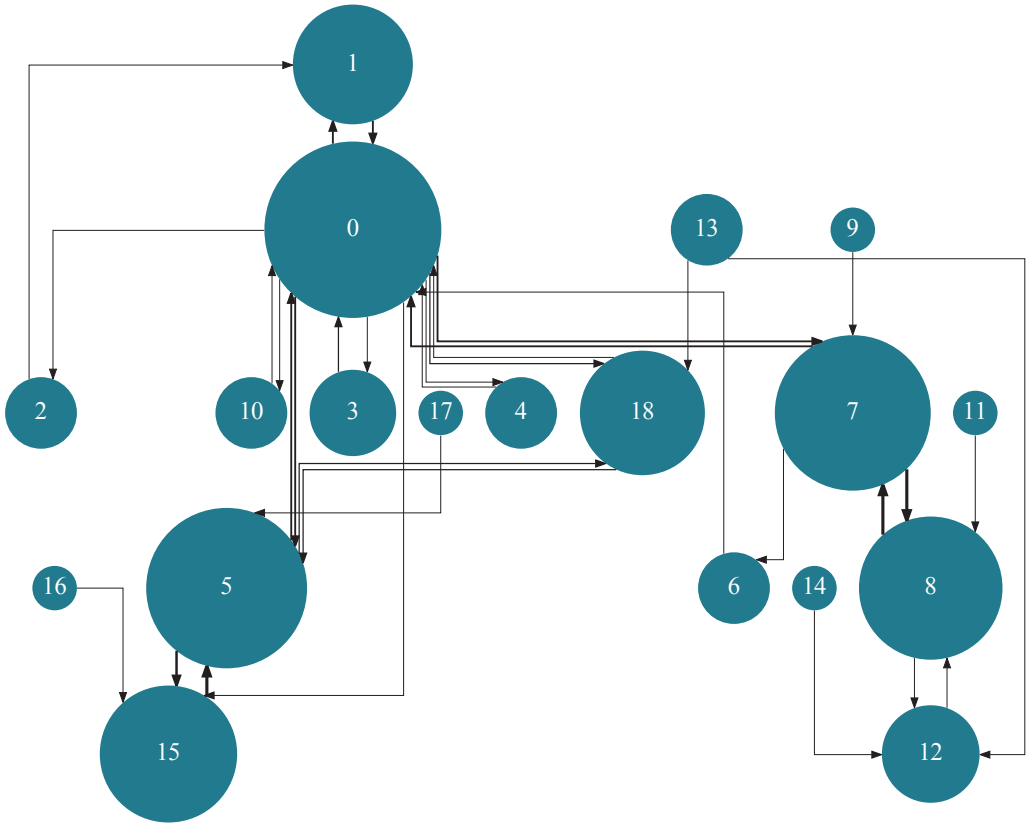

图 **5** 话题 **B** 中的评论者引用关系图

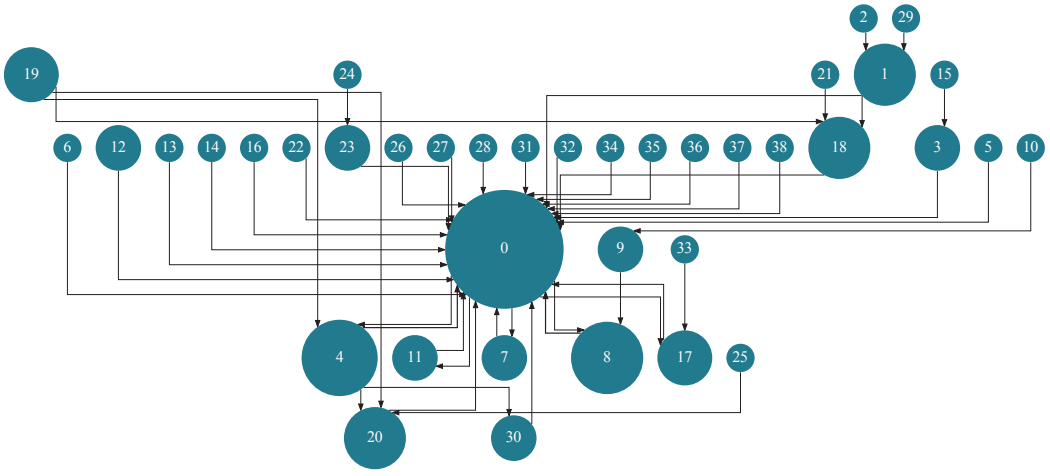

图 **6** 话题 **C** 中的评论者引用关系图

头的节点表示被其他用户引用评论的用户，没有箭头的节点表示引用其他用户评论的用户，边的粗细代表互动的数量，越粗则互动越多。由图 4 ~ 6 可见，话题 A 中用户 ID 为 0,13,14,10 的互动较多，话题 B 中用户 ID 为 0,7,8,5 的互动较多，话题 C 中用户 ID 为 0,4,8,18 的互动较多，这些用户更值得关注。尽管构造了评论者引用网络，但是很多评论者往往发表超过 1 条

评论，而不同评论的价值可能不同，很难做统一的判断，因此基于评论者引用网络只能作为进一步甄别的候选，相比评论引用网络价值要低一些。当然，如果能够长期关注某些用户，获得关于这些用户的更为丰富的刻画，则有可能增加评论者本身的权重。

4.3 评论者行为

评论者可能有多种值得关注的行为，但是从目前抓取的数据出发，只能获得有限的信息，其中有代表性的是评论修改行为。该行为后面蕴含着作者的意图，可能是作者严谨性的代表，也可能是为了掩盖无意中透漏的信息，但无论是哪种都是值得关注的点。通过分析，3 个话题中共有 6 条存在修改行为的评论，如表 12 所示，通过阅读，发现这些里面都是蕴含一定的信息，至于是否是有价值的情报，需要结合情报需求进行进一步甄别。

表 **12** 评论者修改过的评论内容示例

所属话题	评论编号	作者编号	评论内容
话题A	40	17	这样不靠谱肯定受不了，//加税才才几个钱，//这下好了，没人敢买波音，//客机战斗机都完蛋，//在不打仗，估计只有以后等着挨揍。
话题A	46	34	919飞控实现外包给美国公司了。//而且不负责DEBUG。//也就是阿三随便写，有没BUG不管。//中国负责（黑盒？）测试，测不出来活该。
话题B	53	5	海事卫星电话是安装到船上的吧，如果自带电池和天线的话不确定是不是和2战的时候通信兵的背包差不多
话题B	55	0	现在的终端很小了，看起来就是个老式手机
话题C	6	4	看不明白，东西太多了，你做过反向就知道了。//还有一个问题是特种工艺，这需要摸索很久才能搞出来。
话题C	42	19	武直整的还是惟妙惟肖的，//就是发动机没戏，不然彻底嘚瑟起来了

5 基于内外部数据一致性的情报甄别

地理信息是非常重要的信息，可以通过评论的内外部数据进行甄别，论坛评论时会附加 IP 地址，由于涉及国际科技竞争，一般情报人员对本国情况比较了解，通常更多关注非中国大陆境内的 IP 地址，来自这些地址的信息可能会带来更重要的情报。本文中通过对 IP 归属地查询，得到相关信息，目前可以利用的是实时信息，历史信息目前无法获得。

内外部数据的一致值得关注，如某位评论人从美国发表评论，论及美国相关事务，则有可能可信度高一些。一些典型示例如表 13 所示，这些评论可能具有较高的潜在价值。但是这些内外部数据蕴含的信息不一定能直接匹配，而需要利用知识图谱或者其他类型的知识组织体系进行连接，如 737MAX 属于波音，波音是美国公司，Tesla 车是美国特斯拉公司的产品，ARM 是英国公司。但是评论中并没有直接说美国，英国，这是利用地理信息的一个难点，干扰分析处理 IP 地址和地理信息可能的因素在于这些信息可能会变动，比如某位讨论美国事物的评论者，可能长期定居美国，恰好发表评论时在其他国家，或者相反。此外，也可能有各种因素促使评论者使用 VPN，导致实时的 IP 信息不够准确。

doi:10.3772/j.issn.2095-915x.2020.02.003

表 13 内外部地理信息一致性情报甄别示例

所属话题	评论编号	评论内容	评论者IP	IP所属国家
话题A	61	主要是为了737MAX复飞	73.225.94.*	美国
话题B	39	每辆tesla都是一个基站	93.234.243.*	美国
话题B	64	海事卫星一分钟几块钱，星链在粉丝眼里是不要钱的	155.64.38.*	美国
话题C	29	没人愿意踏实去做，主要还是不挣钱，人才都是向赚钱的领域流动。要是像互联网那样，肯定一堆牛人冲进去做起来了。现在高智商人才扎堆互联网，所以互联网行业是世界领先水平，AI也是世界前二。//其实国外一样，亚马逊做服务器芯片，就能从ARM挖很多人，ARM从互联网行业挖人根本不可能。	85.255.236.*	英国

6 结论及下一步工作

本文在科技相关突发事件的评论集上，分别从内部评论内容、外部评论信息以及内外部数据一致性的角度，对 3 个典型话题进行个案分析。经过探索研究，发现评论原创性、评论的长度、偏中性情感、评论主题揭示程度和实体论及程度、方向趋向一致的评论引用、修改行为、基于地理等实体信息的内外部信息一致性构成了有价值的评论的因素；评论者在论坛中的角色、性别、发文数、级别、积分等因素对情报价值不具有明显的影响；评论者引用网络目前还不能充分发挥情报甄别的作用，但是大量信息的收集有可能带来潜在的应用；研究中还发现评论的效用可能随需求不同而发生改变，如附和型或者批判型评论本身可能价值不大，但是对寻找有价值的评论有参考价值。当然以上研究针对的评论数据集体量较小，有一些观察和结论可能是片面或错误的，需要在更大数据集上验证。

在评论类情报甄别过程中复杂网络分析、情感分析、领域主题词标引和实体抽取是重要的技术工具。而且由于论坛中评论往往具备如下特点：（1）篇幅长短不一，大部分评论篇幅比较短，从形式和排版上不太规范；（2）在用语上习惯使用网络用语，简写，代号等，甚至存在一些不文明用语；（3）在发表评论的过程中，可能存在偏离主题的情况，可能针对某一中间评论展开讨论。这些特点为情报甄别分析带来了一定的困难，使用到的技术有必要针对上述特点做改进。知识图谱能够提供常识类知识，在驱动情报甄别方面具有重要的基础性作用，尤其是短文本的情报内在信息不足，知识图谱可以提供外部知识加以补足；同时知识图谱也能够通过适当的推理，改善情报甄别的效果。此外，话题 A 的第一个评论大部分内容是一张引用的图片，而这也并非孤例，在其他领域的评论中也发现了类似的情况且有增加的趋势，因此有必要对图片解析和处理技术做进一步研究。综上，面向评论特点的技术方法、知识图谱构建和运用、情报甄别的多模态扩展都是有价值的研究方向，值得继续探索。

参考文献

[1] 中华人民共和国突发事件应对法 [EB /OL]. (2007-08-30)[2020-01-03]. http: //www.gov.cn/ziliao/flfg/2007-08/30/content_732593.htm.

[2] 曹杰 , 杨晓光 , 汪寿阳 . 突发公共事件应急管理研

究中的重要科学问题 [J]. 公共管理学报 , 2007(2): 84-93, 126-127.

[3] 李纲 , 李阳 . 智慧城市应急决策情报体系构建研究 [J]. 中国图书馆学报 , 2016(3): 39-54.

[4] 刘振 . 融合 CRF 和 SRL 的科技事件抽取研究 [D]. 北京 : 中国科学院大学 , 2017.

[5] 毛凯 , 刘明 , 李志恺 , 等 . 基于 K-means 算法的科技事件影响力评估研究 [J]. 无线互联科技 , 2019, 16(7): 115-118.

[6] 杨艳蓉 . 科技事件的个性化推荐及可视化展示系统设计与实现 [D]. 武汉 : 华中师范大学 , 2019.

[7] Cooke A. A Guide to Finding Quality Information on the Internet: Selection and Evaluation Strategies [M]. London: Library Association Publishing, 2001.

[8] 刘敬学 , 费奇 . 基于指挥决策的情报筛选系统研究 [J]. 情报杂志 , 2005, 24(4): 22-24.

[9] 曾鸿 . 竞争情报与信息甄别 [J]. 图书馆理论与实践 , 2006(4): 40-42.

[10] 徐玉萍 , 刘瑞华 . 竞争情报整理中信息筛选的指标体系研究 [J] . 图书馆学研究 , 2009(1): 60-62.

[11] 张均胜 . 面向情报工程的可信溯源研究 [R]. 中国科学技术信息研究所 , 2020.

[12] Rastogi A, Mehrotra M. Opinion Spam Detection in Online Reviews[J]. Journal of Information & Knowledge Management, 2017, 16(04): 1750036.

[13] Shehnepoor S, Salehi M, Farahbakhsh R , et al. NetSpam: A Network-Based Spam Detection Framework for Reviews in Online Social Media[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(7): 1585-1595.

[14] 张璐 . 虚假商品评论识别的研究与进展 [J]. 计算机工程 , 2019, 45(10): 293-300.

[15] 邓胜利 , 汪奋奋 . 互联网治理视角下网络虚假评论信息识别的研究进展 [J]. 信息资源管理学报 , 2019(3): 73-81.

[16] 吴佳芬 , 马费成 . 产品虚假评论文本识别方法研究述评 [J]. 数据分析与知识发现 , 2019, 3(9): 1-15.

[17] 罗衍 , 王春峰 , 房振明 . 社会互动 , 投资者情绪传染与资产泡沫 —— 基于股票论坛发帖的实证研究 [J]. 运筹与管理 , 2018, 27(2): 124-132.

[18] 程葳 , 钟华 , 孙娇华 . 网络论坛中发帖行为复杂性研究 [J]. 系统工程学报 , 2009, 24(0): 385-391.

[19] 王家箴 , 何美会 , 李美暄 , 等 . 网络论坛用户的行为分析研究 —— 以股吧为例 [J]. 网络安全技术与应用 , 2019(9): 92-94.

[20] 陈卫华 , 徐国祥 . 基于深度学习和股票论坛数据的股市波动率预测精度研究 [J]. 管理世界 , 2018, 34(1): 180-181.

[21] 易洪波 , 李梦璐 , 董大勇 . 投资者情绪与成交量 : 基于网络论坛证据的分析 [J]. 商业研究 , 2016(8): 58-64.

[22] 赖娟娟 . 论坛异常发帖量对沪深 300 股指期货冲击成本的影响研究 [D]. 重庆 : 西南交通大学 , 2016.

[23] 刘三 , 韩雪 , 柴唤友 , 欧阳柏强 . SPOCs 论坛中学习者的交互模式研究 —— 基于回复网络和引用网络的比较 [J]. 中国电化教育 , 2019(11): 73-79.

[24] 李培 . 基于复杂网络理论的讨论区发帖分类推荐算法 [J]. 计算机系统应用 , 2014, 23(7): 180-184.