

面向活动的网络媒体监测与建模分析：IVFC 案例解析

孙星恺^{1,2}, 王晓^{1,3}, 陆浩¹

(1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室, 北京 100190;
2. 中国科学院大学, 北京 100049; 3. 青岛智能产业技术研究院, 山东 青岛 266000)

摘要: 当前, 各类活动(如竞赛、会议等)的举办频次日益增多, 但针对此类活动相关网络媒体内容的数字化监测和分析却缺乏有效的平台。以平行情报中的方法框架与知识自动化理论方法为基础, 提出了一个面向活动的网络媒体监测与建模分析的方法框架, 并以中国智能车未来挑战赛(IVFC)为实际应用案例, 采集 2009 年至 2017 年 3 类主要网络媒体(新闻、微信、微博)数据, 分别从时空、热点、发布源、关键词、主题、语义、实体等维度对网络媒体数据进行解析与可视化。实例分析结果可以有效揭示活动网络媒体的发布规律、关注点、重要发布源以及各活动参与方实体的知识等, 可为活动主办方等相关单位进行策划、宣传、总结等活动前、中、后的各个环节提供数字化辅助决策支撑。

关键词: 网络媒体监测; 多维分析; 中国智能车未来挑战赛; 知识自动化

中图分类号: TP391.1

文献标识码: A

doi: 10.11959/j.issn.2096-6652.201940

Digital monitoring and modeling of activities: the IVFC case study

SUN Xingkai^{1,2}, WANG Xiao^{1,3}, LU Hao¹

1. The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Qingdao Academy of Intelligent Industries, Qingdao 266000, China

Abstract: At present, the frequency of various activities, such as competitions and conferences is increasing day by day. However, there is a lack of an effective platform of digital monitoring and analysis of online media for such events. Based on the method framework in parallel intelligence and knowledge automation theory, a methodological framework for network media monitoring and modeling analysis for events was proposed in this paper, and takes the “Intelligent Vehicle Future Challenge” as an application example. The dataset was collected from three main types of online media (news, WeChat, Weibo) data from 2009 to 2017. Then the online media data were analyzed and visualized from the multiple dimensions including time and space, hotspots, release sources, keywords, topics, semantics, and entities. The results show that the relevant framework can provide effective digital monitoring means and auxiliary decision support for the activity-related organizations.

Key words: digital monitoring, multi-dimensional analysis, Intelligent Vehicle Future Challenge, knowledge automation

1 引言

当前, 随着社会的快速发展与科技的日新月异, 各类活动(如赛事、会议等)的举办频次日益增多, 但针对此类活动的网络媒体数字化监测与分析方法与平台在研究与应用层面较为匮乏。传统的

舆情监测平台也仅停留在浅层次的数据采集与分析层面, 并不能为活动提供针对性的、多维度的数据监测与建模分析。因此, 构建针对社会活动赛事的数字化监测与分析服务平台, 对识别与活动相关的网络关注点与基于数据驱动的辅助决策具有十分重要的指导意义。

收稿日期: 2019-10-15; 修回日期: 2019-11-18

通信作者: 孙星恺, xingkai.sun@ia.ac.cn

社会网络与社会新媒体的快速兴起与普及使得网络新媒体作为一种新的信息传播形式深入人们的日常工作和生活当中。网络虚拟空间对传统社会管理带来了新的问题与挑战，随着大数据、物联网等技术的发展，网络与人类社会无缝联合形成的社会物理信息系统（cyber-physical-social system, CPSS）^[1]将成为未来实现智慧管理的基础。赛事活动作为一种社会化活动，必然会受到网络空间的影响，因此对其在网络空间中的数字化监测与反馈可以更好地指导活动的开展与流程优化。网络媒体大数据时代，海量网络媒体中蕴含着丰富的社会信号，社会化媒体大数据分析与应用研究成为热点。实时、可靠的社会信号是实施有效的社会管理，特别是反馈闭环式社会管理创新的基础^[2]。网络媒体大数据数字化监测与舆情分析的理论与方法基础是社会计算^[3-5]。2005年中科院自动化研究所的王飞跃就已经指出数字网络化时代对社会状态和趋势动态分析的必要性^[6-7]，并提出了基于ACP的社会信号处理与分析的基本框架^[2,8]。通过描述解析、预测解析与引导解析等核心环节，实现对特定社会系统的实时、反馈、闭环式的有效管理，也成为当前针对网络大数据解析的代表性方法框架^[9]。

目前，针对活动网络媒体数据监测与分析的直接研究和应用较少，相关研究主要分布在网络舆情、网络情报与文本挖掘等方向。网络舆情主要通过收集媒体和公众对已发生事件的报道、评论、议论等数据，反映社会的关注度和公众的情绪。由于网络舆情对社会管理、城市管理与企业管理等具有十分重要的意义，已然引起相关学术界与产业界的广泛关注和重视。网络情报的重点是面向不同应用主体和场景，广泛搜集各类相关网络数据，从中发现线索、监测动向等，从而进行更深层次的分析、挖掘与预测、预警。随着Twitter、微博等社交媒体平台的兴起，以及大数据、人工智能等技术的发展，相关技术与方法也逐渐被应用到舆情研究中。有学者对新形势下的网络舆情大数据分析思路^[10]、大数据环境下文本挖掘应用^[11]、舆情系统构建^[12]等方面进行了探讨。目前，大数据环境下对网络舆情分析方向的研究还处于起步阶段，与国外侧重于基础方法、建模分析、预测演化分析等不同，国内学者的研究侧重于舆情监测系统、话题监测、传播机制、理论探讨等方面^[13]。在应用层面，当前国内的舆情监测主要针对单一的突发事件、热门话题等，侧重

于数据的收集与浅层次的数值统计与情感分析。国内外已有的网络舆情系统以监测分析为主，缺乏对舆情事件的建模与仿真、风险评估、预警与引导机制等的系统化研究^[14]。

活动在本质上是一个事件，对活动网络数据的监测与分析既要借鉴网络舆情的基本方法，同时也要融入网络情报研究的方法，从更多维度对活动事件进行分析。近年来，在大数据驱动的智能情报时代背景下，相关学者提出了“情报5.0，即平行情报1.0”的代表性情报理论与方法体系^[9]，其基于虚与实互动、情报与智能合一的平行思想，为智能情报的核心环节/流程（情报解析、情报预测与情报引导）定义了一套可行的框架。此外，知识自动化^[15]作为实现从数据、信息、知识到智慧的转变，从大数据凝练成针对具体问题的“小规则、小知识”的核心技术，是从物理世界的自动化控制转向人类社会本身的智能化管理的基础^[16]。本文以网络大数据情报解析框架为基础，以知识自动化为方法指导，运用社会计算中的基本方法，融合文本挖掘等相关技术，提出了一个面向活动的网络媒体监测与建模分析的方法框架，并以中国智能车未来挑战赛为应用案例对相关网络媒体数据进行了采集与多维度可视化分析。相关结果可以有效揭示活动网络媒体的发布规律、关注点、重要发布源以及活动参与各方的实体知识等，可为活动主办方等相关单位更好地进行活动策划、宣传、评估总结等提供数字化辅助决策支撑。

2 活动网络媒体数据数字化监测与建模分析框架

活动网络媒体数据数字化监测与建模分析整体框架如图1所示。

（1）数据感知获取层

数据感知获取层针对活动的自定义配置，为数据采集设定媒体平台与数据源，采集关键词、时间等数据。系统实时定向采集与活动相关的各项媒体（新闻、微博、微信）数据，为活动数据的处理与解析提供基础数据，是活动进行离线与在线实时监测解析的基础。

（2）数据处理层

数据处理层对数据感知获取层采集的各类网络媒体数据进行预处理，包括数据清洗、数据变换、文本分词、命名实体识别、特征抽取与数据标注等，为数据挖掘解析与建模分析提供统一融合的特定

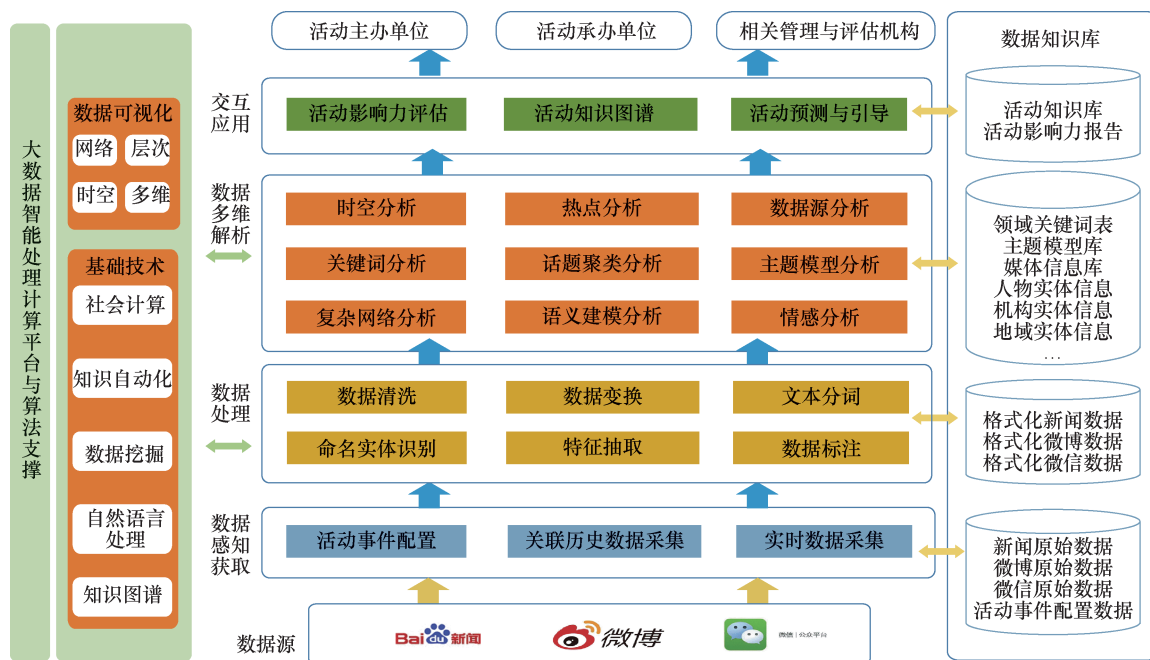


图 1 活动网络媒体数字化监测与建模分析整体框架

格式数据与细粒度数据属性。

(3) 数据多维解析层

数据多维解析层对活动相关网络媒体数据进行时空、热点、数据源、关键词、话题聚类、主题模型、语义建模等多个维度的分析。

(4) 交互应用层

交互应用层通过可视化的方式对各类分析结果进行展示，形成活动网络媒体知识库，为活动主办单位等相关机构开展活动网络媒体评估、宣传、策划、总结等提供辅助支撑。

3 数据集

本文以中国智能车未来挑战赛为例对新媒体在活动网络宣传中的作用进行分析与评估。2009年，中国智能车未来挑战赛由国家自然科学基金委员会创建，是中国在自动驾驶智能车领域最为重要的赛事。以“中国智能车未来挑战赛”与2009—2017中年份名组合作为关键词进行网络媒体数据采集，数据有效时间段为2009年1月1日至2017年12月31日。过滤掉噪声数据后，共采集有效新闻703篇、微信文章588篇、微博614篇。其中，新闻采集字段包括标题、正文、媒体与时间，微信文章采集字段包括“xt”设置的“Marked”标题、正文、公众号与时间，微博采集字段包括发布人、内容、时间、转发次数、评论次数。

4 活动网络媒体数据多维分析

4.1 时间维度分析

对3类媒体数据（不同颜色）的数据量按年份进行统计，历年IVFC网络媒体数据分布如图2所示。

从图2可以看出，IVFC的网络关注度呈现逐年增长的趋势，尤其是2017年，总数据量有明显提升，新闻、微信、微博的数据量均大幅提升。微信由腾讯于2011年推出，数据量从2014年开始逐年提升，2017年较2016年增长幅度高达269%。

对2017年第九届中国智能车未来挑战赛相关数据的发布时间按天进行统计，结果如图3所示。

2017年5月5日为赛事相关数据的首日，从11月9日开始，相关数据量开始逐步提升，并在11月24日至26日3天赛事期间达到最高点。从媒体上看，新闻与微信在赛前与赛中的数据分布趋势较为吻合，赛后新闻在12月6—7日的高峰主要为获奖机构的采访报道，之后的数据以微信公众号文章为主。

对3类媒体数据的发布时间按时刻进行统计，结果如图4所示。

在图4中，不同颜色的气泡代表不同的媒体，大小代表每个时刻的数据量。从图4可以看出网络媒体发布的两个高峰时段为9:00—10:00和16:00—17:00。由于新闻的时效性要求与微博的简短特点，新闻和微博的主要发布时间集中在9:00左

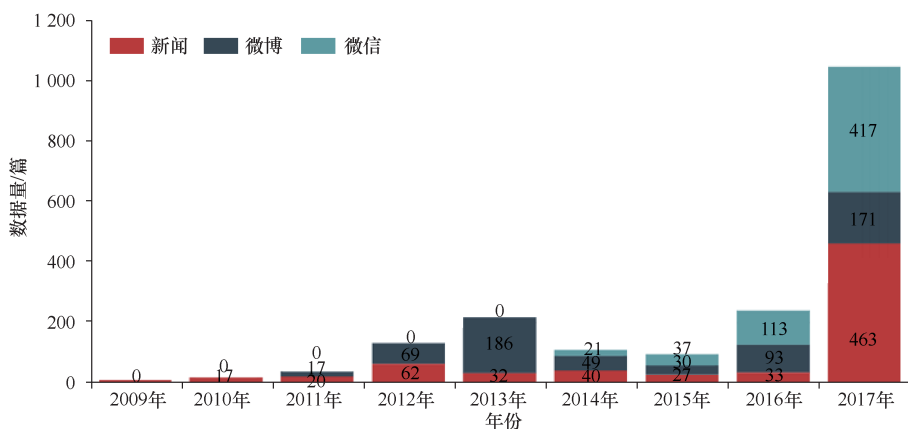


图2 历年IVFC网络媒体数据分布

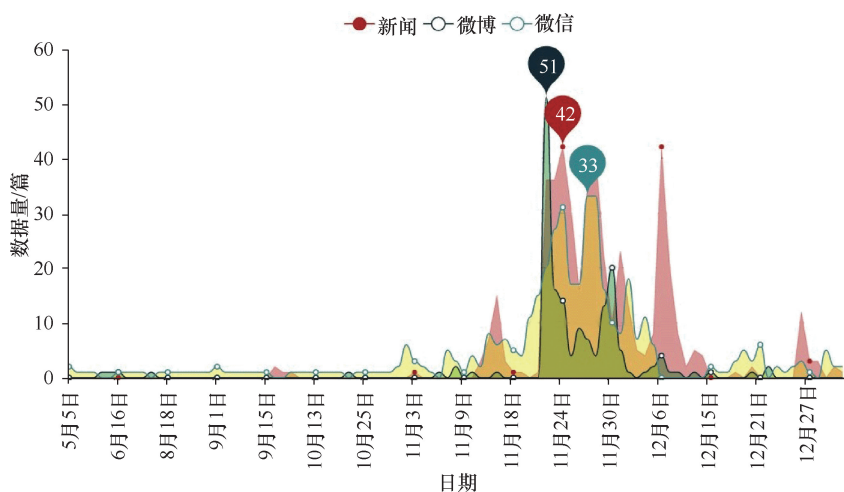


图3 2017年IVFC数据日分布

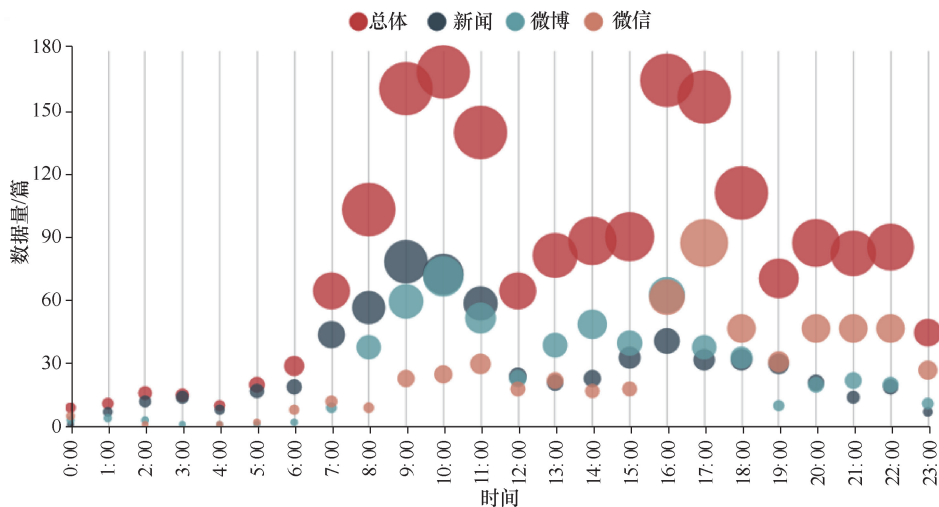


图4 IVFC网络媒体数据的发布时间点分布

右，内容以活动相关实时动态为主。微信公众号文章内容主要基于对赛事活动的阶段总结，一般集中在下午与晚间时段。

4.2 数据源维度分析

IVFC 数据相关媒体数及发文量如图 5 和图 6

所示。平均每个媒体历年发表新闻数约为 3 篇，每个微博用户发布微博数不到 2 条，每个微信公众号发布文章数约为 2 篇。2017 年微博用户平均发文量有所下降，平均每个用户只发布 1 篇，可见集中度较高。

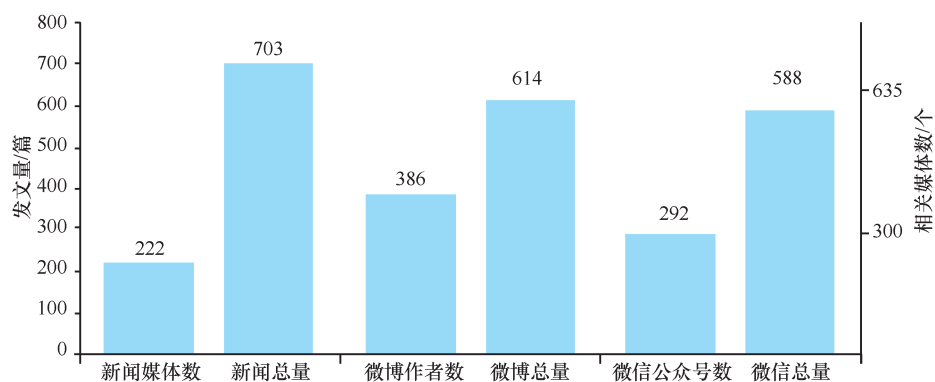


图 5 2009—2017 年 IVFC 媒体数及总发文数

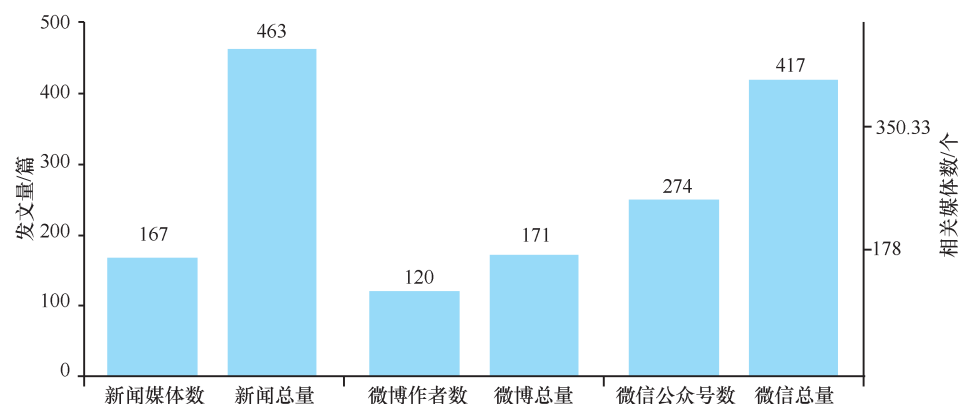


图 6 2017 年 IVFC 媒体数及总发文数

对新闻媒体/微信公众号/微博作者分别进行数量统计，相关结果如图 7~图 9 所示。

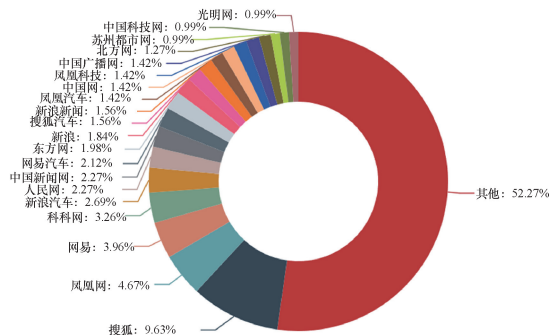


图 7 IVFC 新闻媒体分布

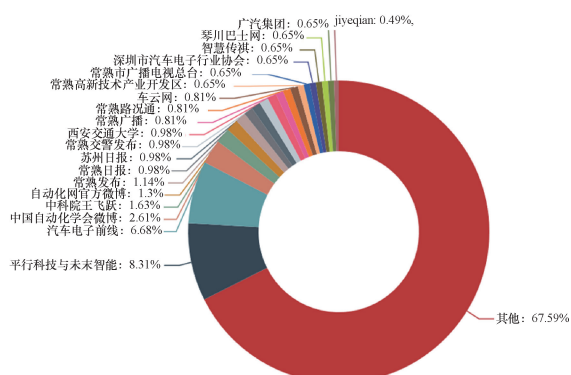


图 8 IVFC 微博作者数量分布

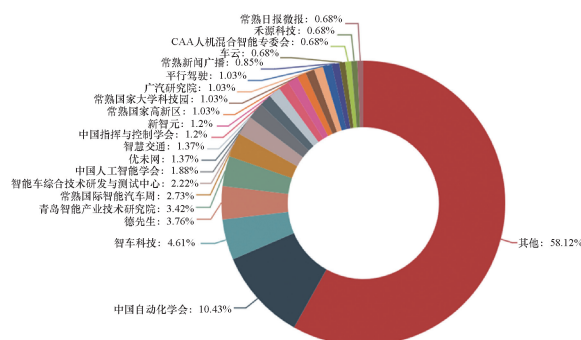


图 9 IVFC 微信公众号数量分布

由图 7~图 9 可以看出，排名前 20 的新闻媒体的发布量约占 50%，主要集中在新闻门户网站及其汽车板块，科学技术站点的发布量也较高，约占 5%。微博发布者中约 67% 是个人用户，排名前 20 的微博发布账号多为赛事相关机构、举办地常熟本地的官方账号及历届赛事中成绩较为突出的高校院所账号。排名前 20 的微信公众号同样集中在赛事相关的单位以及常熟本地公众号，排名前 10 的公众号发文占比高达 35%，较为集中。

4.3 媒体关注热点

通过分析媒体数据的热点可以了解不同媒体

渠道的关注点。热点新闻基于同年度相似新闻数，对每一年的新闻分别进行文本相似度计算。首先进行分词，然后利用 Doc2Vec 对句子进行向量化，最后用余弦相似度计算文本的相似度，相似度大于 0.8 则认为是相似新闻。热点微博基于转发数统计，热点微信基于文章阅读数统计。分别得到排名前 3 的热点数据，相关结果见表 1~表 3。

转发数排名前 3 的微博内容相同，均由与汽车资讯相关的账号发布，获得较高的转发数。微信排名前 3 的热门文章也分别获得了 4 次、3 次与 2 次的微信公众号二次转发。

为了获知网络媒体的热点内容传播情况，本文以“智能汽车：何时让你交出方向盘”这条信息在 3 类网络媒体中的传播情况进行统计，相关结果如

图 10 所示。基于发布时间获取信息源头以确定原创媒体。本条信息最早由《中国青年报》在 2017 年 11 月 28 日刊出，随后在当天依次获得了 20 家网络新闻媒体的二次转发，同时微博与微信平台也分别出现了 3、4 次的转载。

4.4 空间维度分析

为获取 3 类媒体数据中的空间信息，通过命名实体识别工具抽取地理实体信息，再提取数据中包含的中国城市与世界国家信息，并对城市和国家在 3 类媒体数据中出现的篇数进行统计。

结果显示，出现频次较高的城市集中在赛事举办地以及主要参赛队伍所在城市。中国智能车的主要研究力量集中分布在北京、上海、武汉、南京、天津、广州等地。从国家分布来看，由于

表 1 排名前 3 的热点新闻

排名	标题	原创媒体	发布时间	相似新闻数/篇
1	科技部副部长徐南平：我国智能交通将迎来创新发展黄金时代	新华网	2017/11/22 18:11	32
2	智能汽车：何时让你交出方向盘	中国青年报	2017/11/28	28
3	我国智能交通转型明年启动 智能交通将迎创新发展黄金时代	央广网	2017/11/25 9:32	27

表 2 排名前 3 的微博转发数

排名	内容	发布者	发布时间	转发数/次	评论数/个
1	号外！公布一条震撼人心的消息：在 11 月 26 日代表中国无人驾驶智能车最高发展水平的“中国智能车未来挑战赛”上，广汽无人车包揽挑战赛前两名……	时尚潮流汽车	2017/11/30 17:24	1 051	453
2	号外！公布一条震撼人心的消息：在 11 月 26 日代表中国无人驾驶智能车最高发展水平的“中国智能车未来挑战赛”上，广汽无人车包揽挑战赛前两名……	汽车观察编辑	2017/11/30 18:15	733	37
3	号外！公布一条震撼人心的消息：在 11 月 26 日代表中国无人驾驶智能车最高发展水平的“中国智能车未来挑战赛”上，广汽无人车包揽挑战赛前两名……	汽车资讯编辑	2017/11/30 18:15	723	33

表 3 排名前 3 的热点微信文章

排名	标题	发布公众号	发布时间	阅读数/(次)/ 点赞数/(个)
1	广汽无人车斩获 2017 中国智能车未来挑战赛一二名，登上央视《新闻联播》	广汽研究院	2017/11/27	8 096/120
2	首次引入真实交通流，2017 中国智能车未来挑战赛复盘	车云	2017/11/29	6 078/46
3	国家自然科学基金委员会举办 2017 年“中国智能车未来挑战赛”	国家自然科学基金委员会	2017/11/28	4 678/39

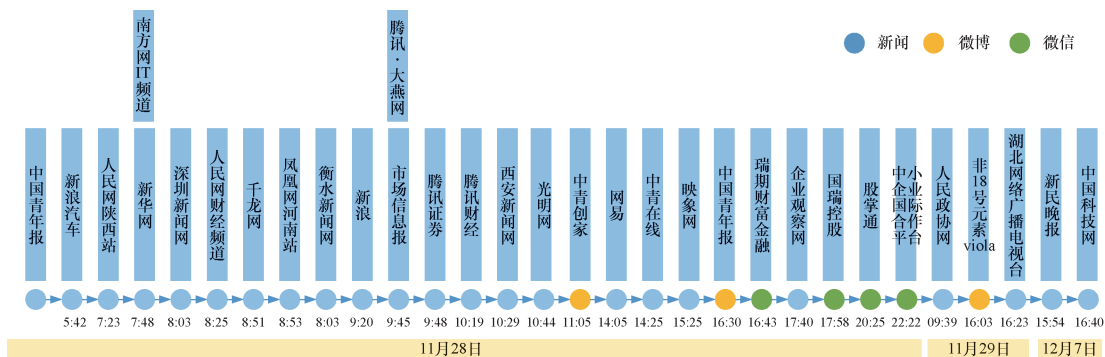


图 10 媒体信息网络媒体传播时间轴（示例）

IVFC 为在中国举办的赛事,“中国”出现在了 1 863 篇媒体数据中;其次,有 298 篇媒体数据中提到了“美国”,侧面反映出美国在无人驾驶智能车领域的领先地位。传统汽车大国德国和日本紧随其后,英国、意大利、韩国、新加坡出现的频次也较高。

基于城市、国家及地区的共现关系,生成对应的共现网络图,如图 11~图 12 所示,可以从更直观的角度看到数据中的城市分布、与国家及地区分布及其关系。

4.5 关键词分析

本文使用基于深度学习模型的 FudanDNN-NLP 进行文本数据的分词,其在序列标注网络结构的基础上,采用快速卷积代替窗口模型,并且引入前向标签与不当状态的交互信息对网络进行进一步优化^[17]。分词前使用正则表达式去除文本中的标点符号、网址等无效内容,并在分词过程中加入停用词词典。对初步分词后的结果进行词频统计,排名前 20 的词汇见表 4。

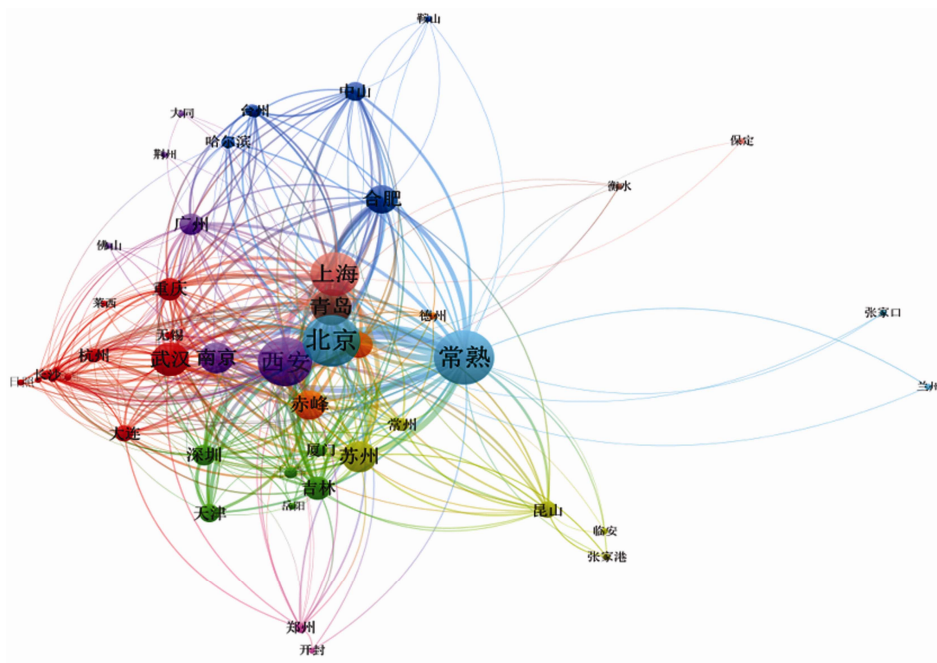


图 11 IVFC 数据城市共现网络

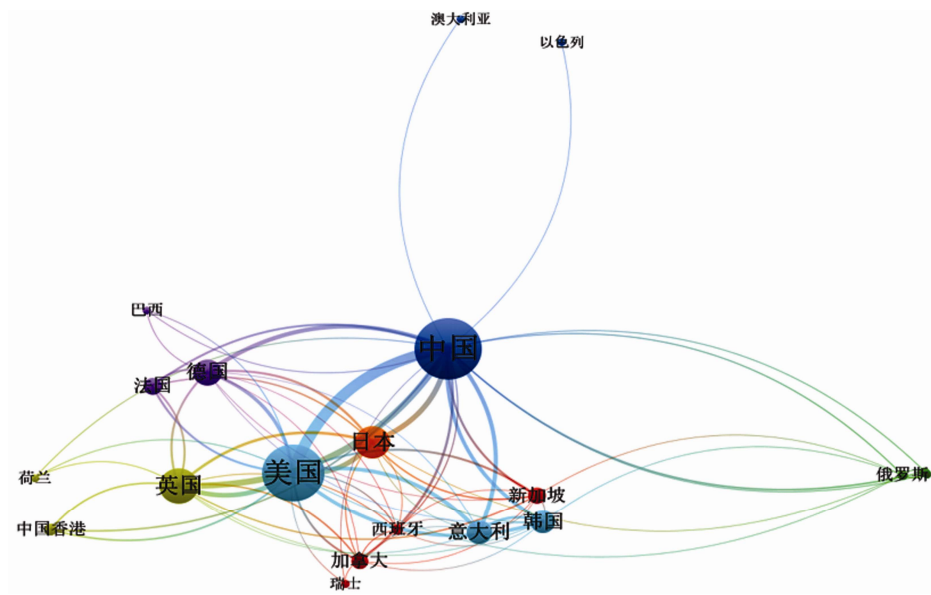


图 12 IVFC 数据国家及地区共现网络

表4 排名前20的初步分词词频统计

排名	关键词	词频/次	排名	关键词	词频/次
1	汽车	4 722	11	无人车	2 604
2	技术	4 163	12	研究	2 327
3	智能车	4 040	13	企业	2 255
4	常熟	3 698	14	车辆	2 210
5	发展	3 641	15	进行	2 157
6	交通	3 285	16	通过	2 146
7	挑战赛	2 970	17	无人驾驶	2 129
8	比赛	2 847	18	一个	2 096
9	中国	2 783	19	我们	2 079
10	测试	2 675	20	系统	2 044

为了有效获取领域词汇，本文基于分词结果使用 N-Gram 算法进行识别。早期，中文领域的新词发现主要基于概率统计技术和关联规则等方法^[18-19]，近年采用基于条件随机场、信息熵、信息传播特性以及融合规则等方法^[20-22]取得了不错的效果。相关实验证明了在基于词和基于字符的新词识别试验中，基于词的 N-Gram 识别效率更高^[19]。识别出的新词及词频见表5。

从表5可以看出，N-Gram 算法能够识别出本领域较高质量的专用词，如“辅助驾驶”“交通标志”“电动汽车”等。同时，准确识别出了领域专家教授提出的新词汇，如“平行驾驶”“平行理论”等。此算法还对命名实体识别中的错误信息进行了纠正，提升了对人名等实体的识别准确率。

表5 排名前20的新词词频

排名	关键词	词频/次	排名	关键词	词频/次
1	中国智能车未来挑战赛	4 722	11	产业发展	307
2	人工智能	4 163	12	智能车研究	253
3	智能汽车	4 040	13	汽车零部件	252
4	中国智能车大会	3 698	14	智能车测试	212
5	无人驾驶汽车	3 641	15	交通标志	208
6	无人驾驶车辆	3 285	16	城市交通	208
7	智能交通	2 970	17	信息科学	198
8	汽车智能化	2 847	18	发展方向	195
9	视听觉信息的认知计算	2 783	19	技术研究	193
10	辅助驾驶	2 675	20	技术水平	182

基于上述分词处理后，对全部数据进行词频统计，并基于词频生成词云，总词云如图13所示。

从图13可以看出，围绕“中国智能车未来挑战赛”，主要词分布在赛事举办地、主办单位、领域专用词、参与单位、主要专家、赛事相关等主题。

为了进一步展示关键词共现情况，本文利用 TFIDF 算法^[23]提取每篇文章的关键词，数量限制为10个。由于 TFIDF 算法能较好地识别出较长的领域词，且对较高频的通用短词有一定的过滤效果，因此在关键词抽取研究早期就被用来抽取中文关键词^[24]。关键词提取完成后进行关键词词频统计，并基于关键词共现关系生成关键词聚类网络，如图14所示。

4.6 话题聚类分析

对于同一个话题，不同的媒体会有不同的侧重



图13 IVFC 总词云

点, 因此通过话题检测对媒体数据进行分析已经成为当前的研究热点。本文分别使用基于潜在语义索引的 Lingo 算法^[25]与典型的基于距离的 K-means^[26]算法进行话题聚类, 并进行对比分析。预处理阶段使用前文的分词结果, 最后得出话题聚类结果使用气泡树 (FoamTree) 的可视化效果, 如图 15~图 16 所示。

FoamTree 聚类视图中气泡块越大, 表示该类别包含的文档数越多。Lingo 算法将 1 905 篇文档划分到了 95 个类别下, 最大的类别为“常熟汽车”, 共包含 447 篇文档。通过聚类结果视图可以看出, 大的话题主要分布在“无人驾驶智能车”“挑战赛赛

事”“智能交通与智能车”等主题。K-means 算法共聚出了 25 个类别, 其不能自动产生类别标签, 以聚类完成后的类别文档高频关键词作为类别标签, 基本可以反映本类别的特征。

对比来看, Lingo 算法较为准确地识别出了一些核心话题, 标签的可读性也较高, 能反映本类别的核心特征; 虽然 K-Means 算法简单易懂、运行速度快, 但是在本文实验部分, 其在类别准确性与话题多样性方面没有 Lingo 算法的效果好, 且出现了较多的话题重合现象。

4.7 主题模型分析

主题模型是对文字隐含主题进行建模的方法,

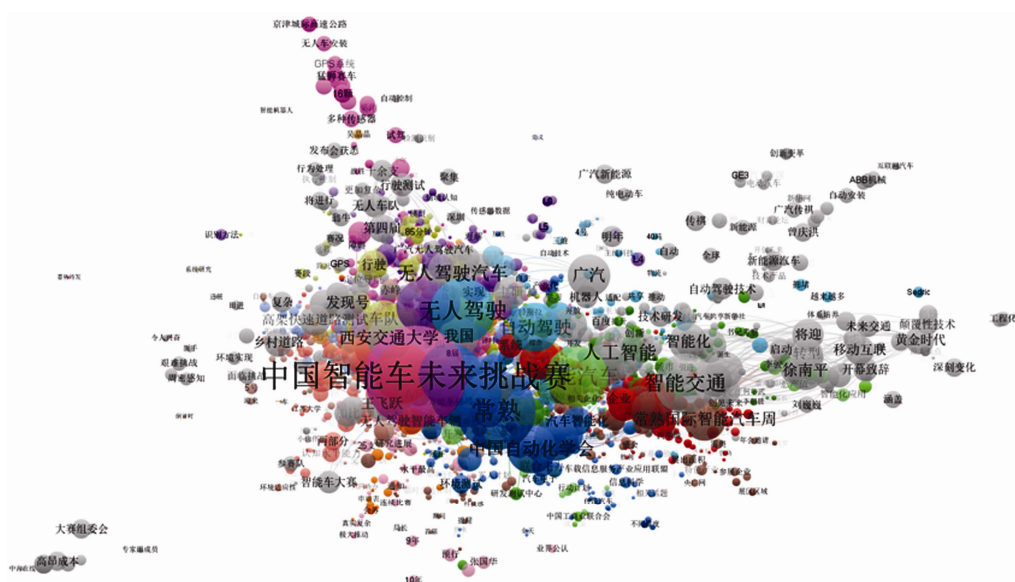


图 14 使用 TFIDF 算法抽取文本关键词共现聚类



图 15 使用 Lingo 算法对 IVFC 网络媒体数据的话题聚类气泡

它能够在海量互联网数据中自动寻找出文字间的语义主题。目前比较常用的基于主题聚类算法有 LDA^[27]和 PLSA^[2]，其中 LDA 相比 PLSA 不容易产生过拟合现象，目前已经成为广泛应用的主题模型代表算法。本文使用 LDA 算法对 IVFC 的数据进行主题模型分析。

设定主题数目的取值范围为[10, 100]，取步长为 5 进行 LDA 主题抽取，通过计算 Perplexity 指标，最终选定稳定拐角点处的值 30 作为最优主题数，并使用 LDAvis^[29]对结果进行可视化分析，结果如图 17 所示。

形成的 30 个主题中主要覆盖：IVFC 无人驾驶



图 16 使用 K-Means 算法对 IVFC 网络媒体数据的话题聚类气泡

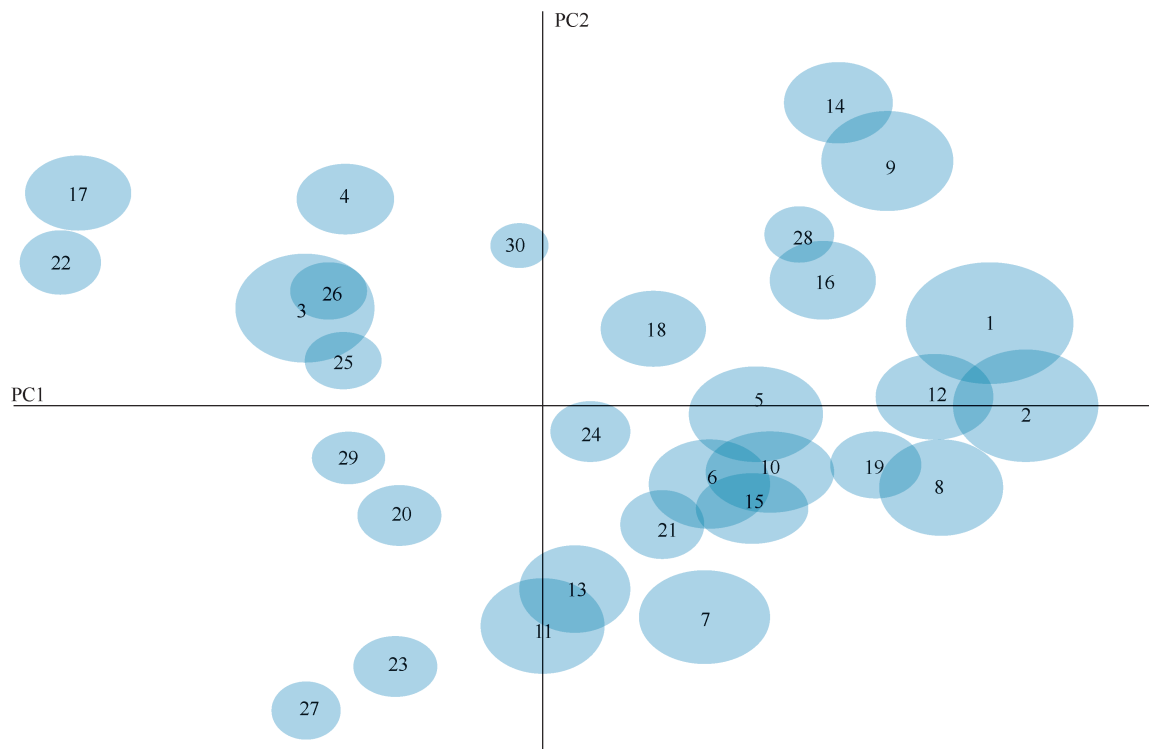


图 17 主题数为 30 的 LDA 聚类效果

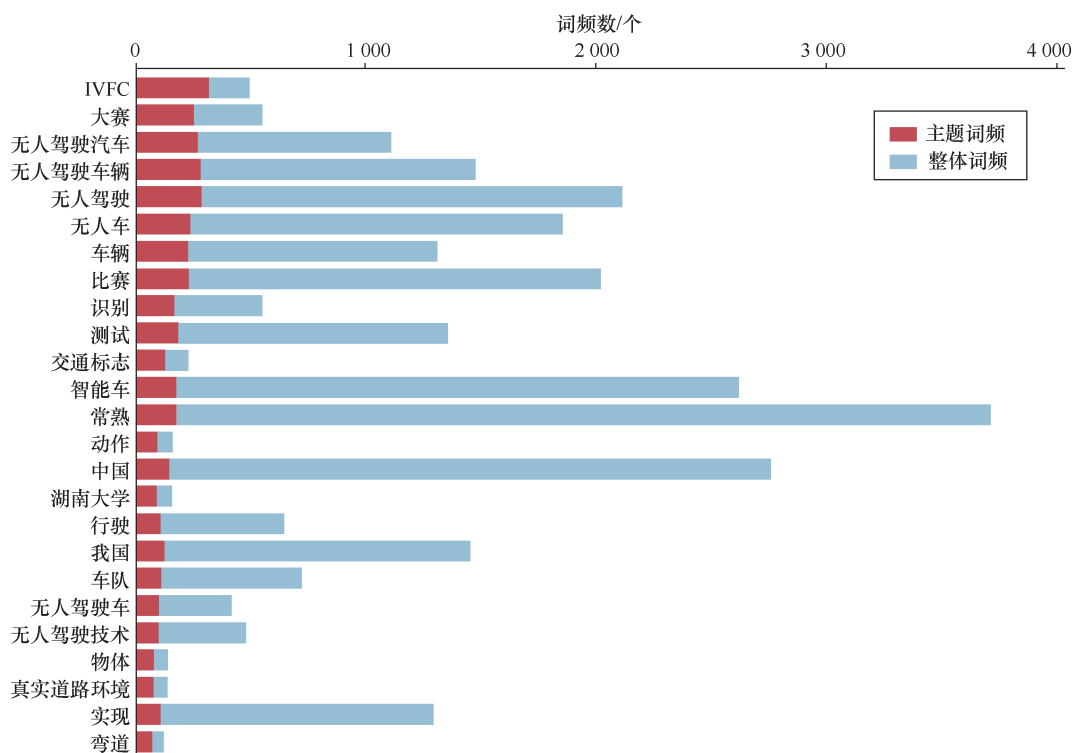


图 18 与主题 1 相关的前 25 个词

车、赛事考察点、西安交大发现号、广汽无人车、美国无人车技术、相关活动、专家观点等。选取主题 1 及前 25 个最相关的词如图 18 所示。

4.8 语义建模分析

Word2Vec 是一种可以进行高效率词嵌套学习的预测模型, 2013 年由 Google 团队提出^[30-31]。Word2Vec 词向量可以较好地表达不同词之间的相似和类比关系。本文使用 Word2Vec 的 Skip-Gram 模型在 TensorFlow 下实现 IVFC 相关数据的 Word2Vec 词向量的训练, 初始词向量维度设定为 300 维。本文选取“王飞跃”作为测试词进行语义效果测试, 使用余弦相似度计算测试词与其他词的语义距离, 本文取前 20 个与测试词语最相关的词语及其距离度量值(保留 3 位小数), 结果见表 6。

由表 6 的结果可以看到, 使用 Word2Vec 的 Skip-Gram 算法训练所得的 IVFC 词向量模型可以较好地抽取测试词的语义关联词, 其中, 对于与“王飞跃”语义相似度排名靠前的词, 基本可以准确地识别出其关联职称、担任的赛事角色、所属机构、关系较为密切的人物、采访中常说的词语等。

表 6 Word2Vec 词向量模型测试词“王飞跃”及前 20 个关联词

排名	关键词	词频/次	排名	关键词	词频/次
1	百姓家	0.904	11	淘汰	0.855
2	直言	0.894	12	时间检验	0.851
3	教授	0.893	13	秘书长	0.849
4	研究员	0.891	14	贺汉根	0.843
5	总裁判长	0.888	15	国家发改委	0.842
6	复杂系统管理控制 国家重点实验室	0.879	16	博士生导师	0.841
7	理论技术	0.876	17	李力	0.829
8	安全性法规	0.875	18	曹东璞	0.828
9	担任	0.875	19	裁判组组长	0.827
10	黄武陵	0.872	20	汽车工程系	0.826

由于训练所得的词向量高达 300 维, 为了更直观地探寻数据的语义关系, 必须对高维数据进行降维, 有效降到二维或三维才能进行词向量可视化。本文分别采用线性降维算法 PCA 与非线性降维算法 t -SNE^[32] 对训练所得的词向量进行降维并通过 TensorBoard 进行可视化, 生成的三维图形如图 19 所示。

从两种降维算法生成的三维图可以看出, 使用 PCA 算法降维后的词分布整体比较均匀, 使用只在局

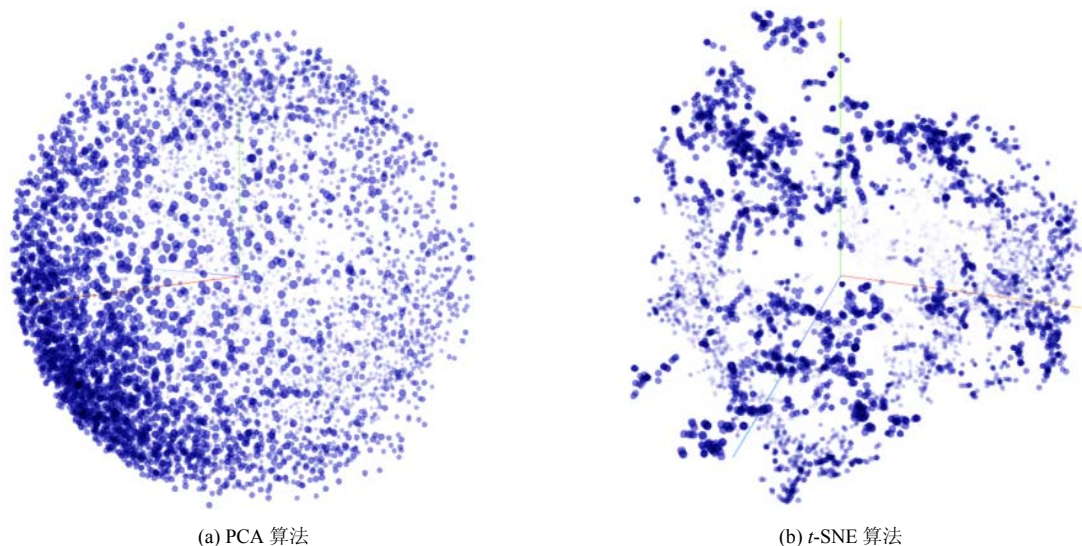


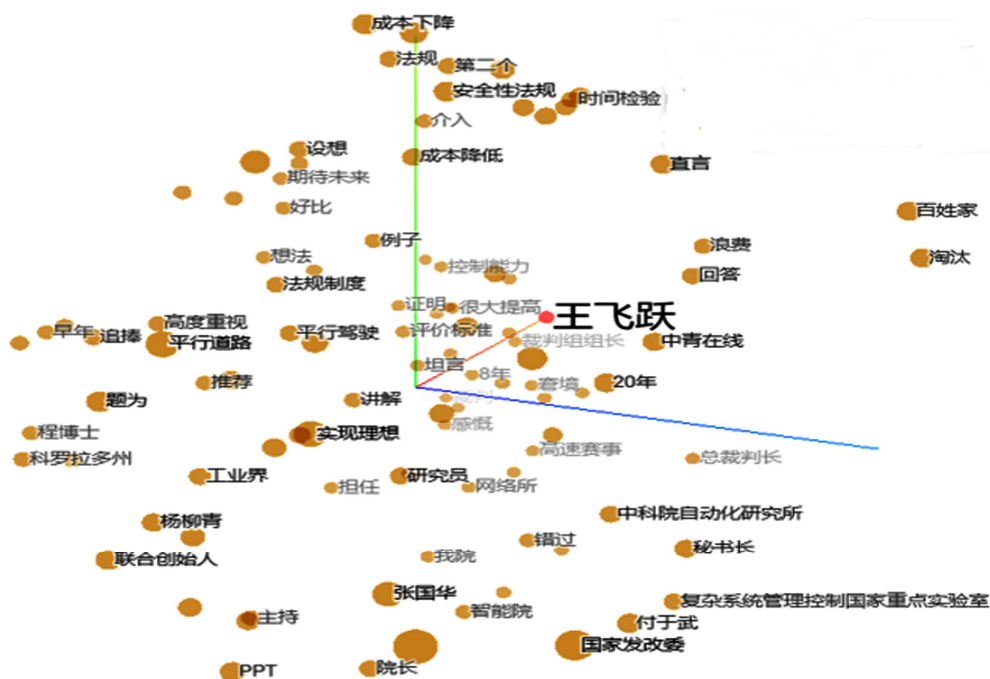
图 19 IVFC 词向量降维三维图

部一面较为密集,而使用 t -SNE 算法降维后的词在三维空间中原有数据的局部特征得到了很好的保留,能明显看出词的簇群。在生成速率上,PCA 速度快,更稳定,而 t -SNE 更好地保留了原有数据的局部特征。以“王飞跃”为测试词对其语义关联度排名前 100 的词分别用两种算法进行分析,如图 20 所示。

4.9 实体统计及网络分析

为了获取网络媒体关注的人物、机构等实体信息,本文从活动关联数据中识别抽取人物与机

构,进行频次篇章统计并基于实体共现关系进行网络聚类分析。首先使用 FudanDNN-NLP 进行初步识别,再通过 N-Grams 方法进行二次消歧与补充,并设置一定的阈值剔除低频无效实体信息。另外需要筛掉新闻中的作者、编辑姓名,对机构进行机构全称与简称的识别与合并。经过处理共抽取机构 151 家、人名 141 个。抽取的机构及关联数据见表 7 (机构后数字为在数据中出现的篇数)。



(a) PCA 算法

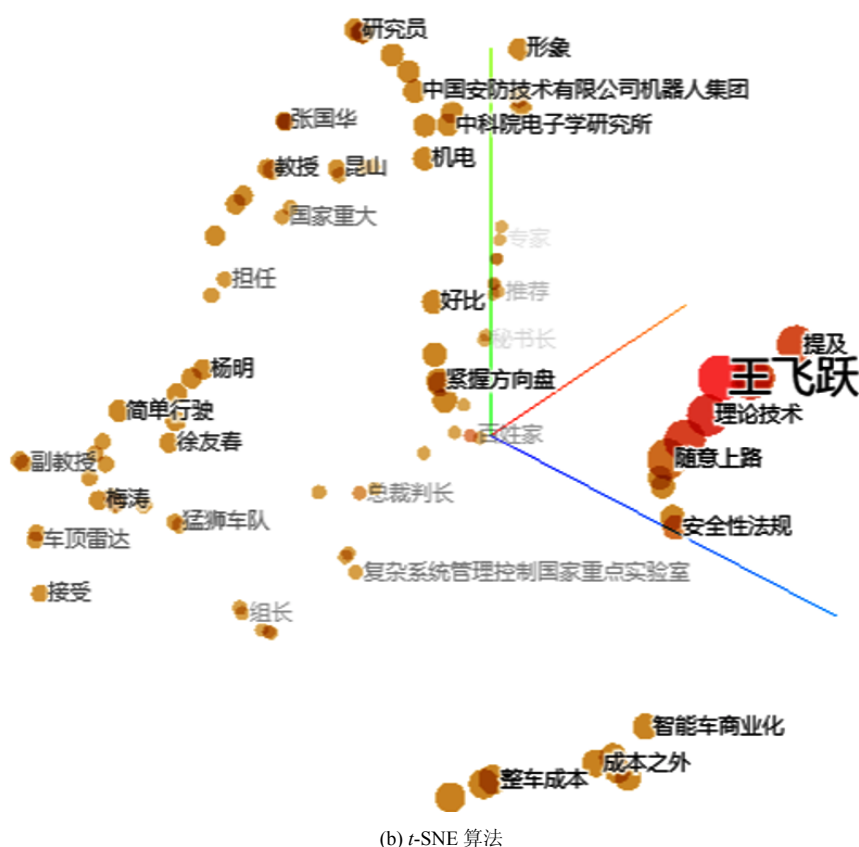


图 20 测试词及其语义关联前 100 词的空间分布

表 7

排名前 20 的 IVFC 网络媒体数据机构

排名	机构名	篇数/篇	排名	机构名	篇数/篇
1	国家自然科学基金委员会	576	11	科技部	233
2	西安交通大学	499	12	军事交通学院	228
3	中国工程院	348	13	国防科技大学	201
4	中国自动化学会	344	14	中国人工智能学会	179
5	清华大学	343	15	中国智能交通协会	168
6	广汽	331	16	中国科学院自动化研究所	157
7	常熟市人民政府	317	17	青岛智能产业技术研究院	157
8	中国智能车综合技术研发与测试中心	297	18	同济大学	157
9	北京理工大学	269	19	谷歌	151
10	丰田	235	20	上海交通大学	149

从表 7 中可以看出,国家自然科学基金委员会作为 IVFC 的主办方出现篇数最多,排名前 20 的主要是大赛参与高校单位、承办单位与协办单位。另外,通过对参与单位的分析发现,机构的媒体关注度与其大赛成绩有较大的关系。谷歌则作为无人驾驶智能车领域的先行者被 151 篇文章提及。

在此基础上,抽取机构间的共现关系,根据机构间的共现频次生成聚类网络,如图 21 所示。

151 家机构被聚成了 10 个簇,主要参赛高校、科研院所等聚成了一类,一些传统的汽车厂商、国外相关大学机构、举办地的相关机构等各自被聚成了不同的簇。

对抽取的人物实体进行统计排名,结果见表 8。

排名前 20 的人物主要为本领域、与本赛事有关的学者、教授和相关部门的官员等,其中郑南宁、王飞跃、李骏、李德毅等均为中国无人驾驶智能车

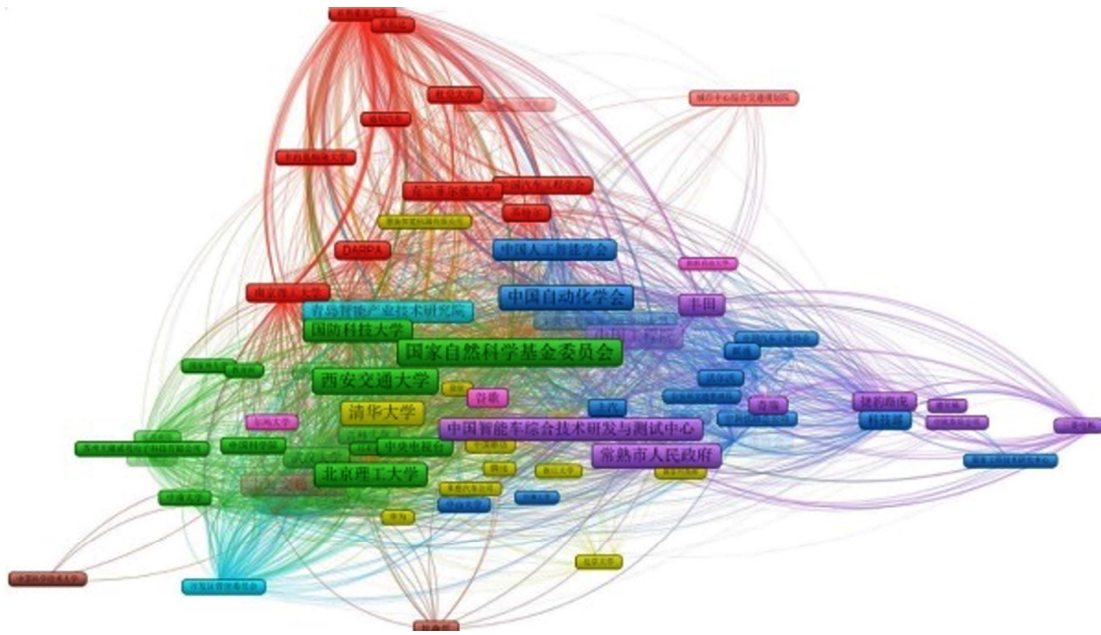


图 21 IVFC 机构关联聚类网络

表 8 排名前 20 的 IVFC 网络媒体数据人物

排名	人名	篇数/篇	排名	人名	篇数/篇
1	郑南宁	315	11	张国华	52
2	王飞跃	228	12	吴忠泽	49
3	徐南平	210	13	徐昕	49
4	李骏	135	14	吴甘沙	48
5	李德毅	105	15	徐友春	47
6	傅志寰	87	16	曾庆洪	46
7	何华武	72	17	梅涛	44
8	李力	71	18	李朝晨	41
9	曹东璞	67	19	杨明	41
10	王飏	58	20	薛建儒	39

领域的著名专家学者与开拓性人物。

在此基础上，抽取人物间的共现关系，根据人物间的共现次数进行聚类分析，通过可视化的方式展现人物间的共现网络，如图 22 所示。

从图 22 可以明显看出核心人物为“郑南宁”“王飞跃”等。最大的群簇包含 23 个人物节点，最小的只包含 2 个人物节点。人物网络中很多人物是 2017 年同时在常熟举办的“中国智能车大会暨国家智能车发展论坛”与“中国智能交通年会”相关的主管领导、知名专家、企业高管及嘉宾等。

4.10 情感分析

本文使用基于情感词典的方式来抽取媒体数据中的情感词，以从侧面分析当前主流媒体对赛

事的情感支持度。目前，英文文本的情感词典中较为成熟的有 SentiWordNet^[33-34]，它提供了词语的积极消极、主观客观与情感强度值；中文情感词典也已提出多种方法进行构建^[35]，现已出现多个版本。本文以 HowNet 情感词、台湾大学的 NTUSD 和清华大学的褒贬义词典三者的并集为基础情感词典，再通过人工方式进行了二次筛选，并进行了少量交通领域情感词的添加，最后形成正面情感词 3 514 个、负面情感词 3 345 个。基于情感词典对 IVFC 相关数据进行情感词抽取，共抽取正向情感词 462 个、负向情感词 133 个。对情感词进行统计，得到前 15 名的正负情感词如图 23 所示。

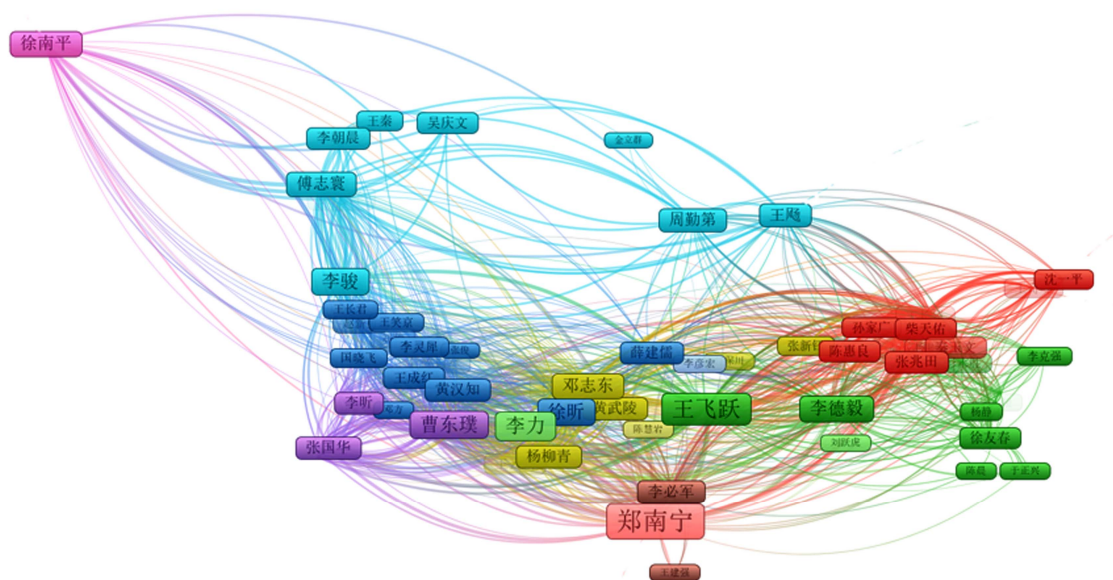


图 22 IVFC 人物关联聚类网络

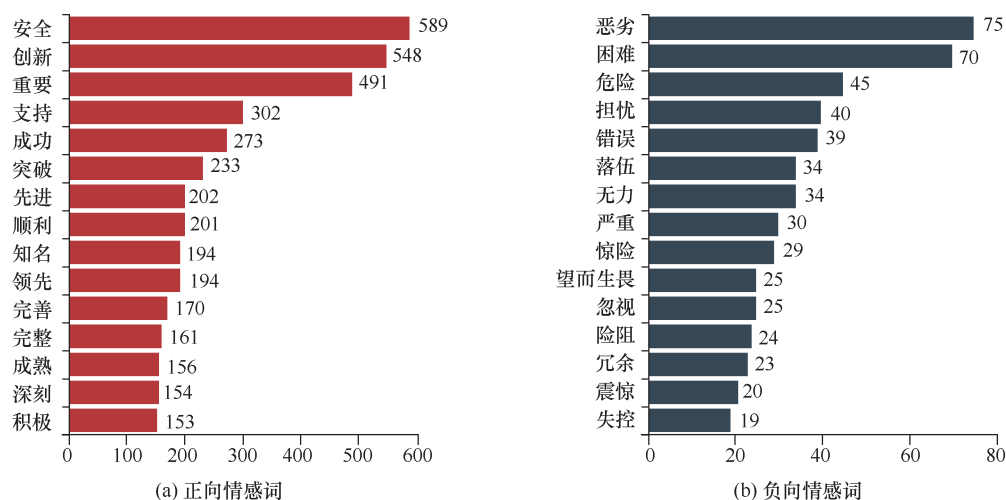


图 23 排名前 15 名的 IVFC 正/负向情感词

通过正负情感词统计及排名可以看出，媒体对智能车挑战赛更多的是持积极肯定的态度，从侧面反映出此项赛事的重要性。负向情感词较少，主要为交通环境、技术困难、车辆表现等描述，这些词反映出智能车挑战赛的必要性，以及当前距离智能车的广泛应用依然存在很多问题。

4.11 知识图谱

知识图谱 (knowledge graph) 旨在描述客观世界的概念、实体、事件及它们之间的关系，在工业界也开展了多方面的应用。为了更好地梳理抽取的各项实体内容，本文基于相关内容构建了一个简单的 IVFC 知识图谱 (如图 24 所示)。通过其可以较为清晰地了解 IVFC 媒体数据中的关键媒体、主要

人物、机构、关键词、城市等信息。

5 结束语

传统的舆情监测平台浅层次的数据采集与分析不能满足对活动提供针对性、全方位、多维度的数据监测与建模分析的需求。鉴于此，本文基于平行情报与知识自动化的理论方法，运用社会计算中的基本方法，融合自然语言处理、机器学习等相关技术，提出了一个面向活动网络媒体数字化监测与建模分析的方法框架，并以中国智能车未来挑战赛为应用示例，对相关网络媒体数据在时空纬度、数据源纬度、关键词纬度、话题聚类、主题模型、语义建模等多个维度进行了数字化描述与建模分析，并基

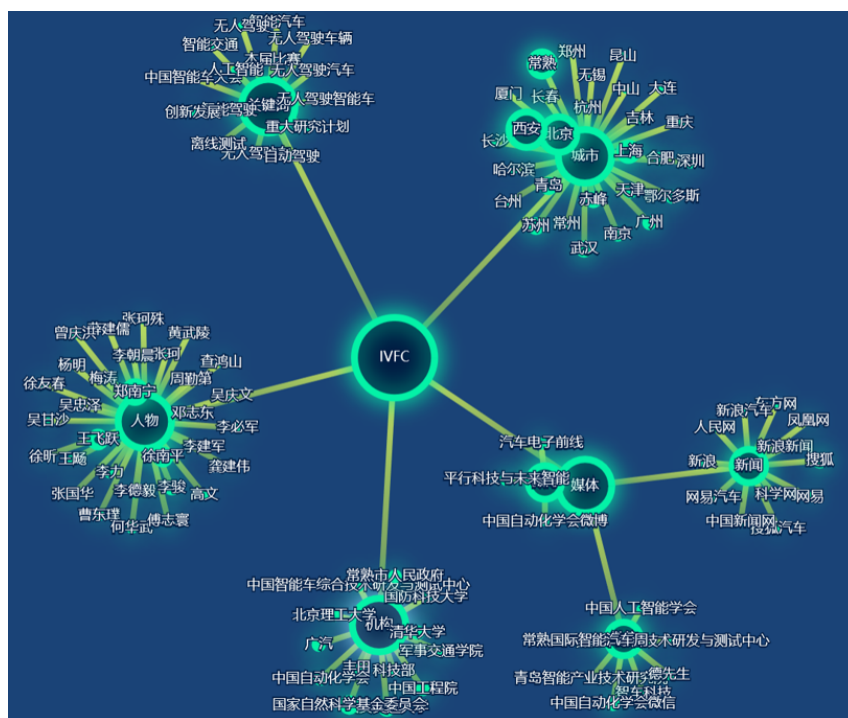


图24 基于网络媒体数据构建IVFC知识图谱

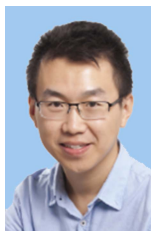
于分析结果构建了与赛事活动相关的知识图谱。结果表明，本文提出的相关方法框架可以为活动关联方提供有效的数字化监测手段与辅助决策支撑。

参考文献：

- [1] WANG F Y. The emergence of intelligent enterprises: from CPS to CPSS[J]. IEEE Intelligent Systems, 2010, 25: 85-88.
- [2] 王飞跃. 社会信号处理与分析的基本框架：从社会传感网络到计算辩证解析方法[J]. 中国科学：信息科学, 2013, 43(12): 1598-1611.
- [3] WANG F Y. A framework for social signal processing and analysis: from social sensing networks to computational dialectical analytics[J]. Scientia Sinica (Informationis), 2013, 43(12): 1598-1611.
- [4] 王飞跃. 从一无所有到万象所归：人工社会与复杂性研究[N]. 科学时报, 2003年3月17日.
- [5] WANG F Y. From nothing to everything: artificial societies and complex system[N]. Science Times, 2003-3-17.
- [6] WANG F Y. Social computing: concepts, contents, and methods[J]. International Journal of Intelligent Control and Systems, 2004, 9(2): 91-96.
- [7] 王飞跃. 关于复杂系统研究的计算理论与方法[J]. 中国基础科学, 2004, 6(41): 2-10.
- [8] WANG F Y. Computational theory and methods on complex system[J]. China Basic Science, 2004, 6(41): 2-10.
- [9] 王飞跃. 社会计算与数字网络化社会的动态分析[J]. 科技导报, 2005, 23, 5(9): 4-6.
- [10] WANG F Y. Social computing and dynamical state analysis of digitalized and networked societies[J]. Science & Technology Review, 2005, 23, 5(9): 4-6.
- [11] 王飞跃, 王珏. 情报与安全信息学研究的现状与展望[J]. 中国基础科学, 2005, 7(2): 24-29.
- [12] WANG F Y, WANG J. Intelligence and security informatics: the state of the art and outlook[J]. China Basic Science, 2005, 7(2): 24-29.
- [13] WANG F Y. Toward a paradigm shift in social computing: the ACP approach[J]. IEEE Intelligent Systems, 2007, 22(5): 65-67.
- [14] 王飞跃. 情报 5.0: 平行时代的平行情报体系[J]. 情报学报, 2015, 34(6): 563-574.
- [15] WANG F Y. Intelligence 5.0: parallel intelligence in parallel age[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 563-574.
- [16] 唐涛. 基于大数据的网络舆情分析方法研究[J]. 现代情报, 2014, 34(3): 3-6.
- [17] TANG T. Research on method of internet public opinion analysis based on big data[J]. Journal of Modern Information, 2014, 34(3): 3-6.
- [18] 李金海, 何有世, 熊强. 基于大数据技术的网络舆情文本挖掘研究[J]. 情报杂志, 2014, 33(10): 1-6, 13.
- [19] LI J H, HE Y S, XIONG Q. On text mining of network public opinion based on big data technology[J]. Journal of Intelligence, 2014, 33(10): 1-6, 13.
- [20] 马梅, 刘东苏, 李慧. 基于大数据的网络舆情分析系统模型研究[J]. 情报科学, 2016, 34(3): 25-28.
- [21] MA M, LIU D S, LI H. Research on the network public opinion analysis system model based on big data[J]. Information Science, 2016, 34(3): 25-28.
- [22] 夏火松, 甄化春. 大数据环境下舆情分析与决策支持研究文献综述[J]. 情报杂志, 2015, 34(2): 1-6.
- [23] XIA H S, ZHEN H C. Public opinion analysis and decision support study under big data surroundings[J]. Journal of Intelligence, 2015, 34(2): 1-6.
- [24] 陈必坤, 王曰芬, 廖海涵. 大数据时代社会舆情分析与决策支持研究的现状及其发展趋势[J]. 情报科学, 2016, 34(10): 8-14.
- [25] CHEN B K, WANG Y F, LIAO H H. Current situation and development trend of social public opinion analysis and decision making based on the background of big data era[J]. Information Science, 2016, 34(10): 8-14.
- [26] 王飞跃. 软件定义的系统与知识自动化：从牛顿到默顿的平行升华[J]. 自动化学报, 2015, 41(1): 1-8.

- WANG F Y. Software-defined systems and knowledge automation: a parallel paradigm shift from Newton to Merton[J]. *Acta Automatica Sinica*, 2015, 41(1): 1-8.
- [16] 王飞跃, 王晓, 袁勇, 等. 社会计算与计算社会: 智慧社会的基础与必然[J]. *科学通报*, 2015, 60(Z1): 460-469.
- WANG F Y, WANG X, YUAN Y, et al. Social computing and computational societies: the foundation and consequence of smart societies[J]. *Chinese Science Bulletin*, 2015, 60(Z1): 460-469.
- [17] ZHENG X Q, CHEN H Y, XU T Y. Deep learning for Chinese word segmentation and POS tagging[C]// The 2013 Conference on Empirical Methods in Natural Language Processing, October 18-21, 2013, Seattle, USA. Stroudsburg: ACL, 2013: 647-657.
- [18] 贾自艳, 史忠植. 基于概率统计技术和规则方法的新词发现[J]. *计算机工程*, 2004, 30(20): 19-21.
- JIA Z Y, SHI Z Z. Probabilistic techniques and rule methods for new word discovery [J]. *Computer Engineering*, 2004, 30(20): 19-21.
- [19] 王立希, 王建东, 汪静. 基于数据挖掘的新词发现[J]. *计算机应用研究*, 2006, 23(12): 195-197.
- WANG L X, WANG J D, WANG J. Approach for lexicon updating based on data mining[J]. *Application Research of Computers*, 2006, 23(12): 195-197.
- [20] 陈飞, 刘奕群, 魏超, 等. 基于条件随机场方法的开放领域新词发现[J]. *软件学报*, 2013, 24(5): 1051-1060.
- CHEN F, LIU Y Q, WEI C, et al. Open domain new word detection using condition random field method[J]. *Journal of Software*, 2013, 24(5): 1051-1060.
- [21] 霍帅, 张敏, 刘奕群, 等. 基于微博内容的新词发现方法[J]. *模式识别与人工智能*, 2014, 27(2): 141-145.
- HUO S, ZHANG M, LIU Y Q, et al. Pattern recognition and artificial intelligence [J]. *Pattern Recognition and Artificial Intelligence*, 2014, 27(2): 141-145.
- [22] 李文坤, 张仰森, 陈若愚. 基于词内部结合度和边界自由度的新词发现[J]. *计算机应用研究*, 2015, 32(8): 2302-2304.
- LI W K, ZHANG Y S, CHEN R Y. New word detection based on inner combination degree and boundary freedom degree of word[J]. *Application Research of Computers*, 2015, 32(8): 2302-2304.
- [23] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. *Information Processing & Management*, 1988, 24(5): 513-523.
- [24] 徐文海, 温有奎. 一种基于 TFIDF 方法的中文关键词抽取算法[J]. *情报理论与实践*, 2008, 31(2): 298-302.
- XU W H, WEN Y K. A Chinese keyword extraction algorithm based on TFIDF method[J]. *Information Studies: Theory & Application*, 2008, 31(2): 298-302.
- [25] MIHALCEAR, TARAU P. TextRank: bringing order into text[C]//The 2004 Conference on Empirical Methods in Natural Language Processing, July 25-26, 2004, Barcelona, Spain. [S.l.:s.n.], 2004: 404-411.
- [26] HARTIGAN J A, WONG M A. Algorithm AS 136: a K-means clustering algorithm[J]. *Journal of the Royal Statistical Society*, 1979, 28(1): 100-108.
- [27] BLEI D M, NG A, JORDAN M. Latent Dirichlet allocation[J]. *Journal of Machine Learning Research*, 2003(3): 993-1022.
- [28] HOFMANN T. Probabilistic latent semantic analysis[C]// The Fifteenth Conference on Uncertainty in Artificial Intelligence, July 30-August 1, 1999, Stockholm Sweden. San Francisco: Morgan Kaufmann Publishers Inc., 1999: 289-296.
- [29] SIEVERT C, SHIRLEY K E. LDAvis: a method for visualizing and interpreting topics[C]// The Workshop on Interactive Language Learning, June 27, 2014, Baltimore, USA. Stroudsburg: ACL, 2014.
- [30] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. *Computer Science*, 2013.
- [31] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// International Conference on Neural Information Processing Systems, December 5-8, 2013, Lake Tahoe, USA. New York: Curran Associates Inc., 2013: 3111-3119.
- [32] HINTON G E. Visualizing high-dimensional data using *t*-SNE[J]. *Journal of Machine Learning Research*, 2008, 9(2): 2579-2605.
- [33] WSULI A, SEBASTIANI F. SentiWordNet: a publicly available lexical resource for opinion mining[C]// The Fifth International Conference on Language Resources and Evaluation, May 24-26, 2006, Genoa, Italy. France: ELRA, 2006: 417-422.
- [34] BACCIANELLA S, ESULI A, SEBASTIANI F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining[C]// International Conference on Language Resources and Evaluation, May 17-23, 2010, Valletta, Malta. France: ELRA, 2010: 83-90.
- [35] 王科, 夏睿. 情感词典自动构建方法综述[J]. *自动化学报*, 2016, 42(4): 495-511.
- WANG K, XIA R. A survey on automatic construction methods of sentiment lexicons [J]. *Acta Automatica Sinica*, 2016, 42(4): 495-511.

[作者简介]



孙星恺(1985—), 男, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室博士生, 主要研究方向为社会计算、科技情报、知识自动化与知识图谱。



王晓(1987—), 女, 博士, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室副研究员, 主要研究方向为复杂系统管理、平行感知。



陆浩(1983—), 男, 博士, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室副教授, 主要研究方为向知识自动化、社会计算与 Web 智能。