数据环境下网络舆情信息的特点及分析技术研究

郭琛[1]

（大连交通大学　辽宁大连　116021）

摘要：在□ 数据时代下，信息越来越容易获得，由此引发的舆情冲击是急需解决的重要问题，本文分析了网络舆情的特征与表现，针对网络舆情的特点进行了技术分析并构建了完整的舆情信息分析方案，为舆情预警预判和及时应对与处理提供了技术支持和有效保障。

Abstract: In the big data age, information is becoming more and more accessible, and the resulting impact on public opinion is an important problem that needs to be solved urgently. This paper analyzes the characteristics and performance of network public opinion, makes a technical analysis based on the characteristics of network public opinion, and constructs a complete public opinion information analysis scheme, which provides technical support and effective guarantee for public opinion early warning and timely response and processing.

关键词：大数据，网络舆情，舆情分析

Key words: big data, network public opinion, public opinion analysis

一、引言

随着大数据技术的不断发展，基于“大数据”的功能性使用越来越多地出现在人们的视野和生活中，大数据既可以用来表达当今时代数据的某种特点，也可以用来表达数据科学研究的对象。大数据一方面指出了数据的规模，另一方面也包括了被称为 4V 的显著特征[1]，即数据量大(Volume)，数据类型复杂(Variety)，数据的高价值、低密度(Value)，数据的实时处理(Velocity)。其处理对象的数据不仅包括了文本、图像、音视频，还包括了人们基于虚拟网络的物理行为活动，诸如评论，留言、发表观点、点击量、关注、点赞，特别是趋于年轻化，简单化的网络用语等充斥着整个社交网络，这些物理行为数据也被纳入了大数据的采集范围。

在现今个性化凸显的自媒体迅猛发展的阶段，个体对于公共事件表达看法是最常见方式，但随之而来的就是不可预判的海量网络舆情出现。网络舆情是指各种社会群体对自己关心或自身利益相关的热点时间和事物所表现出来的具有□ 定影响□ 并带有倾向性的认知、情绪、态度和意见的总和[1]。

因此，要对网络舆情进行准确分析和有效治理，就要进行全面的数据采集和科学的数据处理。而网络舆情的分析和治理能力，很大程度上也取决于大数据的获取能力和处理分析水平。国家早在 2016 年就明确提出了“实施国家大数据战略”的构想，要把大数据作为基础性战略资源，助力产业转型升级和社会治理创新[2]。因此，在大数据时代背景下，运用大数据进行网络舆情的分析和治理，具有时代意义和现实意义。

二、□ 络舆情及其特点

数据环境下多媒体□ 络舆情与传统□ 络舆情相□ ，具有□ 泛性、瞬发性、价值增强性、关联多样性性、主观性、多元覆盖性等特征。具体表现为：

1.基于□ 络的多媒体传播呈海量倍增态势[3]，舆情传播快速，高效。各地的敏感、热点信息，刚刚发生就会通过自媒体网络进行传播，瞬发且快速，各地的人们均可快速获得第一时间信息。

2.低价值低密度的□ 络舆情信息通过多媒体传播后会效应增强，并能挖掘出相关关联的多维空间和时间信息，关联网络可不断扩大。

3.□ 络舆情信息的多样性与媒体的主体表达，会导致对公众认识问题的观念引导，并形成主观的判断和评价[4]。

4.舆情信息采集和获取的渠道多元化，全覆盖化，身边各角落发生的热点信息均可快速获取到。

三、舆情信息分析技术

面对□ 络舆情的特点，基于大数据技术的□ 络舆情分析可分为信息采集、信息预处理、舆情分析、舆情预警四个主要步骤完成。基于此设计的舆情分析系统首先利用网络爬虫技术针对性采集网络舆情信息，利用机器学习技术自动分析海量数据中的情感信息，从而挖掘出有价值的信息。大数据分析的主要技术手段是采用数据挖掘，数据挖掘又称数据库中的知识发现，即指从数据库的大量数据中揭示出隐含的前所未有的并具有潜在价值的信息的价值聚合、提炼的过程。数据挖掘研究拥有强大的技术支撑，基于数据库、人工智能和数理统计等技术交叉融合，能够做到对象的可细分、可价值分析、流失预测、异常发现、预警，也可进行科学发现和改进工作效率，作出具有更强的合理性、准确性、针对性的判断[5]。

数据挖掘拥有以下六种不同功能：关联分析、时序模式、分类、聚类、预测和偏差分析等[5]。这些强大的功能运用于网络舆情的研究十分有价值，它们可以对舆情信息进行针对性的挖掘与分析，准确研判当前网络的舆情动态，对网络的热点、焦点与敏感话题及时做出反应，把握处理危机事件的最佳时机。从而提高网络的监管能力及突发事件的处置能力[6]。

数据挖掘中最成熟的技术之一就是关联分析，它能发现一个事物中某些属性同时出现的规律和模式。通过事物内在的隐含的特征，建立相互关联，大多数关联规则挖掘算法都能够无遗漏发现隐藏在所挖掘数据中的关联关系。

此外，基于聚类的分析功能也是技术分析功能之一，它不但可以将不同的数据按照某一标准或条件整理分成不同的类，还可以建立宏观的概念，从而发现数据的分布模式和可能的数据属性之间的相互关系。这个功能可以很好的应用于网络舆情的研究，可以对互联网中海量的信息进行大致的聚类，也可以对信息的使用者进行聚类，根据信息的使用情况、信息的内容特征等多个方面对信息的使用者进行聚类，概括出每一个聚类的特征，可以便于今后更深一步的分析研究。

分类[7]是数据挖掘应用比较成熟的技术，尤其是在商业的应用中。分类是找出一个类别的概念描述，即该类的内涵描述，它代表了这一类数据的整体信息，使该类与其他数据独立区别。在网络舆情研究中，我们可以根据自身需求对大量的网络信息进行初步的筛选，对各类舆情信息进行分类、分组，如设置“民生问题”、“突发事件”、“公共安全”、“经济危机”等等，为下一步工作做好初步的准备而有针对地进行数据选择并进行数据集合，缩小挖掘的范围，避免盲目搜索，提高数据挖掘的效率和质量从而得到更加精确的、有意义有价值的信息。

预测[8]是利用历史数据中找出的变化规律，建立专用模型，通过此模型对未来数据的种类及特征等其他方面进行预测，得出未来可能出现的结果。预测即是对趋势分析，而时序模式是重要的预测预警方式，是通过对数据库中的数据发生的时间序列进行升序或降序排列整理分析出的重复发生概率较高的模式。预测和时序功能都可以很好地应用于网络舆情的监控和预警，在舆情信息汇集和分析的基础上，对社会运行接近负向质变的临界值的程度所做出初步确定的早期预报。预测和时序功能的应用还能够及时掌握网络舆情动

态，避免很多事件向消极的方向发展，使对不良网络舆情的处理从即时处置型向事前预警型转变。

优秀的互联网舆情分析管理系统利用广泛的互联网信息采集技术和数据挖掘技术，通过自动采集、自动分类、智能过滤、自动聚类、主题检测和统计分析，实现社会热点话题、突发事件、重大案情的快速识别和定向追踪；能够通过网络监控管理，了解预测网民群体的倾向和意愿，提前发现网上不良事件的苗头，及时封堵各类有害信息，同时，还可以在大规模舆情危机爆发之前，根据预测和时序功能，尽早针对热点话题，梳理情况，快速应对，以有效地帮助有关单位快速发现舆情，及时收集到所需的社会网络舆情信息，从而帮助有关单位及时掌握舆情动向，对有较大影响的重要事件快速发现、快速处理，从正面引导舆论和宣传，构建积极向上的主流舆论，并为正确决策提供信息依据。

自动信息搜集功能可以解决人工无法应付海量网络信息收集的困难，自动信息搜集技术主要是通过网络页面之间的链接关系，从网上自动获取页面信息，并且随着链接不断向整个网络扩展，实现网络信息的自动搜集。高效、全方位的网络舆情采集，可最大限度地保证信息的时效性、可用性和全面性，从而为决策分析提供事实依据和数据参考。

数据清理功能用以对信息进行筛选，初步去除无价值的信息，根据不同的舆情调查主题，筛选保留下有价值的信息后对信息有序化处理，对于经筛选后保留下来的大量原始信息，按照其主题、外部形式或内容特征进行有序化处理，从不同的角度对网络舆情信息进行分类。通过聚类和分类功能可以对网络中的敏感话题、热点话题、给定时间段内的热门话题，进行识别，具体可以根据发言时间密集程度、跟帖数量、转贴数量和程度、新闻出处权威度、评论数量等不同参数，进行分类识别，从而实现网络信息的自动分类和聚类。在通过相似搜索对象集合或相关数据库时，可以找到与指定的查询对象相似的数据、对象实例或对象子集。消除掉重复的信息，保留原始出处的信息，消去大量转引的重复信息。此外，舆情信息检索结果可按不同维度展现，包括按内容分类、舆情分类、相关人物、相关机构、相关地区、正负面分类等。每个维度下把搜索结果自动分类统计展示信息，使用户用最短的时间搜索到最精确的信息。

在对上述数据挖掘技术进行分析后，本文设计了基于大数据技术的舆情信息分析方法来进行舆情分析，主要分为三个部分：“热点事件数据采集”、“舆情数据的情感倾向性分析”和“预警管理模块”。第一部分通过配置了 IP 池和 Cookie 池等的分布式爬虫实现。第二部分通过 LSTM 神经网络配合词向量实现。第三部分通过利用 Python 远程操作云数据库，对数据库中的内容进行关联挖掘，及时提取和分析敏感数据及关联网络，及时预警。舆情信息分析技术方案模块如图 1 所示。

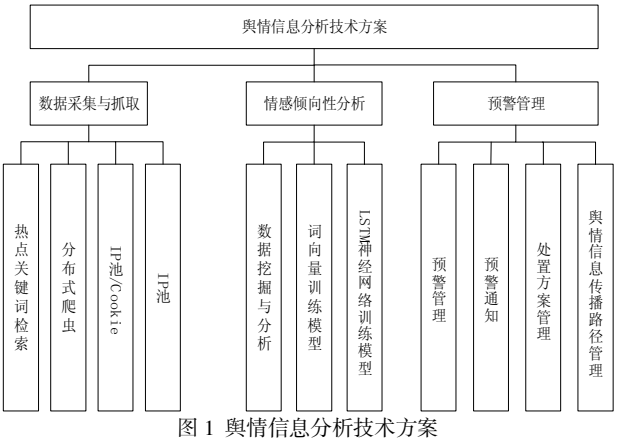

图 1 舆情信息分析技术方案

要进行精准的舆情预警，需要在大数据基础上建立有效的预测模型，才能准确服务好对网络舆情的预处理和快速反映及管理工作。因此，有效预测模型的建立是关键。本文利用数据挖掘后的有效信息数据训练出词向量，再将经过预处理和词向量化处理后的训练集放入 LSTM 神经网络中进行训练，得到情感分类模型。LSTM 神经网络训练模型流程如图 2 所示。

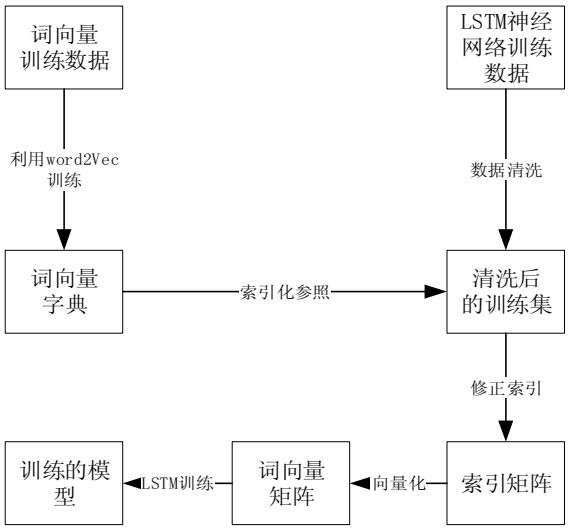

图 2 模型训练流程

通过使用已训练的成熟度模型，系统能够做到对热点事件的及时分析反馈和预警，为及时对舆情事件进行监督和管理实现起到了有效的支撑效用。

四、总结

舆情分析一直以来都是一个重要的话题，随着数据量的快速膨胀，□ 数据时代下的□ 络舆情□ 临各种风险和挑战，因此，基于大数据技术的舆情分析方法和提前预警机制是有效管理和快速解决舆情问题的关键，本文分析了□ 络舆情的现状与特点，给出了基于大数据技术的□ 络舆情分析的方法并构建了完整的舆情信息分析方案，为网络舆情研判，应对和及时处置提供技术支持和有效保障。

参考文献：

[1]赵方亮，刘德路，赵学军.大数据管理研究:概念,应用与挑战.《北方经贸》 2019 年第 11 期 134-136.

[2]刘毅.略论网络舆情的概念、特点、表达与传播[J].理论界,2007(01):11-12.

[3]孙培星. 基于情感倾向性的网络舆情分析及演化预测研究[D].吉林大学,2016.

[4]张科.大数据时代高职院校网络舆情分析及治理刍议[J].太原城市职业技术学院学报,2018(3).

[5]谌志华.基于大数据的网络舆情分析系统[J].《现代电子技术》2017 年第 24 期.

[6]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[7]Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du, Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL 2018.

[8]马哲坤，涂艳.基于知识图谱的网络舆情突发话题内容监测研究[J].情报科学,2019,37(02):33-39.

作者简介：姓名：郭琛 性别 ：男 籍贯：大连 民族：汉 出生年月：1979.4 学位：博士 职称：副教授研究方向：大数据，人工智能 单位：大连交通大学邮编：116021

基金项目：本文系 2019 年度辽宁省社科规划基金项目一般项目（310）研究成果.