基于大数据技术的网络舆情监测分析研究

郭琛[1]

（大连交通大学　辽宁大连　116021）

摘要：大数据时代的互联网成为了现实生活的有效反馈阵地，传统的舆情管理手段已不适应新时代舆情发展的变化，采用新技术分析方法，是对舆情进行有效管理的重要手段。本文从舆情数据的特点出发，介绍了舆情监测分析技术需要具备的功能和舆情分析采用的技术路线，为新时代基于大数据技术的网络舆情监测分析提供助力研究。

关键词：大数据，舆情监测

一、引言

大数据时代，互联网已逐渐成为热门话题和事件讨论的重要平台以及舆情事件的放大器，不论是热点实事还是娱乐八卦等我们身边充斥的各种信息，传播速度都远超我们的想象，现今数据开始转变为一种基础性资源，在极短时间内，就有数万计转发，数百万的阅读，信息传播极为迅速，如此海量信息的爆炸式传播，对社会舆情管理带来了压力。如何更好地利用和管理好大数据，如何能够实时的把握好舆情的方向并作出及时、合理应对，已经成为媒体及学术界普遍关心的话题，也是目前面临的一项重要和关键的工作。因此，无论是媒体还是学术界等都越来越重视运用大数据的方法对网络舆情进行分析，急需对传统舆情进行有效的新型管理，以充分发挥网络舆情信息的价值，为不断创新分析方法创造条件。因此，采用大数据技术分析管理的舆情采集系统恰逢其时。

“大数据”是为了从大容量的、不同类型的数据中获取有价值的信息而设计的新型架构和技术，是指具有更强的洞察力和流程优化能力的海量、多样化的信息。有学者从大数据的外延角度对其进行界定，如根据来源的不同将大数据分为：（1）来自网民在使用互联网以及移动互联网过程中所产生的各类数据；（2）来自各类计算机信息系统产生的数据；（3）来自各类数字设备所采集的数据等[1]。

网络舆情是指民众通过互联网针对自己所关心或与自身权益紧密相关的公共事件、社会现象等做出的主观反映，是多种态度、意见等交互的综合表现。网络舆情具有自由、情绪化、分散、即时、多变等特点，在一些社会热点问题上容易引发较为广泛的社会影响，尤其是负面的影响[1]。

二、舆情数据的特点

目前业界广泛认可大数据的特征为4V特征，即：大量（Volume）、多样（Variety）、高速（Velocity）、价值（Value）。通过对目前网络舆情状况的观察可以看出，互联网的开放性使网民可在网上更为方便地发表自己的意见，导致网络舆情的数据量急剧增长。其次，多媒体的发展使网络舆情的数据形态呈现出多媒体性的特征。再次，现代社会价值观念多元化，各家观点争鸣，舆论不断变化，导致网络舆情快速变化。正是由于以上各种因素的共同作用，使得网络舆情数据越来越呈现出大数据特征[1]。

海量数据的传播，对信息技术的发展提出了更高的要求。大数据使得舆情检测与分析的水平达到了个体级别，社会舆情的描述再也不是整体性的泛泛而谈，而是可以做到针对个体细节的即时跟踪；因此，大数据技术的使用成为了舆情研究的利器。

采用大数据技术处理舆情信息时，舆情信息数据存在以下特点[2]：

1. 数据来源多样，人们身边的各类信息均作为可来源的数据。

2. 数据类型复杂，数据形式不单一，有图片、文字、音频、视频等各种各样数据类型。

3. 数据信息零散，价值密度往往很低，数据类型庞大，数据基数很大，数据之间存在着封闭性与关系断裂性，使得在整理数据获得规律时存在失真或冗余的信息。

4. 网络大数据的时效性高，迭代更新速度快，信息价值随时间变化快速改变。

因此，传统的数据跟踪搜集分析手段已经无法应对现在快速的数据变化，这对数据分析的技术手段提出了更高的要求。

三、舆情监测分析技术需要具备的功能

当前常用的网络舆情分析方法主要有网络调查方法、基于内容挖掘的主题监测方法、基于统计规则的模式识别方法等。网络调查方法是指通过联机网络、计算机通讯和数字交互式媒体，在网络上进行数据收集传输、自动加工处理已实现某一研究目的的统计调查方法。它是传统调查技术与现代网络技术相结合的产物。目前这是应用最为广泛的方法，通常网站都会在相关新闻页面的下方设置新闻评论功能和读者态度倾向调查。在基于内容挖掘的主题监测方面，涉及较多与自然语言处理相关的研究领域，有学者提出监测流程分为 3 步：即网络舆情信息采集与预处理，文本表示与主题发现，网络舆情意见挖掘和观点分析。在基于统计规则的模式识别方面，有学者通过分析某段时间间隔内用户所关注信息点记录，构建了互联网内容与舆情的热点（热度）、焦点（焦度）、重点（重度）、频点（频度）、粘点（粘度）、敏点（敏度）、难点（难度）、拐　点（拐度）、疑点（疑度）和散点（散度）等 10 个分析模式和判据。以此为基础，市场上出现了许多网络舆情监测分析软件，但目前这些舆情监测系统擅长的是抓取新闻网页，在诸如 BBS、QQ 、微信群、博客、微博网络社区中等则效果有效，网络社区中的舆情依然主要依靠人工分析为多[1]。

而新的基于大数据的舆情监测分析技术能够及时发现预测网络舆情，通过构建网络舆情监测与分析系统，可以迅速、准确地了解民意、汇聚民智，及时发现和预测网络舆情，及时掌握舆情事件现状与发展态势，便于及时采取有效措施、解决问题、化解矛盾，避免影响人们的生活秩序和社会稳定，防止不良舆情的扩大[2]；此外，还能够进行自动实时监测，这是由于网络信息传播平台多、网

络信息量大、更新迭代快，靠传统监测手段已经无法满足到全网平台 24 小时监测的需求，因此，自动实现全网平台的追踪与分析，追踪舆情发展态势，分析挖掘全网平台舆情的传播来源、分析情况，即便于采取有针对性的舆情应对措施，还可以节约人力成本。

因此，基于网络舆情分析的技术需要能够实现以下几项功能：

1. 及时发现并关注相关舆情

能够实时分析全网数据，监测各网络平台，根据关键词的设置，及早发现与之相关的舆情。

2. 观点倾向性分析

对不同的观点、倾向性进行统计分析，及时跟踪负面舆情的信息来源、转载量、转载地址、地域分布、信息发布者，进行倾向性与趁势分析，能够很好的把握舆情传播路径，掌控舆情发展态势。

3. 深入分析舆情事件

4. 分析出不同时间段里人们的关注程度，预判未来的发展趋势，及时做好应对措施或可行计划方案。

5. 对突发事件、敏感事件及时预警、通知并给出处置建议方案，为正确引导舆论提供基础支持

6. 预警通知及处置能在第一时间掌握网络舆论动态，能够对关注事件或线索进行持续追踪和多维度分析，以便全面掌握舆情动态，为进行正确的舆论引导提供基础支撑。

四、舆情分析采用的技术路线

群体、媒介和内容是舆情的三个基本要素[2]。对于网络舆情来说，广大网民是网络舆情的传播者，同时也是受影响者；网络是舆情传播的媒体；内容则是被传播的主体。也正是由于网络舆情的传播媒介是网络，相比于传统的舆情，网络舆情具有传播速度快，信息量大的特点。因此在众多舆情分析中网络舆情的分析是最为重要的[3]。

在舆情分析这个方向上，我国的起步比较晚，并且作为一个交叉学科，其发展也不是一蹴而就。这其中经历了从“单一文本关键词提取与倾向分析”向“群体文本热点话题提取与倾向分析”转换并随着网络的普及和数据量的剧增而转向“网络舆情分析”[4]。从分析采用的技术来看，热点话题的发现一般采用聚类算法，其中 Single-Pass 算法因为其速度快，原理简单，尤为受到欢迎。此外，还有很多改进的 Single-Pass 算法来发现网络热点话题，如通过对 Web 文本不同位置特征项进行加权处理等方式，来实现仅需计算新文档与同类别种子文档间的相似度，达到降低漏检率和错检率的目的[5]；如设定唯一聚类质心、不断优化聚类中心方式进行改进等[6]。除 Single-pass 外，还有如融合演化特征的舆情分析方式以及基于知识图谱的舆情分析方式等更多新的分析方式的提出，这些方式都能对舆情的演化进程给出建设性的分析和描述，效果较好。

本文在现有研究基础上分析了基于大数据技术的设计路线。在网络数据的获取渠道中，通过搜索引擎和利用网络爬虫来对信息进行获取，特别是对舆论高发平台的深度获取方法；将待爬取的 URL 加入 URL 队列，Spider 执行具体的数据爬取任务。采集后对云数据库中的数据进行分析，利用 Word2Vec 工具训练词向量，分析热度新闻的各项属性，找出较高的热门评论，根据热点事件的各项属性，计算其热度值；将分析后的结果存入关系型云数据库中，利用 LSTM 神经网络配合词向量训练出情感分类模型，其中对情感分析训练集进行如下预处理：

①分词操作，即将一整句话分成若干个词语。使用的分词工具是“ltp 分词”。ltp 分词后的结果是一个生成器，需要手动转换成一个 list，这样可以方便后来的操作。

②去停用词，目的是减少计算量的同时提高模型的泛化能力。

③索引化，即将训练文本中的词语用词向量词典中对应词语的序列表示。

由于训练集中的文本长度不统一，索引化后长度也不同，为了方便模型的训练，需将所有文本的长度统一。

预处理后，利用训练好的情感分类模型对热点新闻下的热门评论进行分析，分析出网民对热点事件的情感倾向性，将基于训练后的成熟模型应用于舆情分析过程中，得出准确的情感倾向并根据实际进行及时预警，以便于及时应对和处置突然事件的发生，做到及时预警，及时预判，快速处置，取得成效。

五、结束语

随着大数据技术的不断改进以及网络舆情的不断发展，我们必须不断扩展网络舆情的内涵，不断革新舆情的分析方法，保障网络舆情大数据分析方法的可持续开展。这些都将是大数据时代下网络舆情分析的潮流和趋势。而基于大数据的网络舆情分析技术是对新时期网络舆情进行有效管理的重要手段，在运行当中需要收集大量的网络信息数据。对于网络舆情产生的随机性，传播的快速性等问题，对信息及时处理有很高的要求；因此，建立一个高效的网络舆情大数据分析环境，对网络信息进行及时的分析和管理，可以提高对舆情信息处理的工作效率，有效保障对网络舆情信息的管理和正确引导。

参考文献：

[1]燕道成，姜超.大数据时代网络舆情研究综述[J].视听 2015(9):133-136.

[2] 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.

[3] 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.

[4] 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.

[5] BENGIO Y，DELALLEAU O. On the expressive power of deep archi- tectures[C]/ /Proc of the 14th International Conference on Discovery Science. Berlin: Springer-Verlag，2011: 18-36.

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

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.