 DOI: https://doi.org/10.12345/bdai.v5i5.18053

Network Public Opinion Analysis System Based on Big Data Technology

Liying Wang

Guangzhou College of Technology and Business, Guangzhou, Guangdong, 510850, China

Abstract

With the wide application of China's Internet, the rapid development of mobile phone, wireless Internet access and other technologies, the modern Internet has become an important means of human information communication and cultural communication. The network is increasingly infiltrating into our daily life. At the same time, the daily data generated on the Internet grows in geometric times, making it more difficult to analyze public opinion, thus entering the era of big data and increasing the difficulty of public opinion monitoring. Moreover, online public opinion work is the most difficult work in the world, especially in China. If it is not strictly regulated, it may endanger social stability. This paper, first of all, analyzes the current domestic public opinion, and defines the concept and characteristics of big data. It plans to organically integrate public opinion analysis and big data technology to build a set of public opinion analysis system under the big data environment.

Keywords

big data; network public opinion; analysis system; design

基于大数据技术网络舆情分析系统

王丽颖

广州工商学院，中国·广东 广州 510850

摘　要

随着中国互联网的广泛应用，移动电话、无线上网等技术的飞速发展，现代互联网已经成为人类进行信息沟通、文化传播的重要手段。网络正日益渗透到我们的日常生活中,同时，互联网上每日生成的数据以几何倍数增长，使得对舆情进行分析变得更加困难，从而进入了大数据时代，也加大了舆情监测的难度。而且网络舆情工作是全球最困难的工作，特别是中国，如不严加管制，则可能危及社会安定。论文首先对当前国内舆情进行了分析，并对大数据的概念、特征进行了界定，拟将舆情分析和大数据技术有机地融合起来，构建一套大数据环境下的舆情分析体系。

关键词

大数据；网络舆情；分析体系；设计

1 大数据技术和网络舆情的概述

1.1 大数据技术综述

大数据技术的本质就是，数据的数量和规模都比以往要多得多，并且不能通过现有的一些软件来对这些数据进行整理和分析。通过运用大数据技术，政府部门和企业可以有效地对大数据进行分析，并由此制定出一种有效的商业开发规划与管理方式，大数据的规模大，类型多，价值低，密度低，流通速度快。在其他国家，Hadoop 系统已经具备了专门的数据处理技术所无法比拟的可靠、高效和可扩充性，这个平台包含了很多部件，比如多个存储节点，它们能够从一个节点中采集并处理来自各个节点的数据，很多高性价比的电脑都能根据系统的需求增加一个处理器节点。

1.2 网络舆情综述

网络舆情一般是互联网用户在面对社会热点新闻时，尤其是涉及到自身利益和国家政策等方面的情感和观点的一种反映。这类舆情带有某种偏向性，舆情是各种意见、态度的总和，其特征是：具有很强的广泛性，舆情能够在全国范围内很快地扩散开来，参与人的种类也很广；突发事件的特点：在一个区域内，一旦有突发事件发生，就能作为舆情的源头；主观方面表现为：舆情内容、舆情意见主观性强；多元化的特点表现为：参与人对信息的看法、观点的不同，以及对信息的传播、表达等方面的不同。随着时代的进步，舆情的数量与内容也逐年增多，如何及早识别并妥善处理舆

【基金项目】广州工商学院2022年校级科研项目（重点项目）《基于Single-Pass算法增量式聚类的网络舆情挖掘与分析研究》（项目编号：KYYB202228）。

【作者简介】王丽颖（1991-），女，硕士，讲师，从事人工智能、数据挖掘研究。

情，已成为当前政府部门必须面对的重大问题。

2 大数据技术下网络舆情分析系统的设计原则

在大数据的背景下，对舆情分析系统的要求如下。

2.1 信息抓取必须全面

当前，互联网舆情具有涉及面广、种类繁多、网页数目多等特征。所以，舆情的收集需要有能力收集来自各种结构化和半结构化的舆情，建立一种新的储存和网络爬虫器 [1]。

2.2 确保有效的分析

优秀的网络舆情分析软件能够帮助政府部门、企业对网络舆情中的热门主题、热点事件进行及时的追踪，从而对网络舆情作出及时的应对。当出现消极意见时，能够快速地采取恰当的措施来恢复和维持稳定。

2.3 结构设计

①采集模式。并通过网络爬虫技术实现对互联网上的站点的快速、大范围的抓取。系统会为建立表格和黑名单而指定 URL 收集点。基于特定需求，通过搜索、储存“白名单”和“黑名单”站点，可以有效地提高搜索效率，避免不必要的操作。

②信息预处理模式。它的工作是对采集来的原始页面进行转化，再将经拷贝、去噪后得到的格式文本存入数据库，抽取文字信息，对其进行分段，采用中文切分方法产生单项。在此基础上，利用矢量空间、概率等方法，对网络舆情进行特征抽取，形成面向舆情的文本矢量集合。

③舆情分析模式。这是舆情分析中的大数据，其主要作用是实现话题识别，文本趋势分析，热点挖掘等。该方法的基本思想是：利用话题识别技术对文本矢量进行解析、学习、聚集与相同事件有关的海量文本信息，从而发现主题，利用话题追踪方法，对向量化的文本进行检索，并通过相似度来判断话题的相关性，从而实现对新文本的分类。在此基础上，利用微机计算技术对文本中的观点、态度、情感及非事实等信息进行分析，获取恰当的文本语义，为监管部门和政府部门及时发现不良舆情提供依据。通过建立话题的统计，也能识别出一些重要的信息，例如：来源、评论的数量、演讲的时间。

④公众意见报告模式。负责向有关政府部门及监管部门发布调查结果，为其制定政策提供依据。

3 大数据技术下的网络舆情分析系统

3.1 系统功能框架

构建以大数据为基础的网络舆情分析系统。该系统采用分布式文件系统（Distributed File System，HDFS）对数据进行存储与管理。MapReduce 程序设计模型把大型的工作分割成多个小型的工作，这些工作在多台服务器上平行地运行。最后，利用分布式计算框架，归纳各子任务的效果，并进行情感评价与倾向性分析。

3.2 系统功能

3.2.1 数据收集功能

最基本的一个功能就是搜集各种网络上的舆情消息，比如微信、微博等等。在大数据背景下，在利用常规搜索引擎爬虫确保数据完整的基础上，通过有针对性的爬虫来提高信息收集的效率和准确性，还可以使用黑白名单等手段，将合法的链接保留下来，从而更好地实现新的搜索。在对网站进行信息抓取的过程中，首要的工作就是对网站中的文本内容和栏目列表进行采集，在搜集资料的过程中，为下一步的工作打下了良好的基础。

3.2.2 数据处理模块

互联网具有数据量大、时效快和信息多样化等特征，网页的信息结构与内容多种多样，包含文字资料、图片资料、声音资料及影像资料。由于所收集的有关来源数据无法达到直接的解析需求，因此必须对所收集的数据进行预处理与向量化，在互联网舆情的研究中，必须先对互联网舆情进行预处理，然后再进行向量化，对原始文本行集进行去重、去噪，再根据文字的特性选择对原文本集进行量化，从而获得一组文本集。由于所收集到的资料在结构上存在着不同的差别，因此有必要对这些不相干的界面资料进行整理，并且保存界面标题、内容摘要和发布时间等关键信息，在此过程中，为保证数据完整性，需要对遗漏的数据进行删补，论文拟采用 MapReduce、分词工具等方法，对已有格式的文本进行并行处理，抽取出具有代表性的词汇特征，并将其进行聚类，生成相应的文本矢量，并将其存入分布式 HDFS 文档构件中。

3.2.3 舆情数据采集技术

舆情信息收集的方法，就是确定舆情的议题，确定舆情的出发点，舆情数据采集是网络舆情分析中非常重要的一步，可以为以后的数据处理和数据分析提供依据。舆情数据收集技术可以从最初的 URL 中，获得里面的网页信息，之后，把这些网页信息保存在一个本地的系统里，并且通过对该网页的结构和内容的解析，提取这个网页的连接，然后将其转换为一个新的 URL。目前，主流的爬虫技术主要有：主题爬虫和增量爬虫等。每种爬行方式各有其特点与优势，选择爬行方式时应结合网络民意的实际情况 [2]。

3.2.4 舆情数据预处理技术

在使用网页爬行器来获取页面时，其实际的结构和实际的内容存在较大差异，导致大量的数据信息无法满足用户的需要。在这种情况下，为了确保对网络舆情进行有效的分析，必须对其进行预处理。舆情资料的预处理就是要对互联网上的舆情做好准备，排除噪音、重复等问题。例如，在词汇预处理方面，应该充分利用中文分词技术，对采集到的文字进行高效的切分，然后把它转化成包含各种词的集合，能有效地剔除语句中的停顿成分，并对语篇中不同词语的使用情况进行统计，使文本的预处理工作变得更加容易。

3.2.5 舆情智能分析技术

舆情智能化研究是当前互联网舆情研究领域的一个重

要研究方向，其研究内容包括：主题的识别和追踪，热点主题的发现，以及文本倾向性的挖掘。主题辨识则是运用预先设定好的数据向量集，再利用计算机自动完成主题识别。在一个文件中对同类事件进行总结，并确定它们之间的舆情主题，通过 Hadoop 对文本数据进行分块，并将其与核心文件一同传送至 map 函数完成相关运算。Map 可以将小块内的数据分布给离它很近的中间点，然后用键值对的方法，把它传送给 Reduce，这样就可以对计划进行平均化，得到一个新的聚类中心。主题追踪则是基于最近一次更新的向量文本的迅速查找，并通过相似度的计算来实现，并评价其与当前主题的相似程度，如果相似度满足要求，此类文字就会被归类为主题，如果不符合要求，它就会被分类成一个完全不同的主题。在此过程中，要对主题的评论数、转发量和赞数进行分析，再计算受欢迎指标，并与受欢迎指数进行综合，对热点主题进行筛选 [3]。

3.2.6 舆情预测预警技术

舆情预测与预警技术是对舆情进行智能化分析的结果。在对热点新闻、事件或舆情进行监控时，一旦用户到达设置的告警数值，将自动产生相关舆情报告，并以电邮或短信方式向工作人员发送，得到通知，员工就能高效的解决这一问题。

3.2.7 报告应用模块

报告应用模块以对舆情信息资料进行推送，并对其进行智能化的分析，以文本、图表、报告的形式，实现舆情综述、舆情风险监控、热点事件分析、大风险预警可视化等功能。其中，舆情概略完成了用户监控、本地相关信息、本地负面信息、预警信息推送以及本地媒体量的统计展示。舆情风险监控按照主题的热门程度将其划分为不同的类别，并将事件的标题、原始链接、来源、已发布事件、热度值以及发展趋势等信息显示出来，并将其与历史上相同的重要事件或者国家范围内的相似事件进行对比和展示，并通过对其热度与趋势的分析，识别舆情风险。

3.3 系统的设计

在特定模型中，应该考虑如下问题：

①信息的采集必须保证完整。目前，网络舆情呈现出规模大，类型多，网站多的特征。所以，在实际的建模过程中，要充分地收集数据是非常必要的。确保网页、声音、图片、结构化或半结构化的资料，以及最佳储存及网络爬行器优化。

②确保快速响应公众舆情。今天，随着网络的迅猛发展，在互联网上，舆情呈现出了一种灵活多变的特点，而常规的舆情分析模式，已不再适用于目前的情况。特别是在现今的发展环境下，网页与图像中所隐含的信息难以被及时地发现与捕捉，进而，在建模时，需要充分考虑用户对舆情的响应能力，并对隐含的信息进行有效搜集，从而达到准确的网络舆情引导。

③保证分析结果的正确性。一个高效的网络舆情分析软件能够帮助企业及主管部门迅速了解到网络上的热门主题及相关信息，通过这种方式，既能在舆情爆发前得到有效的遏制，又能使舆情的影响力得到最大限度的发挥。因此，在进行模式设计时，必须对舆情信息进行充分了解，才能保证模式的合理性。

4 基于大数据的互联网舆情分析体系的不足之处

在一个具体的应用中，下面三条必须被遵循：

①资料完备性。大数据环境下，互联网舆情具有规模巨大、数量庞大、内容多样等特征，特别是文本、图像、音频、视频等多种应用场景下，因种种因素而未能被及时发现和捕获的隐性信息，为此，如何根据实际情况，不断地创新和优化存储和网络爬虫技术，建立科学合理的数据抓取方法，在保证数据安全的前提下，高效地获得隐藏信息。

②警告的准确性。针对现有的基于数据处理生成的预警系统，特别是在网络舆情的传递中，有许多人工合成的谣言，很难分辨真伪，这时候就必须进行人工干预和判断。

③合理的解析，一个高效、智能化的网络舆情分析工具，能够帮助用户在网络舆情爆发前，高效地对网络上纷繁复杂的信息进行分析，从而发现热点、追踪发展趋势，实现对舆情的科学、有效的管控。所以，在对其进行分析时，要特别注意算法的选取 [4]。

5 结语

大数据本身的特点，使得大数据具有很强的应用价值，大数据能够提供更加精确和全面的信息，能够挖掘出之前没有被挖掘出来的信息的规律与联系，从而帮助使用者快速作出决定。以大数据为基础的互联网舆情分析系统，在对舆情进行预警、预测和报告等方面有着独到的特色，体现出大数据的独特优势。它能有效地采集、处理、分析、展示大量的文本、图像、音频、视频等数据，对互联网舆情的处理起到了至关重要的作用。当然，在具体的实施过程中，必须对数据的准确性、及时性、完整性等方面进行全面的考虑与优化。

参考文献

[1] 代青松,李泽华.基于大数据技术网络舆情分析系统[J].电脑编程技巧与维护,2024(3):72-75.

[2] 陈怡菲,李雨静,肖金兰,等.基于大数据技术网络舆情分析系统[J].网络安全技术与应用,2023(9):62-63.

[3] 张晓飞.基于大数据技术的网络舆情分析系统研究[J].无线互联科技,2021,18(2):17-18.

[4] 孙昊.大数据技术下的网络舆情分析系统研究[J].自动化与仪器仪表,2018(8):26-28.