文章编号：2095-2163(2020)08-0287-03 中图分类号：G202 文献标志码：A

基于大数据的校园舆情热点话题跟踪研究

骆梅柳
（江苏财会职业学院 信息系，江苏 连云港 222061）

摘 要：当今互联网中的热点话题直接反映了师生的舆情动态，成为学校管理和监督的一个重要环节。因此，话题跟踪研究受到更多的关注，诞生了话题跟踪技术。该技术的提出是为了对海量的互联网资源中未知话题的识别和对已知话题的持续性跟踪。但是，网络数据量的增长速度快，传统的话题跟踪研究面对大规模数据时出现了技术上的瓶颈。为了更好提高话题跟踪检测的效率和准确率，本文将大数据的处理技术引入到话题跟踪研究中，通过对师生访问微博数据的采集，针对采集的数据实现话题跟踪研究，及时把握舆情动态，实现突发事件的预警。

关键词：语义网络；权重计算；话题跟踪

Tracking research on Hot topics of Campus Public opinion based on big data

LUO Meiliu
（Department of information, Jiangsu College of Finance& Accouting, Lianyungang, 222062, Jiangsu, China）

【Abstract】 Hot topics on the Internet today directly reflect the public opinions of teachers and students, and become an important part of school management and supervision. Therefore, opic tracking research has attracted more attention, and topic tracking technology has been born. This technology is proposed for the identification of unknown topics in the vast Internet resources and the continuous tracking of known topics. However, with the rapid growth of network data volume, the traditional topic tracking research faces a technical bottleneck in the face of large-scale data. In order to better improve the efficiency and accuracy of topic tracking detection, this paper introduces the processing technology of big data into topic tracking research. By collecting the data of teachers and students´ visits to weibo, the collected data can be used to realize topic tracking research, timely grasp the public opinion dynamics, and realize the early warning of emergencies.

【Key words】 Semantic network; Weight calculation; Topic tracking

0 引 言

近年来，使用网络新闻、微博、微信公众号等平台，以文本类信息为主要载体的评论发布越来越多。通过网络发布意见和评论形成网络热门话题，这些热门话题有生活中的小事，也有国家发生的大事。由于每个人都可以发表自己的观点，当出现一些负面的消息，会对社会稳定和谐发展产生威胁[1]。由于互联网已经深入到各个领域，校园中同样会受到互联网的影响。当学校出现一些负面的舆论时，可能会影响到师生们的情绪，甚至会产生一些突发事件。若在学校对热门话题实现跟踪和检测，则可以帮组学校掌握网络热门话题的变化动态，做一些预先判断，将可能发生的网络热门话题抑制在萌芽状态，预防突发事件的发生，保障学校的安全稳定[2]。

大数据技术的出现加速了互联网的发展，它在各个领域中得到了广泛的应用。使用大数据技术可以处理复杂的数据，从而获取有价值的信息和知识。在大数据背景下，如何快速精准捕获到最受关注的校园内热点信息，成为校园舆情发展的核心问题。因此，发现并分析校园里的热点话题，为学校提供有价值的信息，使学校快速掌握舆情的变化动态，对一些突发事件做到预判。随着网络数据呈现爆发式的增长，使校园的舆情监控受到了极大的挑战，传统的话题检测与跟踪技术在面对爆发式增长的数据出现了处理能力的瓶颈。本文将话题跟踪算法进行了改进，并使用大数据技术实现话题并行化处理。

1 国内外研究现状

话题追踪是对文本实现聚类，它不同于传统的聚类方式，由于数据不断的更新而成为一种实时的聚类方式，数据会不断的增加和更新。卡内基梅隆大学(CMU)通过 KNN 和决策树方法进行基于文本的新闻主题相关事件追踪，分两个阶段解决，与传统单层方法相比效果显著提高[3]；James Allan 提出了 Signle-Pass 算法对新的话题进行检测，该算法简单

基金项目：2018 年江苏省高校哲学社会科学研究基金(2018SJA2019)。

作者简介：骆梅柳(1983-)，女，硕士，讲师，主要研究方向：大数据、复杂网络。

收稿日期：2020-06-24

且容易实现,它使得新检测到的文本与最近时间得到的问题建立联系[4]。文献[5]对 K-means 算法中初始中心点不确定的问题进行改进,提出改进后的算法 IIKM。文献[6]中提出基于 LDA 模型的热点话题跟踪,该模型只选择热点话题进行先验传递,并通过设置同一话题相邻时间片的语义距离来判断话题的状态;文献[7]通过融合用户关系的自适应微博话题跟踪方法,利用改进的 K-means 聚类算法对候选推文集合进行二元聚类,从而划分出相关推文集合,即当前话题目标模型;文献[8]在向量空间模型(VSM)的基础是上提出一种基于话题更新的跟踪算法。

在数据处理技术方面,由于数据的增长速度较快,传统串行话题跟踪方式在处理海量数据时出现处理速度不够的情况。文献[9]中使用 Canopy 算法进行初始化操作,加入传统算法 K-means 对结果不断进行更新,最终引入 Hadoop 技术,构成基于 Hadoop 的融合 Canopy 和 K-means 话题跟踪算法。文献[10]中对 Single-Pass 聚类算法做了改进,设定并不断优化聚类中心,最后形成基于 Hadoop 对中文分词、特征提取和聚类分析进行 MapReduce 优化。以上的研究在一定程度上加快了数据的处理速度。

终上所述,国内外对话题跟踪的研究主要体现在研究话题跟踪的准确性。但近几年数据呈现爆炸式的增长.传统方式已经不能够处理大量的数据,也不能仅仅改进算法的准确度,而将大数据技术引入到话题跟踪中来成为目前研究的重点。本文将以爬取校园师生的微博作为数据来源,针对传统话题跟踪模型存在误差大的弊端,提出改进神经网络的话题跟踪模型,并将改进的模型在 spark 平台上实现并行化处理。

2 相关理论与技术

2.1 文本预处理技术

本文研究的数据,来源于互联网中的微博纯文本数据。由于计算机无法自动地识别文本,因此在对校园舆情热点话题进行跟踪时,需要对文本进行预处理。文本预处理主要包含 3 个步骤:

(1)中文分词处理。将文本内容切分成字、词或者短语,分词方式分别为基于理解、基于统计和基于字符串匹配。

(2)特征选择。通过中文分词的微博文本中包含着无数没有意义的词汇,在文档建模前需要对分词结果进行选择,选取具有代表性的特征词。

(3)构建文档模型。经过前两个步骤后,文档由一系列具有特征的词构成,进而需要构建文本模型,将文本处理成以数字呈现的形式,计算机才能进行识别。

2.2 BP 神经网络

BP 神经网络是最传统的神经网络,它是一种多层的前馈神经网络。BP 神经网络的主要特点是信号是前向传播的,而误差是反向传播的。BP 神经网络由很多神经元组成,它们相互连接构成一定的拓扑结构,最常用的结构包含输入层,隐藏层和输出层。BP 神经网络的训练过程主要包含两个阶段:第一阶段是信号的前向传播,从输入层经过隐含层,最后到达输出层;第二阶段是误差的反向传播,从输出层到隐含层,最后到输入层,依次调节隐含层到输出层的权重和偏置,输入层到隐含层的权重和偏置。改进的 BP 神经网络,就是对神经网络的连接权值和偏置值不断进行优化更新,以达到话题追踪的最优值。

2.3 大数据处理技术

在校园中,每天都会产生大量微博数据,这些数据中蕴藏着很多有价值的信息,如何高效率的从这些数据中进行热点话题追踪成为目前研究热点。目前大数据处理平台应用较多的主要有 Hadoop 平台和 Spark 平台,其中 Spark 是 UC Berkeley AMP lab 所开源的类 Hadoop)Map-Reduce 的通用并行计算框架[10]。Spark 是在 Map-Reduce 基础上实现的分布式计算,不同的是它在每次迭代计算后,并不是将结果保存到文件系统中,而是保存到内存中,这种工作机制让它在数据挖掘中体现出更加优越的性能。

3 改进的神经网络热点话题跟踪模型

构建校园舆情热点话题进行建模之前,首先要从舆情热点话题中提取出反映话题重要信息的特征,当前选择的方式是使用分词和权重表示特征,通过 TF-IDF 完成权重值的计算。通过不断优化权重参数,最终获取最佳的检测效果。具体实现步骤如下:

(1)根据计算的权重参数对校园热门话题特征进行排序,选择出排名靠前的特征描述。

(2)选择出的特征作为神经网络的输入,微博数作为神经网络的输出,构建学习样本。

(3)使用模拟退化算法对权重值进行优化,使 BP 神经网络的网络热门话题检测训练误差根据实际情况达到最优值。

(4)当校园热点话题检测的训练误差达到预期值时,训练完成并确定最终的校园热点话题跟踪模

型。

改进的工作流程如图 1 所示。

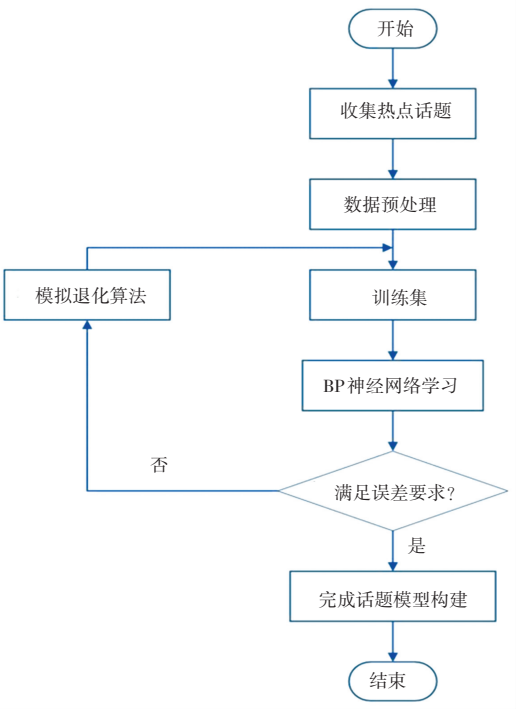

图 1　改进神经网络的校园热点话题跟踪模型流程图

Fig. 1　Flow chart of hot topic tracking model on campus with improved neural network

4　基于 Spark 的神经网络模型并行化实现

校园师生的微博话题数量增加快，如何快速从这些海量信息中准确的获取有价值的信息，可以利用 Spark 的内存计算，实现话题跟踪模型的并行化，加快模型训练阶段的训练速度，提高后续新加入的处理速度，提高校园舆情话题跟踪效率。

基于 Spark 的神经网络模型并行化的实现步骤如下：

(1)通过 textFile 算子从 HFDSZ 中读取初始采集的训练数据集并实现格式转换，接下来通过 Cache 算子保存到 RDD 内存中，最终形成初始化数据。

(2)根据 partion 算子要求将数据随机划分为若干数据模块，每个模块有一个分区与其对应。对每个分区使用 mapPartitions 算子进行改进后的神经网络热点话题跟踪模型训练。

(3)完成第一层训练后，通过 repartition 算子对已经保存的数据实现整合，将整合的数据作为第二层的输入，重复执行步骤(2)操作。

(4)通过步骤(2)和步骤(3)最终实现全局的训练模型。

5　结束语

校园舆情热点话题追踪在校园内具有很重要的价值，为学校安全提供了预警机制。本文在话题模型构建上对传统的神经网络模型做了改进，利用模拟退化算法对权重值不断优化，提高检测的准确度。同时，为了快速的分析校园内舆情热点话题发展趋势，本文在改进的训练模型上引入 Spark，将构建的训练并行化实现。将改进模型在 Spark 平台上实现并行化运算，不仅可以提高校园舆情热点话题跟踪的准确性，而且加快了话题跟踪的运算效率。

参考文献

[1] 洪宇，张宇，刘挺，等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，21(6)：71-87.

[2] 周亚东，孙钦东，管晓宏，等. 流量内容词语相关度的网络热点话题提取[J]. 西安交通大学学报，2007，41(10)1142-1145.

[3] YANG Y，ZHANG J，CARBONELL J，et al. Topic-conditioned novelty detection [C]// Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2002：688-693.

[4] R.Papka，J. Allan. On-line new event detection using single pass clustering[J]. University of Massachusetts，Amherst，1998：37-45.

[5] 雷震，吴玲达，雷蕾，等. 初始化类中心的增量 K 均值法及其在新闻事件探测中的应用[J]. 情报学报，2006，25(3)：289-295.

[6] 陈兴蜀，高悦，江浩，等. 基于 OLDA 的热点话题演化跟踪模型[J]. 华南理工大学学报(自然科学版)，2016，44(5)：130-136.

[7] 柏文言，张闯，徐克付，等. 一种融合用户关系的自适应微博话题跟踪方法[J]. 电子学报，2017，45(6)：1375-1381.

[8] 夏春艳，崔广才，李树平. 话题跟踪方法的研究[J]. 计算机工程与应用，2012，48(15)：129-132.

[9] 李晓瑜，俞丽颖，雷航，等. 一种 K-means 改进算法的并行化实现与应用[J]. 电子科技大学学报，2017，46(1)：61-68.

[10] 王宁，黄敏. 基于 MapReduce 与两层相关性聚类的实体解析方法[J]. 计算机工程，2015，41(9)：81-85.

[11] ZAHARIA M，CHOWDHUIY M，FRANKLIN M J，et al. Spark：cluster computing with working sets[C]//UsenixConference on Hot Topics in Cloud Computing. USENIX Association，2010：1765-1773.