Computer Science and Application 计算机科学与应用, 2019, 9(11), 2045-2051
Published Online November 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.911230

Hans汉斯

Topic Tracking Model and Algorithms for Drug Safety Public Opinion Based on Hadoop

Wenxue Zhang[1], Ying Wang[1], Jing Xu[2]

[1]School of Sciences, Ningxia Medical University, Yinchuan Ningxia
[2]Public Administration Research Center of Ningxia Medical University, Yinchuan Ningxia
Email: wxzhang@163.com

Received: Oct. 25th, 2019; accepted: Nov. 8th, 2019; published: Nov. 15th, 2019

Abstract

Track public opinion of drug safety and from online media news is beneficial to the health departments and pharmaceutical companies judging public opinion and making decisions quickly, accurately and efficiently. The medical news from December 25, 2012 to April 29, 2015 from a medical network is obtained by the Octopus collector. After data cleaning and manual selection, the experiments data are 5667 medical news including 8 categories of drug safety topics. The Hadoop platform and Naive Bayes classification algorithm are used to track drug safety topics. The research results show that the Naive Bayes classification algorithm based on Hadoop platform has better accuracy, poor recall rate and the best overall model when the F1 value is 0.57.

Keywords

Drug Safety and Public Opinion, Topic Tracking, Hadoop, Naive Bayes

基于Hadoop的药品安全舆情的话题跟踪模型与算法

张文学[1]，王　莹[1]，徐　静[2]

[1]宁夏医科大学理学院，宁夏　银川
[2]宁夏医科大学公共管理研究中心，宁夏　银川
Email: wxzhang@163.com

收稿日期：2019年10月25日；录用日期：2019年11月8日；发布日期：2019年11月15日

文章引用: 张文学, 王莹, 徐静. 基于 Hadoop 的药品安全舆情的话题跟踪模型与算法[J]. 计算机科学与应用, 2019, 9(11): 2045-2051. DOI: 10.12677/csa.2019.911230

摘　要

大数据时代如何从网络媒体发布的药品安全事件、药品安全监管及药品安全形势等医药新闻报道中跟踪药品安全舆情，是卫生部门和医药企业研判舆情的关键。本文利用八爪鱼采集器从某医药网获取2012年12月25日到2015年4月29日间9888条医药新闻，经数据清洗和人工识别选取了8类药品安全领域话题，共5667例实验数据。采用Hadoop平台和朴素贝叶斯分类算法实现药品安全话题跟踪。研究结果表明基于Hadoop平台的朴素贝叶斯分类算法的准确率较好、召回率较差、调和平均指标F1值为0.57时模型整体最佳。

关键词

药品安全舆情，话题跟踪，Hadoop，朴素贝叶斯

1. 引言

近年来，与药品有关的事件越来越多，如毒胶囊、问题疫苗、假药等。这些药品事件的事实通过各种媒体报道在网络上飞速传开，引起了越来越多的人的了解和关注，产生了许多不良的舆论。药品安全事件与人们的身体健康密切相关。大众对此关注度高，容易产生消极的非理性情绪。此外，信息传播在时空上呈现出不断变化的复杂趋势[1]，不好的情绪的叠加可能导致更严重的社会危机。有报告指出，近90%的药品安全舆情事件均未取得令人满意的效果[2]，因此，迫切需要对大量的药品舆论信息进行处理并获得敏感信息，达到舆论信息的抓取、解析、观测、示警和跟踪。由于数据太大、信息不集中、数据结构不明显等特点，导致以前的服务器在处理数据上效果不理想，无法有效达到对药品安全舆论分类的目的。Hadoop 技术飞速的发展，提高了对大量数据进行处理的效率。为此，本文研究了基于 Hadoop 的药物安全舆论主题跟踪模型和算法，实现了对药物安全舆论的快速分类，加快对舆论分类的能力，对以后的研究提供方向[3]。

2. 研究现状

刘晓欣[4]在我国药品舆情监管系统数据发掘技术的基础上，又通过采用文献法、基于大数据系统抓取相关数据和案例进行药品网络舆情研究，得出了药品网络舆情特点、对应的引导策略，对政府部门后续处理、媒体报道、企业发展等具有重要的研究意义的结论。

李艳业[5]借鉴危机传播和危机发展理论以及危机处理技术，整理出问题疫苗发展的四阶段，针对具有明显摩擦的对话研究其可能会爆发的时间和持续的时间进行了具体的研究，得出对完善和优化对危机事件传播管控的应对措施和建议：政府需要关注大众需求，对于危机做到良好的修复监管；各类媒体要控制好传播尺度，相辅相成，共同建立良好社会环境；涉医人员要担负起对社会的职责，传播科学的舆论进行引导；基层公众要提高媒体素养，共同抵制舆论的传播的结论。

刘能燕[6]将大数据的相关技术引入政府的舆情监管中，并且以我们国家相关部门采用大数据技术针对舆情进行管理的案例作研究，归纳其对舆情管理的途径，并对当今时代大数据在完善政府职能部门对

于舆情管理的实现途径进行探索研究，得出了现阶段政府职能部门对于舆情管理提出的建议具有可执行性的结论。

李纲等[7]采用文献调研的方式整理了我们国家对于突发事件所采取的预警工作流程，依据工作流程归纳出预警各环节所采用的相关方法的硕果，并针对我们国家政府对于突发事件的监管和预测其带来的不良影响做了研究，得出了已经发生突发事件归纳预警的重点，指出其中可能存在缺陷与后续研究方向，希望可以为后续对突发事件进行预警研究提供参考的结论。

张雄宝[8]提出了一种针对微博突发事件的地域进行分析检测的方法，针对其的传播规律做了研究，结论是：与微博突发事件相关的微博文件分布范围将随着微博突发事件的发展而变化，并将逐步由小到大，最终逐渐缩小。

杨柳[9]针对突发性话题的特征词的选取和聚类以及检测等相关技术进行了研究，提出了图结构模型、突发特征及重要词的研究方法，得出了可以采用聚类和对突发事件进行话题检测的方式选出对应的突发事件的话题，检测结果表明该方法的准确率可达到 85%，召回率可达到 75%，与其他方法相比，其的平均调和值也上升了 14%的结论。

通过上述分析发现关于药品安全舆情的话题跟踪模型与算法已经有很多可行的理论研究，如表 1 所示，但是对基于 Hadoop 的药品安全舆情的话题跟踪模型与算法的研究甚少，在本文中，借助各位学者之前的研究，继续探讨基于 Hadoop 的药品安全舆情的话题跟踪模型与算法。

Table 1. Relevant research literature on drug sensation topic tracking
表 1. 药品舆情话题跟踪相关研究文献

研究点	研究内容	参考文献
研究分析	采用文献法、基于大数据系统抓取相关数据和案例进行药品网络舆情研究	[4]
	完善政府舆情管理实现路径、现有突发事件预警研究重点	[6]、[7]
应对策略	政府需要关注大众需求，对于危机做到良好的修复监管；各类媒体要控制好传播尺度，相辅相成，共同建立良好社会环境；涉医人员要担负起对社会的职责，传播科学的舆论进行引导；基层公众要提高媒体素养，共同抵制舆论的传播	[5]
研究方法	图结构模型、突发特征及重要词的研究方法	[9]
	针对微博突发事件的地域进行分析检测的方法	[8]

3. 基于 Hadoop 和朴素贝叶斯的话题跟踪模型

3.1. 话题跟踪基本概念

话题跟踪(Topic tracking)技术可以帮助人们在社交媒体发布的众多的信息中鉴别出已知话题并对其不断地、准确的跟踪，提供人们所需的信息。它主要包括下述概念：话题，是指某个事件的核心内容或与之关联的事件。事件，是指由特定的原因和条件引发的，包含人物，时间以及地点等。报道，对于某时间段的相关报道，过去主要是指未提及的新闻报道，延伸到社交媒体和社交信息，例如一篇文章可被视为一篇报道。话题检测与跟踪技术是一项非常全面的技术，它要求多种技术相结合，而话题跟踪就是关键。

3.2. 朴素贝叶斯分类

朴素贝叶斯分类(Naïve Bayes, NB)算法十分简单，容易实现；朴素贝叶斯模型分类效率稳定；能处理好小型的数据，它可以同时处理多个分类，还可以增加数据进行训练；对于缺少的数据敏锐度不高，它

经常被用于文本分类。采用朴素贝叶斯(NB)分类算法作为药品安全舆情话题的分类算法，对抓取的数据经过预处理之后将文本依据不同的类别分好类，然后将分好类的部分文本输入，训练话题模型，最后利用模型判定分好类的测试文本的类别。

对于药品安全舆情信息，可以通过朴素贝叶斯(NB)分类算法的步骤一步一步实现药品安全舆情文档的分类，进而实现药品安全舆情话题的跟踪。但是由于药品安全舆情信息的数据量太大，仅仅依靠服务器处理难以满足药品安全舆情信息的跟踪。

3.3. 基于 Hadoop 的朴素贝叶斯分类

Hadoop 作为一个能够处理大量数据的平台就出现了，它通过对药品安全舆情信息进行存储和处理，能够有效的提高数据处理效果。本文的编程模型采用了 MapReduce，可以更有效的提高朴素贝叶斯分类算法对药品安全舆情的分类能力。算法中的 MapReduce，其包含 Map 和 Reduce，首先利用 Map 函数主要是对原始数据进行清洗操作，然后再利用 Reduce 进行数据加工，实现数据合并，获得分类结果；因此可以减少数据移动，并提高算法的处理速度。本文通过其构造基于 Hadoop 的药品安全舆情话题跟踪模型与算法，完成对药品安全舆情信息的跟踪。根据统计药品安全舆情文档中词语出现的概率、各个类别文档的数量等信息输入分类算法进行训练，构建模型分类器，达到药品安全舆情信息的分类处理的目的。

4. 基于 NB 和 Hadoop 的话题跟踪算法

4.1. 计算先验概率和条件概率

本文的朴素贝叶斯算法是通过计算先验概率和条件概率实现对药品安全舆情话题的分类，算法中计算先验概率是通过计算每个类的文档在总训练类中占的比例，即先验概率 P (a) = 类 a 下文档总数/整个用来训练模型的样本的文档总数；计算条件概率是通过计算每类文档中各词语在文档中出现的次数统计加一之和在每类中词语的总统计数与不重复的词语的总统计数之和中占的比例。具体过程如下所示：

Step 1：DocOfClassMap：获取训练模型时统计的训练集的样本数量和每个类下单词的数量，输入处理后的测试数据，测试数据格式<<class doc>, word1 word2 ...>，构建新的<类别，概率>键值对。

Step 2：DocOfClassReduce：对数据进行循环遍历并进行输出，对第一次循环先进行赋值，然后进行比较，当期概率更大时进行更新。

Step 3：GetPriorProbably (String docNum)：根据上述步骤结果计算先验概率，该静态函数计算每个类的文档在总类中占的比例，即先验概率 P(a) = 类 a 下文档总数/整个用来训练模型的样本的文档总数。

Step 4：GetConditionProbably (String wordCount)：根据上述步骤结果计算条件概率，条件概率通过计算每类文档中各词语在文档中出现的次数统计加一之和在每类中词语的总统计数与不重复的词语的总统计数之和中占的比例。

4.2. 算法流程

算法执行过程如下：

Step 1：执行命令。

```
cd /usr/local/run
hadoop fs-rmr/jobs/naive_bayes
hadoop fs-mkdir-p/jobs/naive_bayes/input/data
#进入本地 Excel 文件所在目录，上传 Excel 文件到 hdfs
hadoop fs-put 8 个类别.xlsx/jobs/naive_bayes/input/data/
```

hadoop fs-mkdi-p/jobs/naive_bayes/output

Step 2：预处理数据。

ParseData.parseTrainDataset (dataOrg, trainData)；

ParseData.parseValidateDataset (predictOrg, predictData)；

Step 2.1：parseTrainDataset 解析训练集。

Step 2.2：parseValidateDataset 解析验证集。

Step 3：在 Hadoop 平台利用数据训练的过程，并输出模型。

BayesClassify.main (new String[]{trainData, docNum, wordCount})；

Step 3.1：DocNums_Map：它继承 Mapper，主要实现对训练集数据各类别名的读写；DocNums_Reduce：继承 Reducer，成为自定义的 Reducer，主要业务逻辑就是复写其中的 reduce 函数，实现对训练集各类别下文档遍历并统计每个类对应的文件数量。

Step 3.2：WordCount_Map：继承 Mapper，实现对训练集数据各词语的读写；WordCount_Reduce：继承 Reducer，实现对训练集各类别下词语遍历并统计每个类下词语数量。

Step 4：用输出的模型对测试集文档进行分类测试，输出每个测试文档的分类结果。

Step 4.1：WordMapper：继承 Mapper，实现对测试集数据各文档和词语的读写；WordReducer：继承 Reducer，实现对测试集数据各类别下词语遍历并统计每个类下词语数量。

Step 4.2：DocOfClassMap：输入处理测试数据，测试数据格式<<class_doc>, word1 word2 ...>，设置 HashMap < String, Double > classProbably 为先验概率；HashMap < String, Double > wordsProbably 为条件概率。DocOfClassReduce：循环遍历输出，当概率更大时就更新 tempClass 和 tempProbably。

Step 4.3：GetPriorProbably (String docNum)：计算先验概率，该静态函数计算每个类的文档在总类中占的比例。

Step 4.4：GetConditionProbably (String wordCount)：计算条件概率。

Step 5：利用测试文档的真实类别，计算评价指标。

Step 5.1：OriginalDocOfClassMap：读写原本的文档分类，得到文档分类的类别，Reduce：计算得出初始情况下各个类有哪些文档。

Step 5.2：ClassifiedDocOfClassMap：读取经贝叶斯分类器分类后的结果文档<Doc, ClassName 概率>，并将其转化为<ClassName, Doc>形式，Reduce：计算得出经贝叶斯分类后各个类有哪些文档。

Step 5.3：GetEvaluation：利用所选评价指标评价算法。

5. 实验设计与结果分析

5.1. 实验设计

实验环境。本文在 PC 上搭建整体测试环境，以保证后续系统测试的顺利进行。实验硬件配置：Intel (R) Core (TM) i5-5200U CPU@2.20GHz 2.19GHz，内存 12.0 GB，外部存储硬盘 500 G。实验软件环境：操作系统 Windows8、centos7.3_1，Java 环境 JDK1.8，开发工具 eclipse，Hadoop 版本 hadoop2.7.6。

评测机制。在文本分类中，评估是一个必要的工作，本文采用精确率(Precision)、召回率(Recall)和 F1 来评估话题跟踪的性能，其具体定义为：

Precision：P = TP/(TP + FP)。

Recall：R = TP/(TP + FN)。

P 和 R 的调和平均：F1 = 2PR/(P + R)。

TP 是指统计初始情况下的分类和贝叶斯分类两种情况下各个类公有的文档数目(即针对各个类分类

正确的文档数目)；FN 是指初始情况下的各个类总数目减去结果正确的数目；FP 是指贝叶斯分类得到的各个类的总数目减去结果正确的数目。

5.2. 实验数据

数据集：本文利用八爪鱼采集器从某医药网获取 9888 条医药新闻作为分析数据源，后续的实验数据均由原始数据处理后得到。数据预处理：在数据集中以手工标记的方式从原始数据集中选取 8 类药品安全领域数据共 5667 例作为后续实验数据。手工标记的实验数据类别分布情况如表 2 所示。并利用中文分词技术进行文本的预处理，对标点符号及无意义虚词的剔除，最终形成实验使用的语料库。

Table 2. The experimental data categories of manually labeled
表 2. 手工标记的实验数据类别

序号	记录数	类别
话题 1	377	药品研发
话题 2	500	药品安全
话题 3	511	药品供应
话题 4	635	药品销售
话题 5	936	药品质量
话题 6	1073	疫苗
话题 7	1681	药品生产
话题 8	54	药品法

5.3. 实验及结果

将上述数据集作为输入测试基于 NB 和 Hadoop 的话题跟踪算法，其实验结果如表 3 所示：

Table 3. The test results of topic tracking based on NB and Hadoop algorithm
表 3. 基于 NB 和 Hadoop 的话题跟踪算法测试结果

序号	话题	Precision (%)	Recall (%)	F1 (%)
话题 1	药品研发	54.0	42.6	47.6
话题 2	药品安全	37.3	56.2	44.9
话题 3	药品供应	66.6	42.3	51.8
话题 4	药品销售	56.1	59.9	57.9
话题 5	药品质量	21.6	52.6	30.7
话题 6	疫苗	95.0	3.5	68.3
话题 7	药品生产	14.5	85.1	24.7
话题 8	药品法	92.4	19.1	31.6
average		54.7	45.2	37.0

由表 3 可知，基于 Hadoop 平台的朴素贝叶斯分类算法能够运行并能够实现测试样本的正确分类。但根据的实验结果表明，朴素贝叶斯返回相关实例的能力，即精确度相对较好；而识别所有相关实例的能力，即召回率相对较差；且调和平均指标，即 F1 的整体最佳模型出现在阈值 0.57 处，即话题 4。

6. 结束语

随着网络的高速发展和普及，各种社交媒体每天发布的信息呈指数性增长，面对大量的网络信息，传统人工抓取数据和跟踪药品安全舆情的方式已经不能满足实际舆情工作需求，而 Hadoop 为解决大数据处理和存储问题提供了有效途径。本文以医药新闻为数据源，使用人工标记提取相应的主题，在 Hadoop 平台上使用朴素贝叶斯算法对药品安全话题进行跟踪，虽然取得了一定的效果，但今后还需要提升算法的准确率、召回率以及调和平均指标 F1 值。

基金项目

宁夏自然科学基金(NZ17083)，国家社会科学基金西部项目(17XGL016)，2017 年宁夏医科大学优秀青年后备骨干培育对象(宁医校发[2017]119 号)，宁夏医科大学重点学科建设项目。

参考文献

[1] 袁小量, 李冰倩. 食品药品安全事件网络舆情预警策略研究[J]. 中国市场, 2017(34): 87-88.

[2] 邓飞. 近九成食药安全舆情事件处置效果不理想[N]. 中国经济报, 2014-07-15(A01).

[3] 马宾, 殷立峰. 一种基于 Hadoop 平台的并行朴素贝叶斯网络舆情快速分类算法[J]. 现代图书情报技术, 2015(2): 78-84.

[4] 刘晓欣. 中医药网络舆情分析与对策研究[D]: [博士学位论文]. 北京: 北京中医药大学, 2017.

[5] 李艳业. 冲突与对话: 新媒体语境下公共卫生危机传播研究[D]: [博士学位论文]. 兰州: 兰州大学, 2017.

[6] 刘能燕. 大数据时代政府舆情管理路径研究[D]: [博士学位论文]. 重庆: 西南政法大学, 2016.

[7] 李纲, 王晓, 叶光辉. 国内突发事件预警研究评述[J]. 情报理论与实践, 2017, 40(7): 138-144.

[8] 张雄宝. 基于突发词地域分析的微博突发事件检测方法研究[D]: [硕士学位论文]. 南宁: 广西大学, 2017.

[9] 杨柳. 面向食药安全主题的突发话题检测技术研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2018.